3-20-2023

The Geometry and Topology of Landslides

Kamal Rana
kr7843@rit.edu

The Geometry and Topology of Landslides

A dissertation submitted in partial fulfillment of the requirements towards the degree of Doctor of Philosophy in Imaging Science

by

Kamal Rana

Chester F. Carlson Center for Imaging Science

College of Science

Rochester Institute of Technology

Rochester, New York

March 20, 2023

Ph.D. Program Director: _____

CHESTER F. CARLSON CENTER FOR IMAGING SCIENCE
COLLEGE OF SCIENCE
ROCHESTER INSTITUTE OF TECHNOLOGY
ROCHESTER, NEW YORK

CERTIFICATE OF APPROVAL

Ph.D. DEGREE DISSERTATION PROPOSAL

The Ph.D. Degree Dissertation proposal of *Kamal Rana* has been examined and approved by the dissertation committee as satisfactory for the dissertation requirement for the Ph.D. degree in Imaging Science.

Dr. Nishant Malik, Dissertation Advisor

Dr. John Kerekes, Committee Member

Dr. Anthony Vodacek , Committee Member

Dr. Matthew Hoffmann , Committee Member

Dr. John McCluskey , External Chair

Date

The Geometry and Topology of landslides

by

Kamal Rana

Submitted to the Chester F. Carlson Center for Imaging Science in partial fulfillment of the requirements towards the Doctor of Philosophy Degree in Imaging Science at the Rochester Institute of Technology

Abstract

Landslides are often catastrophic, causing loss of life and destruction of property and infrastructure. Landslide susceptibility and hazard modeling help mitigate these losses by finding regions prone to landslides and providing probabilistic forecasting of landslide occurrence in a region. However, these landslide susceptibility and hazard models' efficacy depends on the quality of existing databases that often lack crucial information, like the underlying trigger and failure mechanism of a landslide. In this Ph.D. project, we developed methods to identify landslide triggering and failure mechanism information using their geometric and topological properties. For identifying landslide trigger information, we developed three different methods based solely on landslide polygon shapefiles containing landslides' two-dimensional (2D) polygon shapes. The first method uses geometric properties of landslide polygons as a feature space for a machine learning classifier–random forest. In the second method, we transformed these 2D shapes into three-dimensional (3D) point clouds by incorporating the digital elevation data and then extracting landslides' topological properties by topological data analysis (TDA) of these 3D points clouds; and to classify landslides; we treat these topological properties as the feature space of a random forest classifier. The third method uses images of landslides as input to a convolutional neural network (CNN). We tested all three methods using two different testing schemes on six known trigger inventories spread over the Japanese archipelago. In the first scheme, we combine all inventories and then split the dataset into various combinations of training and testing. We train the method on five known triggered inventories in the second

scheme and test it on the sixth inventory. Moreover, we implemented each method on an inventory without triggering information to showcase a possible real-world application. The TDA-based method is consistently the most accurate in the above analyses, ranging between 84% to 98% accuracy.

To determine the failure mechanism, we explored various geometric and topological properties of landslide shape and found that topological properties are excellent predictors for identifying landslide failure mechanisms. Therefore, we developed a method for determining landslide failure types using landslide topology. First, we extracted the topological features of the landslide 3D shape using Topological Data Analysis and then fed these features as an input to the machine learning algorithm–random forest. We implemented the developed method on the Italian and the US data separately. The method achieves above 95 and 80 accuracies for each landslide failure type for the Italian and US data sets.

The methods presented in this Ph.D. dissertation show strong performance in identifying landslide triggers and failure mechanisms. The methods are easy to use as they depend on landslide polygon as input and are transferrable to different regions of the world with adequate training data from areas with similar tectonic and climatic properties. We anticipate that the landslide community and modelers will find our method useful in determining landslides' trigger and failure mechanism. Moreover, we expect that the developed method will enhance the efficiency of landslide predictive models, such as landslide susceptibility and hazards models.

Acknowledgement

I completed this Ph.D. thesis with support from several people. However, the most crucial help came from Dr. Nishant Malik, my advisor. I would not have been able to complete this thesis without his continuous support and guidance. He was the backbone of my Ph.D. I feel grateful to him for spending an enormous amount of time improving my writing, communication, and creative thinking skills, which helped me to become a better researcher. I am also thankful to him for providing me an opportunity to spend three semesters in Germany and work with top researchers from the GFZ German Research Centre for Geosciences, Germany, and the University of Potsdam, Germany. Working with world-class researchers in Germany was a fantastic learning experience and helped me to achieve my Ph.D. objective. Dr. Malik's never-give-up attitude was a constant source of inspiration for me. His friendly, caring, and humble personality motivated me to give my best when faced with challenges. I want to say thank you to Dr. Malik from the depth of my heart for everything and for showing trust in me.

I am grateful to Dr. Ugur Ozturk (scientist at GFZ German Research Centre for Geosciences, Germany) for introducing me to the landslide domain and helping me whenever I needed him. Furthermore, I want to thank him for being an excellent host in Germany and helping me adjust to the new working and living environment in Germany.

Working with excellent collaborators and friends: Kushanav Bhuyan (Ph.D. student at the University of Padova, Italy), Joaquin (Ph.D. student at the University of Potsdam, Germany), and Adam Giammarese (Ph.D. student at RIT) was among the most fun parts of the Ph.D. In addition, I am incredibly grateful to Kushanav Bhuyan for late night discussion in his timezone during the last few months of my Ph.D.

During my Ph.D., I met incredible friends with whom I had a great time and who always supported me. I thank my imaging Science friends, especially Surya, Harsh, Biswa, Manish, Snehal, and Prateek. I am grateful to meet good friends in Germany, particularly Galan, Anushree, London, Jane, Efe, Sho, and Emma. Moreover, I would like to thank my childhood

friend Atul and my friend Sachin who supported me during my entire Ph.D.

I want to thank all my teachers who motivated me to pursue higher education and research, especially Dr. Deepak Jain (University of Delhi, India), Dr. Dalip Mehta, and Dr. Joby Joseph from the Indian Institute of Technology, Delhi.

Finally, I want to thank my parents for their sacrifice and support throughout my education. I am so grateful to my mother for teaching me to dream and work hard to achieve them. I am also thankful to my elder brother and sister, who always encouraged me to do well academically.

Publications and Software

Publications

1. **Rana, K.**, Ozturk, U., & Malik, N. (2021). Landslide geometry reveals its trigger. *Geophysical Research Letters*, 48(4), e2020GL090848.

2. **Rana, K.**, Malik, N., & Ozturk, U. (2022). Landsifier v1. 0: a Python library to estimate likely triggers of mapped landslides. *Natural Hazards and Earth System Sciences*, 22(11), 3751-3764.

3. **Rana, K.**[1], Bhuyan K.[1], Ferrer J.V., Cotton F., Catani F., Ozturk, U., & Malik, N. (2023). Landslide topology uncovers its failure mechanism *(under review in Nature Communication)*

Software

1. Landsifier v1. 0: https://pypi.org/project/landsifier/. Contributors: **Rana, K.**, Ozturk, U., & Malik, N. Released under the `OSI Approved :: MIT License`.

[1]These authors contributed equally.

Contents

Chapter 1

Introduction

1.1 Motivation

Landslides are natural hazards responsible for thousands of deaths and billions of dollars in annual economic losses (Froude and Petley, 2018). Predictive models like Landslide susceptibility and hazard models aim to mitigate these losses by locating landslide-vulnerable regions, evaluating the potential risks and consequences, and forecasting landslide occurrences. However, the efficacy of these models depends on the quality of existing landslide databases that often lack triggering and failure mechanism information, a much-required information for landslide predictive modeling. Furthermore, even state-of-the-art landslide mapping methodologies, including those reliant on satellite technology, routinely overlook the inclusion of triggering events and failure mechanism information, thereby impeding the optimal performance of predictive models.

Earthquakes and rainfall-induced landslides have different geological and geotechnical properties as two distinct physical processes trigger them (Taylor et al., 2018b; Öztürk, 2018). Similarly, distinct failure mechanisms have unique geotechnical properties due to their contrasting mechanical and kinematic behavior. Therefore, separate predictive models must be built for each trigger and failure mechanism-induced landslide. Consequently, using earthquake landslides for predictive rainfall models or vice-versa and grouping all failure mechanisms into one group leads to biases in predictive modeling (Reichenbach et al., 2018; Fressard et al., 2014)

Moreover, these biases might be catastrophic, for example, the inability to locate landslide-prone regions or predicting a low probability of landslide in a high landslide-risk zone. Hence, determining the triggering events and failure mechanism of landslides is a critical problem for landslide predictive modeling, and we decided to address and solve these problems in this Ph.D. thesis.

The landslide database contains information about the landslide's shape, time of failure, trigger mechanism, failure types, area, and volume of the failed slopes. However, the amount of information available in databases varies a lot, and most of the databases contain only information about landslide shape, especially in economically emerging countries. So, we planned to determine landslides' trigger and failure mechanisms solely based on their shape. Our approach is motivated by existing studies that showed that physical processes are embedded in the general shape of the landslides (Varnes, 1996). Inspired by these studies and the availability of polygons in landslide databases, we wanted to investigate how far the trigger and failure mechanism is ingrained in the landslide's shape.

The presented work focuses on determining landslide triggers and failure mechanisms based solely on landslide shape by exploring its geometric and topological properties. The methods presented in this work employ state-of-the-art techniques, such as Topological Data Analysis (TDA), Convolutional Neural Network (CNN), and decision tree-based classification algorithms.

1.1.1 Objectives

This Ph.D. dissertation is based on several objectives, which are explained below:

1. To explore geometric properties of landslide polygons in order to identify their triggering mechanisms.

    (a) To investigate geometric properties that distinguish between the earthquake and rainfall-induced landslide polygon.

    (b) To develop a method to determine the landslide's trigger using geometric properties of landslide shape.

2. To employ topological properties of landslide shape to identify its triggering mechanism

   (a) To use the 3D shape of landslides to capture more information about their shape

   (b) To extract topological properties of the 3D shape of landslides to enhance triggers classification performance

3. To develop a python based library to estimate the likely triggering mechanism of the landslides.

   (a) To include the presented method for identifying triggering mechanisms in an open-source Python library.

4. To explore landslide topological and geometrical properties to determine its failure mechanism.

   (a) To compare geometric and topological properties of landslides and identify the superior predictors in separating failure mechanisms.

   (b) To develop a method for determining the failure mechanism using the best predictors based on the landslide shape.

1.2 Landslides

Landslides are the movements of rocks, debris, and soil under gravity's influence and often have disastrous consequences. Each year landslides cause thousands of deaths, injuries, and billions of dollars in economic losses (Froude and Petley, 2018). Moreover, landslides are responsible for 17% of all deaths among all-natural disasters. At times, landslides can be cataclysmic; for example, the landslides that occurred in the aftermath of the 1920 Haiyuan earthquake in China led to $100,000$ deaths (Tianchi et al., 1992). Landslides are prevalent in almost all parts of the world. In the United States alone, landslides cause around $25-50$ deaths yearly, and around 1 to 4 billion dollars in annual economic losses (Dai et al., 2002). In Japan, the annual losses surpass the US and are around 4 to 6 billion dollars. In European countries like

Italy, France, Switzerland, and Austria, the annual economic loss due to landslides is around 1 to 5 billion dollars (Geertsema et al., 2009).

Furthermore, globally, around 5 percent of the world population and about 3.7 million square kilometers is vulnerable to landslides, and around 66 million people live in high-risk landslide areas (Dilley, 2005). Also, with the increase in urbanization leading to expansion of population and infrastructure, there are increasing numbers of people that are vulnerable to landslides.

1.3 Causes of Landslides

Various factors like natural causes, human activities, and a combination of natural and human actions cause landslides by disturbing slope stability. Human activities, such as deforestation, which can initiate landslides, as flora help hold soil and stabilize slopes. Moreover, constructing infrastructure without sufficient slope grading causes surface soil instability, which contributes considerably to the occurrence of landslides (Kjekstad and Highland, 2009). Another human activity that contributes to landslides is mining in landslide-prone areas.

The natural causes of landslides are earthquakes, heavy or prolonged rainfall, rapid snow melting, and volcanic eruptions. These natural causes are sudden events that imbalance the stable slope and lead to failure. Heavy rainfall and earthquakes are the most common natural causes of landslides. Heavy rainfall increases the soil weight and decreases the shear strength and cohesion between soil particles, leading to surface failures (Tacher and Bonnard, 2007). In contrast, earthquakes are caused by the movement of tectonic plates, leading to seismic waves that often cause landslides; that is why earthquake-triggered landslides are common in tectonically active zones and mountain ranges.

The effects and size of landslides triggered by natural causes depend on factors like lithology, topography, and trigger intensity (Öztürk, 2018). Properties of rocks called lithology affect landslide movements related to the landslide effect. For example, clay layers behave like a fluid with a sufficient water content that minimizes the gravitation pull on the clay particles (Stark et al., 2017). Whereas karst layers form holes between rock layers, leading

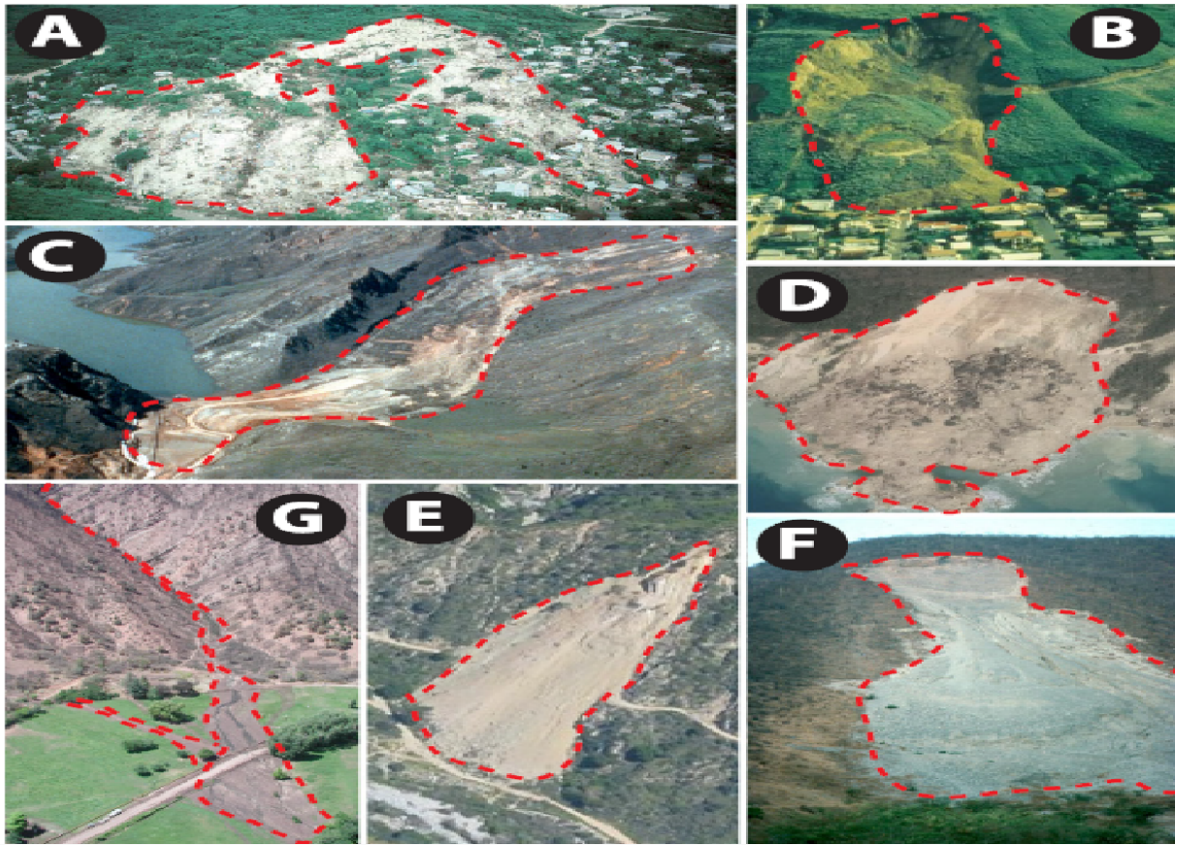

Figure 1.1: The landslides are shown in the image with the landslide polygon as the red dotted line outlining the landslides (This image is reproduced with the permission of the authors of the paper (Landslide shape, ellipticity and length-to-width ratios) (Taylor et al., 2018b).

to slope instability and surface failure (Martinotti et al., 2017). Apart from lithology, the region's topographic properties play a critical role in causing landslides. The slope is the essential topographic property that determines the landscape stability through the gravitation force (Imaizumi et al., 2017). Regions with high slopes are more prone to surface failure than regions with low slopes. Variation in slope called curvature is equally crucial in determining surface failure.

1.4 Landslides Susceptibility and Hazard Models

The economic and life losses that occur due to landslides can be mitigated by landslide predictive models such as landslide susceptibility and hazard models. Landslide susceptibility

models provide the probability of a region prone to slope failure without the expected failure time (Hervás and Bobrowsky, 2009). These models utilize the existing landslide databases with their triggers and other landslide-controlling factors to locate the landslide-prone regions (Hervás and Bobrowsky, 2009). Moreover, these models assume that past landslides include information about the conditions that lead to failure, which will also lead to future failures (Nadim et al., 2006; Hong et al., 2007). Susceptibility models can be categorized based on physical and statistical models. Usually, physical models for landslide susceptibility include lithological information like soil texture, water infiltration rate, and topographic factors like slope and curvature of landscape (Trigila et al., 2015). In contrast, statistical landslide susceptibility models primarily depend on landslide-triggering factors that can locate landslide-prone areas. These statistical models assume that landslides depend more on landslide-triggering mechanisms than a region's topographic and lithological properties (Lee et al., 2018).

Landslide hazard modeling provides the probability of landslide occurrences in a specific time interval with potential landslide area, and volume (Crozier and Glade, 2005). Some models can even calculate where and how the failure material will travel and deposit. Landslide hazard models assume that future landslides will occur in similar climatic, geomorphological, and hydrogeological conditions that cause the past landslides. This model efficacy depends on the landslide databases quality, mainly containing triggering information, as there are separate hazard models for distinct triggering mechanisms (Ozturk et al., 2021a). For example, hazard models for an earthquake require only past earthquake landslide datasets.

Apart from susceptibility and hazard models, there are other ways that help in mitigating landslide losses—for example, avoiding construction in steep slopes and existing landslide areas. Monitoring the slope stability in construction and populated areas is another way to mitigate landslide losses through the use of active sensors like ground-based interferometric synthetic aperture radars (InSAR) or satellite-based InSARs. In addition, diverting roads and highways from the rockfall landslide regions and providing drivers with early warning help mitigate losses.

1.5 Types of Landslides

Landslides are classified according to the type of movement, material, and trigger. Depending on the material involved, landslides are classified as rock, sand, or soil (Öztürk, 2018). Based on the type of movement, landslides are mostly classified as slide, flow, complex, and fall (Hungr et al., 2014a). Based on the triggers, landslides are classified as earthquake-triggered, rainfall-induced, a combination of earthquake and rainfall, and anthropogenic landslides. In this Ph.D. project, we developed methods to classify landslides based on their trigger and failure mechanism, as triggering and failure information is much-needed information for landslide predictive modeling.

1.5.1 Based on their Trigger Mechanism

Triggers are the external stimulus that leads to the instability of slope failures by rapidly decreasing the strength of the slope materials. Earthquakes, prolonged or heavy Rainfall, snowmelt, volcanic eruption, and heavy constriction near the steep slope are examples of landslide triggers. Earthquakes and Rainfall causes most of the landslides, so in this Ph.D. project, we only focused on classifying them.

The sudden movement of tectonic plates created seismic activity leading to earthquakes that often destabilize steep slopes and trigger landslides. For example, the earthquake (MW 7.9) in 2002 Denali and 2008 Wenchuan earthquake triggered around 1,500 and 60,000 landslides (Jibson et al., 2004; Cui et al., 2011).

Rainfall induces landslides by rising water tables that lead to a decrease in shear strength, and cohesion, an increase in soil weight, and causes the erosion of soil. Both rainfall intensity and amount can induce landslides. However, high-intensity Rainfall usually induces shallow landslides, whereas high rainfall amounts are also responsible for reactivating deep-seated landslides. For example, intense and prolonged Rainfall during hurricane Mitch induced tens of thousands of landslides of different sizes (Devoli et al., 2007); 800 mm of Rainfall in the Fukuoka region in Japan in a day induced around 2000 shallow landslides.

1.5.2 Based on their Failure Mechanism

Landslides are classified into four types based on their failure mechanism–slide, flow, fall, and complex. Slide failure mechanism involves the movement of soil and rock sliding through almost a planar or curved concave surface (Varnes, 1978). Slides usually occur in steep slopes having cohesive rock or soil. In contrast, falls is the rapid movement of rock down detached from the cliff of a steep slope (Bourrier et al., 2013). Usually, the removed material strikes at a lower angle slope, leading to breaking or rolling rocks until it reaches a flattened area.

A flow is a continuous spatial movement in which shear surfaces are short-lived, closely spaced, and rarely preserved. The displacing mass of a flow's component velocities is similar to those of a viscous liquid as they propagate in a viscous kinematic state contouring the landscape. Whereas complex failures exhibit characteristics of a combination of more than one failure mechanism that occur simultaneously or later (Cruden, 1996; Hungr et al., 2014b). For example, a combination of slide and flow is considered a complex landslide.

1.6 Landslides Dataset

Detailed information about past occurrences of landslides is stored in landslide databases. These databases contain information about the landslide polygon, which outlines the landslide (Fig.1.1), time of failure, trigger mechanism, failure types, slope angle, rainfall patterns, and vegetation in the failure location. The quality of databases is an essential factor for landslide predictive models such as landslide susceptibility and hazards and understanding landslide causes, effects, and impacts. However, most of these past landslide databases lack trigger and failure mechanism information, much-needed information for predictive modeling. Also, newly acquired landslide databases lack triggering and failure information due to the automated data-capturing process. This Ph.D. thesis is focused on developing methods to identify landslide triggering and failure mechanisms for better quality landslide databases and their importance for better landslide predictive modeling.

1.7 Past Studies on Triggering and Failure Mechanism

Past studies suspect landslide polygons to reveal their triggering mechanism (Varnes, 1996). (Taylor et al., 2018a) showed that earthquake-triggered landslides have longer accumulation zone than rainfall-induced landslides and thus have a lower length-to-width ratio than rainfall-induced landslides. Moreover, few other studies attempt to categorize landslides based on their triggering mechanism and use perimeter, area, and similarity of landslides to circle or ellipsoid (HR.Pourghasemi et al., 2014; Samia et al., 2017; Taylor et al., 2018a; Milledge et al., 2014). However, none of the past studies classify landslides based on triggering mechanisms due to irregularities in landslide polygons (O.Marc and Hovius, 2015).

Regarding the identification of failure mechanisms, there have only been a few research that can be divided into two categories: knowledge-based and data-driven methods. Knowledge-driven approaches to identifying failure mechanisms are limited to small areas as these studies are based on spectral, topographic, and morphometric parameters that vary from region to region, because of which they cannot be adapted dynamically to other regions (Barlow et al., 2006). To the best of our knowledge, there exists only one data-driven method for determining failure mechanism (Amato et al., 2021); however, this method uses only a 2D landslide polygon and misses crucial information like kinematic and mechanic information of landslide failure. Kinematic and mechanic information on landslide failure is much more important information for distinctive different failure mechanisms, and disregarding this information reflects in the performance of the data-driven method.

1.8 Layout of the Thesis

This Dissertation consists of 8 chapters containing the introduction, results, and conclusion of the work done during the Ph.D. period. Chapter 1 provides the motivation, objectives, and introduction of the Ph.D. project. Chapters 2 and 3 address the first objective of the Ph.D. project, which is using geometric properties to identify the trigger mechanism of landslides. Chapter 4 addresses the second objective of the Ph.D. project, using topological properties

to enhance landslide triggers classification accuracy. Chapter 5 includes the third objective of the Ph.D. about developing a library to estimate likely triggers of landslides. Chapter 6 and 7 contains the last objective of the Ph.D. about using geometrical and topological properties to determine the failure mechanism of the landslides. Finally, Chapter 8 contains the conclusion of the work.

Chapter 2

Landslide Geometry Reveals its Trigger

This chapter is based on Objective 1 of the Ph.D. thesis about exploring the geometric properties of landslide polygon to identify its triggering mechanism. Here, we explored the various geometric properties of the landslide polygon and found seven geometric properties that are excellent predictors in identifying landslide triggering mechanisms. The work presented in this chapter is published in:

Rana, K., Ozturk, U., & Malik, N. (2021). Landslide geometry reveals its trigger. *Geophysical Research Letters*, 48(4), e2020GL090848.

2.1 Abstract

Electronic databases of landslides seldom include the triggering mechanisms, rendering these inventories unusable for landslide hazard modeling. We present a method for classifying the triggering mechanisms of landslides in existing inventories, thus, allowing these inventories to aid in landslide hazard modeling corresponding to the correct event chain. Our method uses various geometric characteristics of landslides as the feature space for the machine learning classifier *random forest*, resulting in accurate and robust classifications of landslide triggers. We applied the method to six landslide inventories spread over the Japanese archipelago in several

different tests and training configurations to demonstrate our approach's effectiveness, and we achieved mean accuracy ranging from 67% to 92%. We also provide an illustrative example of a real-world usage scenario for our method using an additional inventory with unknown ground truth. Furthermore, our feature importance analysis indicates that landslides having identical trigger mechanisms exhibit similar geometric properties.

2.2 Introduction

Landslides pose a constant threat to human life, individual property, and infrastructure in rugged terrains globally. Thus, algorithms that can learn patterns from past landslides and provide early-warning signals are highly sought-after (Osanai et al., 2010). However, the performance of such algorithms critically depends on the quality of the existing landslide inventories. Some of these inventories are compiled by mapping landslides that cause damages, e.g., along roads (Pittore et al., 2018) and merging these mappings with existing datasets to create a knowledge base (Havenith et al., 2015). More contemporary mapping includes repeated satellite monitoring of landslide-prone regions to create complete landslide databases (Behling et al., 2014; Tanyas et al., 2017). Although event-based inventories include clear traces of the triggering mechanisms of landslides (von Specht et al., 2019), many other inventories, such as satellite-based, lack crucial information linking a given landslide to a specific triggering mechanism (Behling et al., 2014). The missing information about triggering mechanisms decreases the efficacy of these inventories in landslide hazard analyses, as this could introduce biases, for instance, inadvertently using earthquake-triggered landslides to assess landslide hazard for extreme rainfall (Ozturk et al., 2020). Hence, there is a need to identify triggers of landslides in exiting databases to make them usable in hazard models.

Geometric features of landslide polygons are suspected to reflect their trigger (Varnes, 1996). Coseismic landslides tend to have a lower length-to-width ratio than the rainfall-induced landslides due to their extended accumulation zone (Taylor et al., 2018a). Topographic site effects amplify seismic signals at higher altitudes causing coseismic landslides to cluster around mountain ridges (Meunier et al., 2008; Rault et al., 2019). On the contrary, rainfall-

induced landslides happen to connect to the local drainage network, which could be deceptive considering their extended accumulation (Marc et al., 2018; Ozturk et al., 2018). Accordingly, existing studies try to categorize landslide based on their triggers using (i) perimeter area indices (HR.Pourghasemi et al., 2014); (ii) similarity of a landslide planform to a circle or an ellipsoid (Samia et al., 2017; Taylor et al., 2018a); and (iii) scaling relationships between landslide dimensions and area (Milledge et al., 2014). However, it is challenging to apply these metrics for automatic classification of landslides due to the irregularities in landslide polygons and amalgamation of several types of landslides in the databases (see, e.g.,(O.Marc and Hovius, 2015)).

Here we present a machine learning-based approach that uses the geometric features of landslide polygons to identify the underlying triggering mechanisms of landslides. It is the geometry of the landslides that provides physical insight about their trigger. The trigger mechanisms get encoded into the geometry of the landslide polygons (e.g. the outline of debris field). To illustrate the effectiveness and accuracy of our approach, we apply it to several inventories spread over the Japanese archipelago with known triggers. We anticipate that our robust yet straightforward approach will be transferable to build a landslide knowledge base. For future practical deployment of the approach, we also demonstrate the applicability of our method on a landslide inventory without any triggering information. Apart from providing a useful technique for landslide trigger classification, our study also highlights that the information of trigger mechanisms is embedded in the geometric features of landslides.

2.3 Data

In this work, we analyze seven landslide inventories that belong to six different regions of Japan (see Fig. 7.2). We know the trigger mechanisms in six of the seven inventories, and we will employ these inventories with known triggers to test the efficiency of our method. For the seventh inventory, the triggering mechanism is unknown, and we will use it to demonstrate the practical implementation of the method as this scenario represents the most probable real-world usage of our method. We used the coseismic landslides associated with following earthquakes:

the 2018 Hokkaido Eastern Iburi (M_W6.6); the 2008 Iwate–Miyagi Nairiku (M_W6.9), and the 2004 Niigata (M_W6.6). The Geospatial Information Authority of Japan (GSI) is the source of landslide inventory from the Hokkaido Eastern Iburi earthquake, while the source of the other two coseismic inventory is the repository created by (Schmitt et al., 2018b). GSI also provides rainfall-induced landslide inventories of the Fukuoka (July 2017) and Saka (July 2018) regions. We employ two more inventories from the Kumamoto region provided by the National Research Institute for Earth Science and Disaster Resilience (NIED) of Japan. In one of these inventories, the underlying trigger is documented as rainfall; however, the second inventory lacks any triggering information. We will refer to this second inventory as "unspecified."

The unspecified Kumamoto inventory locates in the Kumamoto region of Japan but is geographically far from the rainfall-triggered Kumamoto inventory. The landslides in the unspecified Kumamoto inventory are mapped along the rims of the Aso Caldera, which is an volcanic active region.

2.4 Method

The landslide planforms (polygons) are one of the primary information in landslide inventories, and geometric features of these polygons are a rich source for understanding physical mechanisms underlying a particular landslide (Kasai and Yamada, 2019; Milledge et al., 2014; HR.Pourghasemi et al., 2014; Samia et al., 2017; Taylor et al., 2018a). Therefore, we explored various geometric properties of landslide polygons and identified a subset to form the feature space for machine learning-based automatized classification of landslides into two categories: earthquake-induced and rainfall-induced landslides.

We started by exploring a broad set of measures to quantify the geometric shapes of two-dimensional polygons extracted from the landslide planforms. In Table 3.1, we provide a complete list of these geometric measures. Using a combination of feature selection approaches (described in detail in section 3.2 and 3.3 of the chapter 3), we were able to select seven features that lead to the best classification accuracy (Ambroise and McLachlan, 2002; Friedman et al., 2001; Chandrashekar and Sahin, 2014; Scott, 2015a). These seven geometric features were area

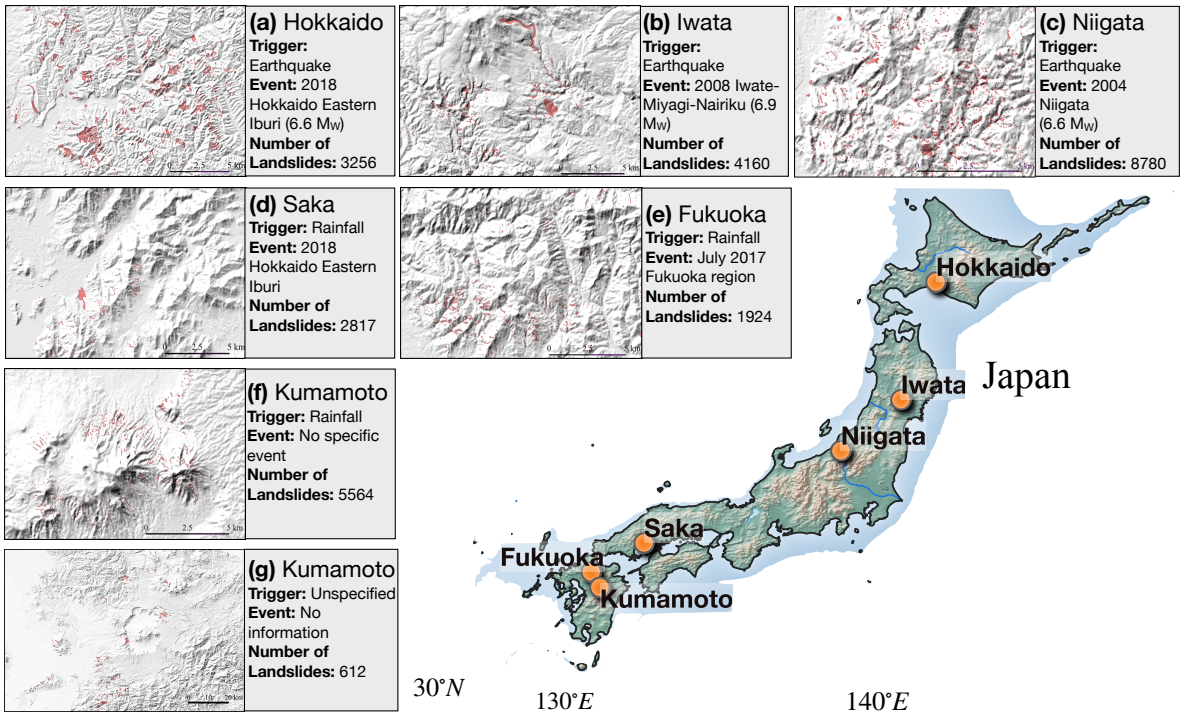

Figure 2.1: The map of Japan shows the geographical locations of the seven landslide inventories used in this work. (a-g) Shows the Digital Elevation Model (DEM) of these inventories, and the adjoining panels list the region, trigger type, the origin of the event, and the number of landslides. The red color overlays on the DEM are a subset of landslide polygons of each inventory. Japanese Geospatial Information Authority (GSI) is the source of data in (a),(d), and (e); National Research Institute for Earth Science and Disaster Resilience (NIED) is the source of (f) and (g) and data in (c) and (b) is from (Schmitt et al., 2018a).

A, perimeter P, convex hull based measure $C_h = \dfrac{A}{A_c}$, where, A_c is the area of the convex hull fitted to the polygon (hereafter, we will refer C_h as convex hull measure), the ratio of area and perimeter $\dfrac{A}{P}$, width of the minimum area bounding box W, minor axis s_m, and eccentricity of the fitted ellipse e (Fig. 2.2). As each feature has a different range of measurement values, we standardized the data by calculating z-scores of each feature.

For classifying landslides, we employed *random forest*, an ensemble-based learning method (Liaw et al., 2002; Breiman, 2001; Biau and Scornet, 2016; Biau, 2012; Barnett et al., 2019; Kursa, 2014; Roy and Larocque, 2012; Rodriguez-Galiano et al., 2012). Unlike the standard tree-based methods, where all the attributes are used for the best split of a node, in a random forest, only a random subset are used, and each decision tree is constructed using different

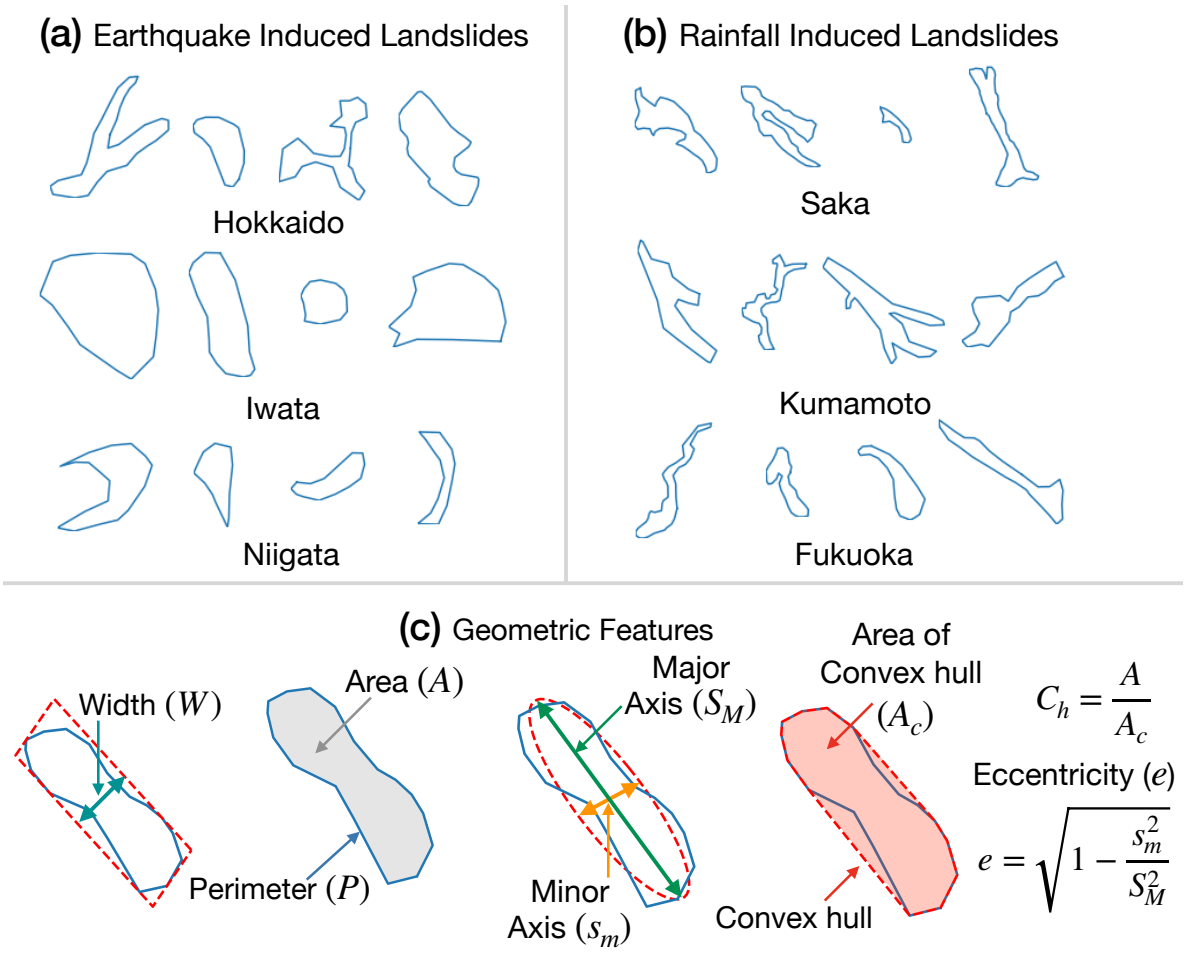

Figure 2.2: Sample landslide planforms from (a) Earthquake triggered (b) Rainfall triggered inventories (c) Geometric features (left to right): Width (W) of minimum area bounding box fitted to the landslide polygon, area (A) and perimeter (P) of the landslide polygon, minor (s_m) and major axis (S_M) lengths of an ellipse fitted to the polygon and convex hull based measure $C_h = \frac{A}{A_c}$ where A_c is the area of convex hull fitted to the polygon.

bootstrap samples of the data (Liaw et al., 2002; Breiman, 2001). For the testing sample, each tree predicts the class independently, and the class having the majority vote is the class prediction of the sample. Below we briefly describe the steps in implementing random forest for our binary classification problem (Liaw et al., 2002; Breiman, 2001; Zhang and Ma, 2012a; Friedman et al., 2001).

We build a decision tree T_b for each bootstrap sample b of the training set; let $\mathbf{X}_i = [x_{ij}]_{j=1}^{p}$ represent a p dimensional feature vector for data point i. Next, we recursively repeat the

following two steps for each node of the tree: (i) select m variables among p and (ii) split the node into daughter nodes that best separate the classes. As we carry out binary classification, i.e., classifying landslides into earthquake or rainfall triggered, the parent node q is split into two daughter nodes left l and right r. For this splitting, we employ the Gini index method, where Gini index for l and r are $G_l = 1 - p_{l1}^2 - p_{l2}^2$ and $G_r = 1 - p_{r1}^2 - p_{r2}^2$ respectively. Note p_{lj} is the probability of data points of class j in node l. For each split of a node, the Gini index of the sub-nodes should be less than the parent node, and can be achieved for a split s_q of node q if we maximize the decrease in the quantity $\Delta\theta(s_q) = G_q - \rho_{rq}G_r - \rho_{lq}G_l$; where ρ_{rq} (ρ_{lq}) are the ratio of the number of data points in daughter nodes r (l) to the total number of points in the parent node q (Zhang and Ma, 2012a; Kuhn and Johnson, 2013). The two recursive steps (i) and (ii) continue until a predefined criterion is satisfied. For example, we are left with one data point each in the daughter nodes, and no further splitting is possible.

Now if $C_b(\mathbf{Y}_i)$ is the class prediction of the b-th random forest tree for $\mathbf{Y_i}$ point in the training set, then the class prediction of $\mathbf{Y_i}$ is given by the majority vote in the set $\{C_b(\mathbf{Y}_i)\}_1^B$; where B is the total number of bootstrap samples. In this setting, the importance of the k-th feature in predicting the training set is given by $I(k) = \dfrac{1}{B}\displaystyle\sum_{T_b}\sum_{q\in T_b: D(s_q)=k} \rho_q\theta(s_q)$; where ρ_q is ratio of number of points in the q-th node to total number of points in training data and $D(s_q) = k$ implies that the feature involved in the split s_q is k-th feature. The $I(k)$ in above formulation measures the average of weighted impurity decrease $\rho_q\theta(s_q)$ over all the splits in the ensemble of random forest trees. As $\sum_k I(k) = 1$, we express the feature importance in percentage as $\dfrac{I(k)}{\sum_k I(k)} \times 100$.

2.5 Results

For the first numerical experiment, we constructed testing and training set by combining the six inventories with known triggers to validate our approach (Fig. 2.3). In this experiment, the total number of samples in the combined dataset were $n_{total} = 26501$, with $n_{rainfall} = 10305$ and $n_{earthquake} = 16196$. We randomly resample the data to take an equal number of the earthquake and rainfall samples ($n_{rainfall} = n_{earthquake} = 10305$) to avoid any class imbalance.

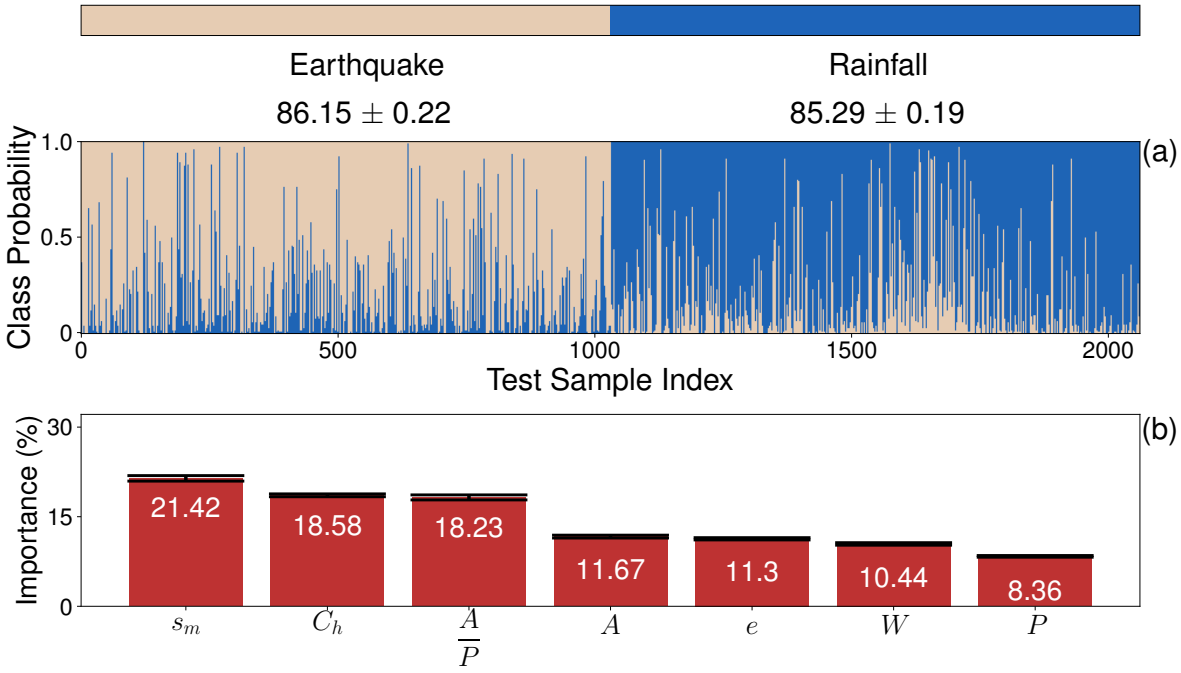

Figure 2.3: (a) Earthquake and rainfall triggered class accuracies using random forest classifier on geometric features of landslide polygons. The underlying data is a concoction of all the six inventories with available ground truth, where the training set has 18,548, and the test set has 2,062 samples. Within the test set, 1,031 samples are earthquake driven, and 1,031 are rainfall driven. Using 1000 runs of10-fold cross-validation, we identified $86.15 \pm 0.22\%$ earthquake-triggered landslides and $85.29 \pm 0.19\%$ rainfall-triggered landslides correctly. The plot in (a) is the output from one of the runs of the random forest classifier. The class probability represents the proportion of votes for a class in the ensemble of trees. The x-axis is the index of the sample in the test set. (b) The importance of geometric features used in (a). The percentage corresponding to each feature represents the mean decrease in the tree leaf impurity over the full random forest such that the total percentage sums to 100.

Thus, we apply the algorithm on $n = 2 \times 10305 = 20610$ samples. We also employed 1000 runs of 10-fold cross-validation to swap training and testing sets to avoid the likelihood of results influenced by overfitting and smaller standard deviation in each case indicate more stable class performances. Dividing 20610 landslide samples into 10-folds with an equal number of landslide samples in each fold leads to uneven numbers of earthquake and rainfall samples (20610/10 =2061; an odd number). To tackle this issue, the five-folds out of ten-folds have 2062 samples (both earthquake and rainfall have 1031 samples), and the remaining five-folds have 2060 landslide samples (each earthquake and rainfall have 1030 samples). Thus, each iteration of 10-fold cross validation has either $n_{train} = 18548$, $n_{test} = 2062$ (1031 each rainfall-induced and

coseismic) or $n_{train} = 18550$, $n_{test} = 2060$ (1030 each rainfall-induced and coseismic). Here n_{train}, n_{test} are number of training and testing samples.

In the experiment with combined data, using $\mathbf{X} = [A, P, C_h, \frac{A}{P}, W, s_m, e]$ as the feature vector, we achieved the mean classification accuracy of 85.73 $\pm$ 0.16%, where 86.15 $\pm$ 0.22% earthquake-triggered and 85.29 $\pm$ 0.19% rainfall-triggered events were classified correctly (see Fig. 2.3 (a)). Among all the geometric features, minor axis length s_m has the highest feature importance of 21.42%, followed by the convex hull measure C_h (18.58%) (see Fig. 2.3 (b)). To check for the possible bias in feature importance due to multi collinearity of the geometric predictors. We use two different approaches and end up with the same conclusion that minor axis length s_m , convex hull measure C_h and area to perimeter($\frac{A}{P}$) are the most important predictors and perimeter is the least important feature (described in detail in section 3.6 of the chapter 3).

In section 3.6 of the chapter 3, we provide further detailed analysis of these results, including additional metrics evaluating the algorithm's performance. Also, see Figs. 3.6-3.7 and Table 3.2 in the chapter 3.

In the second experiment, we applied to our approach to individual inventories; we train the algorithm on five of the six inventories and predict the trigger of landslides in the sixth inventory. That is, the training data has no information on the test data inventory—a situation similar to the one we anticipate this method will be used in the real world. We use the same set of geometric features as in Fig. 2.3, and to avoid class imbalance, we keep the number of rainfall and earthquake samples the same by resampling the data. The method achieved over 85% classification accuracy for the Saka ($n_{train} = 14976, n_{test} = 2817$), and Niigata ($n_{train} = 14832, n_{test} = 8780$) region. 83.63% accuracy for the Kumamoto region ($n_{train} = 9482, n_{test} = 5564$), 75.59% for the Iwata region ($n_{train} = 20610, n_{test} = 4160$), and 66.62% for the Hokkaido ($n_{train} = 20610, n_{test} = 3256$), and 69.40% for Fukuoka ($n_{train} = 16762, n_{test} = 1924$) regions (see Fig. 2.4). In this experiment, the model performed better with classifying the rainfall triggered inventories than the earthquake triggered inventories. Performance drops as low as 67% in the case of Hokkaido. We repeat the run 1000 times with

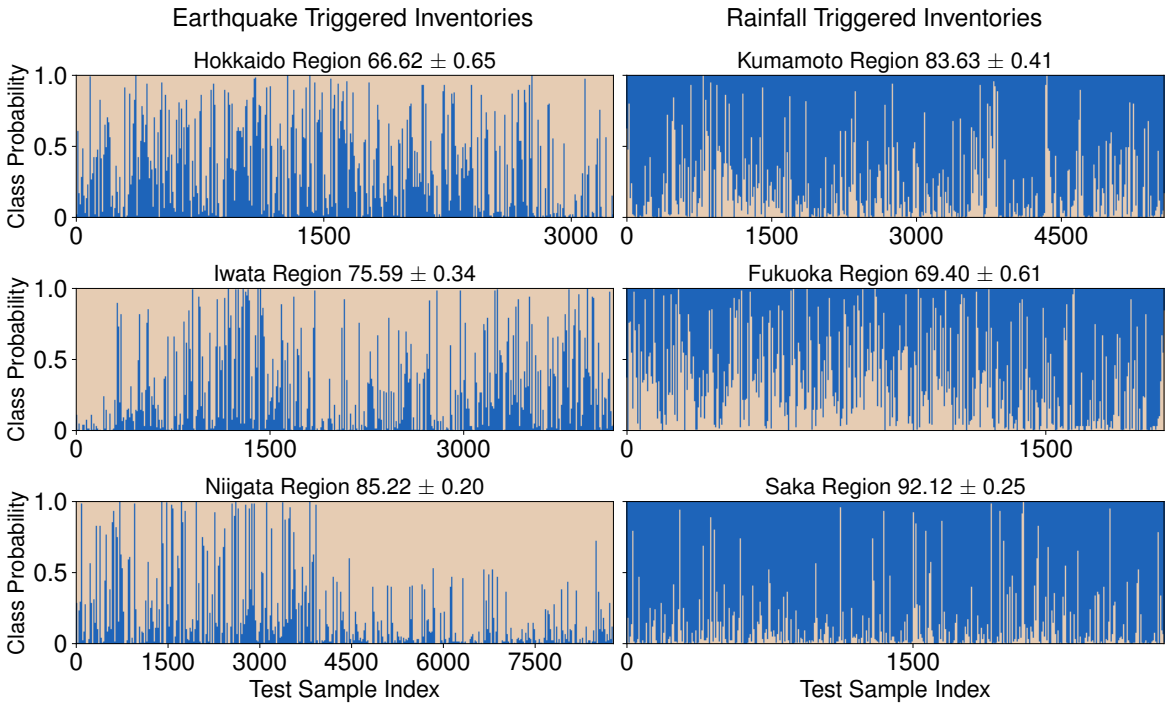

Figure 2.4: Predicting one of the six inventories while the random forest classifier was trained on the rest of five inventories, i.e., the classifier has no information on the data being predicted. The geometric features used are the same as in Fig. 2.3. The deviation quoted in the accuracy percentage was calculated by running the classifier algorithm 1000 times (note cross-validation is not possible to implement in this configuration of testing and training data). Also, see Figs. S8 and S6 in the SI.

random selection of equal earthquake and rainfall landslide samples for training. The results are stable with change in training samples, as the standard error is small and greatest being 0.65% for the Hokkaido region. For further analysis of this experiment see Sec. 3.5 in the chapter 3.

In the third experiment, we applied our method to the unspecified trigger inventory from the Kumamoto region (inventory (g) in Fig. 7.2). Out of 612 landslides in the test case, 604 were classified as earthquake-induced and 8 as rainfall-induced. We present a detailed analysis and discussion on this inventory and our results in section 3.4 of the chapter 3.

2.6 Discussion

We showed that using the random forest algorithm in conjunction with geometric features of landslide planforms is a robust technique for classifying landslides trigger mechanisms, and it can achieve excellent classification rates in a variety of settings. For example, while using the aggregate data from different regions of Japan with varied geology and topography, this scheme achieved an accuracy of 85%. Whereas in the individual analysis of different regions, the classification accuracy went up as high as 92%. Moreover, we identified seven geometric features of landslide polygons that appear to be the best predictors of the underlying trigger mechanisms. Our results indicate that although there is diversity in the physical mechanisms producing landslide events, there is also a universality in these mechanisms that get embedded in the geometry of landslide planforms.

A possible reason for the emergence of this universality could be that both rainfall-induced and coseismic landslides in our databases consist of a mostly shallow landslide, e.g., debris flows (von Specht et al., 2019; Watakabe and Matsushi, 2019; Kasai and Yamada, 2019). Although coseismic rockslides tend to cut the mountain ridge and slide through concave slopes likely on a lithological failure plain with relatively short accumulation zones (Havenith, 2002), the number of such failures is limited in our databases. Shallow landslides, such as debris flows, tend to flow through convex slopes, following the local morphology with extended accumulation that could increase the source area tenfold (Uchida et al., 2013; Wang et al., 2015; Hungr et al., 2014b). Additionally, debris flows are prone to form concave planforms due to flow divergence along valley bottoms. Hence more extended accumulation areas will result in elevated convex hull and area–perimeter ratio (Fig. 2.2c) in our analyses. Although these shallow slope failures could be classified into the same landslide type (Varnes, 1996; Hungr et al., 2014a), our results indicate that the coseismic and rainfall-induced landslides form divergent planforms (Taylor et al., 2018a). Hence length-to-width and area-perimeter ratios, together with the semi-minor axis, contribute the most to the classification (see Fig. 2.3), highlighting that rainfall and earthquake driven landslides have distinct geometry.

To further demonstrate the practical aspects of our approach, we also applied it to an

unspecified trigger inventory from the Kumamoto region (inventory (g) in Fig. 7.2). The landslides in this inventory are mapped along the rims of the Aso Caldera; the active volcano Mount Aso shakes the surrounding area, frequently triggering landslides within its vicinity (Saito et al., 2018). Therefore, the majority of the unspecified landslides may be seismically triggered. In our analysis, we find the same, out of 612 landslides in the test case, our algorithm classified 604 as earthquake-induced and 8 as rainfall-induced. Given most of the landslides classified as coseismic, we are confident in flagging this inventory as seismically triggered due to the volcanic activity.

The size of the databases we used is relatively small, i.e., it is challenging to apply more sophisticated image classification algorithms, such as the convolutional neural networks (CNN) to the problem at hand, as these algorithms require extensive training sets. In contrast, as illustrated above, the random forest was able to achieves good enough accuracy to apply the method in real world application even when training sets were rather small. . Also, the random forest is comparatively computationally inexpensive, its complexity in our case was $\mathcal{O}(n \log n)$, where n is the number of samples. Furthermore, the random forest is highly portable, as nowadays many machine learning packages include an elaborate implementation of this algorithm (Pedregosa et al., 2011). Given these advantages, we anticipate that the landslide modeling community will find our scheme useful.

2.7 Conclusion

Historic landslide inventories rarely include the triggering mechanisms of the observed landslides, a critical piece of information for landslide hazard and susceptibility models. We developed a method that can fill this missing information by classifying existing landslides in digital databases of landslides. Our method uses geometric characteristics of landslide polygons as features for random forest classifier. The resulting algorithm is highly portable and accurate and can be applied to any region of interest with adequate training data from regions with similar tectonic and climatic features. We also identified seven geometric features of landslide planforms that appear to capture some universal patterns in the landslide trigger mechanisms.

Furthermore, we applied our scheme to several different tests and training set configurations of the available data from Japan, and our results indicate this method is versatile, robust, and can classify landslide triggers with high accuracy. Envisioning that this method will be applied to individual inventories with unknown triggers in practice, we prepared such a test application and demonstrated that the model classifies it as an earthquake driven inventory, a highly plausible classification based on the geographic location of the inventory.

Our study is limited to the Japanese archipelago consisting of landslides samples triggered by either earthquake or rainfall. We hope research community will use our method to datasets inventories from various geographical regions with different mechanism.

Chapter 3

Additional Evidences showing landslide geometry exhibit its triggering information

This chapter is based on objective 1 of the Ph.D. thesis about exploring the geometric properties of landslide polygon to identify its triggering mechanism. Here, we showed a details analysis of selecting seven geometric properties for identifying the triggering mechanism. The work presented in this chapter is published in the **supplementary material** of the Journal manuscript:

Rana, K., Ozturk, U., & Malik, N. (2021). Landslide geometry reveals its trigger. *Geophysical Research Letters*, 48(4), e2020GL090848.

3.1 Introduction

This supporting information (SI) to the manuscript titled: "Landslide Geometry Reveals its Trigger," describes the approaches we followed for selecting the geometric predictors used in the landslide triggers classification. It also contains an in-depth analysis of these geometric attributes' probability distributions quantifying the dissimilarities between earthquake and rainfall triggered landslide polygons. With the help of the probability density of geometric predictors, we provide additional evidence that the earthquakes are the triggers for the un-

specified Kumamoto inventory (inventory with unknown ground truth). In this SI, we have also included an elaborate discussion on the performance of our algorithm.

3.2 Geometric Attribute Selection For Landslide Classification

Selecting the optimum number of features for the model is crucial as it reduces the overfitting, generalization error, and computational time. A model with fewer predictors is much more interpret-able than a model with a large number of predictors(Ambroise and McLachlan, 2002)(Friedman et al., 2001)(Chandrashekar and Sahin, 2014). In our work, we used 17 attributes(refer to Table. S1) of landslide polygons quantifying the different geometric properties of a landslide shape, i.e., area, perimeter, complexity, compactness, and circularity. Out of these 17 geometric attributes, some of the features are strongly correlated, as a few predictors represent the same geometric property of a shape. More than one feature representing the same geometric property in a model will not significantly improve the generalization accuracy, and the possibility of overfitting will increase. Removing one of the two strongly correlated features is a common practice (unsupervised feature selection)(Friedman et al., 2001), and we used Spearman rank-order correlation, a measure of the monotonic relationship between the two variables to identify strongly correlated features(Kuhn and Johnson, 2013).

Spearman rank-order correlation, ρ vary from -1 (perfect negative correlation) to 1 (perfect positive correlation). Out of the two strongly correlated predictors, one was removed from the feature space if $|\rho| = 1$ between them. For instance, the eccentricity of the fitted ellipse has $|\rho| = 1$ with the ratio of major and minor axis, complexity measure, compactness calculated using iso-perimetric quotient and Richardson complexity measure. Elongation and aspect ratio of the fitted minimum bounding box are also strongly correlated with $|\rho| = 1$. Therefore, among all these features, we kept the eccentricity and elongation for further feature selection and removed the rest of them. After an initial unsupervised selection of features, the feature importance of the remaining eleven relevant predictors is calculated with the Gini index-based method using a random forest algorithm (see manuscript for a brief description; for a more detailed description, see (Breiman, 2001)). For estimating error bounds on feature importance

and due to class imbalance (briefly explained in manuscript), we employed 100 realizations of K-fold cross-validation(K=10) that divides the data into K-groups(folds). Each iteration of cross-validation chooses K-1 folds for training and rest one fold for the testing purpose such that none of the iterations have the same subset of fold for testing the model (Kuhn and Johnson, 2013). The feature with the least feature importance was removed. We repeat the whole process of removing one feature at a time until the error rate remains constant. We found that the error rate is constant until we are left with seven features. We chose these seven features in the model and removed the less relevant predictors as they were not significantly increasing accuracy and only contributing to increasing the model's complexity. The seven features selected are minor-axis, convex hull measure, area, perimeter, width, eccentricity, and the ratio of area and perimeter.

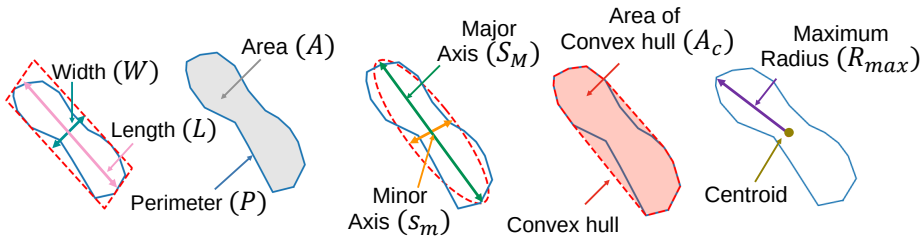

Figure 3.1: Geometric features (left to right): Width (W) and length(L) of minimum area bounding box fitted to the landslide polygon, area (A) and perimeter (P) of the landslide polygon, minor (s_m) and major axis (S_M) lengths of an ellipse fitted to the polygon, convex hull fitted to landslide polygon and area of convex hull, and maximum radius of landslide polygon.

3.3 Probability Density of Geometric attributes

In this section, we visualize the probability density functions (PDF) of the geometric properties used in the classification of earthquakes, and rainfall triggered landslides. We estimate these PDFs' using Kernel Density Estimation (KDE), which gives a non-parametric estimate of the

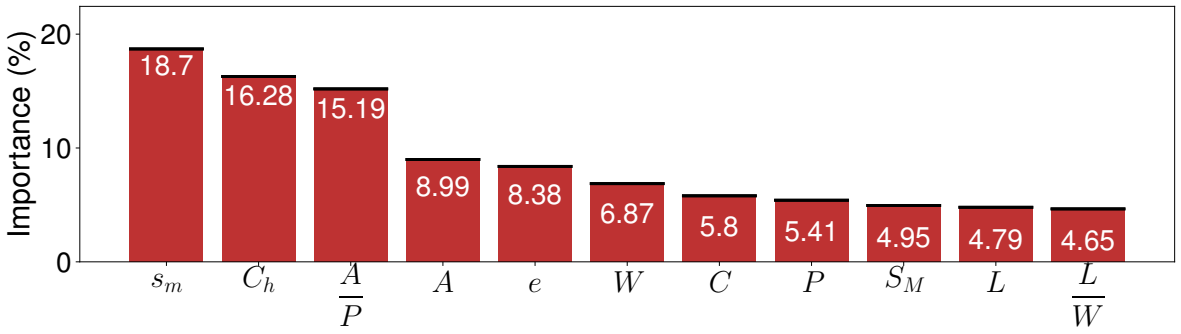

Figure 3.2: The importance of geometric features used after the removal of strongly correlated predictors. The percentage corresponding to each feature represents the mean decrease in the tree leaf impurity over the full random forest such that the total percentage sums to 100. The Y-axis shows the feature importance values and x-axis shows geometric attributes. The abbreviations used in the figure for geometric attributes are listed in Table 3.1.

PDF (Friedman et al., 2001; Scott, 2015b). These PDFs' (shown in Fig. 3.3) revealed that predictors with high feature importance have relatively high dissimilarity between the PDFs of two landslide triggering classes: rainfall and earthquake. For instance, s_m, the minor axis of an ellipse fitted to rainfall landslide polygons have a relatively high probability of having a value below 25 meters and fewer chances for having above 50 meters. However, observing earthquake landslide polygons with s_m above 25 meters and below 75 meters is high.

For the convex hull measure C_h, the earthquake polygons have a very sharp peak around $C_h = 1.0$; whereas, this peak is more subdued in rainfall. The PDF of C_h reveals that the earthquake polygons are more compact than rainfall, as the convex hull measure quantifies the compactness of polygon. Similarly, the value of eccentricity e of the fitted ellipse for rainfall-triggered landslide has a sharp peak near $e = 1.0$, in contrast, distribution for earthquake lacks such a peak (see Fig. 3.3). For the ratio of area to perimeter $\frac{A}{P}$, the probability of observing rainfall landslides polygon with $\frac{A}{P} < 10$ is higher than earthquake polygons (see Fig. 3.3). The PDFs of area (A), width (W), perimeter (P), length (L), and major axis (S_M) shows that the probability of observing earthquake polygons with relatively higher values of these variables is more than the rainfall-triggered landslide polygons.

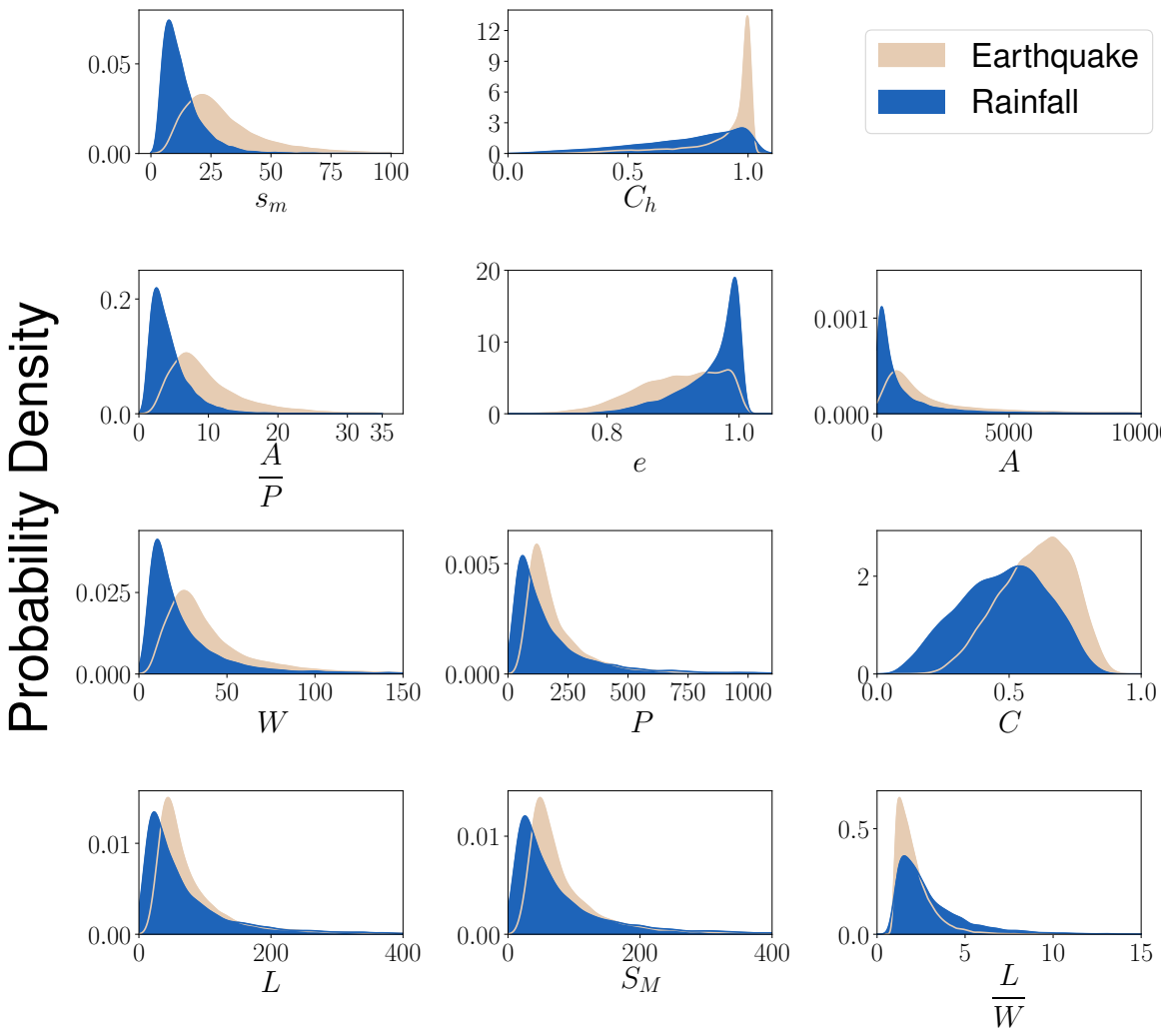

Figure 3.3: Probability density functions of all the geometric attributes used after the initial removal of highly correlated predictors of earthquake and rainfall triggered polygons. The y-axis shows the probability density values and x-axis shows the value of geometric attributes. The probability density calculated using Kernel Density Estimate (KDE) method using gaussian kernel and bandwidth selected using Scott's rule (Scott, 2015b).

3.4 Unspecified Kumamoto Landslide Polygons classification

The PDFs of geometric features used for classifying landslides polygons of the unspecified Kumamoto inventory (inventory (g) in Fig. 1 in the paper) has no resemblance to the rainfall triggered landslide polygons(see Fig. 3.4). However, for properties C_h and e there is a considerable

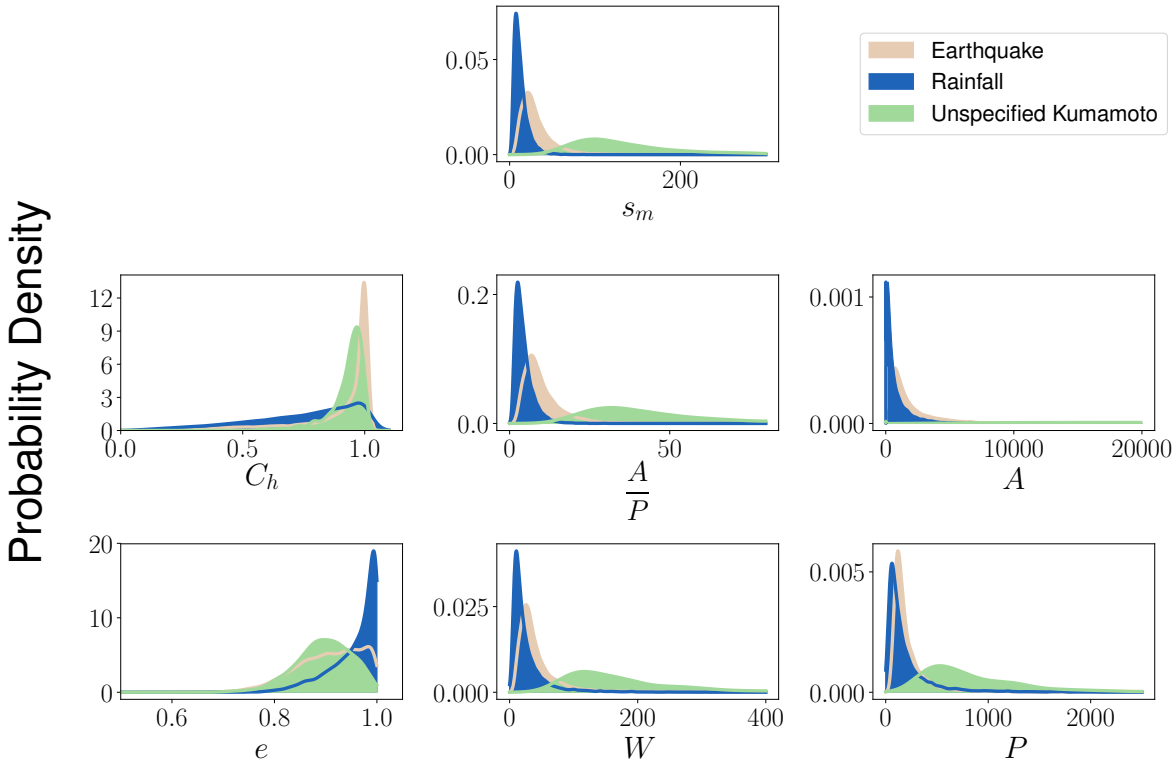

Figure 3.4: Probability density functions of the relevant predictors of unspecified kumamoto inventory compared with probability density functions of earthquake and rainfall landslide polygons. The y-axis shows the probability density values and x-axis shows the value of geometric attributes. The probability density calculated using Kernel Density Estimate (KDE) method using gaussian kernel and bandwidth selected using Scott method.

visual resemblance between PDFs of the unspecified inventory and the earthquake-triggered landslide polygons. It is evident from the PDFs of geometric attributes that landslides in the unspecified Kumamoto inventory are not triggered by rainfall; there is some evidence in these PDFs that these landslides are triggered by earthquake.

Moreover, we calculated the average landslide trigger class probability of each landslide sample in unspecified Kumamoto using 1000 iterations(training on the equal number of earthquake and rainfall samples from the six known triggered inventories)(see Fig. 3.5). We observed that around 90% of samples have earthquake class probability from 90 to 100%, and approximately 95% of samples have earthquake class probability from 80 to 100% (see Fig. 3.5). High earthquake class probabilities for most of the samples provides additional evidence

that earthquakes triggered unspecified Kumamoto inventory landslides. As we have observed in other cases, when the algorithm is confident (assign a high possibility to a landslide trigger class for a test sample), almost all the time, it is correct.

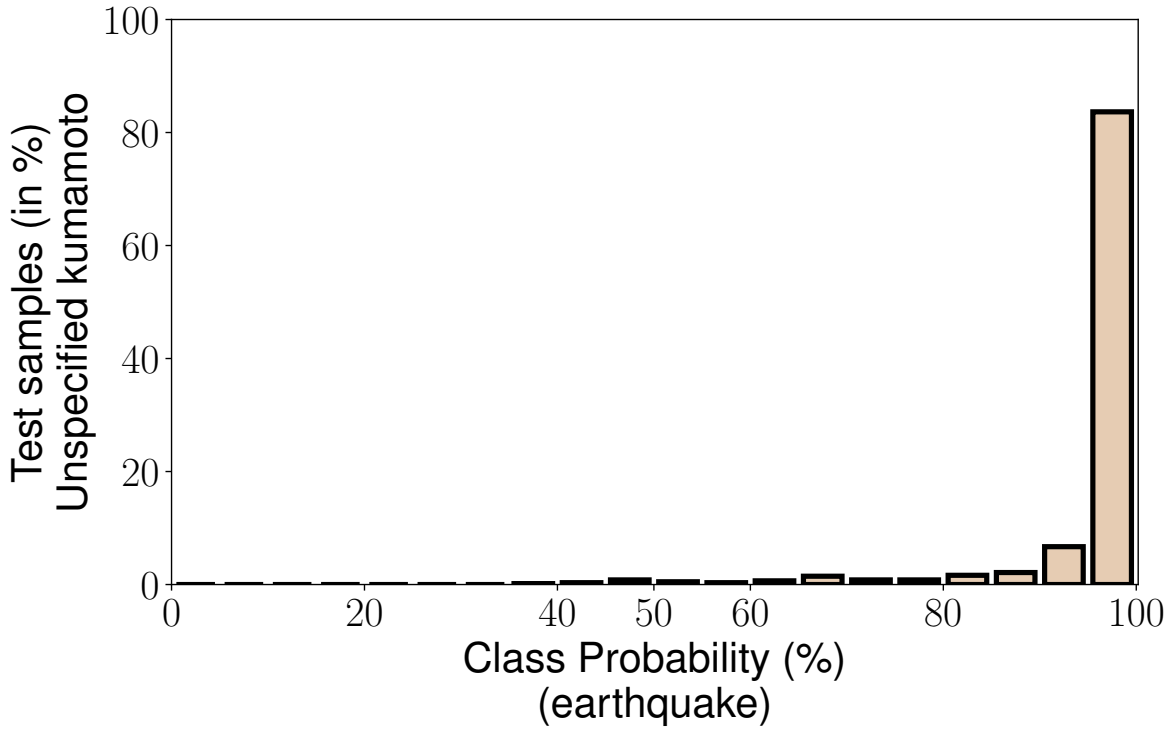

Figure 3.5: The plot shows the percentage of testing landslides samples that fall within a particular interval of earthquake class probability. The random forest classifier was trained on six known triggered inventories and tested on the unspecified Kumamoto inventory. The geometric features used are the first seven variables in Table 3.1. Observe that 95% of samples have earthquake class probability from 80 to 100%, highlighting algorithms high confidence in classifying majority of landslides in this inventory as earthquakes.

3.5 Other measures to evaluate model performance

We applied various other measures to evaluate the performance of our method. Assuming the earthquake as the positive class and rainfall as the negative class, we first measured the true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) values and then using them, we calculated sensitivity, specificity, precision, negative predictive value, and F1 score. Furthermore, we also constructed the confusion matrix for each of our numerical

(a)

	Actual Positives	Actual Negatives
Predicted Positives	True Positives (TP)	False Positives (FP)
Predicted Negatives	False Negatives (FN)	True Negatives (TN)

(b)

875 (TP)	150 (FP)
156 (FN)	881 (TN)

(c) Hokkaido Region (d) Iwata Region (e) Niigata Region

Earthquake Triggered

2174 (TP)	* (FP)
1082 (FN)	* (TN)

3133 (TP)	* (FP)
1027 (FN)	* (TN)

7500 (TP)	* (FP)
1280 (FN)	* (TN)

(f) Kumamoto Region (g) Fukuoka Region (h) Saka Region

Rainfall Triggered

* (TP)	933 (FP)
* (FN)	4631 (TN)

* (TP)	583 (FP)
* (FN)	1341 (TN)

* (TP)	235 (FP)
* (FN)	2582 (TN)

Figure 3.6: Confusion matrices for various numerical experiments, we assumed earthquakes as the positive class and rainfall as the negative class. (a) Table explaining the terminologies used in the confusion matrices. (b) Confusion matrix for one iteration of 10 fold cross-validation for the numerical experiment when landslide samples are from all the six inventories with known ground truth. Fig. 3 in the paper corresponds to this experiment. (c-h) Confusion matrix for one realizations for the case where the algorithm's training is done on five of the six inventories and tested on the remaining one inventory. (c-e) for earthquake-triggered inventories (f-h) rainfall triggered inventories. Fig. 4 in the paper corresponds to this experiment. Note, the asterisk symbol represents the case where we cannot calculate a particular element of the confusion matrix due to the presence of only one class in testing samples.

experiments (see Fig. 3.6).

The sensitivity ($\frac{TP}{TP+FN}$) is the positive class classification rate, whereas specificity($\frac{TN}{TN+FP}$) is the negative class classification rate. Precision ($\frac{TP}{TP+FP}$) is the proportion of samples designated correctly positive among all predicted positive values by the algorithm. Similarly, negative predictive value ($\frac{TN}{TN+FN}$) is the proportion of samples assigned accurately negative among all predicted negative values. F1 score ($\frac{2TP}{2TP+FP+FN}$) is the harmonic mean of pre-

cision and sensitivity. Accuracy ($\frac{TP+TN}{TP+TN+FN+FP}$) is calculated as the ratio of total samples correctly classified to the total samples. We used the average TP, FP, FN, and TN estimated over 1000 10-fold cross-validation run (see Table 3.2).

All the measures calculated are over 85% and have a standard deviation of less than 0.25% (Table 3.2). The high accuracy and low standard deviation of all the measures show that the presented method can efficiently predict landslide triggers. We can only calculate sensitivity for earthquake-triggered inventories and specificity for negative-triggered inventories for the experiment, where we trained on five of the six inventories and tested on the remaining one inventory. As in this experiment, we have either earthquake or rainfall triggered test samples.

For each testing landslide sample, the random forest algorithm assigns the probability (p_e, p_r) to each landslide trigger class (earthquake, rainfall). Hereafter we will express these class probabilities (p_e, p_r) in percentage (note: $p_e = 1 - p_r$). We divided the class probability percentage range of earthquakes p_e from 0 to 100 into 20 equal bins to calculate the variation of accuracy with class probability intervals. In Fig. 3.7 we show variation of accuracy with class probability intervals for the numerical experiment which includes all the data with ground truths accessible to us. We observed that for more than 50% of testing samples, the method gives a very high probability of over 90% for being earthquakes or rainfall. The approach predicts landslide triggers class accurately in more than 95 percent cases in this probability range. There is relatively less classification accuracy for the intervals where the algorithm assigns class probabilities between 50 to 70%. However, the proportion of testing samples with this probabilities range are relatively low compared to instances where the algorithm assigns class probabilities of around 80 to 100%. We found that the algorithm is more confident (assigning higher probability for any class) when a high probability class is the landslide's actual trigger. Next, we carried out the same analysis for each inventory where the algorithm is trained on five of the six inventories and tested on the remaining one inventory. We found similar observations; most correctly classified samples have a higher probability of actual trigger class(see Fig. 3.8). For instance, most of the correctly classified samples in earthquake-triggered inventories have 80 to 100% earthquake class probability. Similarly, correctly classified samples

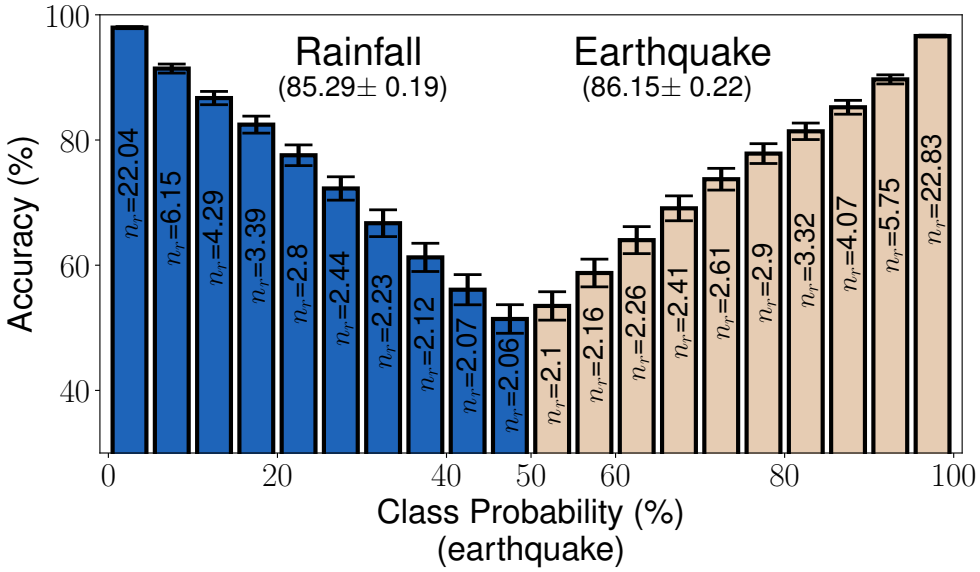

Figure 3.7: This plot corresponds to the numerical experiment where we have used all the inventories with known ground truth (i.e., Fig. 3 in the main paper). n_r is the proportion of testing landslide samples that fall within the given class probability interval. The y-axis shows the classification accuracy, and the x-axis shows the percentage class probability interval for the earthquake. Here, a landslide with class probability greater than 50 is classified as triggered by an earthquake, and a landslide with class probability less than 50 is classified as triggered by rainfall. The seven geometric features used are the first seven variables listed in Table 3.1. Error bars were calculated using 1000 runs of 10-fold cross-validation.

of rainfall trigger inventories have 0 to 20% earthquake class probability (80 to 100% rainfall class probability).

Our motivation for carrying out the above calculations was to show that our algorithm's high-class probabilities also correspond to highly accurate results. In real-world applications, this fact can be manipulated to identify a subset of landslides in an inventory for which the algorithm is highly confident about the underlying trigger.

3.6 Approach to examine bias in feature importance

Random Forest selects a random subset of features for node split that decreases the bias in feature importance due to collinearity. However, it does not decrease the bias completely. To minimize this bias, we already removed strongly correlated features. Note we started with 17

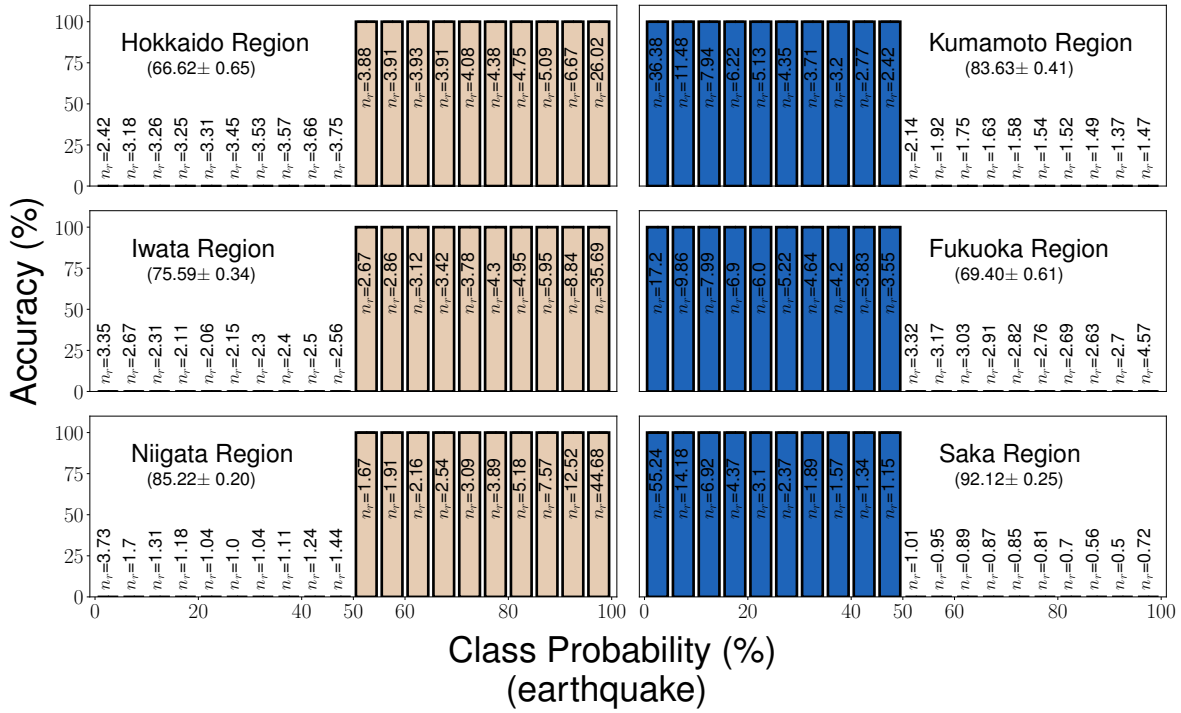

Figure 3.8: This plot corresponds to the numerical experiment where we trained on five of the six inventories and tested on the remaining inventory (i.e., Fig. 4 in the main paper). n_r is the proportion of testing landslide samples that fall within the given class probability interval. The y-axis in each subfigure shows the classification accuracy, and the x-axis shows the percentage class probability interval for the earthquake. Here, a landslide with class probability greater than 50 is classified as triggered by an earthquake, and a landslide with class probability less than 50 is classified as triggered by rainfall. The seven geometric features used are the first seven variables listed in Table 3.1.

features and removed 6 of them due high correlations between them and used only 11 of them for the further analysis. Please see section S2 for details.

We implemented PCA to get another view on the feature importance of predictors. The contribution to variance in principal components is almost equal by all predictors except convex hull measure and eccentricity, which have a minimum contribution to variance. However, we did not end up to any conclusions as it shows most of the features are equally important. Also, it is not always true that features that contributed the maximum to variance are the most important features, and PCA works only for linear data.

To check the bias in the feature importance of predictors, we used another approach. We took all possible subset of seven predictors and calculate feature importance for all the cases.

In each case, we observed that minor-axis, convex hull measure, and area to perimeter ratio have relatively higher feature importance than rest predictors. Area, eccentricity, and width have comparable feature importance, and the perimeter has the least important feature among all predictors. The order of relative feature importance of predictors found in all cases supports the order of predictor feature importance using all seven predictors.

We also project features data into a lower dimension for classification using PCA, and even we used sophisticated dimensionality reduction techniques like manifold learning. However, there was no improvement in the classification accuracy of landslides. Other issues with these techniques are that once we project our geometric features into a lower dimension, the predictors' geometric interpretation will be lost.

Variables	Description
A	Area of the polygon.
P	Perimeter of the polygon.
A/P	Ratio of the area and perimeter of the polygon.
C_h	Convex hull measure ($C_h = A/A_c$); here A_c is the area of the convex hull fitted to the landslide polygon.
s_m	The minor axis of the fitted ellipse (with area A and perimeter P) to the polygon,($s_m = \sqrt{P^2 - \sqrt{P^4 - 16\pi^2A^2}}/\pi$)
S_M	The major axis of the fitted ellipse to the polygon ($S_M = 4A/\sqrt{P^2 - \sqrt{P^4 - 16\pi^2A^2}}$)
L	The length of the minimum area bounding box fitted to the polygon.
W	The width of the minimum area bounding box fitted to the polygon.
L/W	The elongation of the minimum area bounding box fitted to the polygon.
W/L	The aspect ratio of the minimum area bounding box fitted to the polygon.
C	The circularity measure; it is calculated using the ratio of the radius of a circle having the same area as of polygon to the maximum radius of a polygon($C = \sqrt{A/\pi}/R_{max}$), here R_{max} is the maximum distance from the centroid to vertex of a polygon
e	The eccentricity of the fitted ellipse to the polygon ($e = \sqrt{1 - s_m^2/S_M^2}$)
s_m/S_M	The ratio of minor axis to major axis of the fitted ellipse
S_M/s_m	The ratio of major axis to minor axis of the fitted ellipse
C_{p1}	Compactness of landslide polygon, calculated using Iso-Perimetric Quotient ($C_{p1} = 4\pi A/P^2$)
C_{p2}	Compactness of landslide polygon, calculated using Richardson complexity measure ($C_{p2} = 2\sqrt{\pi A}/P$)
C_v	Complexity of landslide polygon, calculated using Schwartzberg complexity measure ($C_v = 0.5P/\sqrt{\pi A}$)

Table 3.1: List of all the variables (their abbreviations and formulas) used in this work. The variables listed here were used to identify the most relevant predictors for landslide trigger classification, and we identified seven features (first seven variables in the table) as the most relevant predictors.

Measures	Average Values (in %)	Standard Deviation values(in %)
Sensitivity	86.15	0.22
Specificity	85.29	0.19
Accuracy	85.73	0.16
Precision	85.42	0.17
False Negative Rate	86.03	0.20
F1 Score	85.78	0.16

Table 3.2: List of all the measures calculated to evaluate the model for landslide triggers classification. We are assuming an earthquake as the positive class and rainfall as the negative class. The average value and error bounds of each measure are calculated using 1000 runs of 10-fold cross-validation. These results correspond to the case shown in Fig. 3 and Fig. 3.6; that is where we include all the available data into the testing and training set.

Chapter 4

Using Landslide topology for determining its trigger mechanism

This chapter is based on objective 2 of the Ph.D. thesis about exploring the topological properties of landslide shapes to enhance their triggering mechanism performance. Here, we computed the topological properties of the landslide 3D shape to identify its triggering information. The work presented in this chapter is published in:

Rana, K., Malik, N. & Ozturk, U. (2022). Landsifier v1.0: a Python library to estimate likely triggers of mapped landslides. *Natural Hazards and Earth System Sciences*, 22(11), 3751-3764.

4.1 Abstract

Landslide hazard models aim at mitigating landslide impact by providing probabilistic forecasting, and the accuracy of these models hinges on landslide databases for model training and testing. Landslide databases at times lack information on the underlying triggering mechanism, making these inventories almost unusable in hazard models. We developed a Python-based unique library, `landsifier`, that contains three different Machine-Learning frameworks for assessing the likely triggering mechanisms of individual landslides or entire inventories based on landslide geometry. Two of these methods only use the 2D landslide planforms, and the third utilizes the 3D shape of landslides relying on an underlying Digital Elevation

Model (DEM). The base method extracts geometric properties of landslide polygons as a feature space for the shallow learner—Random Forest (RF). An alternative method relies on landslide-planform images as an input for the deep learning algorithm—Convolutional Neural Network (CNN). The last framework extracts topological properties of 3D landslides through Topological Data Analysis (TDA) and then feeds these properties as a feature space to the Random Forest classifier. We tested all three interchangeable methods on several inventories with known triggers spread over the Japanese archipelago. To demonstrate the effectiveness of developed methods, we used two testing configurations. The first configuration merges all the available data for the k-fold cross-validation, whereas the second configuration excludes one inventory during the training phase to use as the sole testing inventory. Our geometric features-based method performs satisfactorily, with classification accuracies varying between 67% and 92%. We have introduced a more straightforward but data-intensive CNN alternative, as it inputs only landslide images without manual feature selection. CNN eases the scripting process without losing classification accuracy. Using topological features from 3D landslides (extracted through TDA) in the RF classifier improves classification accuracy by 12% on average. TDA also requires less training data. However, the landscape autocorrelation could easily bias TDA-based classification.

Finally, we implemented the three methods on an inventory without any triggering information to showcase a real-world application.

4.2 Introduction

Landslides are gravitational movements of rock and debris that pose a severe threat to the human environment (Depicker et al., 2021). Hazard models are developed to forecast landslides or to aid in understanding landslide processes to mitigate their undesired consequences (Lombardo et al., 2020). These models commonly rely on mapped landslides to assess the relevant landslide causes in combination with landslide triggers, i.e., earthquake and rainfall (Lombardo and Tanyas, 2021; Ozturk et al., 2021b; Marin et al., 2020). However, many historical landslide inventories lack information about the triggering mechanism decreasing their potential utility

in models (Bíl et al., 2021; Martha et al., 2021). More recent semi-automated satellite-based landslide mappers also often disregard the triggering information (Behling et al., 2014; Behling et al., 2016; Ghorbanzadeh et al., 2019), except the event-based inventories—landslide mapping campaigns following a precursory triggering event such as a strong earthquake (Stumpf and Kerle, 2011; Gorum et al., 2014). Using landslide inventories with missing triggers could introduce biases as it is possible to accidentally use an earthquake-triggered inventory to assess rainfall-induced landslide hazards and vice-versa. Hence, classifying the trigger of entire landslide inventories or mapped individual landslides would enhance the usability of newly acquired and historical inventories in landslide models (Guzzetti et al., 2012).

Landslide planforms are used to estimate the mobilized landslide volume, for example, estimating the potential sediment budget of a large landslide triggering events (Malamud et al., 2004; Fan et al., 2012). This type of scaling relationship between the area of landslide planforms to mobilized landslide volume allows comparing the impact of different landslide triggers, such as human versus earthquakes, in terms of the landslides triggered influence on landscape (Tanyaş et al., 2022). However, this area-volume scaling depends on the triggering mechanism of landslides. For example, an earthquake-triggered landslide has a different area-volume relationship than a rainfall-induced landslide. Hence, extracting the landslide triggers information could enhance the estimation capacity of landslide volumes (Moreno et al., 2022) and also help predict the size of co-seismic landslides for a given earthquake (Lombardo et al., 2021). Also, when the exact trigger is known, observed landslides help assess earthquakes' ground motion patterns when no seismic observation is available (Lombardo et al., 2019).

Landslides with the same trigger morphologically cluster, for example, covering narrowly the available statistical variability of hillslope angles in a study region (Jones et al., 2021) and, thus, could have characteristic shapes reflecting their triggering mechanism, for instance, by having similar area and perimeter ratio, or size (Taylor et al., 2018a; Samia et al., 2017). We developed a binary classifier that groups landslides either as earthquake-triggered or rainfall-induced based on this hypothesis (Rana et al., 2021). This initial model demonstrated that the landslides with an identical trigger indeed exhibit similar geometric properties. Thus, finding

the trigger of landslides is a classification problem, and one can employ machine learning tools to carry out automated classification of landslide triggers. In each classification problem, the principal idea is to construct a classifier based on training samples and evaluate its performance on testing samples. The classifier predicts the class y corresponding to the input sample x. These input samples x can be one-dimensional vectors or images; for instance, in a soil classification problem (Bhattacharya and Solomatine, 2006), x is a one-dimensional vector, and in any image classification problem, x is an image (2D or multi-dimensional matrix) (Domingos, 2012).

Our preliminary model (Rana et al., 2021) can classify landslide triggers by only using the geometric properties of landslide polygons. Here, we introduce two additional methods for landslide trigger classification. In one new method, we treated landslide polygons as images, and these images are fed as the sole predictor to a deep learner—Convolutional Neural Networks (CNN). Treating landslide polygons as images eases the workflow as an image already resembles some of the geometric features of the first method. Both these methods rely on two-dimensional (2D) landslide planforms, ignoring the three-dimensional (3D) shapes of real-world landslides. In another approach, we included the 3D shapes of landslides by incorporating the elevation of landslides via a Digital Elevation Model (DEM). In this approach, we extracted the topological features of these 3D shapes using a recently developed technique known as Topological Data Analysis (TDA). These topology-based features are input to the decision tree-based shallow learner as in the first method. We included the TDA-based model considering its potential to handle other relevant classification problems in future versions of our tool, e.g., classifying landslide types (Cruden and Varnes, 1996; Varnes, 1978). Above listed methods could be used independently following similar script streams.

This study also introduces a new Python library, `landsifier`, that classifies the trigger of landslides, individually or as a whole, in an inventory, where the landslide source mechanism is undocumented. `Landsifier` is the first-ever library built for estimating likely triggers of mapped landslides, the methods used in this library to find landslides' triggers are new. Two of these methods are introduced in this paper for the first time, while the third was

published in our preliminary work (Rana et al., 2021). The library consists of three different machine learning-based methods mentioned above; we elaborate on these methods in section 4.4. Various functionalities of the library are described in section 5.2 of the chapter 5; where we also list several supporting functions to calculate landslide polygons' geometric properties, convert landslide polygons' shape to a binary-scale image, download a Digital Elevation Model (DEM) corresponding to inventory location, and evaluate the diagnostic performance of the final classification. To demonstrate the efficacy of the developed methods, we apply each to six landslide inventories with known triggers spread over the Japanese archipelago and document our findings in section 4.5. In section 4.7, we further highlight the weaknesses of each method to ease choosing the suitable classifier for the various applications.

4.3 Data

In this work, we used seven landslide inventories spread over the Japanese archipelago (Figure 4.1). The trigger mechanism of six out of seven landslide inventories are known (Figure 4.1a–f), whereas the last inventory has no documented triggering information (Figure 4.1g). We use the last inventory to demonstrate the practical deployment of the final model as this case represents the model's real-world usage. Out of six landslide inventories, three inventories are earthquake-triggered (Figure 4.1d–f) that are associated with the 2018 M_W6.6 Hokkaido Eastern Iburi (3256 landslides); the 2008 M_W6.9 Iwate–Miyagi Nairiku (4160 landslides), and the 2004 M_W6.6 Niigata (8780 landslides). The remaining three are rainfall-induced (Figure 4.1a–c), and these are associated with the 2017 Fukuoka-northern Kyushu torrential rainfall disaster (1924 landslides), the 2018 Saka-Japan floods (2817 landslides), and Kumamoto inventory (5564 landslides) that is collected over 1992–2012—not associated with any particular event.

The Geospatial Information Authority of Japan (GSI) is the source of the Hokkaido Eastern Iburi earthquake (September 2018), Fukuoka rainfall (July 2017), and Saka rainfall (July 2018) inventories. The source of the other two coseismic inventories—Iwata and Niigata—is the global repository created by (Schmitt et al., 2018b). The remaining two inventories from the Kumamoto region are provided by Japan's National Research Institute for Earth Science and

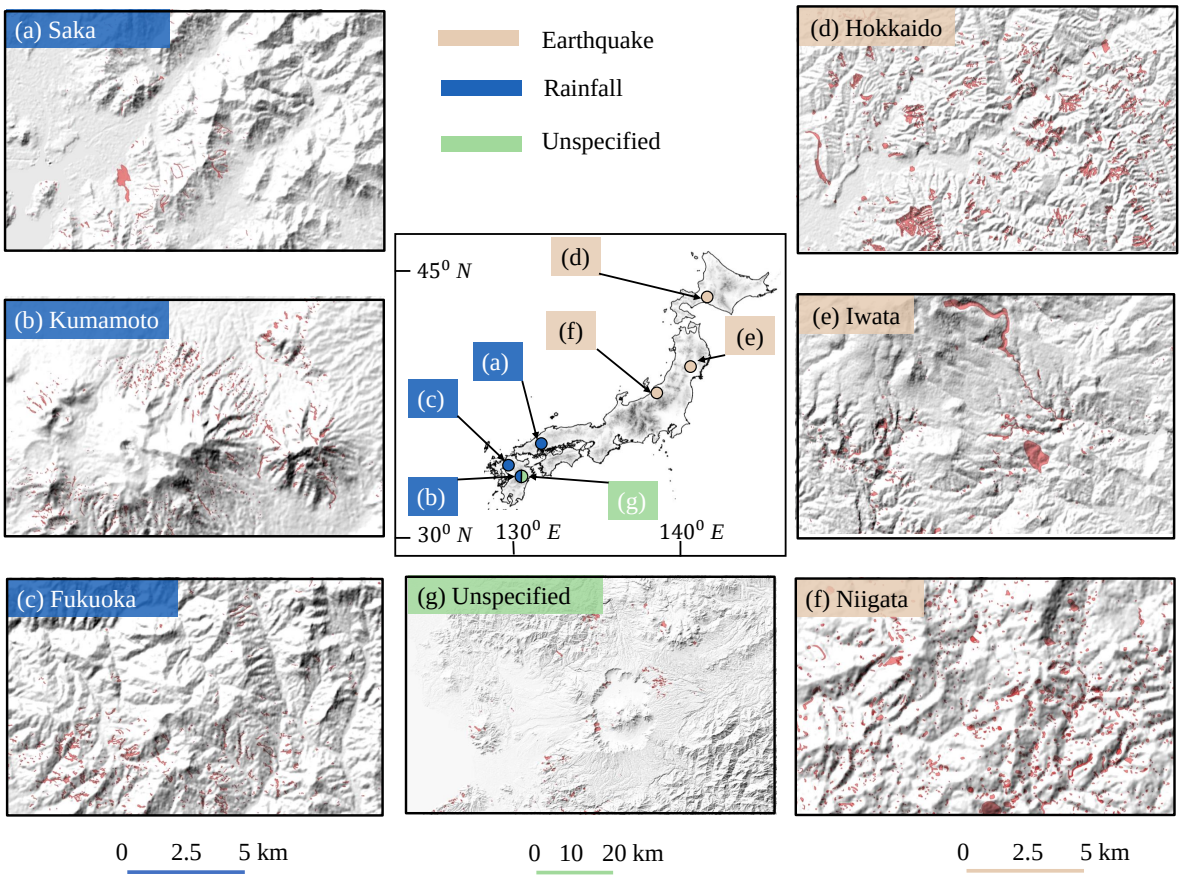

Figure 4.1: The seven landslide inventories used in this work are spread over Japan, and their geographical locations are shown on the Country's map at the center of the figure. (a)–(g) Shows the subset of landslide polygons highlighted by red color on local hillshades. (a)–(c) Rainfall-induced inventories; (d)–(f) coseismic inventories; (g) undocumented "Kumamoto unspecified" inventory.

Disaster Resilience (NIED). The first inventory from Kumamoto is associated with rainfall (Figure 4.1b), whereas the second inventory is without any triggering information (Figure 4.1g). From hereafter, we refer to this second inventory as "Kumamoto unspecified" (it consists of 612 landslides with unknown triggers).

The TDA-based method uses elevation data to obtain the 3D shapes of landslides from their 2D planforms. We use the *Shuttle Radar Topography Mission* (SRTM) Digital Elevation Model (DEM) data that comes with a spatial resolution of approximately 30 meters. The SRTM data is freely available from `https://www2.jpl.nasa.gov/srtm/` by manually selecting the

tiles which correspond to topographic quadrangles. Each tile covers 1 degree of both latitude and longitude region. The `landsifier` library automatically downloads the corresponding tile(s) covering the region of the used landslide inventory (explained further in section S2 of the Supplementary information).

4.4 Methods

In our preliminary study (Rana et al., 2021), we introduced a method that can classify landslide triggers by only using geometric features of landslide planforms. This initial model constitutes the first method in `landsifier` library, and for continuity, we briefly describe it in section 4.4.1. In this paper, we further diversify our initial model and introduce two new methods, one based on the topological features of 3D shapes of landslides computed using TDA; described in section 4.4.2. The other new method uses CNN to carry out an image-based classification of landslide triggers; see section 4.4.3. We anticipate that the variety of methods and corresponding Python library presented here would allow researchers to perform this analysis seamlessly.

4.4.1 First method: geometric features based classification

In the first method, we used the geometric properties of 2D landslide polygons for the classification. We explored several geometric properties of landslide polygons (e.g., Figure 4.2). Using a combination of feature selection methods and feature importance analysis, for instance, removing highly correlated features, we choose the seven geometric properties of polygons that lead to optimum results. These geometric features are area A, perimeter P, convex hull based measure $C_h = \frac{A}{A_c}$, where, A_c is the area of the convex hull fitted to the polygon (hereafter, we will refer C_h as convex hull measure), the ratio of area to perimeter $\frac{A}{P}$, the width of the minimum area bounding box W, minor axis s_m, and eccentricity of the fitted ellipse e having area A and perimeter P. All these seven geometric features are calculated using the Python library, `shapely` (Gillies, 2013). The feature vector ($[A, P, C_h, W, s_m, \frac{A}{P}, e]$) is input variable to machine learning algorithm—random forest (described in the chapter 5 section 5.1). Further

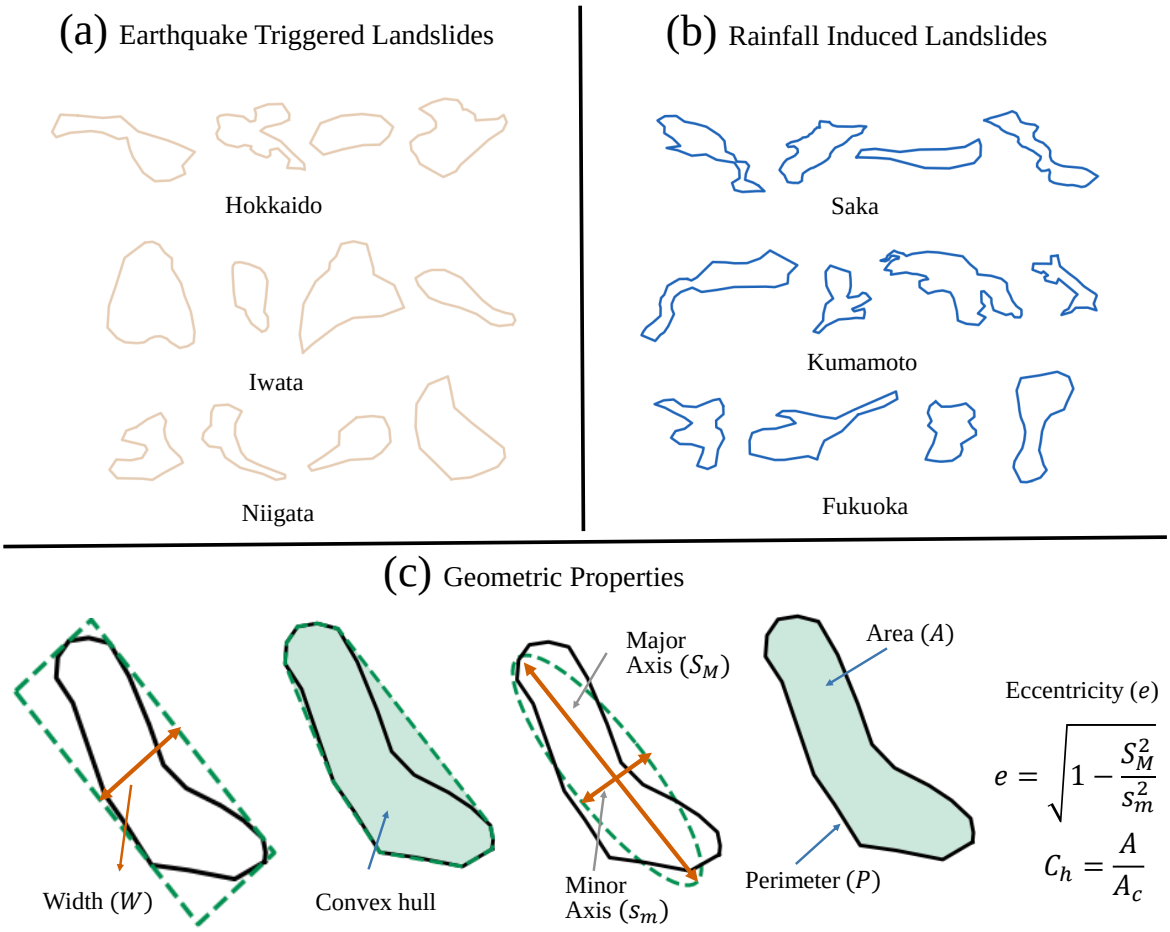

Figure 4.2: Sample landslide planforms from all six known triggered inventories (a) Earthquake triggered inventories, (b) Rainfall induced inventories. (c) geometric properties of landslide polygon (from left to right): width (W) of the minimum area bounding box fitted to polygon, convex hull based measure (C_h), minor(s_m), and major axis(S_M) of an ellipse fitted to polygon having area A and perimeter P, area (A) and perimeter (P) of the polygon.

details of the method can be found in (Rana et al., 2021).

In (Rana et al., 2021), we analyzed the distributions of geometric properties of the earthquake and rainfall polygons and found geometric dissimilarities between earthquake and rainfall polygons' shapes. Earthquakes polygons are more likely to have a compact shape (as measured by convex hull-based measure) than rainfall polygons. Moreover, earthquake polygons have more chances to have a larger area (A), perimeter (P), the ratio of the area to the perimeter ($\frac{A}{P}$), and minimum width (W) than rainfall polygons. In contrast, rainfall polygons have a larger eccentricity (e) than earthquake polygons of an ellipse fitted to the polygon.

Rainfall polygons are more sinuous in shape leading to the smaller minor axis and larger major axis leading to the larger eccentricity of the ellipse fitted to the polygon (Rana et al., 2021).

4.4.2 Second method: topological features based classification

In the second method, we used the 3D shapes of landslides by incorporating the elevation data of the landslide regions. We extracted geometrical and topological properties of a landslides' 3D shapes using Topological Data Analysis (TDA) and then used these properties as a feature space for the machine learning algorithm—random forest (described in section 5.1 of the chapter 5). The topological properties of the landslide's 3D shape extracted using DEM provide additional insights into the landslide triggers, which might further improve the accuracy of the landslide trigger classification. We converted the 2D landslide polygons to 3D landslide polygons using interpolation of 30 meters' elevation data (DEM) around the bounding box of landslides. We took only the elevation data within the landslide polygons to preserve the geometric shape of the landslides (Figure 4.3). We explored various TDA features to quantify the 3D shapes of landslides using the Python library, `giotto-tda` (Tauzin et al., 2021). Using random forest feature importance analysis, we selected the top ten most relevant features, as irrelevant features increase the complexity of the model and are ineffective in improving the classification results. These selected relevant features constitute the input variables for the random forest classifier.

Topological Data Analysis (TDA) provides a gamut of metrics to quantify the multidimensional shape of data by applying techniques of algebraic topology (Carlsson, 2009). These metrics could also serve as a feature space for machine learning algorithms to solve classification problems, e.g., the classification of manifolds or complex geometric shapes. The central idea of TDA is persistent homology that identifies persistent geometric features in the data; it uses simplicial complexes to extract topological features from the point cloud data. A simplicial complex is a collection of simplexes and building blocks of higher dimensional counterparts of a graph. For example, a point is a 0-dimensional simplex, an edge which is a connection between two points is 1-dimensional simplex, a filled triangle formed by connecting three non-linear points is a 2-dimensional simplex. In general, an n-dimensional simplex is formed by

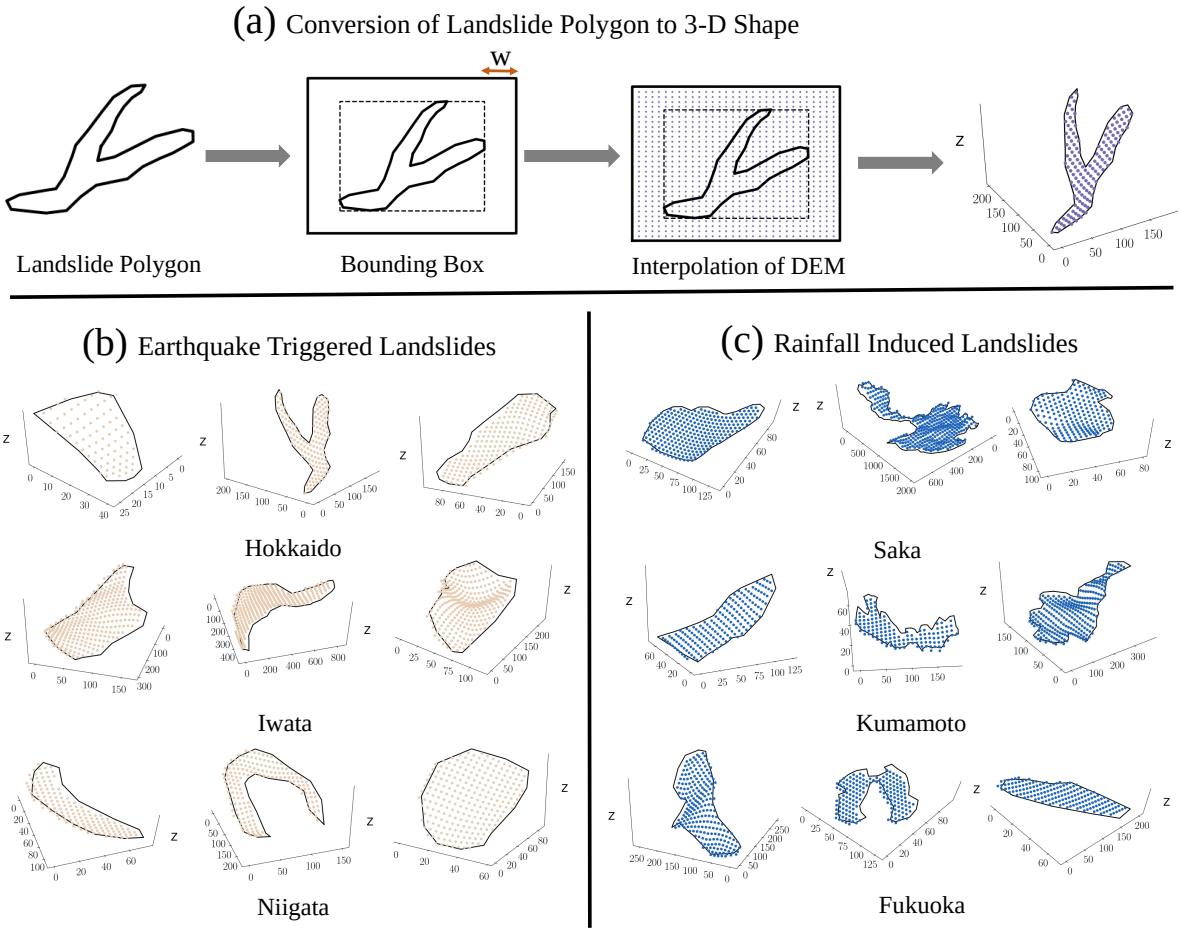

Figure 4.3: Sample $3D$ landslides from six known triggered inventories, (a) flow chart of conversion of $2D$ landslide planforms to $3D$ landslide shape. (b) Earthquake triggered $3D$ landslide samples, (c) rainfall induced landslide $3D$ samples. The $2D$ landslide planforms converted to $3D$ landslide shapes by using the elevation of landslides through a Digital Elevation Model (DEM).

connecting n+1 affinely independent points (Munch, 2017; Garin and Tauzin, 2019).

Generally, in TDA, one constructs a simplicial complex by the Vietoris-Rips complex method, where one chooses a parameter $\epsilon > 0$ to find the structure present in the data. For each pair of points (x, y) in the point cloud data, add an edge between x and y if euclidean distance (d) between x and y is less than ϵ. For a n-dimensional simplex, distance between each pair of $n+1$ affinely independent points should be less than ϵ $(d(x, y) < \epsilon)$. Each value of ϵ provides a set of simplexes representing a data structure. Different values of ϵ could lead to a different structure in data. To get the complete information about the structures present in

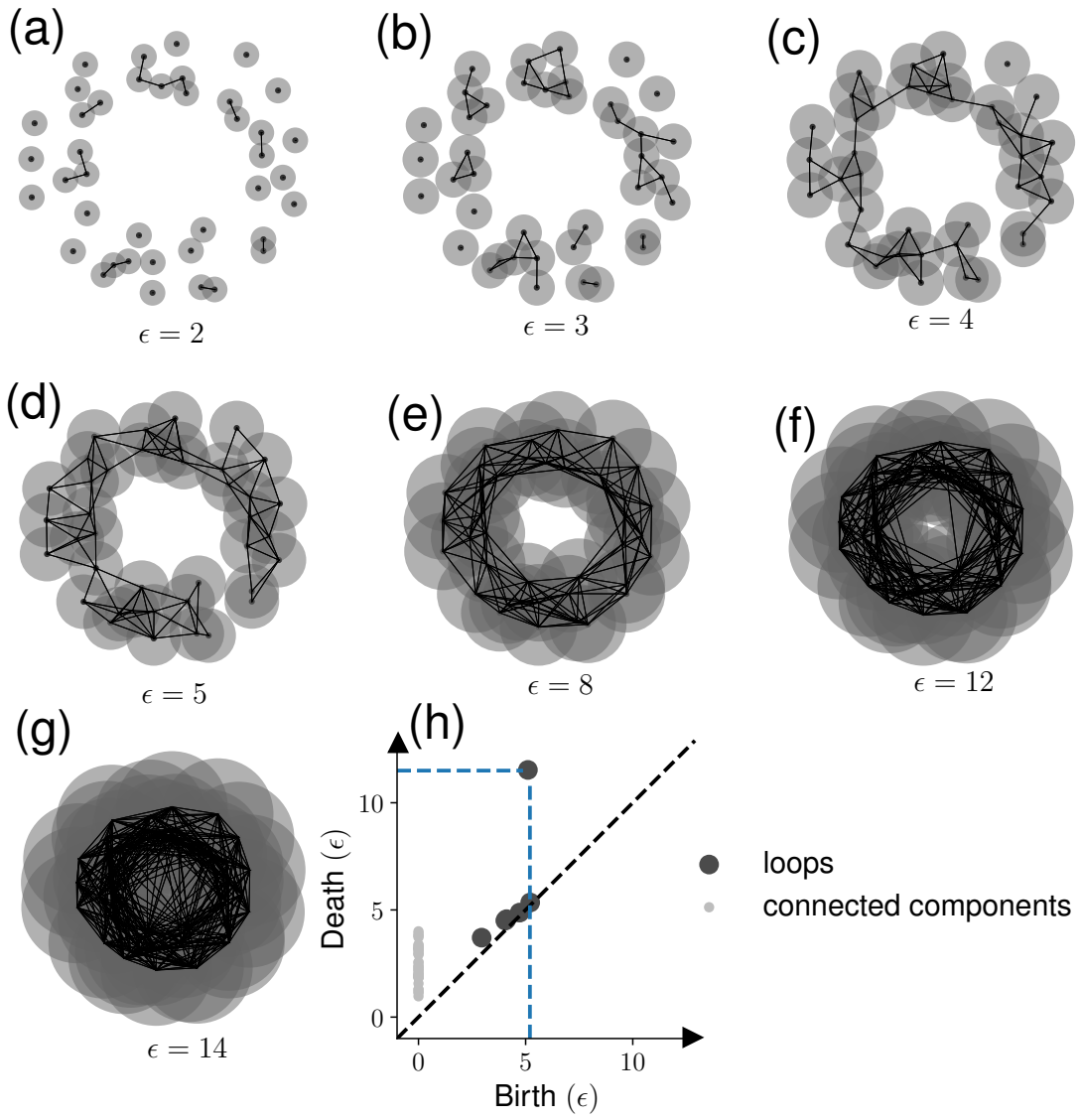

Figure 4.4: An example of using persistence homology: the data points are sampled from a noisy circle. (a)-(g) As the disk's radius increases ($\frac{\epsilon}{2}$), persistence homology captures various structures in the data. (h) The origin (birth) and disappearance (death) of loops and connected components is shown in the persistence diagram. The biggest loop in the noisy circle data is captured by the data points shown with the blue dotted line in (h).

the data, all the possible values of ϵ are used, creating a sequence of simplicial complex (this process is called filtration, Figure 4.4a-g).

Homology measures particular structures present in the data providing valuable information about the geometrical and topological properties of the data. For example 0-dimensional

homology captures connected components or clusters, 1-dimensional homology measures loops, 2-dimensional homology measures voids (Munch, 2017; Hensel et al., 2021). Structures like connected components, holes, and voids originate (birth) and disappear (death) with a change in the value of ϵ. A persistence diagram, shown in Figure 4.4(h); documents the birth and death information of these structures. Using the birth and death information of clusters, holes, and voids present in the persistence diagram, we can calculate several topological features of the data. We used various topological features to quantify the shape of data such as persistence entropy, average lifetime, number of points, betti curve-based measure, persistence landscape curve-based measure, Wasserstein amplitude, Bottleneck amplitude, Heat kernel-based measure, and landscape image-based measure. Each topological metric considers different homology dimensions separately.

The above mentioned topological features can be explained using two objects, one the set of $\{(b_i, d_i)\}_{i=1}^{i=N}$ birth-death pair in the persistence diagram; where i and N are the birth-death pair index and the total number of birth-death pairs respectively, and two the elements of lifetime vector $[l_i]_{i=1}^{i=N}$, calculated as difference between death and life of (b_i, d_i) pair ($l_i = d_i - b_i$). Then the number of points is the length of the lifetime vector, whereas Wasserstein and Bottleneck amplitudes are p-norm and ∞-norm of lifetime vector, respectively. Average lifetime and persistence entropy are average and Shannon-entropy of lifetime vector.

Betti and persistence landscape curves based features are calculated from p-norm of discretized betti and persistence landscape curves. Betti curve is a function $B(\epsilon)$ that maps persistence diagram to an integer-valued curve, $B(\epsilon) : \mathbb{R} \to \mathbb{Z}$, it counts the number of (birth, death) pairs at ϵ that satisfy the condition $b_i < \epsilon < d_i$ (Garin and Tauzin, 2019). Whereas, persistence landscape curve is a function $\lambda(k, \epsilon) : \mathbb{R} \to \mathbb{R}_+$, where $\lambda(k, \epsilon) = k_{max}\{f_{b_i,d_i}(\epsilon)\}_{i=1}^{i=n}$, k_{max} is k-th largest value of set of functions defined by $f_{b_i,d_i}(\epsilon) = max\{0, min(\epsilon - b_i, d_i - \epsilon)\}$ for each (b_i, d_i) pair (Bubenik and Dłotko, 2017).

The heat kernel-based feature is calculated using the p-norm of the 2D function discretization obtained using the heat kernel on the persistence diagram. Heat kernel transforms the persistence diagram to a function on $\mathbb{R}^2$ obtained by placing a Gaussian kernel with standard

deviation σ to each (birth, death) pair and negative of Gaussian kernel with same standard deviation in the mirror image of (birth, death) pairs across the diagonal (Reininghaus et al., 2015). Whereas persistence image-based measure is calculated using the p-norm of 2D function discretization obtained using the weighted Gaussian kernel on the birth-persistence diagram. Weighted Gaussian kernel transforms birth-persistence diagram to a function on $\mathbb{R}^2$ obtained by placing a weighted Gaussian kernel with standard deviation σ to each (birth, death - birth) pair in birth-persistence diagram (Adams et al., 2017). In the birth-persistence diagram, the y-axis represents the lifetime (death-birth) information of each (birth, death) pair.

4.4.3 Third method: image based classification

In the third method, we used landslide planform images as input to Convolutional Neural Networks (CNN) for the classification. We converted landslide polygons into binary images in a way that preserves the relative shape and structure of the polygons (Figure 4.5). Then using CNN for landslide triggers classification is straightforward via a simple CNN architecture with 3 convolutional layers and 2 fully connected layers. The input to CNN is a 64×64 binary pixel image, and the output is the probability of the input image belonging to one of the landslide trigger classes.

Convolutional Neural Networks (CNNs) are a class of artificial neural networks that are effective for various applications, such as image classification and object detection (Li et al., 2014; Guo et al., 2017; Albawi et al., 2017). The CNN architecture for classification problems consists of the input, hidden, and output layers (as shown in Figure 4.6). The input layer consists of the input data to CNN, an image of a landslide polygon in our application. The hidden layer primarily contains convolutional layers, max-pooling, and fully connected layers. Finally, the output layer provides the probability of input data belonging to an output class—rainfall-induced or coseismic.

Convolutional layers are the fundamental component of CNN that uses kernels (matrix of learnable parameters) to perform convolutions operations on the input. The resulting output of the convolution operation is called a feature map that learns the feature representation of the input data (Yamashita et al., 2018). Each neuron in a feature map captures the antecedent

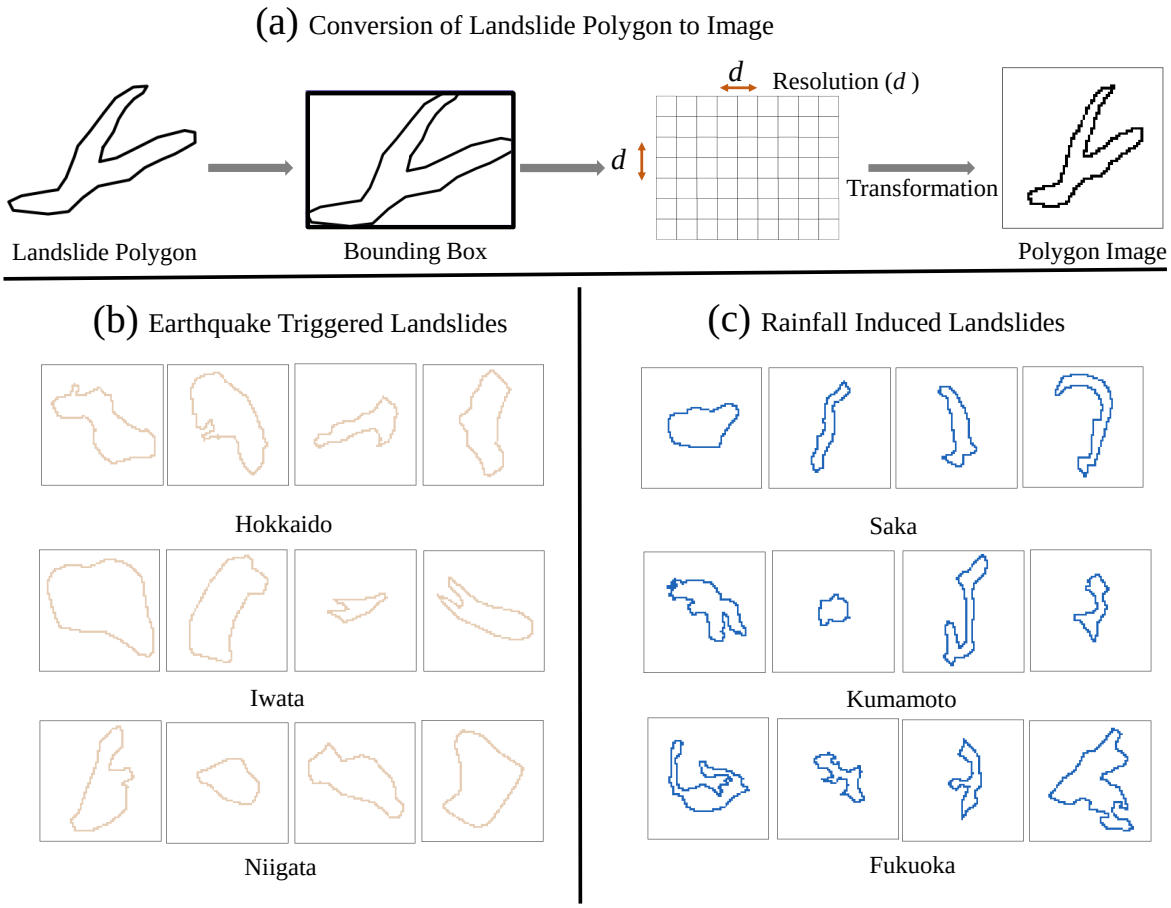

Figure 4.5: Sample input images for the image-based classification. (a) Flow chart of converting landslide planforms to a landslide polygon image. (b) Earthquake-triggered landslide image samples. (c) Rainfall-induced landslide image samples.

layer's local characteristics by convolution of kernels with the previous layer's feature maps (Guo et al., 2017). However, increasing convolutional layers could lead to over-parametrization and increase model complexity and, thus, over-fitting. One of the ways to avoid the issue is to use pooling layers that reduce feature maps dimension and the number of neurons in the output layer of CNN's (Yamashita et al., 2018; Guo et al., 2017). We used max-pooling layers of $n \times n$ ($n = 2$) size that takes a patch of size $n \times n$ from a feature map and produces one-value corresponding to that patch, and the pooling layer itself is free from parameters (Li et al., 2014).

Activation functions in CNN's capture the non-linear relationship between the input data

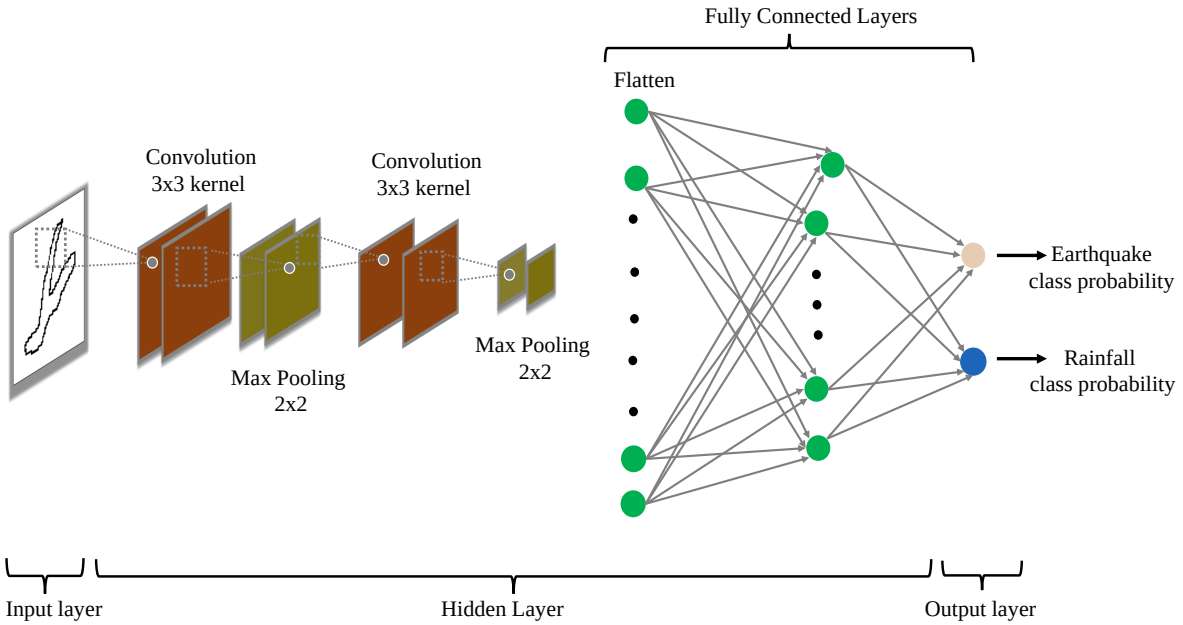

Figure 4.6: The figure shows the Convolutions Neural Network (CNN) architecture used in the image-based method. The input of CNN is a binary scale landslide image, and the output of CNN is the probability of a landslide image belonging to an earthquake or rainfall-induced class.

and its output class. We used ReLu for the hidden layer neurons activation functions as past studies have proved that ReLu improves classification results and learning speed (Li et al., 2014; Krizhevsky et al., 2012). The output of ReLu activation function is $f(x) = max(0, x)$, here x means the output of a neuron (Li et al., 2014). For the output layer, we used the softmax activation function. The softmax activation function calculates the output probabilities of the input sample belonging to each class in the last layer of CNN. The class probabilities are calculated as

$$P_i = \frac{\exp z_i}{\sum_{j=1}^{j=m} \exp z_j}, \tag{4.4.1}$$

where z_i is the output from last layer of CNN corresponding to i class and m is the number of classes (in our case, $m = 2$).

Fully connected layers (FCC) work as a classification layer for CNNs and comes after the convolutional layers. All layers in FCC are fully connected which means each neuron in a layer is connected to every neuron in the next layer of FCC (Albawi et al., 2017; Guo et al.,

2017). In classification problems, the last layer of the FCC layer gives the probabilities of the input image to belong to one of the output classes with the help of the softmax activation function (Eq. 4.4.1). The output predicted probabilities of the input sample are used in a loss function that evaluates how well the model works for classifying the class of the input image dataset. We used the cross-entropy loss function that measures the difference between actual and predicted probability distribution. The cross-entropy loss function for a sample is defined as: $-\sum_{i=1}^{i=m} y_i log(\hat{y}_i)$, where m is the total number of classes, y_i ($\hat{y}_i$) is actual (predicted) probability corresponding to class i. If i is actual class of the input sample then $y_i = 1$, otherwise $y_i = 0$. In the case of binary classification $m = 2$. The sample's output probabilities are a function of parameters used in convolution kernels and FCC layers to connect neurons in one layer to the next layer. These parameters are altered iteratively using the back-propagation algorithm and stochastic gradient method to increase the probability of samples belonging to the actual class and thus, minimize the loss (Aurisano et al., 2016).

4.5 Landsifier model evaluation

We used two different testing configurations to evaluate the efficacy of our methods. Finding the triggers of individual landslides irrespective of their inventories is the first testing configuration. Here, we combined all the known trigger landslides from all six known triggered inventories and then split the combined landslides data into various training and testing sets following the k-fold cross-validation framework. In this testing configuration, landslides in each training and testing set are from all six landslide inventories. The second testing configuration finds the trigger of landslide inventories itself. We used all the possible combinations to train the algorithm on five known trigger inventories and test it on the sixth inventory. In this second testing configuration, landslides in the testing set are from a single inventory. Note that there are seven inventories in the analyzed data set, and six have known triggers. The analysis of this seventh inventory (Kumamoto unspecified) with unknown triggers is presented in the section 4.7.

4.5.1 Evaluation of the first method (geometric features based classification)

Combining all the landslide inventories with known triggers lead to $26,501$ samples (n_{total}), out of which $16,196$ are earthquake-triggered landslides ($n_{earthquake}$) and $10,305$ are rainfall-induced landslides ($n_{rainfall}$). As the number of earthquake-triggered landslides is much larger than the number of rainfall-induced landslides, we use equal numbers of each trigger class to avoid any class-imbalance problems. To avoid selection bias and overfitting, we apply 10-fold cross-validation. k-fold cross-validation splits the combined classes dataset into k random subsets where each iteration of cross-validation, $k-1$ folds are used for training and the remaining fold for testing. We use $20,610$ samples ($n_{rainfall} = n_{earthquake} = 10,305$) for cross-validation and to get generalizable results we employ 1000 runs of cross-validation. In each run of cross-validation we randomly select $10,305$ earthquake samples from $16,196$ earthquake landslides. We achieved $86.15 \pm 0.22\%$ classification accuracy for earthquake, $85.29 \pm 0.19\%$ for rainfall, and $85.73 \pm 0.16\%$ as the mean classification accuracy.

For the second split configuration, we trained the random forest classifier on five inventories and tested it on the sixth inventory. For earthquake triggered inventories the method achieved classification accuracy of $66.62 \pm 0.65\%$, $75.59 \pm 0.34\%$ and $85.22 \pm 0.20\%$ for the Hokkaido ($n_{train} = 20,610, n_{test} = 3,256$), Iwata ($n_{train} = 20,610, n_{test} = 4,160$) and Niigata ($n_{train} = 14,832, n_{test} = 8,780$) inventories (for geographical locations of these inventories see Figure 4.1). For rainfall induced inventories, we achieved classification accuracy of $83.63 \pm 0.41\%$, $69.40 \pm 0.61\%$ and $92.12 \pm 0.25\%$ for Kumamoto ($n_{train} = 9,482, n_{test} = 5,564$), Fukuoka($n_{train} = 16,762, n_{test} = 1,924$) and Saka ($n_{train} = 14,946, n_{test} = 2,817$) region. In each one of the the case we took equal number of earthquake and rainfall triggered landslide samples to avoid any class imbalance issues ($n_{earthquake} = n_{rainfall}$). The low standard deviation in classification accuracy shows that results are stable with change in training samples.

4.5.2 Evaluation of the second method (topological features based classification)

In the first test and training set split configuration, as in section 4.5.1, we used n_{total}= 20,610 (total number of samples), $n_{earthquake}$=10,305 (number of earthquake-triggered samples) and $n_{rainfall}$ =10,305 (number of rainfall-induced samples), keeping numbers of each trigger class equal to avoid class imbalance. We first identify the top ten relevant topological features out of thirty features, employing 1000 runs of 10-fold cross-validation of random forest. Using these top ten relevant topological features as the feature space for the random forest classifier, we carry out 1000 runs of 10-fold cross-validation to get generalized classification accuracy. The method achieved above 94% classification accuracy for earthquake, rainfall, and mean class classification.

In the second split configuration, this method achieves above 90% accuracy for the Iwata, Niigata, Kumamoto, and Saka inventories. For the Hokkaido and Fukuoka region, the method achieves above 80% classification accuracy (see Figure 4.7). The number of training and testing samples for each case is the same as in section 4.5.1.

4.5.3 Evaluation of the third method (image based classification)

As explained above in section 4.4.3 we removed large landslides from the analysis leading to $n_{total} = 24,311$, $n_{earthquake} = 14,892$, and $n_{rainfall} = 9,419$. We used an equal number of training samples of the coseismic and rainfall-induced landslides to avoid any class imbalance issues. We used 100 runs of different test and training sets instead of different runs of 10 fold cross-validation as convolutional neural networks are computationally expensive. The method achieved above 85% classification accuracy for earthquake, rainfall, and mean class classification.

For the second split configuration, the method achieved above 80% accuracy for the Saka region ($n_{train} = 13,738, n_{test} = 2,550$). For the Niigata ($n_{train} = 12,780, n_{test} = 8,502$), Kumamoto ($n_{train} = 8,276, n_{test} = 5,281$) and Fukuoka ($n_{train} = 15,662, n_{test} = 1,588$) region the method achieves accuracy of above 70%. The Method achieves 67% accuracy for the

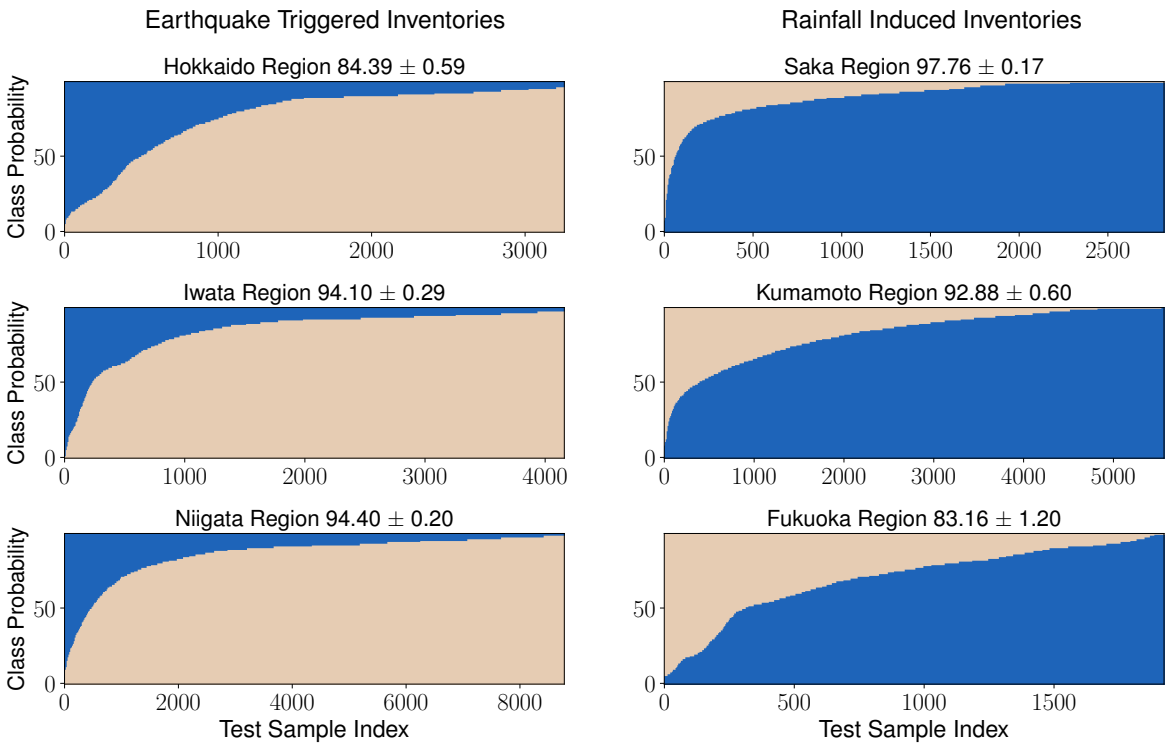

Figure 4.7: The topological features-based method (second method) accuracies for all the six known triggered inventories. The model is trained on five inventories in each case and tested on the sixth inventory. The y-axis in the plot shows the probability of landslides belonging to the earthquake and rainfall class, and the x-axis shows the sample index of landslides.

Hokkaido inventory ($n_{train} = 18,838, n_{test} = 2,431$). In each one of the the cases, we took equal number of earthquake and rainfall induced landslide samples to avoid class imbalance issues.

4.6 `Landsifier` library

One of the main aims of this paper is to introduce `Landsifier`, a Python library we built to provide the landslide research community with a user-friendly computational package to implement the methods described above. At the moment, we have made the code available on the corresponding author's GitHub: `https://github.com/kamalrana7843/Landsifier.git`. Furthermore, we published the `Landsifier` library under an open-source license in a way

that it is accessible via the python terminal using the import command. In section 5.2 of the chapter 5 we provide details of the library and brief descriptions of the available functionalities. Apart from three different methods for landslide trigger classification, the library also contains other useful functions like calculating geometric properties of landslide polygons, converting polygons to binary scale images, downloading DEM corresponding to an inventory region, and converting 2D landslide polygon to 3D landslide shape (see Figure 3a and 5a). Please refer to section 5.2 of the chapter 5 for further details about the library functions (Figure 5.1). Each of the three methods used in the library is simple to use and only requires polygon shapefiles as input. Also, the computation process is relatively fast; for example, the geometric, image, and topological features-based method takes less than 5, 15, and 45 minutes for training on 20,000 landslides (equal earthquake and rainfall samples) on a windows machine with 16 GB of RAM (Random-Access Memory) using only landslide shapefiles as input. Moreover, none of the methods requires a GPU (Graphics Processing Unit).

4.7 Discussion

The geometric properties of landslides can provide information about their trigger (Taylor et al., 2018a). Our preliminary work on landslide trigger classification demonstrated that landslides with identical triggers share similar geometric properties, which could be exploited to classify landslide triggers—see the publication (Rana et al., 2021) and briefly reproduced results here in sections 4.4.1, 4.5.1 and chapter 5 section 5.1. In this work, we further expanded our initial approach by adding two additional methods for landslide triggers classification and a Python library `Landsifier` to implement them. One of these two new methods uses 3D shapes of landslides for their trigger classification by incorporating the elevation information. We compute topological features of these 3D shapes using Topological Data Analysis (TDA) and use the features as an input to a machine learning-based algorithm—random forest. The other method uses binary scale landslide polygon images as an input to Convolutional Neural Networks (CNN) for the classification. Using six landslide inventories spread over the Japanese archipelago, we showed that each method exhibit strong performance to classify landslide

triggers. However, each method has its strengths and limitations that primarily depend on training and testing landslide data quality, quantity, and location. We explained each methods' strengths and limitations in different conditions in this section. Before providing some hints about potential future work and opportunities that could arise from using `Landsifier` library, here, we also present and discuss the results of each three method on the seventh Kumamoto unspecified inventory.

The landslide data quality depends on the data acquiring technique; e.g., landslide data obtained using aerial or satellite images are much higher quality than the data acquired via field campaigns. Geologists collect landslide data acquired via field campaigns, and by nature, such inventories tend to fail to represent the smaller landslides and cover the larger landslides (Ozturk et al., 2020). Whereas landslides inventories acquired via aerial or satellite images cover both small and larger landslides and are called complete inventories as they adequately capture landslides of various sizes in their respective study area, e.g., see (Schmitt et al., 2018b). The performance of developed methods depends on landslides data quality and without similar data quality in training and testing set the accuracy of classification techniques could be insufficient to conclude the trigger of landslide inventory and also might lead to biases. Training the geometric feature-based and image-based methods on landslide planforms with landslide data acquired via satellite or aerial images and testing on data acquired via field campaign or vice-versa could lead to biases in landslide classification results. The methods based on landslide planforms shape consider the area and perimeter as the most important features and rely on the information that coseismic landslides are generally larger than rainfall-induced landslides (Rana et al., 2021) (e.g., Taylor et al., 2018; Tanyas et al., 2021). So, a testing inventory triggered by rainfall but lacks smaller landslides due to field campaign acquisition technique could be classified as earthquake-triggered—given that training inventories are satellite or aerial image-based. We recommend using similar field campaign acquired inventories with known triggers to train the models for more accurate classification in such a scenario. Another option is to sample landslides from the satellite or aerial image-based inventories that resemble the size distribution of the testing data acquired via field campaign. This

shortcoming motivated us to offer another alternative solution relying on topological analysis of 3D shapes of landslides.

Landslides are 3D shapes; thus, using 3D shapes of landslides instead of 2D could provide additional information related to the landslide morphology. Consequently, a 3D landslide shape-based method might elevate classification accuracy, especially in regions without proper training and testing data of similar quality. We use TDA, a method rooted in algebraic topology, to compute topological features of a landslides' 3D shapes to classify landslide triggers. TDA based method extracts topological information along with geometric information of landslide shape. Whereas, geometric features based method and likely Image based method use only geometric information of the landslide shape for landslide classification. We expect TDA based method will provide best landslide trigger classification results. In Table 4.1 one can observe that the TDA-based method indeed performs better than the other two methods. However, TDA-based measures encode landslide morphology; hence, if testing and training inventories share similarities in the geomorphology of the studied regions (spatial autocorrelation) (Oksanen and Sarjakoski, 2005), then the trigger prediction is highly influenced by training inventory. Geometric features and image-based methods are less sensitive to the geomorphological similarities between the training and testing landslide inventories, as these only use the 2D landslide planforms. Although the image-based performs satisfactorily only when adequate large training data is available. Hence, we recommend using geometric or topological features-based methods in inventories with limited landslide counts.

We applied each method to classify landslides triggers in the Kumamoto unspecified inventory having an undocumented trigger to demonstrate the real-world application of the `Landsifier` library. Out of 612 landslides in the inventory, the geometric feature-based and topological feature-based classified 604 and 612 landslides as earthquake-triggered. In comparison, the image-based method uses 164 landslides after removing landslides having width and length greater than 180 meters (see section 4.4.3 for more details) and classified all of the landslides as seismically triggered. As each method classifies the majority of the landslides as earthquake-triggered, we are confident that earthquake is the most likely trigger for most of

Inventory Region	Geometric features based method (%)	Topological features based method (%)	Image based method (%)
Hokkaido	67	**84**	68
Iwata	76	**94**	67
Niigata	85	**94**	77
Saka	92	**98**	88
Kumamoto	84	**92**	78
Fukuoka	69	**83**	70

Table 4.1: The table shows landslide classification results using the three methods. The model is evaluated on all possible training set combinations of the five inventories and tested on the sixth inventory.

the landslides in this inventory. Moreover, the Kumamoto unspecified inventory documents landslides along the rims of the Aso Caldera, the active volcano Mount Aso shakes the surrounding area frequently triggering landslides within its vicinity (Saito et al., 2018). Hence, it is very likely that this inventory is consistent of landslides of cosesimic origin.

Considering the above discussions, in future work, we plan to explore further the sensitivity of our trigger classification methods to spatial autocorrelations. We will also examine the influences of landslide size distributions on each method. Specifically, we plan to classify the trigger of large landslides (Area $> 90,000$ square-meters) as they are the most dangerous landslides and effect huge area by training each method on large landslides training dataset.Moreover, we will consider model transferability to different regions by extensively testing these methods on national landslide inventories, e.g., India, Nepal, Taiwan, and the USA. Our methods could also provide other opportunities. For example, assessing landslide-prone regions as an alternative to landslide susceptibility measure using TDA. Also, TDA could be used to classify landslide types, according to the types described in (Cruden and Varnes, 1996) and (Varnes, 1978). Landslide type information plays a crucial role in landslide risk assessment which is usually missed in landslide databases (Loche et al., 2022). We plan to further develop the current version of the `Landsifier` by incorporating a landslide type classifier in the next version (Amato et al., 2021). This method will be able to find the analogy between an observed landslide and a generic landslide type by (Cruden and Varnes, 1996).

4.8 conclusions

The landslide triggering mechanism is crucial information to develop landslide hazard models, e.g., a landslide hazard model for extreme rainfall incidents requires landslide inventories related to rainfall events only. However, modern automated landslide mappers for continuous monitoring and historical landslide inventories rarely report the landslide triggering mechanism. Missing triggers in the landslide inventories decrease their efficacy for landslide hazard models. In this work, we developed a Python library, `Landsifier`, containing three methods for landslide trigger classification by exploiting landslide planforms and 3D shapes. To develop the first two of these methods, we combined geometric and topological features with machine learning, and in the third method, we used deep learning. The latter two methods are new, i.e., we are reporting them here for the first time.

We use seven landslide inventories spread over the Japanese archipelago. Six among these seven inventories have known triggers, while the seventh inventory has a missing trigger. We applied each method to all possible sets of five training inventories and one testing inventory using six known triggered inventories. Moreover, we took different training and testing sets of landslides by mixing all known triggered landslides inventories following the k-fold cross-validation. The achieved results demonstrate that the methods are robust and capable of classifying triggers of landslide inventories with high accuracy (70%–95%). To demonstrate the real-world application of our toolbox, we also applied the three methods to the seventh inventory without any triggering information and classified it as an earthquake-triggered inventory.

Python-based `Landsifier` library provides a user-friendly computational package to implement the methods described above to the landslide research community. Two of the three methods included in the library are new and introduced here for the first time, while the third method is published in our previous work. To our best of knowledge `Landsifier` is the first python tool developed for landslide triggers classification, and also such a tool does not exist in other programming languages. We anticipate that the landslide research community will find the `Landsifier` library helpful in finding the trigger mechanism of inventories or

individual landslides. The presented methods and the library could be deployed in any region of the world with adequate training data from areas with similar climatic and tectonic features. `Landsifier` library also contains useful functions like finding geometric properties of landslides polygon, downloading DEM corresponding to an inventory region, and converting landslide polygon to landslide 3D shape, these elements could be useful for the landslide research community."

Furthermore, methods in `Landsifier` library are easy to use as they require only shapefiles of landslide polygons as input. `Landsifier` is a modular software, we hope the landslide community will further improve the offered tool and expand the available functions for new applications such as classifying landslide types, assessing landslide-prone regions, and other possible usage are listed in the discussion section. At the moment, we have made the code available on the corresponding author's GitHub: `https://github.com/kamalrana7843/Landsifier.git`. Moreover, we published the `Landsifier` library under an open-source license in a way that it is accessible via the python terminal using the import command. In section S2 of the Supplementary information, we provide details of the library and brief descriptions of the available functionalities.

Chapter 5

Landsifier v1.0: a Python library to estimate likely triggers of mapped landslides

The work presented in this chapter is published in the **supplementary material** of the paper: **Rana, K.**, Malik, N. & Ozturk, U. (2022). Landsifier v1.0: a Python library to estimate likely triggers of mapped landslides. *Natural Hazards and Earth System Sciences*, 22(11), 3751-3764. This chapter also describes the landsifier library functions and their usage in detail. We developed this library as a software product as part of this thesis, for more information, see: https://pypi.org/project/landsifier/.

5.1 Random forest

Random forest (RF) is a decision-tree based ensemble-learning method, a proven and powerful technique for classification and regression (Barnett et al., 2019; Biau, 2012; Biau and Scornet, 2016; Breiman, 2001; Kursa, 2014; Chaudhary et al., 2016; Rodriguez-Galiano et al., 2012). The random forest classifier consists of multiple classifiers, where each classifier bootstraps the training data samples (Breiman, 2001; Liaw et al., 2002). Bootstrapping in each random forest classifier is done by selecting N samples randomly from training samples of size N with replacements. For N training samples bootstrapping N times leads to the approximate

selection of 2/3 of training samples (Azar et al., 2014; Belgiu and Drăguţ, 2016). Hence, each tree in a random forest classifier is trained independently using around 2/3 of the training samples selected using bootstrapping.

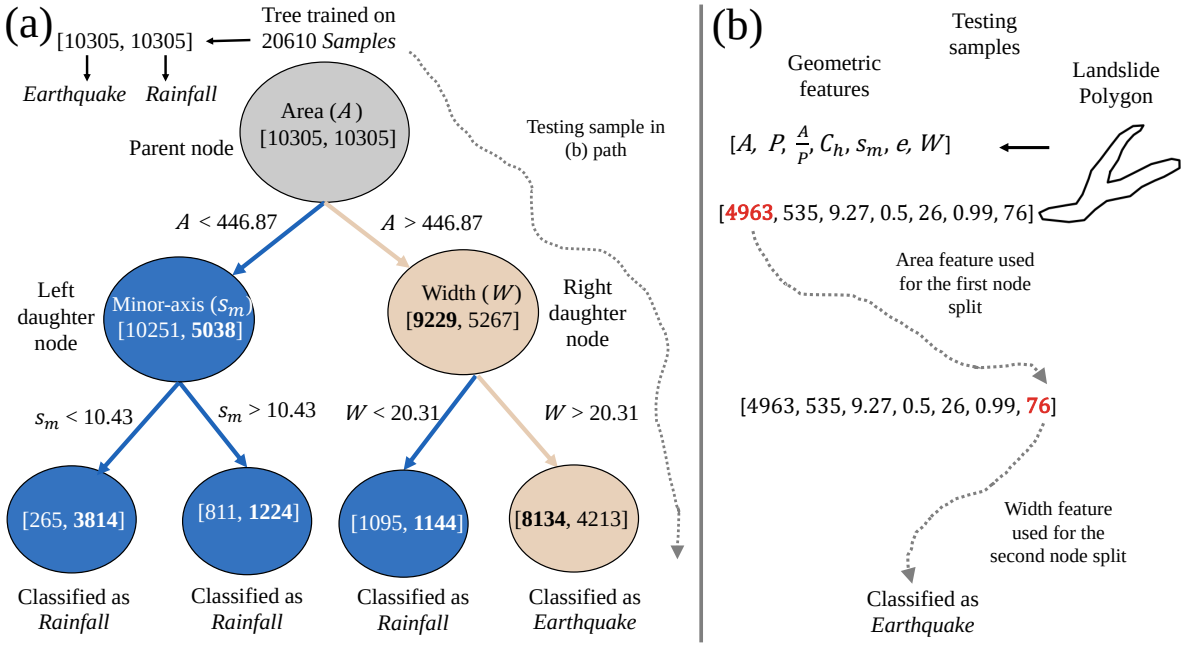

Figure 5.1: (a) The sample architecture of one of the trees of random forest classifier. The tree is trained on 20610 landslides samples with 10305 each earthquake and rainfall trigger class. Feature vector ($[A, P, C_h, W, s_m, \frac{A}{P}, e]$) represents landslide geometric property corresponding to each landslide sample. For illustration purposes, the tree is grown to only depth three. (b) testing sample of landslide tested on the tree shown in (a). The sample landslide polygon is classified as an earthquake.

In a binary classifier, as in our case, each parent node q splits into two daughter nodes: right r and left l. Instead of selecting all the p features for node split, a subset of features m (m=$\sqrt{p}$) is selected randomly for each node split (Azar et al., 2014; Okun and Priisalu, 2007). Among m features, one of the features selected for the node split is based on optimizing a criterion. The criterion is called the 'Gini Index,' which measures the features' impurity to

the classes. The Gini index of right r and left l daughter nodes are calculated as:

$$G_r = 1 - P_{r1}^2 - P_{r2}^2 \tag{5.1.1}$$

$$G_l = 1 - P_{l1}^2 - P_{l2}^2, \tag{5.1.2}$$

where P_{rj} (P_{lj}) is the probability of samples in the right (left) daughter nodes having class j. The Gini index is calculated for each predictor in the subset of predictors m, and the features that maximize the change in Gini index is chosen for node split. Change in Gini-index is calculated as:

$$\Delta\theta(s_q) = G_q - \rho_{rq}G_r - \rho_{lq}G_l, \tag{5.1.3}$$

where ρ_{rq} (ρ_{lq}) are the ratio of the number of data points in daughter nodes r (l) to the total number of points in the parent node q (Kuhn et al., 2013; Zhang and Ma, 2012b). The process of splitting nodes continues until a stopping criterion is met, e.g., when no further samples are remaining, or the Gini-index of parent nodes is lower than the daughter nodes.

The steps for constructing trees in the random forest are as follows:

(i) Select bootstrap samples of size N from training samples of size N.

(ii) Randomly select m variables among p variables for the node split.

(iii) Choose one variable among m variables that best split the node according to the Gini-index criterion.

(iv) Continue repeating steps (i) to (iii) until the stopping criterion is met.

For testing, each tree classifier predicts the class of testing sample independently, and the class with majority votes is the final outcome of random forest (Kuhn and Johnson, 2013; Pal, 2005; Arabameri et al., 2021; Belgiu and Drăguţ, 2016).

In random forest, bootstrapping training samples selection and random selection of features for a node split reduces the correlation between trees. This technique has proven to improve the predictive power of ensemble learning (Azar et al., 2014). In addition, random forest assigns each feature a score that provides its relative importance (Qi, 2012; Friedman et al.,

2001). Features with low relative scores should be discarded as they are neutral to the model accuracy and increase the model complexity.

5.2 Details of `Landsifier` library

`Landsifier` is a Python library we built with version 3.6 of Python and the code is available on GitHub: `https://github.com/kamalrana7843/Landsifier.git` (we published the Landsifier library under an open-source license in a way that it is accessible via the python terminal using the import command). On this link, prospective users can also find the list of Python packages used in the library. `Landsifier` contains three methods for landslide trigger classification, and these methods only use shapefiles from landslide inventories (two of these methods use 2D polygon shapes of landslides, while the third method uses the 3D shapes of landslides). This section describes various functions provided in the `Landsifier` library to implement the above methods and Figure 5.2 summarizes these functions in form of a flowchart. Also, Figure 5.3 shows a sample output of the `Landsifier`.

5.2.1 Functions for geometric features based classification

Below we list functions to implement the geometric features-based classification, details of the method can be found above in section 3.1 of the main paper and in our publication (Rana et al., 2021). Note below we have described functions in a form that this method can be used for inventories with unspecified triggers, i.e., unknown ground truth.

`latlon_to_eastnorth(latlon_polydata)`: This function takes polygons data in $(longitude, latitude)$ coordinates as an input and provides polygon data in $(easting, northing)$ coordinates as output to the function. This function is used to get landslide polygons in $(easting, northing)$ coordinates when polygon data in shapely files are in $(longitude, latitude)$ coordinates.

`Calculate_geometric_properties (polygon_shapefile)`: As the name suggests, this function calculates the geometric properties of each of the landslide polygons present in shapefile.

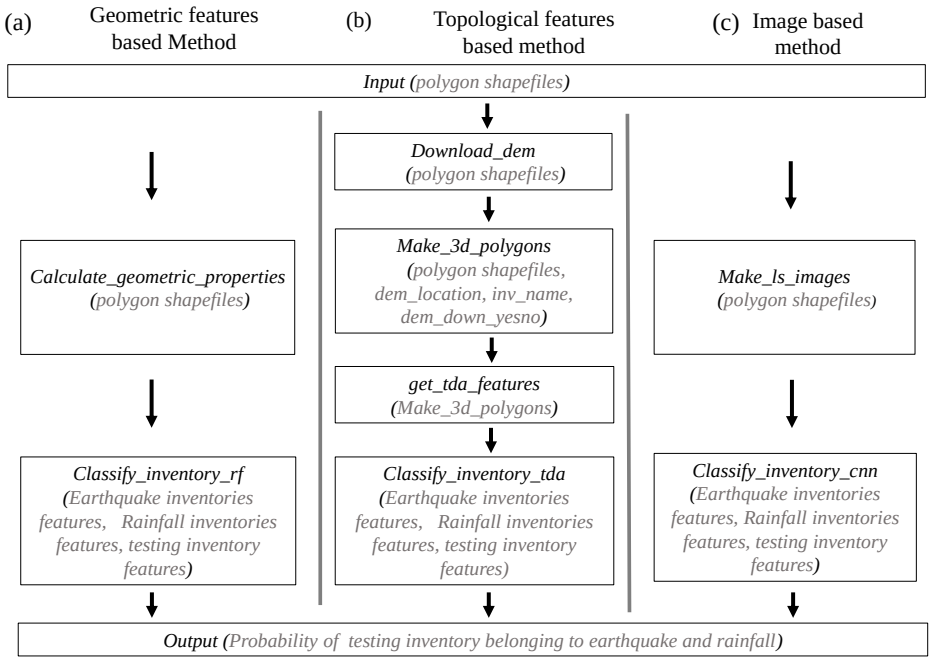

Figure 5.2: The figure shows the flowchart of implementations of all the three methods using functions and their variables used in the `landsifier` library. All three models use polygon shapefiles as an input to the model and provide the probability of landslide belonging to earthquake and rainfall as an output (a) geometric features based method (b) topological features based method (c) image-based method.

This function takes polygons shapefiles

(`polygon_shapefile`) as input, converts polygon data into $(easting, northing)$ coordinates if required using the

`latlon_to_eastnorth` function, and then provides the geometric properties of polygons as output to the function. For each landslide polygon it calculates a vector$([A, P, C_h, W, s_m, \frac{A}{P}, e])$ containing polygon geometric properties as output to the function. All the geometric properties of the landslide polygon are calculated using the *shapely* package in Python.

`classify_inventory_rf (earthquake_inventory_features, rainfall_inventory_features, test_inventory_features)`: This function takes the earthquake-triggered inventories $(earthquake\_inventory\_features)$, rainfall-induced inventories $(rainfall\_inventory\_features)$ and testing inventories $(test\_inventory\_features)$ geometric features as the input. Within

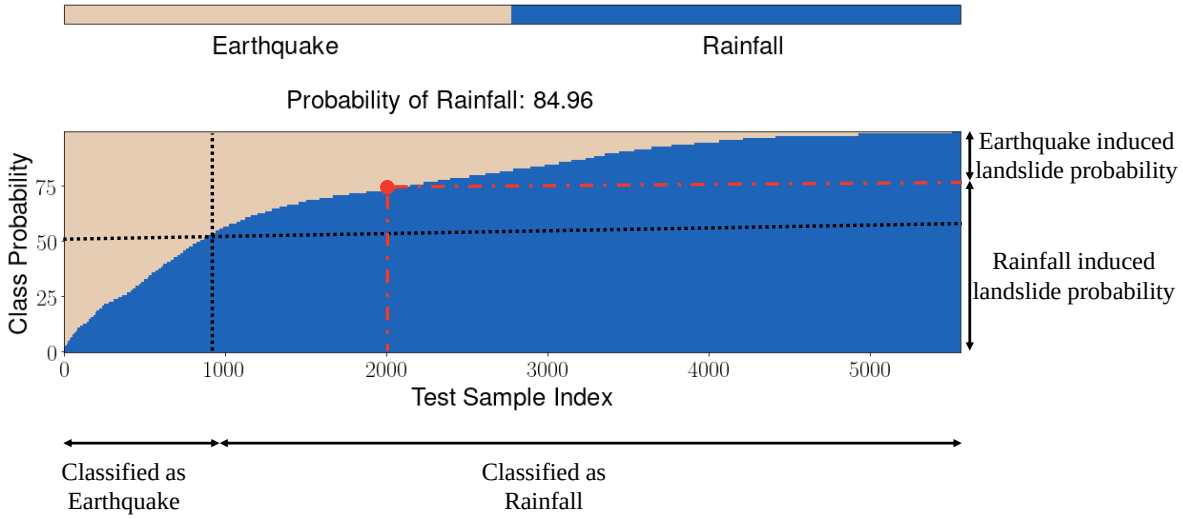

Figure 5.3: The output of the geometric feature-based method when Kumamoto inventory is testing inventory and the rest five inventories are used as training inventories. Each method in the `landsifier` will produce similar outputs. The y-axis in the plot shows the probability of landslides belonging to the earthquake and rainfall class, and the x-axis shows the sample index of landslides. For each landslide in the testing inventory, all the models give a probability of landslides belonging to earthquake and rainfall-induced classes. The predicted trigger of most of the testing landslides is the probable trigger of the testing inventory.

the function, it trains the random forest algorithm on training data containing equal samples of the earthquake and rainfall-induced class. The output of the function is the probability of testing landslides belonging to each trigger class.

5.2.2 Functions for topological features based classification

Below we list functions to implement the topological features-based classification, details of the method can be found above in section 3.2 of the main paper. Note below we have described functions in a form that this method can be used for inventories with unspecified triggers, i.e., unknown ground truth.

`download_dem (polygon_shapefile)`: This function takes shapefile of landslide polygons as

an input and downloads the *Shuttle Radar Topography Mission* digital elevation model (DEM) of resolution 30 meters corresponding to inventory region (Farr et al., 2007). It takes the bounding box over the entire inventory location and calculates the `(minimum latitude, minimum longitude)` and `(maximum latitude, maximum longitude)`. Using the elevation package in Python, it downloads the DEM data of a region bounded by `minimum latitude, minimum longitude, maximum latitude, and maximum longitude` coordinates. This function downloads all the tiles (one tile constitutes $1^\circ \times 1^\circ$ region in both latitude and longitude) of the inventory region and combines all the tiles into one file corresponding to one inventory.

`make_3d_polygons (polygon_shapefile, dem_location, inv_name, dem_down_yesno )`: This function takes landslide polygon shapefiles (`polygon_shapefile`), DEM path location (`dem_location`), inventory name (`inv_name`) and Boolean parameter (`dem_down_yesno`) as input and provides 3D point cloud data of landslides as output. This function carries out several tasks. First, it downloads the DEM data corresponding to the whole inventory region in path location (`dem_location`) with inventory name (`inv_name`) using `download_dem` function if `dem_down_yesno` is True. If users already have DEM corresponding to inventory in path location (`dem_location`) with inventory name (`inv_name`) then (`dem_down_yesno`) is False. Then corresponding to each landslide polygon it interpolates the DEM data around the bounding box of the polygon. Using the shapely package, the function removes all the interpolated data outside the outline of the landslide polygon and takes elevation data only within the landslide.

`get_tda_features (three_d_data)`: This function takes the 3D shape of landslides point cloud data (`three_d_data`) as an input and provides machine learning features corresponding to each 3D landslides as an output to function. This function calculates the persistence diagram using Vietoris Rips persistence for each 3D landslide, and then using the persistence diagram, it calculates the following TDA metrics: persistence entropy, average lifetime, number of points, betti curve based measure, persistence landscape curve based measure, Wasserstein amplitude, Bottleneck amplitude, Heat kernel-based measure, and landscape image-based measure corre-

sponding to each homology dimension–0, 1, and 2. These TDA metrics are used as a feature space for the machine learning algorithms.

`classify_inventory_tda (earthquake_inventory_features, rainfall_inventory_features, test_inventory_features)`: This function takes training earthquake inventory's `(earthquake_inventory_features)`, training rainfall inventory's `(rainfall_topological_features)` and testing inventory's `(test_inventory_features)` TDA based features as input to function. Inside the function, it first selects the top 10 features with the highest feature importance using training data. It then combines an equal number of training earthquake and rainfall samples to avoid any class imbalance problem. It trains the random forest algorithm on training data and predicts the probability of testing landslides belonging to each trigger class.

5.2.3 Functions for image based classification

Below we list functions to implement the image-based classification, details of the method can be found above in section 3.3 of the main paper. Note below we have described functions in a form that this method can be used for inventories with unspecified triggers, i.e., unknown ground truth.

`increase_resolution_polygon (poly_data)`: This function takes a single polygon coordinates data (`poly_data`) in $(easting, northing)$ coordinates as input and increases the number of points between any two adjacent vertexes of the polygon within the function. This function is useful in creating smooth binary scale landslide polygon images. The output of the function is landslide polygons coordinates data having multiple points between the adjacent vertex of polygons.

`make_ls_images (polygon_shapefile)`: This function takes polygon shapefiles (`polygon_shapefile`) as an input and provides landslide polygon images as an output to the function. It creates $N \times N$ ($N = 64$ in our case) pixel image with binary values of 0 or 255 for each pixel. This func-

tion first increase the number of data points in polygons using `increase_resolution_polygon` function and then takes a bounding box of polygon and transforms the coordinates of polygons by subtracting polygon $(minimum\_easting, minimum\_northing)$ value from each point in the polygon. Then divide each point in polygon $(easting, northing)$ value by resolution of pixels (desired spatial distance between any two adjacent horizontally or vertically pixels) and convert them into nearest integers. Then for each pixel (x, y) the value of the pixel is 255 if there exists a point in the polygon with coordinates (x, y) otherwise the value of the pixel is 0. This function also removes those landslides having length and width of bounding box greater than 180 meters as the image of a polygon has some restrictions on maximum landslide polygon it can have (resolution of pixels (3 meters) $\times$ N = 192 meters).

`train_augment (train_data, train_label)`: This function takes input training landslides data `(train_data)` and training labels `(train_label)` as input. The main idea behind using `train_augment` function is to augment the training data as CNN is data extensive algorithm. It rotates each image by 90°, 180°, 270°and flip the image vertically and horizontally to increase the number of training samples. The output of the function is augmented training data and labels.

`classify_inventory_cnn (earthquake_inventory_images, rainfall_inventory_images, test_inventory_images)`: This function takes training earthquake inventory images `(earthquake_inventory_images)`, training rainfall inventory images `(rainfall_inventory_images)` and testing inventory images `(test_inventory_images)` as input to the function. Within the function, it combines an equal number of training earthquake and rainfall samples to avoid any class imbalance problem and then augments the training data by using the `train_augment` function. Then it trains the CNN algorithm on augmented training data and predicts the probability of testing landslides belonging to each of the trigger classes.

Chapter 6

Landslide Topology Uncovers failure Mechanism

This chapter is based on objective 4 of the Ph.D. thesis about developing a method for determining the failure mechanism of landslides. We explored various geometric and topological properties of landslide shape to identify failure mechanism and found that topological properties are excellent predictors for identifying failure mechanism. The work presented in this chapter will be submitted soon to Nature Communication Journal:

Rana, K.[1], Bhuyan K.[1], Ferrer J.V., Cotton F., Catani F., Ozturk, U., & Malik, N. (2023). Landslide topology uncovers its failure mechanism *(under review in Nature Communication)*

6.1 Abstract

The death toll and monetary damages from landslides continue to rise despite advancements in predictive modeling. The predictive capability of these models is limited as landslide databases used in training and assessing the models often have crucial information missing, such as underlying failure mechanisms. Here, we present an approach for identifying failure mechanisms—flows, falls, slides, and complex—by leveraging 3D landslide topology. We observe topological proxies reveal prevalent signatures of mass movement mechanics embedded in the

[1]These authors contributed equally.

landslides' morphology or shape, such as unfolding mechanisms within complex landslides. We find identical failure mechanisms exhibit similar topological properties, and by using them as predictors, we are able to identify failure mechanisms in Italy and the US's Pacific Northwest region with 80-94% accuracy. These new insights can considerably improve the performance of landslide predictive models and impact assessments. Moreover, our work introduces a new paradigm for studying landslide shapes to understand underlying processes through the lens of landslide topology.

6.2 Introduction

Every year, landslides cause economic damages worth 20 billion US dollars (Klose et al., 2016), and between 2004 and 2019 non-seismic landslides alone caused about 70,000 fatalities worldwide (Froude and Petley, 2018). Within the first two months of 2023, we have seen reports of devastating landslides in São Paulo, Brazil (Simoes and Araujo,), Southern Peru (Aljazeera,), and New Zealand (Craymer and Feast,), injuring many and killing approximately 70 people. Adding to this, recent studies count over one million landslide occurrences with annual volumes estimated at fifty-six billion cubic kilometers globally (Broeckx et al., 2020), presenting a risk to sixty million people (Ozturk et al., 2022). With the increase in urbanization, global climate change, and environmental change trends, the frequency of landslides and the associated risks will keep increasing globally over time (Ozturk et al., 2022). In line with this, landslides are anticipated to evolve and remobilize with increased frequency under changing climatic conditions on a decadal scale (Fan et al., 2021; Gariano and Guzzetti, 2016). Our ability to identify hazards from emerging landslides and dynamically assess impact areas is essential in averting risk to rapidly urbanizing communities and adapting to changing environmental conditions (Lima et al., 2023; Ozturk et al., 2022).

To address the rising landslide risk, predictive models for hazard, risk, and early warning systems are developed which assist in forecasting landslide occurrences and locating landslide-prone regions to mitigate the associated impacts (Corominas et al., 2014). However, the efficacy of these models is contingent on the quality of the underlying landslide databases.

These databases often lack the much-needed information about failure mechanisms of the mapped landslides (Guzzetti et al., 2012). Baffling predictive models (e.g., Huang et al.(Huang et al., 2023)), these databases instead rely on a broader definition of landslides that covers all types of gravitational mass wasting processes, such as slides, flows, and falls. Typically, each landslide failure mechanism exhibits different geological, geometrical, and geotechnical properties (see Figure 6.1). For instance, slides have conspicuous primary scarps and collapse along the planar or rotational surfaces (Varnes, 1978), flows such as mudflows exhibit visco-plastic or viscous/fluid kinematics caused by excess pore water pressure (Bradley et al., 2019), and rock falls entails the free falling of fragmented rocks from steep slopes (Bourrier et al., 2013) (see Supplementary Section S2 for detailed explanations).

Practitioners usually combine these different failure types into one group, despite their different properties (Guzzetti et al., 2006; Lombardo and Mai, 2018; Rossi et al., 2010), since categorizing them manually requires comprehensive surveys (remote and field) and standardized classification protocols (Guzzetti et al., 2012), which are laborious and time-consuming. Consequently, predictive models start to harbor significant levels of uncertainty and bias (Huang et al., 2023), hence failing to match empirical observations, especially when moving from local levels to regional and global scales (Kirschbaum et al., 2010; Reichenbach et al., ; Fressard et al., 2014). For instance, they may predict a low probability of landslide occurrence in a high landslide-prone region. Therefore, identifying landslide failure mechanisms is fundamental to improving predictive modeling.

Preliminary attempts at identifying failure mechanisms have considered both knowledge-driven and data-driven approaches. While the former are region-specific, bounded by expert-based rules, and constrained to small areas (Martha et al., 2010; Barlow et al., 2006), the latter addressed these problems with supervised learning and have successfully identified landslide failure types in the Italian context (Amato et al., 2021). However, the existing solutions are still limited in their prediction capabilities, as the failure information is derived from geometric properties of two-dimensional (2D) landslide polygons (outlining the landslide planforms). Owing to the inherent limitations of 2D landslide polygons, crucial kinematic and mechanical

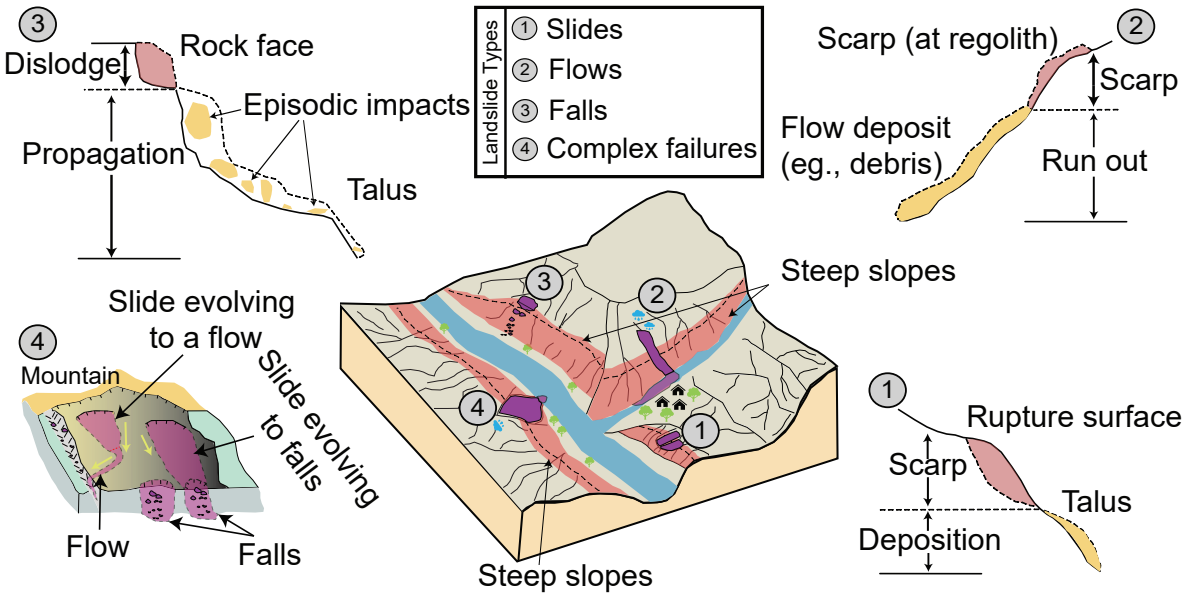

Figure 6.1: The schematic illustrates different landslide failure mechanisms and their associated mechanical and kinematic behavior. For example, Type 1 refers to slide-type failures that can constitute deep ruptures, where a cohesive unit of soil or rock slides down a slope following a well-defined rupture plane. Type 2 refers to flow-type failures where soil, rock, or other material flows down a slope as a dense, fluid-like mass with a flow-like motion. Type 3 is a fall-type of failure where a body of rock detaches from a steep slope or cliffs and exhibits free falling and episodic impacts as they propagate down the slope. Type 4 refers to the complex interaction and effect of numerous geomorphic processes transpiring in a single failure event, where processes start as one mechanism and evolve into another; such as a slide-type failure evolving to a flow-type failure.

details embedded in the landslides' three-dimensional (3D) morphology are overlooked, such as the style of kinematic progression, deformation patterns and structures, and debris deposits at the talus. Furthermore, the kinematic evolution of one or more failure types may culminate in the convergent evolution of landslide shapes, wherein landslides starting as completely different movements may evolve to follow similar planar outlines. This phenomenon complicates the discernment of mechanisms based solely on 2D representations or simple topographic measures. We posit that such morphological and kinematic information is rooted in the landslide's morphology which can be extracted via topology.

Topology is a sub-discipline of mathematics explored in many fields that concern the study of shapes (Lum et al., 2013), such as in protein structures, data modeling, complex networks, and signal processing (Luo, 2023; Carlsson, 2020; Han and Bao, 2022; Zangeneh-Nejad and

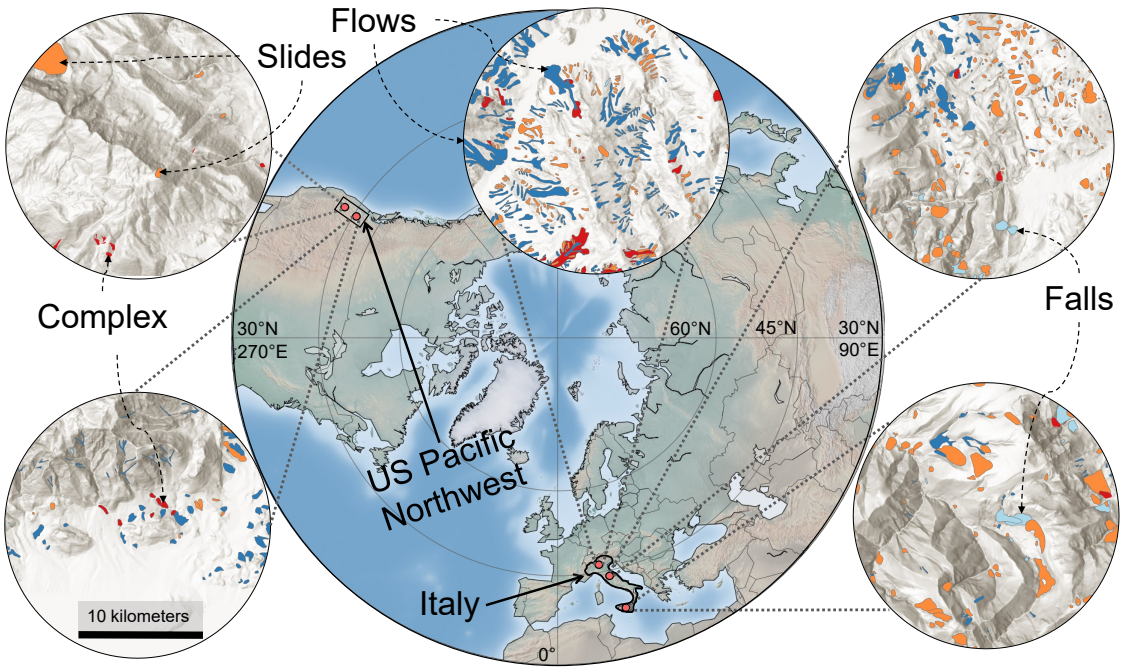

Figure 6.2: The diagram shows the geographical regions: Italy and the US Pacific Northwest, whose data we analyzed in this work. The inset in the circular shape shows snippets from diverse regions with landslide polygons of different failure mechanisms on top of the World Hillshade Map. Map credits: World Shaded Relief-ESRI (ESRI, 2020)

Fleury, 2019). We explore advanced data analysis tools rooted in topology, known as topological data analysis (TDA), which captures critical structures present in the data's shape (in our case, the landslide's 3D shape). We hypothesize that key features of landslide kinematics are embedded in the 3D topology of the involved landforms and that TDA properties can capture the kinematics and mechanics as a proxy for identifying landslide failure

In this study, we introduce an approach for uncovering landslide failure mechanisms by examining the topological properties inherent in the 3D shapes of landslides. We demonstrate that this method offers a more comprehensive understanding of the underlying failure processes compared to traditional analyses based on 2D polygonal geometry. To validate the effectiveness and applicability of our approach, we apply it to landslide data sets from two distinct geomorphological settings: Italy and the United States (US) Pacific Northwest region (see Figure 6.2). Additionally, we utilized the topological properties of complex landslides to reveal the underlying physical processes behind their formations. Here, we showcase with our

findings that the proposed method (1) is user-friendly, exclusively requiring only the landslide polygonal shape and a Digital Elevation Model (DEM) as input, (2) exhibits high performance in discerning failure mechanisms, (3) is transferable across various geomorphological regions, and (4) shows strong performance and remains robust, even when the availability of samples is limited, indicating its applicability in data-scarce regions. By offering a deeper understanding of landslide failure mechanisms, this approach has the potential to enhance the accuracy and reliability of landslide susceptibility, hazard, and risk assessment models, by providing valuable insights to the predictive modeling community.

6.3 Results

6.3.1 Landslide topology as a proxy to identify failure mechanisms

The underpinning of topological data analysis (TDA) is rooted in structures in the data's shape, such as connected components and holes. Holes represent the empty spaces in the data's shape, and connected components represent the connection of the data's points linked by a continuous path. Using the holes and connected components, we can calculate various topological properties to quantify a shape. For this, we perform the TDA on landslide shapes to compute topological properties which can then be used as a proxy to investigate the underlying failure mechanisms. It is important to note that we employ 3D point clouds (containing geographical latitude, longitude, and elevation information) from the landslide's outline (see Figures 6.3 and 7.2) obtained via the landslide polygon and the Digital Elevation Model (DEM). The landslide polygon provides the best available approximation of the landslide boundaries in the geographic space, as derived by standard surveying methods with suitable accuracy.

The degree of compactness in a landslide shape is essential when identifying failure mechanisms (Amato et al., 2021). For instance, slides are characterized by a more cohesive material that tends to remain as a single component (e.g., slides with clay-rich soil (Kenigsberg et al., 2020)), leaving behind a more compact-shaped footprint as they fail. In contrast, flows involve more fluid and fragmented materials deposited at the talus and display viscous/fluid

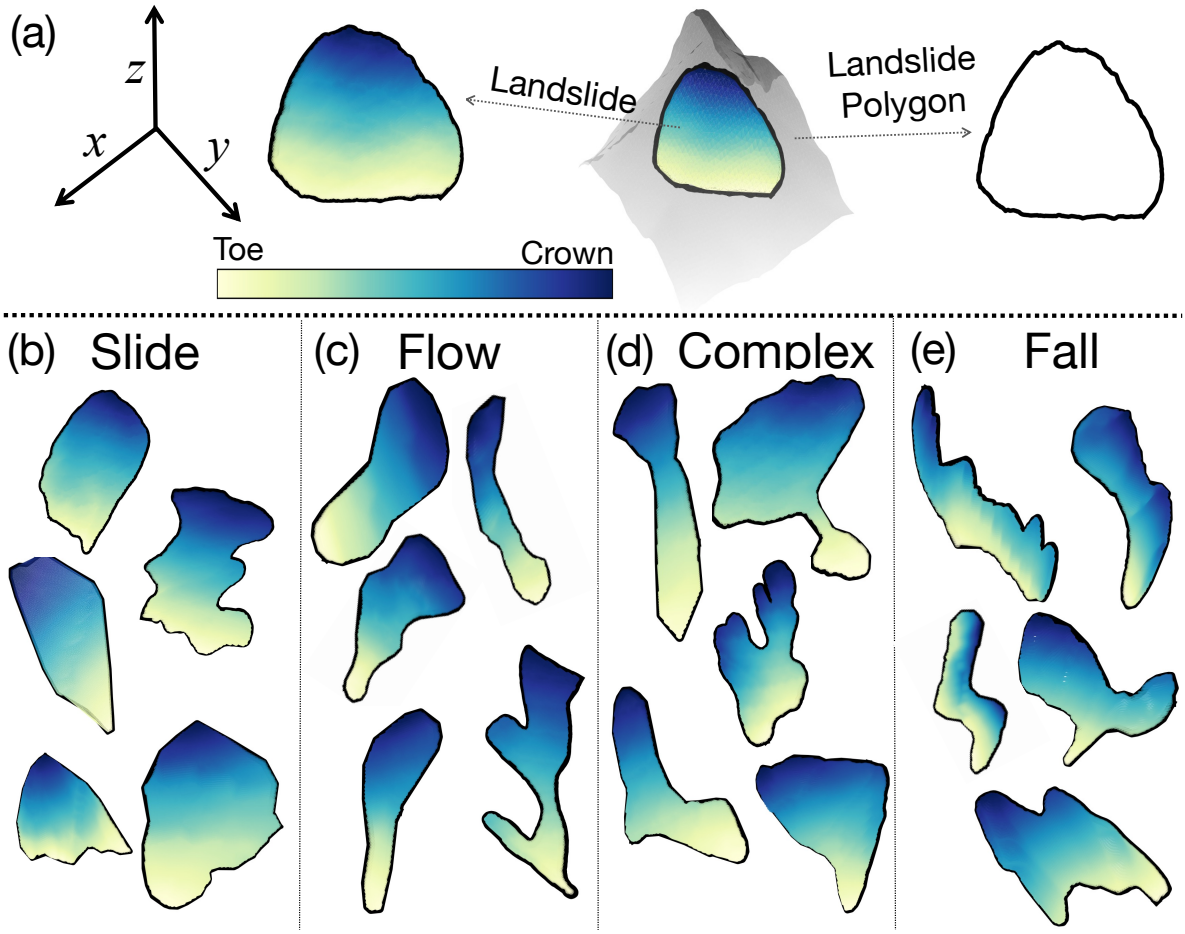

Figure 6.3: (a) An example of a landslide failure in a terrain with a steep slope. The diagram also shows the 3D landslide polygon, which outlines the landslide shape. (b-e) 3D landslide samples for different landslide failure types, namely, the slide, flow, complex, and fall types.

kinematics that follows the contours of the natural landscape and are less compact. On the other hand, falls consist of fragmented materials that roll/bounce off steep cliffs with a rather shorter and straighter run-out path compared to flows which leave behind a footprint that has an intermediate degree of compactness between the slide and flow-type failures. We use the amount of empty space inside the footprints of the landslide shape outline to quantify its compactness. To give a simple example, a higher amount of empty space in a landslide shape outline is associated with a higher degree of compactness (e.g., for slide-type failures as seen in chapter 7 Figure 7.1-a). Representing the average lifetime of holes, AL_H (one of the topological properties) computes the hole's average size and estimates the information pertaining

to the empty space, and thus the compactness of a landslide's shape. So, landslide shapes with a longer AL_H are more compact than shapes with a shorter AL_H. Based on the Probability Density Function (PDF) of the AL_H, our analysis reveals that slides are more compact than flows, falls, and complex landslides because they have a longer AL_H (see Figure 6.4). Moreover, we observe the AL_H PDF curve for falls to lie between those of slides and flows, showing that empty spaces generated in falls do not survive long since materials detach from steep slopes and travel a short distance, thereby leaving behind a footprint that represents an intermediate level of compactness. Also, the PDF of the AL_H for complex landslides shows an intermediate level of compactness, credited to their amalgamated behavior as a combination of slides, falls, and flows.

Another critical property for diagnosing failure mechanisms is the sinuosity of the transport zone, which describes the landslide's path or kinematic propagation as it progresses downslope. Of all failure types, flows are the most sinuous, following the contours of the landscape, owing to the fluid and mobile nature of the materials involved, while slides are the least sinuous as their material is rarely channelized and remains on the open slope, resulting in a relatively straight and uniform path. Fall-type failures are comparatively less sinuous than flows but still exhibit some degree of sinuosity, as they too follow the landscape's contours. Sinuosity defines the existence of numerous curves in the landslide's shape attributed to the landscapes' contours, leading to the generation of partitions within the landslide outlines by the TDA and hence, generating multiple empty spaces with shortened lifetime (see Methods section). This information on the sinuosity of landslide shapes is inferred from the combination of two topological properties–the bottleneck amplitude of holes, BA_H, and the average lifetime of holes, AL_H. The BA_H represents the maximum lifetime of holes in the landslide shape, which quantifies the maximum empty space in the 3D space occupied by the landslide. As sinuous shapes result in numerous smaller empty spaces with shorter lifetimes, the AL_H drops without significantly impacting the largest empty space as determined by the BA_H (see Figure 6.4).

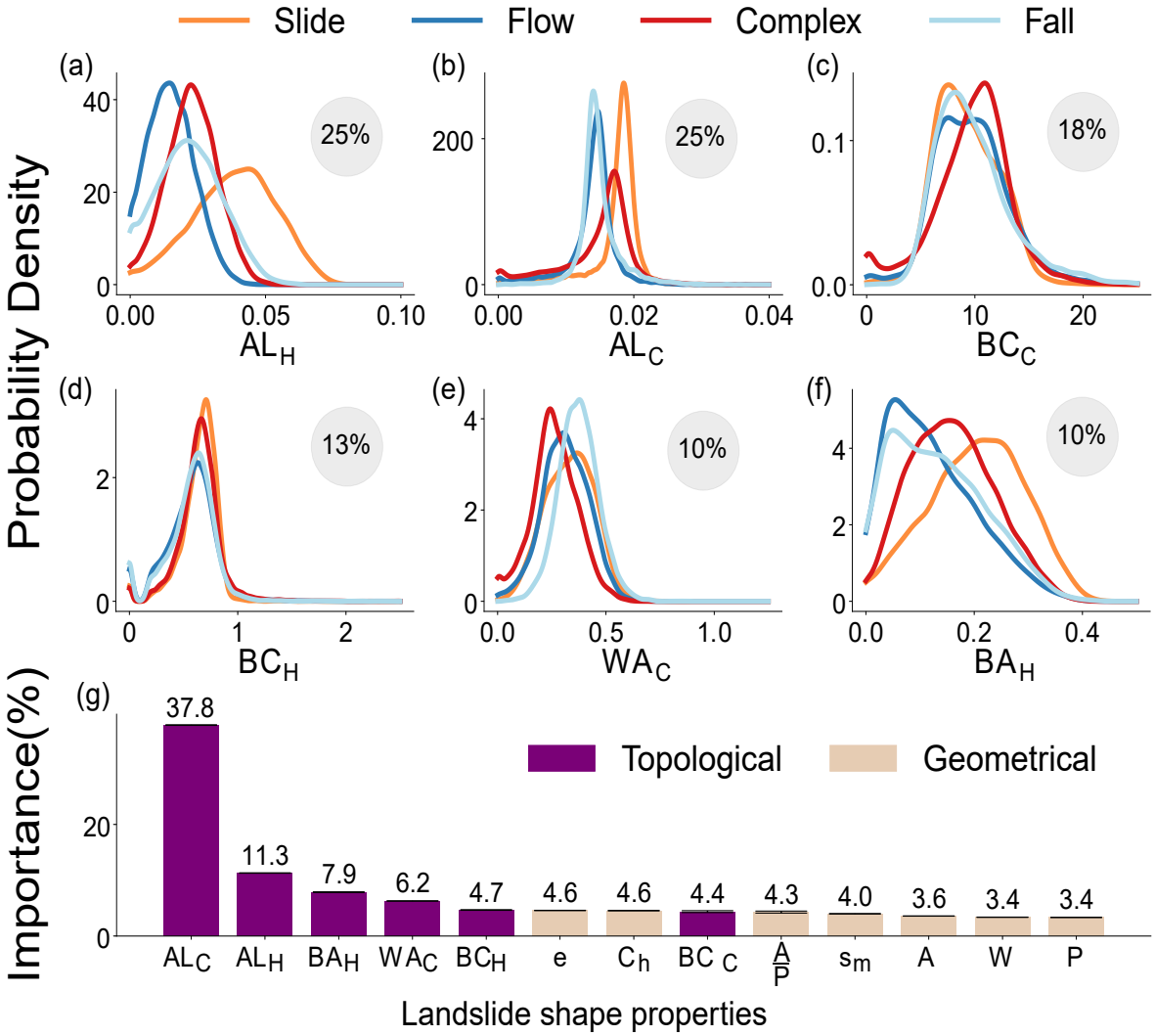

Figure 6.4: Plots (a-f) show the probability distribution functions of the six most optimal topological properties used in classifying the failure mechanisms for slides, flows, complex, and falls in Italy. The y-axis shows the probability density values (calculated using kernel density estimation), and the x-axis shows the value of topological attributes. The topological properties in plots (a-f) are: Average lifetime of holes (AL_H), Average lifetime of connected components (AL_C), Betti-curve based feature of connected components (BC_C), Betti-curve based feature of holes (BC_H), Wasserstein amplitude of holes (WA_H), and Bottleneck amplitude of holes (BA_H) (the computations of these properties are explained in detail in chapter 7 Section 7.3 and Figure 7.2). The percentage values in the gray circular disk in each figure indicate the topological feature's importance (in %), as estimated by the random forest-based classification procedure. Sub-plot (g) shows the joint computed feature importance of topological and geometric properties by the random forest model. The geometric properties from top to bottom are: area (A), perimeter (P), the ratio of area to perimeter $\frac{A}{P}$, convex hull-based measure (C_h), minor(s_m), and width (W) of the minimum area bounding box fitted to the polygon.

Consequently, a landslide shape with a relatively higher BA_H and a shorter AL_H is indicative of increased sinuosity. In light of these observations, our findings indicate that flow-type landslides indeed display a higher degree of sinuosity compared to other failure mechanisms. This is evident from the fact that flows exhibit a similar BA_H to falls but a shorter AL_H in comparison. This is expected, as flows, being the most sinuous, cause multiple small holes or empty spaces (see flow-type failure in chapter 7 Fig. 7.1-b) that end up shortening the AL_H. Conversely, slides display both longer BA_H and AL_H, reflecting their minimal level of sinuosity.

We are also interested in the role that slope variations play, as they significantly impact the stability of the slope and influence the type of landslide that occurs. For example, falls and slides have a more significant slope transition in their profiles compared to flows, which propagate with a nearly constant slope (Catani et al., 2005). This slope variation is captured by the lifetime of the connected components. A sharp change in slope causes the points outlining the landslide to be spaced vertically further apart, leading to a longer lifetime of the connected components. Two topological properties—the Wasserstein amplitude of the connected components, WA_C, and the average lifetime of the connected components, AL_C—help capture information about this slope variation in a landslide's profile. The WA_C quantifies the set of longer lifetimes of the connected components, quantifying the most significant slope change in the landslide outline. This is nicely illustrated in the PDF (Figure 6.4) of WA_C, which shows that slide and fall failures underwent more drastic slope changes compared to flows. Yet, falls possess a shorter AL_C than slides. This is due to the lower portion of the shape's outline (at the talus) displaying a flatter terrain (representing the area where materials accumulate) and attributing negligible slope change, which ultimately shortens the AL_C. In contrast, flows display the minimum AL_C, as they more or less propagate on constant slopes.

Several topological properties, like the Betti curve-based feature (BC), capture more intricate landslide shape properties and help in discerning landslide failure mechanisms. The Betti curve-based feature represents the total number, lifetime, and presence of the structures (holes and connected components) emerging simultaneously. We hypothesize that it encompasses a

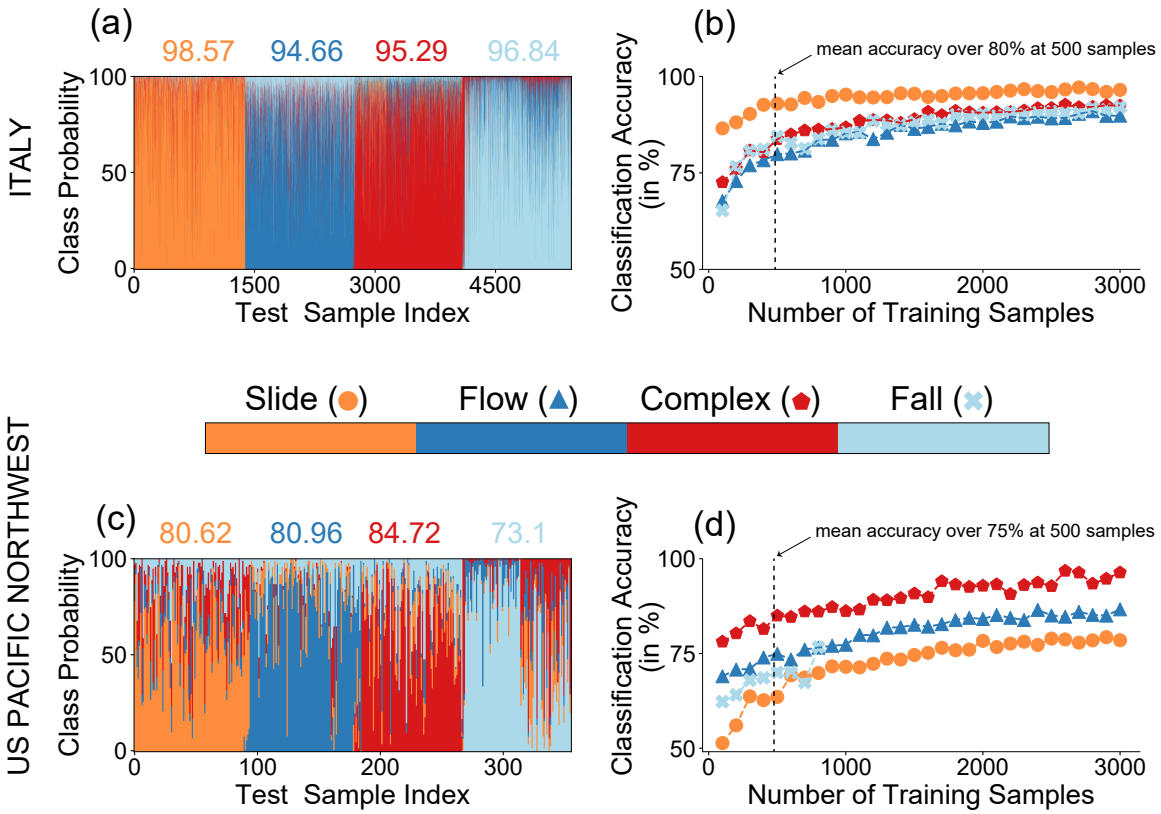

Figure 6.5: Plots (a and c) show the classification accuracy (in %) for each failure class in Italy and the US Pacific Northwest. The x-axis of the plots shows the testing sample's index, and the y-axis shows the class probability corresponding to each failure class. Plots (b and d) show the classification accuracy (in %) corresponding to each failure mechanism with the number of training samples. The x-axis shows the number of training samples from each class used to train the model, and the y-axis shows the classification accuracy (in %) corresponding to each class. At 500 samples, the mean classification accuracy reaches over 80% in Italy, and over 75% in the US Pacific Northwest

combination of compactness, sinuosity, slope variations, and similar structures within a given landslide shape. However, the exact connection to the underlying physical mechanisms is not clear due to the complex nature of this topological property. We anticipate that such properties consider higher-order information about the landslide shape that is not immediately apparent to us.

Through our analysis, we discovered that common topological properties such as AL_H, AL_C, and BA_C govern the general movements of distinct failure types. These properties act as proxies for the diverse kinematic and mechanical characteristics, which are essential to

consider when identifying the different failure mechanisms. This finding simplifies and brings coherence to our understanding of landslide behaviors from a topological lens, offering a more effective approach to discerning and predicting these complex natural phenomena.

6.3.2 Comparison between landslide geometry and landslide topology

Traditional geometric descriptors of landslide shape, including properties such as area, perimeter, and convexity, are derived from 2D representations of the landslide body. As a consequence of this inherent simplification, these 2D-based geometric properties may not adequately capture crucial information, such as failure depth or internal deformations, associated with the landslides' 3D configuration. To address these limitations and provide a more comprehensive understanding of the landslide dynamics, we computed topological properties that are derived from the landslides' 3D configurations (more detailed information can be found in chapter 7 Section 7.3). We postulated that the topological properties would prove more meaningful in decoding the characteristics of the landslides and their underlying failure mechanisms than the traditional geometric counterparts. To test this, we used a set of seven well-known geometric properties that are commonly employed in the literature (Rana et al., 2021; Taylor et al., 2018b; Stark and Guzzetti, 2009) along with six topological properties (for the justification of using six topological properties, please refer the Methods section) to determine the failure mechanisms in the Italian context.

We jointly computed the feature importance of geometrical and topological properties using the Gini-index feature importance method in the random forest algorithm (see Figure 6.4-g). After running over 100 iterations on the Italian data set, our findings consistently demonstrated that topological properties exhibited higher feature importance than the traditional geometric counterparts (achieving Micro F1-scores of 94%, ∼65% respectively), yielding superior predictive capabilities for identifying failure mechanisms. Additionally, we observed that even the least important topological property (BC_C) has similar feature importance as the other geometric ones, while the former conveys unique information about the landslide shapes as discussed in the previous section. Moreover, we calculated the Probability Density Function (PDF) for both geometric and topological properties and observed that the latter

had greater dissimilarity than the former among different failure types (see chapter 7 Figure 7.3). These two findings demonstrate that topological properties are stronger predictors for identifying failure mechanisms. The reason for this can be attributed to the enhanced capacity of topological properties to encapsulate important information pertaining to landslide kinematic progression, failure depth, sinuosity, compactness, and slope variation.

6.3.3 Determining failure mechanisms with TDA and machine learning

Next, we employed Topological Data Analysis (TDA) to compute a diverse array of topological properties/features. Subsequently, we conducted a correlation analysis and feature importance assessment to identify the six most optimal properties out of thirty. Our evaluation unveiled that several TDA properties were redundant, amplifying the model's complexity and undermining its predictive potential. Consequently, we opted to eliminate the irrelevant ones. Leveraging the six best features, we applied the random forest algorithm to discern the failure mechanisms. We independently scrutinized the performance of our approach using various training and testing sets for both Italy and the US Pacific Northwest without merging the data sets.

The cumulative number of samples (n_{total}) for Italy equates to 254,916, encompassing 100,419 slide-type samples (n_{slides}), 89,316 flow-type samples (n_{flows}), 51,511 complex-type samples ($n_{complex}$), and 13,610 fall-type samples (n_{fall}). To counteract potential class imbalances, we used an equal number of samples from each landslide type for training, with $n_{slides} = n_{flows} = n_{complex} = n_{fall} = 13,610$. To alleviate overfitting and bias, we executed 10-fold cross-validation, iterated 1,000 times on 54,440 samples ($4 \times 13,610 = 54,440$), and we achieved a Micro F1-score—a performance metric—surpassing 94% for each failure class (see Figure 6.5-a), with a performance standard deviation below 0.2%. This illustrated the robustness of our methodology in handling variations among training samples across Italy. We examined various other metrics (chapter 7 Section 7.7), such as the True Positive Rate (TPR) and True Negative Rate (TNR), to evaluate the method's performance. These metrics consistently exhibited high scores across all classes, thereby ascertaining the model's classification ability.

In parallel, the aggregated samples from the US Pacific Northwest amounted to $n_{total} = 36,686$, with $n_{slides} = 11,826$, $n_{flows} = 20,805$, $n_{complex} = 3,165$, and $n_{fall} = 890$. We utilized an equal number of samples from each class ($n_{slides} = n_{flows} = n_{complex} = n_{fall} = 896$) and conducted 1,000 iterations of 10-fold cross-validation to mitigate class imbalance, bias, and overfitting. We attained an average Micro F1-score exceeding 80% with a standard deviation reported below 1% (see Figure 6.5-c). Strong performances with respect to other metrics were also observed, similar to those in the Italian context (see chapter 7 Figure 7.5).

Additionally, we examined model transferability by training the model in Italy and testing it on the US Pacific Northwest dataset, thereby gauging its performance and robustness when training and testing the methods in geographically disparate regions (see Supplementary Figure S4). However, the method performs poorly in this scenario which was anticipated as these regions have different internal (e.g., geological and topographical) and external (e.g., rainfall and seismic shaking intensities) conditions. Such contrasts in the geophysical systems will harbor unique properties even in similar failure mechanisms. More details are provided in Supplementary Section S6.

6.3.4 Method performance with limited landslide samples

The efficacy of data-driven approaches is dependent upon the number of training samples (Devarakonda, 2022), which, in our study, are samples with known failure mechanisms. A methodology exhibiting robust performance, even when confronted with limited training data, is of paramount importance in real-world scenarios. To assess the method's performance in data-scarce circumstances, we conducted several assessments, incrementally increasing the number of training samples for each failure type starting from 100 samples in 100-step intervals (Figure 6.5-b, d). Remarkably, even with a meager 100 training samples from each class, the method achieved a mean classification accuracy exceeding 50% in both Italy and the US Pacific Northwest. The method's performance was enhanced with the increment of training samples, attaining a mean classification accuracy above 80% and 75% with just 500 training samples from each class.

Our findings show that the method performs well even with fewer training samples (>500

from each type) and can benefit regions with scarce samples of known failure types. In areas devoid of documented samples with failure data, practitioners can manually annotate a few instances to train the algorithm, enabling them to discern the failure mechanisms for the remaining landslide population.

6.3.5 Deciphering complex landslides in the US Pacific Northwest

Complex landslides typically occur as amalgamations of numerous processes or failure types that appear successively, such as slides to flows (LaHusen et al., 2020). Because complex landslides include multiple failures, it is challenging to investigate their behavior for the purposes of predictive modeling. Topological properties are capable of capturing intricate information between different failure processes (as seen in the previous sections), and hence, we further explored its capability to understand the underlying physical processes that lead to complex failures. We utilize 428 complex landslides from the US Pacific Northwest data set to discern the combination of failure types present in them. Out of 428 complex landslides, 198 of them are documented as "Translational rock slides followed by rock falls" and the rest are documented as "Rotational slides followed by flows" (as reported by the Statewide Landslide Information Database for Oregon, SLIDO(Franczyk et al., 2019)).

To identify the failure mechanisms within these complex failures, we trained our method with three classes (i.e., slides, flows, and falls) and forced the model to predict the class probability corresponding to each failure type. For 198 complex landslides documented as "Translational rock slides followed by rock falls", our model predicts slide-type failures with the highest probability followed by falls (see chapter 7 Figure 7.6-a). Similarly, for the remaining 230 "Rotational slides followed by flows" complex landslides, our model predicts slide-type failures with the highest probability followed by flows (see chapter 7 Figure 7.6-b). Among these 428 landslides, slides are predicted as the most dominant failure type, which is also evident when observing the resemblance of the slide and complex failure topological properties (such as AL_H, BC_H, and BA_H) in the PDF plots (see chapter 7 Figure 7.2). These findings demonstrate that topological properties are able to decipher more than just one physical process in a given landslide.

6.4 Discussion

In this work, we attempt to determine the failure mechanisms through the lens of landslide topology. Our key findings elucidate the connection between the topological properties as a proxy to identify underlying failure mechanisms. We observe that identical landslide types harbor similar topological properties, indicating the presence of common morphological characteristics that govern the general movement of the failures. Also, we observe that topological properties offer a more profound capacity to distinguish between failure types than traditional geometric properties. This finding can be attributed to the fact that topological properties inherently capture critical information related to landslide kinematic progression, failure depth, sinuosity, compactness, and variations in slope. In contrast, geometric properties tend to oversimplify the complex spatial, kinematic, and mechanical relationships that govern the behavior of landslides and are hence less effective in helping to differentiate between various failure mechanisms. Additionally, we utilized topological properties to dive deeper into the complex failures and identified the underlying processes that contribute to their formation. Our findings suggest that topological properties can reveal more than one physical process in a given landscape. Given these advantages, we anticipate that our method will present new avenues for future research, particularly in the landslide modeling community.

Understanding the failure type of a landslide can provide insight into its triggering mechanism (Rana et al., 2021). Therefore, by analyzing the landslide failure types, we can diagnose the temporal frequency of the associated trigger that causes them. This is particularly important for forecasting the remobilization of unconsolidated deposits of post-seismic landslides that evolve into frequent rainfall-triggered flow-type landslides that interface with risk on a decadal scale (Tanyaş et al., 2021; Fan et al., 2021). Forecasting the increased frequency and identifying the magnitude of impacted areas under changing climate conditions rely on our ability to support scenarios of physically-based hazard models with landslide-type-specific triggering information (Lima et al., 2023; Gariano and Guzzetti, 2016). Furthermore, identifying failure mechanisms increases our capability of extracting quantitative data on sediment budgets and on dominant geophysical cycles at continental and global scales, with direct effects

on our knowledge of forecasted climate change impacts (Francis et al., 2022).

Building on this, it is crucial to acknowledge that predictive models that simulate type-specific behaviors require multiple input parameters tailored to each failure type (Katz et al., 2014). For example, sliding mechanisms triggered by seismic activity require additional parameters such as seismic loading, the internal friction angle of the soil, and peak ground acceleration. In contrast, flow-type failures such as debris flow require parameters like the precipitation amount, pore water pressure, and infiltration rate. Failing to differentiate between these distinct landslide types leads to an oversight of the parameters which drive their behavior and can lead to inaccurate predictions when simulating them; for instance, locations or slopes susceptible to seismically-induced slides may be underestimated if a model is calibrated with precipitation data to simulate rainfall-induced flow-type failures, ultimately resulting in erroneous predictions and ineffective hazard assessments of seismic-slides.

Additionally, the failure-type information has notable potential to improve landslide risk assessment and associated hazard models (Huang et al., 2023). The level of damage to infrastructures and the risk of human casualties vary depending on the intensity of the landslide, which differs for each failure type (Varnes, 1978). For example, a slow-moving deep-seated rotational landslide (1.5 m/year to 16 mm/year) may not pose an immediate threat to the population, but it can cause extensive structural damage to buildings over a prolonged period (Sundriyal et al., 2023; Dille et al., 2022). In contrast, flow-type failures, such as debris flows, have rapid mobility and can result in significant casualties and infrastructure damage simultaneously (Vega and Hidalgo, 2016; Perkins, 2012). Similarly, episodic impacts in fall-type failures can cause massive damage to infrastructures in a matter of seconds due to their high energy (e.g., impact pressure measured in kilopascals, kPa) (Dietze et al., 2017). We can infer from these broad examples that the availability of failure-type information is crucial for different aspects of predictive modeling and that incorporating it benefits the landslide community as it enables the development of accurate landslide predictive models.

The potential of the proposed method reaches beyond just understanding the complex interplay between landforms, their shapes, and the underlying geophysical processes responsible

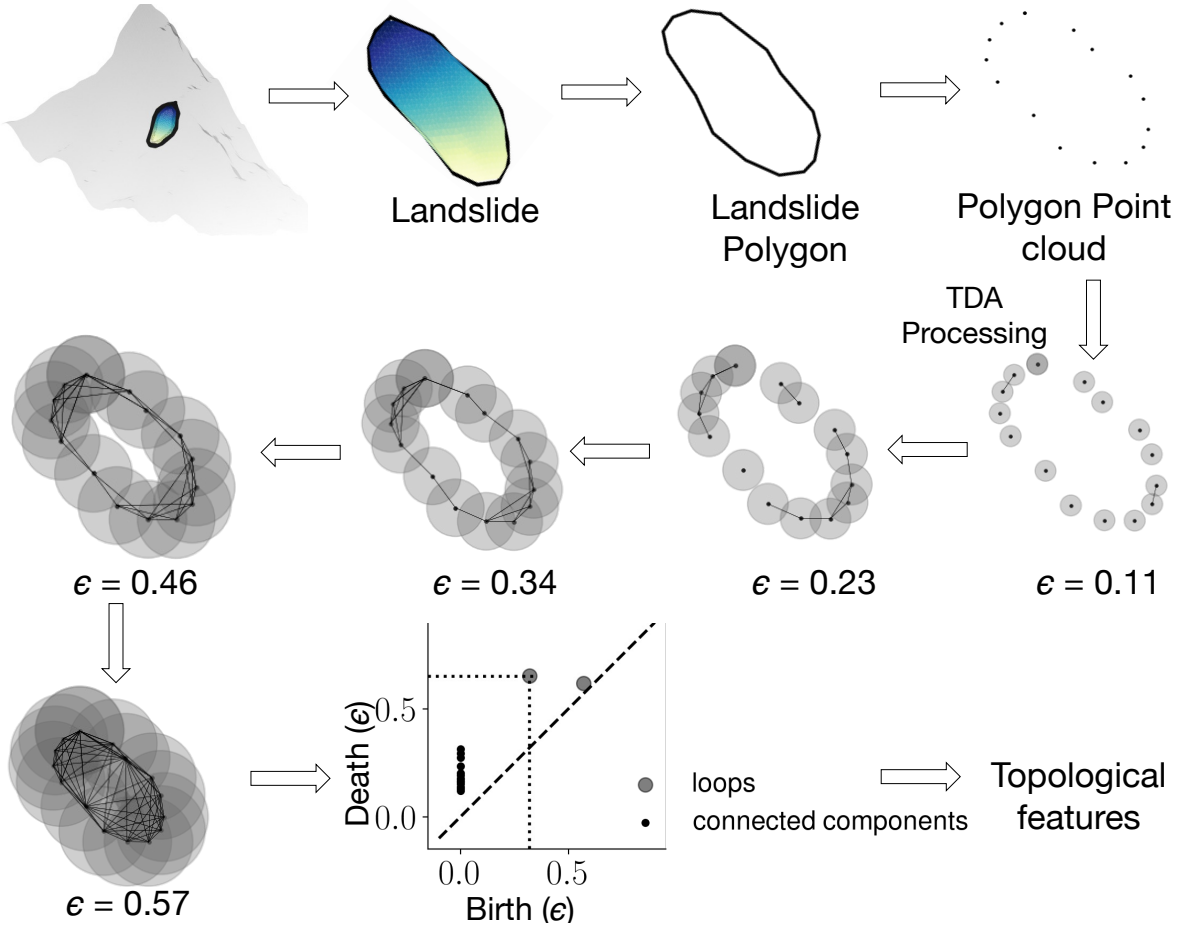

Figure 6.6: The diagram illustrates the procedure of computing topological features corresponding to a landslide shape. The flowchart shows the use of persistent homology in capturing various structures of the landslide shape by using an evolving disk size (ϵ) around each point in the point cloud. With the increase in ϵ, various structures like connected components and holes emerge in the data's shape which is captured by the persistence diagram. Using this information, we can calculate the topological properties of the landslide's shape. Please note that when processing the TDA features, we display the flowchart using a 2D illustration for simplicity and better visualization.

for their formation; they also serve as a subject captivating interest across various geophysical disciplines. The ability to acquire knowledge about the processes generating complex landforms based solely on their shapes suggests a rich presence of signatures imprinted on the landscapes. Our method leverages their topological properties to effectively extract this information. Envisioning compelling applications beyond landslides, we can explore other geophysical processes such as permafrost-borne retrogressive thaw slumps in Arctic regions (Nicu

et al., 2021) and sub-surface processes, e.g., submarine landslides (Frey-Martínez et al., 2006), which commonly occur in a typical data-scarce environment such as the sea bottom, where geological and geotechnical information are almost absent. These processes also give rise to unique landforms displaying distinct shapes and configurations, and employing topology can aid in gauging the mechanisms governing their occurrences.

While the proposed solution demonstrates notable success in identifying failure mechanisms, there are inherent limitations that warrant attention. The effectiveness of the method relies on the quality and geographical location of the training and testing data used in the model. Manual annotations of failure mechanisms can lead to bias since various mappers will have different perspectives (mapping on aerial or satellite imagery versus geomorphological field mapping can display distinct perceptions) when annotating the landslides and their failure types. Also, due to the ambiguous nature of complex-type failures, they could include slides and flows simultaneously, which can impact the overall performance of the model. The method's reliance on a DEM for converting 2D polygons into 3D shapes also presents potential challenges. DEM quality in the training and testing regions can bias the results, particularly for smaller landslides, as coarser DEM resolutions may struggle to capture the profiles of these smaller-scale events.

In the future, we plan to develop a complete transferability method that learns from data-rich regions such as Italy and implement it to identify failure mechanisms in data-scarce landslide-affected regions of the world, such as Africa, central and south Asia, and the Caribbean island countries. We also envisage including this complete transferability method in a Python library for the landslide community. Moreover, we will also diversify the landslide failure types to accommodate the community's needs in adapting to other classification systems (such as Cruden and Varnes (Cruden, 1996)), including failure mechanisms such as spreads, earth flows, debris flows, and rotational and translational slides. We hope that the landslide community will find our method useful in their line of research, and appreciate its contribution to understanding/identifying failure mechanisms, and make an effort to improve landslide predictive modeling in different parts of the world.

6.5 Methods

6.5.1 Topological Feature Engineering

In the proposed method, landslide polygons serve as the primary input. These polygons represent the 2D outline of the landslide body on the ground and are commonly found in landslide databases. Each vertex of the landslide polygon comprises geographical latitude and longitude coordinates. Utilizing the Digital Elevation Model, the landslide polygons are transformed into normalized 3D shape outlines, wherein each vertex encompasses latitude, longitude, and elevation information. Topological Data Analysis (TDA) is employed to extract the geometrical and topological characteristics of a landslide's 3D shape outline (see Figure 7.2). This information is subsequently used as input for a machine learning algorithm, specifically the random forest. The Python library Giotto-TDA is leveraged to extract an assortment of TDA properties/features from the 3D shape of landslides (Tauzin et al., 2021). To ascertain the most pertinent features for landslide-type classification, a correlation test is conducted between TDA features, and those with high correlation are removed. The remaining, less correlated features are then assessed, and the least important ones are iteratively eliminated until six robust predictors remain. The exclusion of additional predictors results in decreased performance, while incorporating more than seven yields comparable outcomes. Utilizing fewer predictors facilitates the development of a more generalizable model. The six features thus form a feature space for the random forest classifier.

Topological Data Analysis (TDA) quantifies the multidimensional shape of data using algebraic topology techniques. TDA offers a variety of metrics for capturing the geometric and topological properties of data shape (Carlsson, 2009). These metrics could be used as a feature space for the machine learning algorithm to solve various classification and regression problems, such as shape classification. TDA's central idea is persistent homology, which identifies persistent geometric features by using simplicial complexes to extract topological features from point cloud data. Simplicial complexes are a collection of simplexes that are the building blocks of higher-dimensional counterparts of a graph. An n-dimensional simplex

is formed by connecting n+1 affinely independent points (Munch, 2017; Garin and Tauzin, 2019). For example, a point is a 0-dimensional simplex, an edge that connects two points is a 1-dimensional simplex, and a filled triangle formed by combining three non-linear points is a 2-dimensional simplex. A Vietoris-Rips complex indicates the simplicial complex in the data's shape using a parameter ϵ. The main idea of the Vietoris-Rips complex is to connect any two points in the point cloud data set whose distance is less than ϵ. These connections of data points create structures in the data that change with the parameter ϵ. Therefore, to get complete information about all the structures in the data, the idea is to use all $\epsilon > 0$ values.

Only specific structures in the data shape provide crucial information about the geometrical and topological properties of the data. Homology measures these unique structures in the data, where e.g., 0-dimensional homology captures connected components or clusters, 1-dimensional homology measures loops, and 2-dimensional homology measures voids (Munch, 2017; Hensel et al., 2021). These crucial structures emerge and die with changes in ϵ, and this information is captured in the persistence diagram. With the help of a persistence diagram, we can calculate various measures quantifying the topological properties of the shape– persistence entropy, average lifetime, number of points, Betti curve-based measure, persistence landscape curve-based measure, Wasserstein amplitude, Bottleneck amplitude, Heat kernel-based measure, and landscape image-based measure (Bubenik and Dłotko, 2017; Reininghaus et al., 2015; Adams et al., 2017). We have explained all these topological features in detail in chapter 7, section 7.3. Finally, we used all these measures as input in the machine learning method–random forest.

6.5.2 Machine learning model: Random Forest

Random forest is an ensemble-based learning method that has shown promising results in various classification and regression problems (Barnett et al., 2019; Biau and Scornet, 2016; Breiman, 2001; Kursa, 2014; Chaudhary et al., 2016). Random forest classifiers consist of multiple classifiers trained independently on bootstrapping training samples. Bootstrapping N training samples leads to $\frac{2N}{3}$ independent samples, so each tree in the random forest is

constructed from a distinct subset of training samples (Azar et al., 2014; Belgiu and Drăguţ, 2016). Moreover, each tree in the random forest predicts the output class of the testing sample independently, and the class with the majority votes is the final decision of the random forest (Arabameri et al., 2021; Belgiu and Drăguţ, 2016).

Each random forest tree divides a parent node into two daughter nodes, right (r) and left (l). For each node split, the random forest chooses p features from the m total features of the samples (Azar et al., 2014; Okun and Priisalu, 2007). Among p features, the random forest selects a single feature for a node split based on the "Gini-index" criterion. The Gini-index for each right and left daughter node can be calculated as: G_r=1-$\Sigma_{j=1}^{j=N} P_{rj}$ and G_l=1-$\Sigma_{j=1}^{j=N} P_{lj}$. Here, $P_{rj}(P_{rlj})$ and N are the probability of the samples in the right (left)nodes having class j and the total number of the classes. The features that maximize the change in the Gini-index that is calculated as follows: $\Delta\theta(s_q) = G_q - \rho_{rq}G_r - \rho_{lq}G_l$ is used for the node split (Kuhn et al., 2013; Zhang and Ma, 2012b). Here, ρ_{rq} and ρ_{lq} are the proportion of samples in the right and left daughter nodes. The process of splitting nodes continues until a stopping criterion is met, such as when no more samples are available for splitting, or when the Gini-index of parent nodes is lower than that of daughter nodes.

6.6 Data availability

The dataset we utilized in this study to classify the failure mechanism of landslides was obtained from the Inventario dei Fenomeni Franosi (Inventory of Landslide Phenomena) in Italy (IFFI) (Trigila et al., 2010). The IFFI project catalog (`www.progettoiffi.isprambiente.it`) was created in 1999, with the aim of mapping and identifying landslides in Italy, and holds information on over 250,000 usable landslide polygons. Aerial image interpretation, historical sources, and field surveys were used to acquire and validate this catalog, while the classification protocol/scheme referred to that of Varnes (1978) (Varnes, 1978) and Cruden and Varnes (1996) (Cruden, 1996). In our work, we chose the polygonal landslide data from this catalog and also carried out post-processing to correspond to the spatial extent and resolution of the 25-meter EU-DEM (Bashfield and Keim, 2011).

The second study area is the Pacific Northwest region of the United States. The data set from the US Pacific Northwest consists of inventories from Oregon's Statewide Landslide Information Database for Oregon (SLIDO-4.4 updated 10/29/2021; Franczyk et al.(Franczyk et al., 2019)), mapped by the Oregon Department of Geology and Mineral Industries (DOGAMI), and the Washington State Landslide Inventory Database (WASLID updated 2018/08/01; Slaughter et al.(Slaughter et al., 2017)), mapped by the Department of Natural Resources, Washington Geological Survey (WGS). The combined inventories comprise $47,653$ landslides from the US Pacific Northwest region. The inventories contain LiDAR-derived landslide polygons guided by protocol to capture the movement types with spatial information on the scarps, head scarps, toes, and deposits (Burns and Madin, 2009; Burns and Mickelson, 2016; Slaughter et al., 2017). Since this data is categorized using a combination of Cruden and Varnes (Cruden, 1996) and Hungr et al. (Hungr et al., 2014b) (i.e., slides, flows, complex, and falls), we modified the Italian data correspondingly to maintain uniformity in the taxonomy of the failure mechanisms.

The EU-DEM for Italy was downloaded from `https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1` and the DEM for the US Pacific Northwest was downloaded from `https://www.opentopography.org/`.

Chapter 7

Additional evidence for landslide topology uncovering its failure mechanism

This chapter is based on objective 4 of the Ph.D. thesis about developing a method for determining the failure mechanism of landslides. Here, we showed additional evidence that the landslide topological properties are better than the geometrical properties for identifying its failure mechanism. The chapter is supplementary material of the paper that will be submitted soon to Nature Communication Journal:

Rana, K.[1], Bhuyan K.[1], Ferrer J.V., Cotton F., Catani F., Ozturk, U., & Malik, N. (2023). Landslide topology uncovers its failure mechanism *(Under review in Nature Communication)*

7.1 Introduction

This supporting information (SI) to the manuscript titled "Landslide topology uncovers failure mechanisms" includes a detailed analysis of landslide topology and its importance in finding the failure mechanisms of landslides. The SI includes an in-depth analysis of the topological features and their probability distributions, quantifying the differences among the different fail-

[1]These authors contributed equally.

ure mechanisms and the connection between the landslide topology and the physical processes. Moreover, we include a section for a detailed evaluation of the model and its transferability analysis. We also show that landslide topological information provides more information about landslide shape than classical geometric information like area, perimeter, and ellipticity, and therefore can be helpful in other landslide research.

7.2 Behavior of different failure mechanisms

The inherent differences between failure mechanisms, notably their kinematic and mechanical behaviors, contribute prominent intricacies to the topography (see surface profiles in Figure 1 in the main manuscript). These intricacies are attributed to slope deformity, interior deformation, kinematic width of failures while propagating down-slope, main scarp deformation, run-out length represented by the debris/earth/soil transportation, and accumulated debris at the talus, and others are captured by topology. The followings are some of the most common failure mechanisms and their various behaviors.

The profile of rotational slides is marked by a conspicuous primary scarp and a distinctive back-tilted bench at the head, but little interior deformation (a schematic view can be seen in Figure 1 in the main manuscript). They are typically slowly moving a large portion of the weak rock mass. At the same time, kinematically rapid planar sliding is marked by the sliding of a rock mass on a planar rupture surface with little to no internal deformation, where the scarp might be separated from the stable rock at deep vertical tension cracks. Typically, they exhibit very compact shapes. Cohesion, c plays an important role in slides, as the degree of internal strength between the particles in a block of material determines the strength and stability along the slip plane. Translational landslides, like the ancient Seimareh slide in Iran's Zagros Mountains, are among the largest and most destructive on Earth (Roberts and Evans, 2013).

Flows are characterized by very rapid movements consisting of saturated granular material on moderate slopes, including liquefaction of materials (in the context of co-seismic triggers) or excess pore pressure (in the context of rainfall triggers) originating from the landslide source.

When the internal friction angle, φ, is low (due to the mixture of solid and fluid particles), less external force is required to instantiate a failure because they are displaced quite easily. Kinematically flow-ish movements are observed with channelized streams and a bulk deposit of debris at the talus (deposition zone), representing highly elliptical, elongated bodies (Evans et al., 2001).

Usually limited in volume, falls (particularly rock falls) exhibit ballistic movements (high velocity, energy, and momentum) that are massively destructive. They detach from cliffs and move at high velocities, either by rolling, falling, or bouncing due to the influence of gravity. The run-out of a rock fall is often shorter and is more likely to travel along a straight path, whereas the run-out of debris flows is longer and can meander and spread out over a wider area (Bourrier et al., 2013).

Complex failures are very hard to describe, as there is an amalgamation of different failure mechanisms occurring at the same time or subsequently, and they can therefore can exhibit multiple characteristics of other mechanisms. For example, irregular debris slides evolving into a debris flow or any other combination of slides, flows, and falls eventually evolving into another movement style can be considered examples of complex failure mechanisms (Cruden, 1996; Hungr et al., 2014b).

Such morphological and geometrical information for each distinct mechanism is theorized to be captured in the topological space by the topological properties, which are then utilized in the machine-learning model to identify the failure mechanisms.

7.3 Topological Features

Persistence diagrams capture the life-death information of structures like connected components, holes, and voids. The persistence diagram consists of a set of $\{(b_i, d_i)\}_{i=1}^{i=N}$ pairs corresponding to each structure type; here, i and N are the indexes of birth-death pairs and the total number of the birth-death pairs. Using the set of $\{(b_i, d_i)\}_{i=1}^{i=N}$ pairs, we can calculate various topological features such as persistence entropy, average lifetime, number of points, Betti curve-based measure, persistence landscape curve-based measure, Wasserstein ampli-

tude, Bottleneck amplitude, Heat kernel-based measure, and landscape image-based measure.

Some of the above topological features can be explained using a lifetime vector that is calculated using a set of $\{(b_i, d_i)\}_{i=1}^{i=N}$ pairs. The lifetime vector $[l_i]_{i=1}^{i=N}$ is calculated as the difference between death and life of the (b_i, d_i) pair ($l_i = d_i - b_i$). The number of points, average lifetime, and persistence entropy are the length, average, and Shannon entropy of the lifetime vector. In comparison, topological features like Bottleneck and Wasserstein's amplitudes quantifying the magnitude of the lifetime vector are p-norm (p=2) and ∞-norm of the lifetime vector, respectively.

The Betti curve-based feature is a p-norm of a 1D discretized betti curve, which is a function ($B(\epsilon) : \mathbb{R} \to \mathbb{Z}$) mapping the persistence diagram to an integer-valued curve, and it counts the number of birth-death pairs at a given ϵ, satisfying the condition $b_i < \epsilon < d_i$ (Garin and Tauzin, 2019). Similarly, a persistence landscape curve-based feature is a p-norm of a 1D discretized persistence landscape curve defined as $\lambda(k, \epsilon) : \mathbb{R} \to \mathbb{R}_+$, where $\lambda(k, \epsilon) = k_{max}\{f_{b_i,d_i}(\epsilon)\}_{i=1}^{i=n}$, k_{max} is the k-th largest value of a set of functions defined by $f_{b_i,d_i}(\epsilon) = max\{0, min(\epsilon - b_i, d_i - \epsilon)\}$ for each (b_i, d_i) pair (Bubenik and Dłotko, 2017).

The heat kernel-based feature is a p-norm (p=2) of the discretized 2D function obtained using the operation of the heat kernel on the persistence diagram. Heat kernel uses a gaussian kernel (σ) and a negative of the gaussian kernel (σ) for each (b_i, d_i) pair and mirror of (b_i, d_i) pair across the diagonal (Reininghaus et al., 2015). In contrast, the persistence image-based feature is a p-norm (p=2) of the discretized 2D function obtained using the operation of the weighted Gaussian kernel on all $(b_i, d_i - b_i)$ pairs in the birth-persistence diagram (Adams et al., 2017). The birth-persistence diagram consists of $(b_i, d_i - b_i)$ pairs where the x-axis shows the birth information, and the y-axis shows the lifetime of the (b_i, d_i) pair.

7.4 Probability Density of Topological Features

The probability density (PDF) of topological features shows that a distinct failure mechanism usually leads to a similar landslide 3D shape each time; for example, slides tend to be more compact than flow failure types. Topological features are calculated using the persistence di-

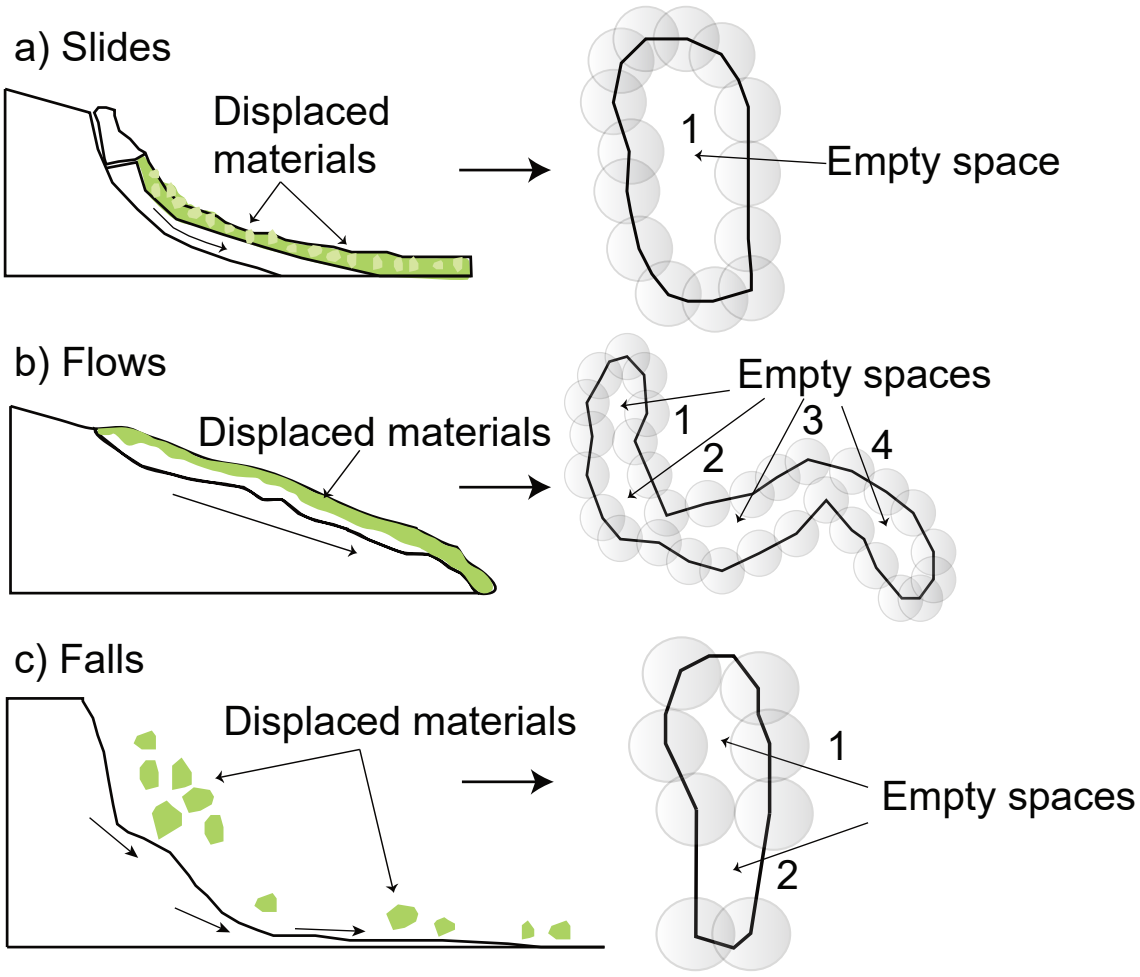

Figure 7.1: The diagram shows a schematic outlook of how empty spaces are created in different landslide types. This illustration is shown using a simplified 2D transformation of a rather complex 3D topological phenomenon for ease of understanding. Sub-plots (a–c) refer to the possible configurations of empty spaces created in the typical polygons of each failure mechanism. Slides tend to have the fewest empty spaces or holes due to their compact shapes, followed by falls. Flow-type failures tend to have multiple numbers of empty spaces due to the sinuous shapes they conjure as they follow the landscape's contour.

agram containing the lifetime information on connected components, holes, and voids. With this, we get information such as the degree of compactness, which is calculated by measuring the amount of empty spaces (holes) in a landslide's morphological footprint (see Figure 7.1).Using a persistence diagram, we can calculate topological features that capture various landslide morphological properties, like compactness, sinuosity, and variations in slope. We visualized the probability density distribution of topological features important for classifying the land-

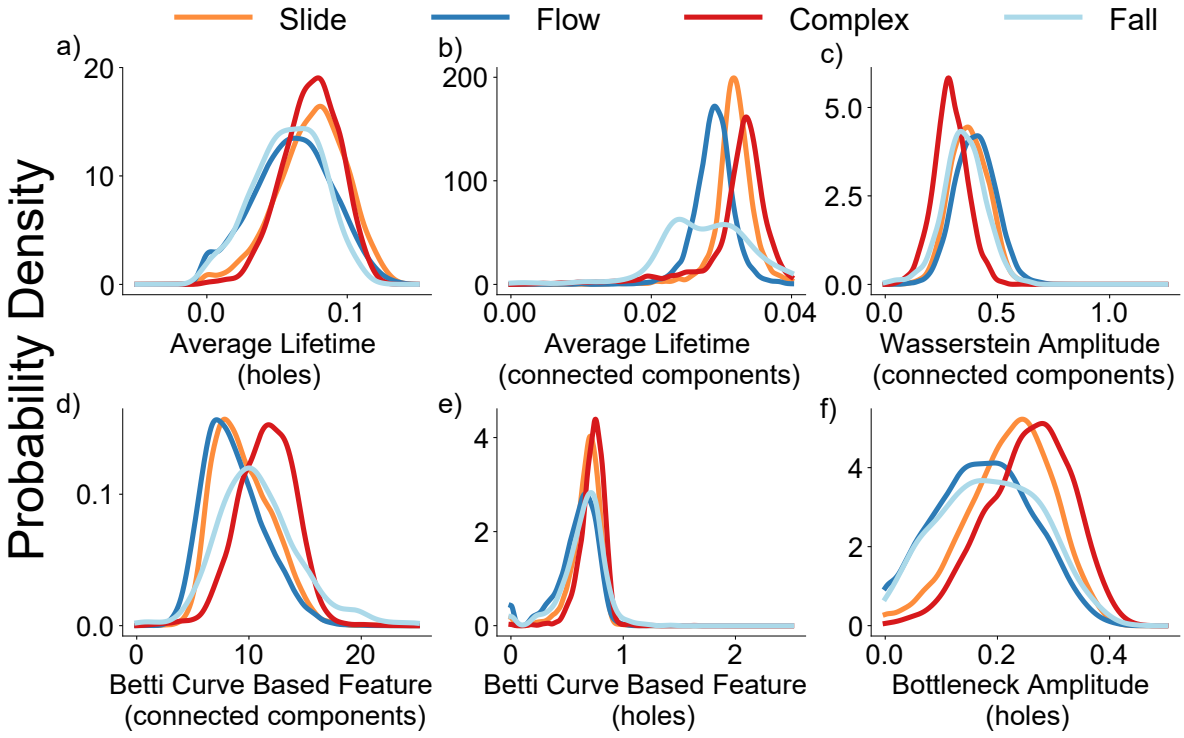

Figure 7.2: The diagram shows the probability distribution functions of the topological features used for classifying failure mechanisms for slides, flows, complex, and falls in the US Pacific Northwest. The y-axis shows the probability density values (calculated using kernel density estimation), and the x-axis shows the value of topological attributes.

slide failure mechanism using the Kernel Density Estimation (KDE) method. The topological features' probability distribution functions (PDF) reveal the differences between the failure mechanisms. Below, we explain some topological features that capture the characteristics of landslide morphology in the US Pacific Northwest and how it is connected to the general movements of different failure mechanisms.

Bottleneck amplitude is the L-∞-norm of the lifetime vector, which is the maximum lifetime of the (birth, death) pair in the persistence diagram. The PDF of the bottleneck amplitude (holes) shows that the probability of getting a bottleneck amplitude (holes) above 0.2 is much higher for slide than for flow, fall, or complex. The PDF of the bottleneck amplitude (holes) shows that slides have a bigger circular shape in the landslide shape than do flows and falls, as can be seen in Figure 7.2-f. However, the PDF peak of the complex landslides for the bottleneck amplitude (holes) appears to be bigger than that of slides, which can be indicative of complex

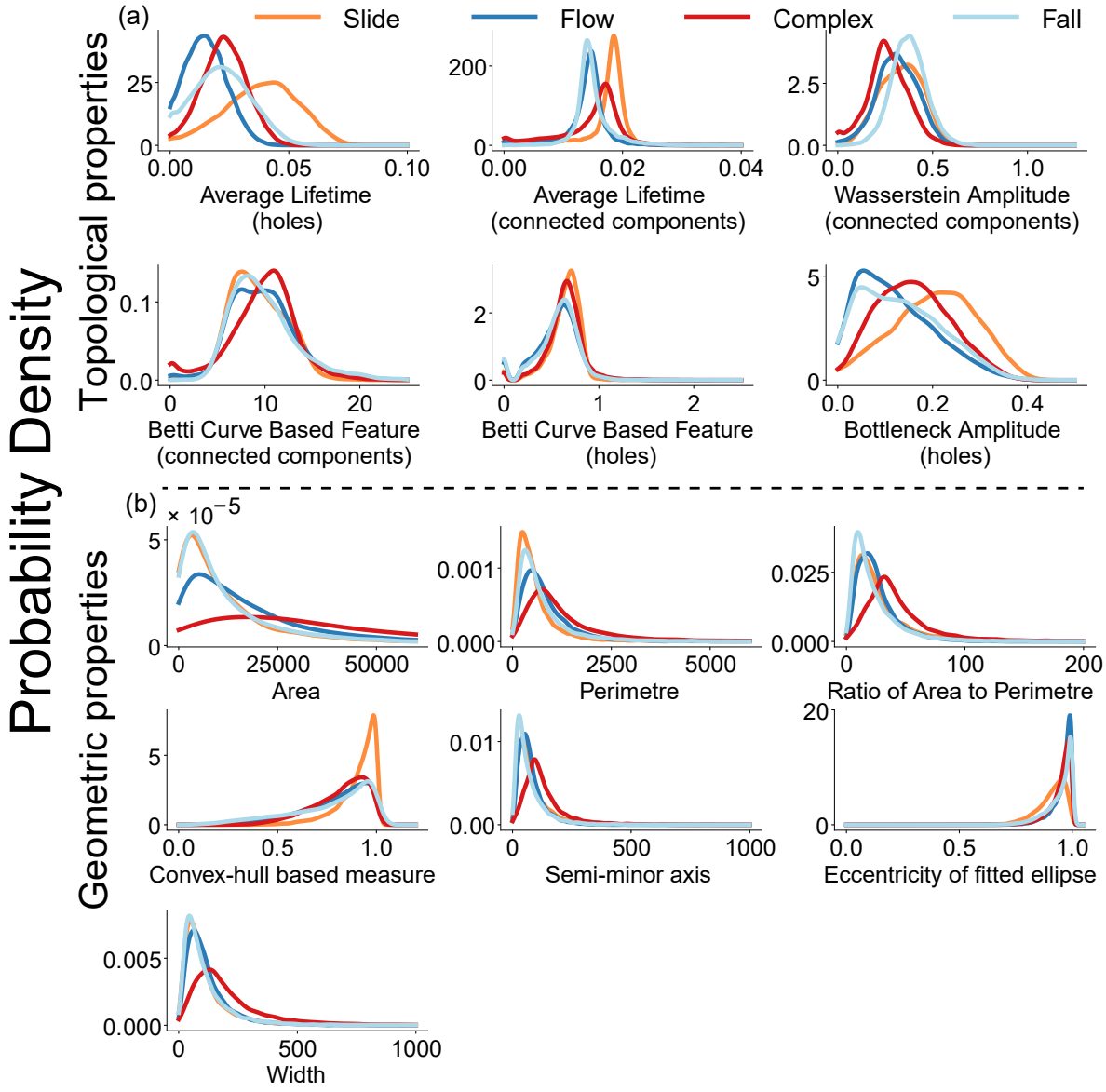

Figure 7.3: Probability distribution functions of the geometrical and the topological features for each failure mechanisms type– slides, flows, complex, and falls–in Italy. The y-axis shows the probability density values (calculated using kernel density estimation), and the x-axis shows the value of topological or geometrical attributes. The geometric features from top to bottom are: area (A), perimeter (P), the ratio of area to perimeter $\frac{A}{P}$, convex hull-based measure (C_h), minor(s_m), and width (W) of the minimum area bounding box fitted to the polygon.

failure types evolving primarily or at the least, predominantly as a sliding mechanism and developing into either a flow or fall in the US Pacific Northwest. Furthermore, as indicated

in the main manuscript and demonstrated in the example of slide and complex failures in the US Pacific Northwest (Figure 7.2-f), they possess longer lives, indicating the presence of more compact morphologies in their footprints. Similarly, the Wasserstein amplitude, which is the p-norm (p==2) of the lifetime vector, shows a similar pattern for the PDF of the different failure mechanisms. The Wasserstein amplitude is 2-norm, which is more influenced more by the largest holes present in the landslide's morphology.

Average life (holes) is the average life of (birth, death) pairs of holes and will be influenced by the size (small or large) and the number of holes. The PDF of the Average life (holes) shows that slides and complex have, on average, larger holes in the landslide shape compared to flows and falls. Also, the PDF peak of the flow and fall failures for Average lifetime (holes) decays quickly, which shows that all the holes in the flow are smaller. This can also be attributed to their rather sinuous footprint which stems from their kinematic progression as they propagate following the landscape's contours. In particular, as flows propagate in a meandering fashion across the landscape, they conjure multiple shorter lifetime holes (refer to Figure 7.1-b) that indicates high degree of sinuosity.

7.5 Geometric versus Topological Features

Geometric properties define an object's shape and size, but topological properties explain the connections and topological interactions among its parts. Geometric properties such as area, perimeter, convexity, and ellipticity define the physical dimensions of a landslide, whereas topological properties such as average lifetime of holes, Betti curve, and landscape curve describe the connections and interactions of the soil and rock masses, the width of kinematic propagation, and depth of failure in a landslide. Geometric properties are, however, sensitive to any changes made to the original shapes of the geometry and therefore, more susceptible to drastically changing the geometric property values. For example, any change to a landslide's boundary/body would inadvertently change each of the values of the geometric properties like area, perimeter, convexity, etc., but the same cannot be said for topology as it relies on the number of voids that are generated based on the overall shape of the landslide body.

This is even more pronounced upon investigating the landslides in 3D. Since these geometric properties cannot be broadcasted to 3D, much information related to variational changes in the topography (attributed by elevation and slope) is lost. As TDA captures this 3D information and utilizes it when engineering topological features, intricate information on landslides such as depth of failure, deformations pattern, the width of kinematic progression are well recorded. Therefore, to assess and evaluate the differences between the classical and topological properties, we compare them in this section. This comparison was based on KDE plots that represent the PDFs of the samples for each failure mechanism. We also plotted box plots to compare the median values and distribution of said values between the geometric and topological properties. As we see in Figure 7.3-b, the PDFs of the failure types are very similar to each other, specifically when looking at the ellipticity, semi-major axis, perimeter, and width. However, when comparing them to the topological properties Figure 7.3-a, we observe that the PDFs of the failure types are more dissimilar to each other under each property (e.g., the average lifetime of holes, bottleneck amplitude of holes, Wasserstein amplitude of holes). This can be the reason why the random forest models show promising results, as the PDFs are dissimilar enough to find evident differences between each failure type when using the topological properties/features.

7.6 Model Transferability Analysis

As we see in Figure 7.4, sharing samples from two different study areas could not drastically increase the overall performance compared to simply training the model independently in a particular area (like the US Pacific Northwest). We achieved an overall F1-score of ∼65% in identifying the landslide failure mechanisms in the US Pacific Northwest. This is reasonable, given the differences in climatic, geomorphometric, and topographical conditions between these systems, which are not only because they are geographically separated but also because landslides in these two parts of the world behave differently even while having similar failure mechanisms.

This is actually quite a comprehensible problem, as both regions have distinct and di-

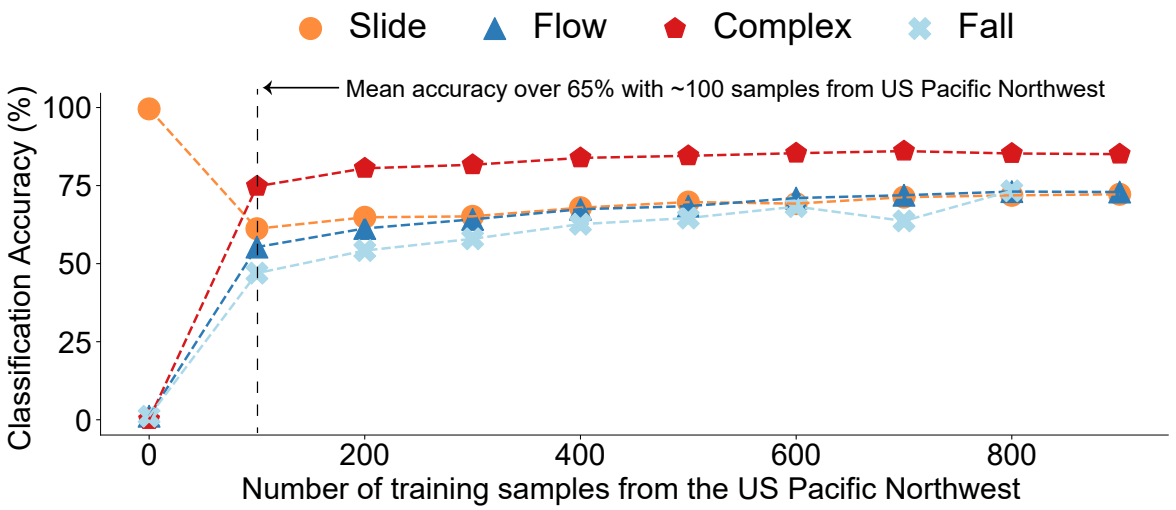

Figure 7.4: The figure shows the classification accuracy when the model is trained on just the Italian data and tested to identify failure mechanisms in the US Pacific Northwest. The performance is analyzed by investigating first with zero samples and then incrementally increasing sample populations from the US Pacific Northwest data.

verse geologies; for example, the US Pacific Northwest is distinguished by young volcanic rocks and strong tectonic activity, whereas Italy is distinguished by steep hilly terrain and old rock formations, as well as high tectonic activity. These geological differences can influence soil and rock stability and, therefore, the probability of landslide occurrence in these two contexts.

Now, let us consider the elevation as one factor from a topographical point of view, and let us take a look at similar failure mechanisms occurring in the same geography but at different elevations. Debris flows at different elevations can vary as the size and content of the sediment change accordingly, which can impact the density and viscosity of the flow and hence define its mechanical and kinematic properties (Roelofs et al., 2022). A debris flow behavior can be affected by factors such as slope angle, sediment size and composition, and water content, which can change with elevation. In the context of the Himalayas, debris flows at higher elevations may have a higher amount of fine glacial sediment, such as clay and silt, rendering the flow more thick and viscous, and hence more difficult to erode and transport, whereas debris flows on gentler slopes may have a greater amount of coarse alluvial sediment, such as gravel and sand, making the flow more laminar and less thick and viscous, and thus, simpler

Italian Dataset

Predicted Class

Actual Class	Slide	Flow	Complex	Fall
Slide	**98.60**	0.00	1.18	0.22
Flow	0.0	**94.77**	1.70	3.53
Complex	0.66	1.99	**96.39**	0.96
Fall	0.15	2.80	0.96	**96.11**

Metrics	Slide	Flow	Complex	Fall
TPR	98.50	94.69	96.27	96.00
TNR	99.69	98.38	98.67	98.40
FPR	0.30	1.61	1.32	1.59
FNR	1.49	5.30	3.73	3.97
F1-Score	**98.79**	**94.90**	**96.16**	**95.64**

USA Dataset

Predicted Class

	Slide	Flow	Complex	Fall
Slide	**85.06**	6.90	4.60	3.45
Flow	11.63	**82.56**	0.0	5.81
Complex	3.45	0.0	**86.20**	10.34
Fall	4.60	3.49	17.24	**74.72**

	Slide	Flow	Complex	Fall
TPR	83.18	80.76	84.62	73.42
TNR	93.13	96.00	92.10	92.76
FPR	6.87	4.0	7.89	7.24
FNR	16.82	19.24	15.38	26.58
F1-Score	**81.64**	**83.79**	**81.25**	**75.75**

Figure 7.5: The figure shows the confusion matrix and the associated accuracy metrics of the random forest model for the data of both Italy and the US Pacific Northwest.

to erode and transport. This can impact the debris flow's characteristics, such as its velocity, turbulence, and propensity to erode the channel and transfer sediment. These changes will be captured in the propagation of the debris flows, which changes their topology, thereby contributing to different properties that are picked up by the TDA. Such apparent differences guided by the topography can lead to differences in similar failures between Italy and the US Pacific Northwest.

Another example can be based on climatic conditions. Italy is characterized by the Mediterranean climate, which features high precipitation in autumn and winter seasons and dry summers. Torrential downpours in the autumn and winter months can affect the soil saturation, causing the soil and rock to be less stable and prone to landslides. This is especially true in

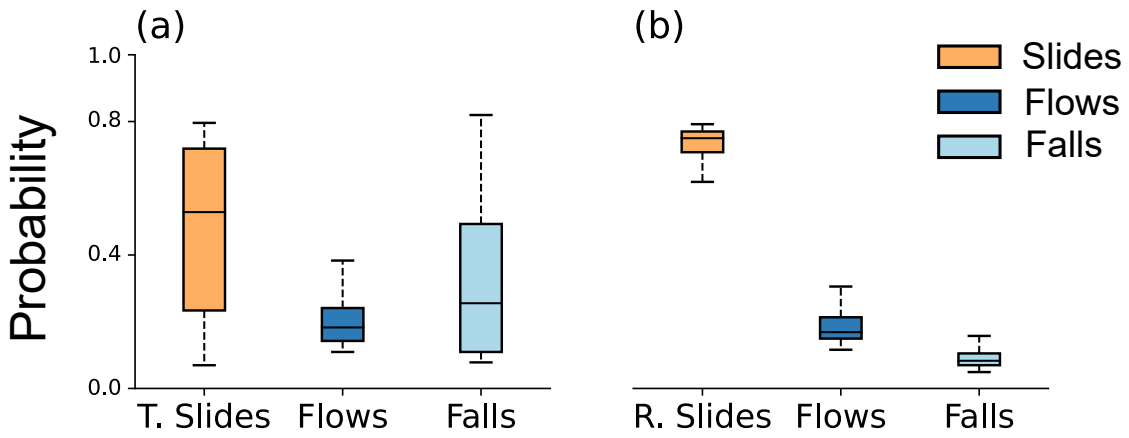

Figure 7.6: The diagram shows the probability of complex landslides belonging to each of the failure types class (slide, flow, and fall) as predicted by the model. Box plot (a) shows the complex landslide samples that occur from "Translational rock slides followed by rock falls" as documented in the Statewide Landslide Information Database for Oregon (SLIDO). Model predictions indicate sliding mechanisms to be the predominant type of failure, which are most likely translational slides according to SLIDO. Similarly, box plot (b) shows the complex landslide samples that occur from "Rotational slides followed by flows" as recorded in SLIDO. Model predictions indicate sliding mechanisms to be the predominant type of failure, which most likely rupture rotationally. Beige bars illustrate sliding mechanisms while bars with darker and lighter shades of blue illustrate flows and fall mechanisms, respectively. Note the number of annotated complex failures with behavioral definitions by SLIDO in box plot (a) constitutes 198 samples and box plot (b) constitutes 230 samples.

locations where the soil is already saturated from the rainy summers, as the soil will be unable to take any further water. High precipitation in the autumn and winter months can also form thick, saturated topsoil that is susceptible to landslides. When soil becomes saturated, it loses strength, and soil particles move more easily, resulting in landslides. The Pacific Northwest has a maritime climatic system with mild winters and close proximity to the Pacific Ocean. This proximity to the ocean can lead to increased coastal erosion and landslides, particularly in areas where the coastline is steep and composed of unstable rock. Therefore, the conditions of the triggering mechanism and their inherent differences can lead to different kinematic and mechanical properties of similar failures, which constitute different behaviors.

Considering these observations, we can gauge that properties of similar failures can bear different characteristics, especially if they are geographically separate, wherein climate and topography (as examples) can massively influence their kinematic propagation.

7.7 Other measures to evaluate model performance

In order to evaluate the performance of the method, we also calculated the confusion matrix and other accuracy metrics like the True Positive Rate, True Negative Rate, False Positive Rate, False Negative Rate, and the F1-score.

The True positive rate (also known as Sensitivity, Recall) (equation 1) and true negative rate (also known as Specificity) (equation 2) are performance metrics that are used to assess a model's accuracy in accurately detecting positive and negative instances. The number of genuinely negative instances identified as positive by the model are known as false positives (FP) (equation 3). The number of cases that are genuinely positive but are categorized as negative by the model are known as false negatives (FN) (equation 4).

$$TPR\ (= Recall) = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{7.7.1}$$

$$TNR = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \tag{7.7.2}$$

$$FPR = \frac{False\ Positives}{True\ Negatives + False\ Positives} \tag{7.7.3}$$

$$FNR = \frac{False\ Negatives}{True\ Positives + False\ Negatives} \tag{7.7.4}$$

The F1-score (equation 6) is the harmonic mean of precision (equation 5) and recall (equation 1), and it is used to balance the precision-to-recall trade-off. Precision is the number of correct positive predictions produced by the model out of all correct positive predictions made by the model, and recall is the number of correct positive predictions made out of all correct positive occurrences.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{7.7.5}$$

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{7.7.6}$$

In Figure 7.5, we see the confusion matrix and the respective scores of the TPR, TNR,

FPR, FNR, and F1-score of both Italy and the US Pacific Northwest.

Chapter 8

Conclusion and future work

8.1 Conclusion

Landslide predictive models, such as susceptibility and hazards, aim to mitigate losses by locating landslide-prone regions and forecasting landslides. However, the effectiveness of these models relies on the quality of the landslide databases, which often miss triggering event and failure mechanism information, crucial information for landslide predictive modeling. Also, newly acquired landslide databases lack landslide trigger and failure information due to the automated data-capturing process.

Combining different failure mechanisms into one group leads to biases as each failure type has different geological and geotechnical properties. Similarly, using earthquake-triggered landslides for predictive rainfall models or vice-versa leads to biases in the model as they have distinct geological and geotechnical properties and are triggered by different physical processes. These biases could be dangerous; for example, missing high landslide-prone regions or predicting a low probability of landslide occurrences in high landslide-prone areas. This crucial but missing information in landslide databases motivated us to develop methods to identify landslides' triggering and failure mechanisms.

Landslide databases often contain polygons information, which outlines the landslide's 2D shape. Existing studies show that physical processes are embedded in the general morphology of the landslides. Inspired by these studies and the availability of polygons in landslide

databases, we wanted to investigate how far the trigger and failure mechanism is ingrained in the landslide morphology. For identifying landslide trigger information, we developed three different methods for landslide trigger classification based solely on landslide polygon shapefiles containing landslides' two-dimensional (2D) polygon shapes. The first method uses geometric properties of landslide polygons as a feature space for a machine learning classifier–random forest. In the second method, we transformed these 2D shapes into three-dimensional (3D) point clouds by incorporating the digital elevation data and then extracting landslides' topological properties by topological data analysis (TDA) of these 3D points clouds; and to classify landslides; we treat these topological properties as the feature space of a random forest classifier. The third method uses images of landslides as input to a convolutional neural network (CNN). Each of the methods showed above 80% performance in six landslide databases spread over the Japanese archipelago.

To identify the failure mechanism, we explored various geometric and topological properties of landslide shape and found that topological properties are excellent predictors for classifying landslide failure mechanisms. Therefore, we developed a method for determining landslide failure types using landslide topology. First, we extracted the topological features of the landslide 3D shape using Topological Data Analysis and then fed these features as an input to the machine learning algorithm–random forest. Finally, we separately implemented the developed method on the Italian and US data. The technique achieved above 95 and 80 accuracies for each landslide failure type for the Italian and US data sets.

The presented landslide-triggering methods are easy to use as they depend solely on the landslide polygon and show robust performance. Furthermore, the geometric-based method shows that triggering information is embedded in a landslide polygon shape. We also found seven geometric properties, which are excellent predictors for identifying trigger information. Moreover, we found that using the topological properties of the landslide 3D shape provides better results than methods based on the landslide polygon shape, which is the 2D outline of the landslide shape. Even though the triggering mechanism methods are limited to the Japanese archipelago, our methods are portable to different world regions provided training

and testing data are from similar climatic and tectonic areas. We included all these methods in a python library, "`landsifier`," that provides all the methods for finding triggers and other valuable functions like downloading DEMs corresponding to an inventory region. We made the `landsifier` library publicly available with detailed documentation and implementation of the library.

The proposed landslide failure mechanism methods solely depend on landslide polygon and are portable to different regions. By leveraging the topology of landslides, the method can efficiently classify landslide failure types in existing and new databases and support experts in generating quality inventories. The study shows that landslide topology provides more information than landslide geometry for identifying the failure mechanism. We also connect topological properties to the physical process, which could be helpful in other landslide research related to its shape. Furthermore, the method shows strong performance even with fewer training samples and will be helpful in regions with fewer training samples. Even in areas without training samples, the landslide experts can manually label a few landslide samples to be used in the developed method to predict the failure mechanism of more extensive databases.

8.2 Future Work

In this Ph.D. thesis, we developed methods to determine the triggering and failure mechanisms of landslide databases. The presented methods show strong performance and robustness in the testing regions with numerous annotated landslide samples with triggering and failure information. However, the transferability of the techniques by training and testing the approach in geographically far apart areas has yet to be explored. Such a case is of immense importance in regions lacking annotated training samples to train the technique. In such scenarios, we can use annotated landslide samples from high-quality, data-rich areas to train the methods and test them on the regions lacking triggering and failure information.

Our work identifies the trigger mechanism of the earthquake and rainfall-triggered landslides and excludes anthropogenic landslides. In future work, one can include anthropogenic landslides as another triggering mechanism type. Identifying the anthropogenic landslides will

help understand landslides caused by human intervention on the landscape and help develop separate predictive models for them. Moreover, the developed method could identify the large landslides' (area $>$ 90,000 meters-square) trigger and failure mechanism, which are the most deadly because of their vast impact area.

The efficacy of predictive models depends on crucial information, such as the triggering mechanism, which is often missing in the landslide database. Separate predictive models are required for earthquake and rainfall landslides as distinct physical mechanisms cause them. For example, rainfall pattern in a region is crucial information for rainfall landslide predictive models, whereas seismic zone information is essential for earthquake-triggered landslides. Our methods will aid in identifying the trigger mechanism of past landslide databases that will make these databases useful for predictive modeling and enhance their efficiency. Identifying the trigger mechanism will reduce the bias and uncertainty in predictive models that can occur by accidentally using earthquake landslides for predictive rainfall-fed landslide models and vice-versa. In addition, identifying the failure mechanism can also aid in developing more finer predictive models for different failure and triggering mechanisms, .e.g, predictive models for earthquake-triggered slide landslides. In future work, landslide researchers can implement our methods to identify the trigger and failure mechanisms of landslides and develop finer predictive models with higher performance.

Our methods have the potential to identify the regions that are prone to landslides, as topological data analysis-based methods can also capture the variations in the slope of a given landscape. In the future, we envision the use of topological information of the regions to determine their proneness to landslides by training the methods on both landslide and non-landslide regions. Moreover, our methods can potentially identify other geophysical processes like badlands, permafrost-borne thaw slumps, and marine landslides.

Our presented methods have huge potential for improving the efficacy of landslide predictive models and are helpful for various landslide research topics like large landslides. Therefore, we anticipate that the landslide community and modelers will find our methods helpful in their line of research in understanding the landslide process and in determining the trigger and fail-

ure mechanism of the landslides.

Bibliography

Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L. (2017). Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18. 4.4.2, 6.5.1, 7.3

Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. Ieee. 4.4.3, 4.4.3

Aljazeera. Deadly landslide strikes southern peru amid heavy rainfall. *AL JAZEERA AND NEWS AGENCIES*. 6.2

Amato, G., Palombi, L., and Raimondi, V. (2021). Data–driven classification of landslide types at a national scale by using artificial neural networks. *International Journal of Applied Earth Observation and Geoinformation*, 104:102549. 1.7, 4.7, 6.2, 6.3.1

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the national academy of sciences*, 99(10):6562–6566. 2.4, 3.2

Arabameri, A., Chandra Pal, S., Rezaie, F., Chakrabortty, R., Saha, A., Blaschke, T., Di Napoli, M., Ghorbanzadeh, O., and Thi Ngo, P. T. (2021). Decision tree based ensemble machine learning approaches for landslide susceptibility mapping. *Geocarto International*, pages 1–35. 5.1, 6.5.2

Aurisano, A., Radovic, A., Rocco, D., Himmel, A., Messier, M., Niner, E., Pawloski, G., Psihas, F., Sousa, A., and Vahle, P. (2016). A convolutional neural network neutrino event classifier. *Journal of Instrumentation*, 11(09):P09001. 4.4.3

Azar, A. T., Elshazly, H. I., Hassanien, A. E., and Elkorany, A. M. (2014). A random forest classifier for lymph diseases. *Computer methods and programs in biomedicine*, 113(2):465–473. 5.1, 5.1, 5.1, 6.5.2

Barlow, J., Franklin, S., and Martin, Y. (2006). High spatial resolution satellite imagery, dem derivatives, and image segmentation for the detection of mass wasting processes. *Photogrammetric Engineering and Remote Sensing*, 72(6):687 – 692. Cited by: 102; All Open Access, Hybrid Gold Open Access. 1.7, 6.2

Barnett, I., Malik, N., Kuijjer, M. L., Mucha, P. J., and Onnela, J.-P. (2019). Endnote: Feature-based classification of networks. *Network Science*, 7(3):438–444. 2.4, 5.1, 6.5.2

Bashfield, A. and Keim, A. (2011). Continent-wide dem creation for the european union. In *34th international symposium on remote sensing of environment. the GEOSS era: Towards operational environmental monitoring. sydney, australia*, pages 10–15. 6.6

Behling, R., Roessner, S., Golovko, D., and Kleinschmit, B. (2016). Derivation of long-term spatiotemporal landslide activity—A multi-sensor time series approach. *Remote Sensing of Environment*, 186:88–104. 4.2

Behling, R., Roessner, S., Segl, K., Kleinschmit, B., and Kaufmann, H. (2014). Robust automated image co-registration of optical multi-sensor time series data: Database generation for multi-temporal landslide detection. *Remote Sensing*, 3:2572–2600. 2.2, 4.2

Belgiu, M. and Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114:24–31. 5.1, 5.1, 6.5.2

Bhattacharya, B. and Solomatine, D. P. (2006). Machine learning in soil classification. *Neural networks*, 19(2):186–195. 4.2

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095. 2.4, 5.1

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227. 2.4, 5.1, 6.5.2

Bíl, M., Raška, P., Dolák, L., and Kubeček, J. (2021). Childa–czech historical landslide database. *Natural Hazards and Earth System Sciences*, 21(8):2581–2596. 4.2

Bourrier, F., Dorren, L., and Hungr, O. (2013). The use of ballistic trajectory and granular flow models in predicting rockfall propagation. *Earth Surface Processes and Landforms*, 38(4):435–440. 1.5.2, 6.2, 7.2

Bradley, K., Mallick, R., Andikagumi, H., Hubbard, J., Meilianda, E., Switzer, A., Du, N., Brocard, G., Alfian, D., Benazir, B., et al. (2019). Earthquake-triggered 2018 palu valley landslides enabled by wet rice cultivation. *Nature Geoscience*, 12(11):935–939. 6.2

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. 2.4, 3.2, 5.1, 6.5.2

Broeckx, J., Rossi, M., Lijnen, K., Campforts, B., Poesen, J., and Vanmaercke, M. (2020). Landslide mobilization rates: A global analysis and model. *Earth-Science Reviews*, 201:102972. 6.2

Bubenik, P. and Dłotko, P. (2017). A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation*, 78:91–114. 4.4.2, 6.5.1, 7.3

Burns, W. J. and Madin, I. (2009). Protocol for inventory mapping of landslide deposits from light detection and ranging (lidar) imagery. 6.6

Burns, W. J. and Mickelson, K. A. (2016). Protocol for deep landslide susceptibility mapping. 6.6

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308. 4.4.2, 6.5.1

Carlsson, G. (2020). Topological methods for data modelling. *Nature Reviews Physics*, 2(12):697–708. 6.2

Catani, F., Casagli, N., Ermini, L., Righini, G., and Menduni, G. (2005). Landslide hazard and risk mapping at catchment scale in the arno river basin. *Landslides*, 2:329–342. 6.3.1

Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28. 2.4, 3.2

Chaudhary, A., Kolhe, S., and Kamal, R. (2016). An improved random forest classifier for multi-class classification. *Information Processing in Agriculture*, 3(4):215–222. 5.1, 6.5.2

Corominas, J., van Westen, C., Frattini, P., Cascini, L., Malet, J.-P., Fotopoulou, S., Catani, F., Van Den Eeckhaut, M., Mavrouli, O., Agliardi, F., et al. (2014). Recommendations for the quantitative analysis of landslide risk. *Bulletin of engineering geology and the environment*, 73:209–263. 6.2

Craymer, L. and Feast, L. New zealand declares national emergency as cyclone gabrielle wreaks havoc. *Reuters*. 6.2

Crozier, M. J. and Glade, T. (2005). Landslide hazard and risk: issues, concepts and approach. *Landslide hazard and risk*, pages 1–40. 1.4

Cruden, D. M. (1996). Cruden, dm, varnes, dj, 1996, landslide types and processes, transportation research board, us national academy of sciences, special report, 247: 36-75. *Landslides Eng. Pract*, 24:20–47. 1.5.2, 6.4, 6.6, 7.2

Cruden, M. D. and Varnes, J. D. (1996). Landslide Types and Processes. In *Landslides: Investigation and Mitigation*, number 247 in Special Report(National Research Council (U.S.) Transportation Research Board), pages 36–75. National Academy Press, Washington, D.C. 4.2, 4.7

Cui, P., Chen, X.-Q., Zhu, Y.-Y., Su, F.-H., Wei, F.-Q., Han, Y.-S., Liu, H.-J., and Zhuang, J.-Q. (2011). The wenchuan earthquake (may 12, 2008), sichuan province, china, and resulting geohazards. *Natural Hazards*, 56:19–36. 1.5.1

Dai, F., Lee, C. F., and Ngai, Y. Y. (2002). Landslide risk assessment and management: an overview. *Engineering geology*, 64(1):65–87. 1.2

Depicker, A., Jacobs, L., Mboga, N., Smets, B., Van Rompaey, A., Lennert, M., Wolff, E., Kervyn, F., Michellier, C., Dewitte, O., et al. (2021). Historical dynamics of landslide risk from population and forest-cover changes in the kivu rift. *Nature Sustainability*, 4(11):965–974. 4.2

Devarakonda, M. V. (2022). Data-driven approaches to selecting samples for training neural networks. In *System Dependability and Analytics: Approaching System Dependability from Data, System and Analytics Perspectives*, pages 327–345. Springer. 6.3.4

Devoli, G., Strauch, W., Chávez, G., and Høeg, K. (2007). A landslide database for nicaragua: a tool for landslide-hazard management. *Landslides*, 4:163–176. 1.5.1

Dietze, M., Mohadjer, S., Turowski, J. M., Ehlers, T. A., and Hovius, N. (2017). Seismic monitoring of small alpine rockfalls–validity, precision and limitations. *Earth Surface Dynamics*, 5(4):653–668. 6.4

Dille, A., Dewitte, O., Handwerger, A. L., d'Oreye, N., Derauw, D., Ganza Bamulezi, G., Ilombe Mawe, G., Michellier, C., Moeyersons, J., Monsieurs, E., et al. (2022). Acceleration of a large deep-seated tropical landslide due to urbanization feedbacks. *Nature Geoscience*, pages 1–8. 6.4

Dilley, M. (2005). *Natural disaster hotspots: a global risk analysis*, volume 5. World Bank Publications. 1.2

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87. 4.2

ESRI (2020). Arcgis - world shaded relief. 6.2

Evans, S., Bovis, M., and Hutchinson, J. (2001). Landslides of the flow type. *Environmental & Engineering Geoscience*, 7(3):221–238. 7.2

Fan, X., van Westen, C. J., Korup, O., Gorum, T., Xu, Q., Dai, F., Huang, R., and Wang, G. (2012). Transient water and sediment storage of the decaying landslide dams induced by the 2008 wenchuan earthquake, china. *Geomorphology*, 171:58–68. 4.2

Fan, X., Yunus, A. P., Scaringi, G., Catani, F., Siva Subramanian, S., Xu, Q., and Huang, R. (2021). Rapidly evolving controls of landslides after a strong earthquake and implications for hazard assessments. *Geophysical Research Letters*, 48(1):e2020GL090509. 6.2, 6.4

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., et al. (2007). The shuttle radar topography mission. *Reviews of geophysics*, 45(2). 5.2.2

Francis, O., Fan, X., Hales, T., Hobley, D., Xu, Q., and Huang, R. (2022). The fate of sediment after a large earthquake. *Journal of Geophysical Research: Earth Surface*, 127(3):e2021JF006352. 6.4

Franczyk, J., Burns, W., and Calhoun, N. (2019). Statewide landslide information database for oregon release-4.0, slido 4.0. 6.3.5, 6.6

Fressard, M., Thiery, Y., and Maquaire, O. (2014). Which data for quantitative landslide susceptibility mapping at operational scale? case study of the pays d'auge plateau hillslopes (normandy, france). *Natural Hazards and Earth System Sciences*, 14(3):569–588. 1.1, 6.2

Frey-Martínez, J., Cartwright, J., and James, D. (2006). Frontally confined versus frontally emergent submarine landslides: A 3d seismic characterisation. *Marine and Petroleum Geology*, 23(5):585–604. 6.4

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York. 2.4, 3.2, 3.3, 5.1

Froude, M. J. and Petley, D. N. (2018). Global fatal landslide occurrence from 2004 to 2016. *Natural Hazards and Earth System Sciences*, 18(8):2161–2181. 1.1, 1.2, 6.2

Gariano, S. L. and Guzzetti, F. (2016). Landslides in a changing climate. *Earth-Science Reviews*, 162:227–252. 6.2, 6.4

Garin, A. and Tauzin, G. (2019). A topological" reading" lesson: Classification of mnist using tda. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1551–1556. IEEE. 4.4.2, 4.4.2, 6.5.1, 7.3

Geertsema, M., Highland, L., and Vaugeouis, L. (2009). Environmental impact of landslides. In *Landslides–disaster risk reduction*, pages 589–607. Springer. 1.2

Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S. R., Tiede, D., and Aryal, J. (2019). Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sensing*, 11(2):196. 4.2

Gillies, S. (2013). The shapely user manual. 4.4.1

Gorum, T., Korup, O., van Westen, C. J., van der Meijde, M., Xu, C., and van der Meer, F. D. (2014). Why so few? landslides triggered by the 2002 denali earthquake, alaska. *Quaternary Science Reviews*, 95:80–94. 4.2

Guo, T., Dong, J., Li, H., and Gao, Y. (2017). Simple convolutional neural network on image classification. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pages 721–724. IEEE. 4.4.3, 4.4.3, 4.4.3

Guzzetti, F., Mondini, A. C., Cardinali, M., Fiorucci, F., Santangelo, M., and Chang, K.-T. (2012). Landslide inventory maps: New tools for an old problem. *Earth-Science Reviews*, 112(1-2):42–66. 4.2, 6.2

Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., and Galli, M. (2006). Estimating the quality of landslide susceptibility models. *Geomorphology*, 81(1-2):166–184. 6.2

Han, Y. and Bao, X. (2022). Topological mapping of complex networks from high slope deformation time series for landslide risk assessment. *Expert Systems with Applications*, 206:117816. 6.2

Havenith, H., Torgoev, A., Braun, R. S. A., Torgoev, I., and Ischuk, A. (2015). Tien shan geohazards database: Landslide susceptibility analysis. *Geomorphology*, 249(1):32–43. 2.2

Havenith, H.-B. (2002). Site effect analysis around the seismically induced ananevo rockslide, kyrgyzstan. *Bulletin of the Seismological Society of America*, 92(8):3190. 2.6

Hensel, F., Moor, M., and Rieck, B. (2021). A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4:52. 4.4.2, 6.5.1

Hervás, J. and Bobrowsky, P. (2009). Mapping: inventories, susceptibility, hazard and risk. In *Landslides–disaster risk reduction*, pages 321–349. Springer. 1.4

Hong, Y., Adler, R., and Huffman, G. (2007). Use of satellite remote sensing data in the mapping of global landslide susceptibility. *Natural hazards*, 43(2):245–256. 1.4

HR.Pourghasemi, Moradi, H., Aghda, S., EA, S., Jirandeh, A., and Pradhan, B. (2014). Assessment of fractal dimension and geometrical characteristics of the landslides identified in north of tehran, iran. *Environmental earth sciences*, 71(1):3617–3626. 1.7, 2.2, 2.4

Huang, F., Xiong, H., Yao, C., Catani, F., Zhou, C., and Huang, J. (2023). Uncertainties of landslide susceptibility prediction considering different landslide types. *Journal of Rock Mechanics and Geotechnical Engineering*. 6.2, 6.4

Hungr, O., Leroueil, S., and Picarelli, L. (2014a). The varnes classification of landslide types, an update. *Landslides*, 11(2):167–194. 1.5, 2.6

Hungr, O., Leroueil, S., and Picarelli, L. (2014b). The varnes classification of landslide types, an update. *Landslides*, 11(2):167–194. 1.5.2, 2.6, 6.6, 7.2

Imaizumi, F., Hayakawa, Y. S., Hotta, N., Tsunetaka, H., Ohsaka, O., and Tsuchiya, S. (2017). Relationship between the accumulation of sediment storage and debris-flow characteristics in a debris-flow initiation zone, ohya landslide body, japan. *Natural Hazards and Earth System Sciences*, 17(11):1923–1938. 1.3

Jibson, R. W., Harp, E. L., Schulz, W., and Keefer, D. K. (2004). Landslides triggered by the 2002 denali fault, alaska, earthquake and the inferred nature of the strong shaking. *Earthquake spectra*, 20(3):669–691. 1.5.1

Jones, J. N., Boulton, S. J., Bennett, G. L., Stokes, M., and Whitworth, M. R. (2021). Temporal variations in landslide distributions following extreme events: Implications for landslide susceptibility modeling. *Journal of Geophysical Research: Earth Surface*, 126(7):e2021JF006067. 4.2

Kasai, M. and Yamada, T. (2019). Topographic effects on frequency-size distribution of landslides triggered by the hokkaido eastern iburi earthquake in 2018. *Earth, Planets and Space*, 71(1):89. 2.4, 2.6

Katz, O., Morgan, J. K., Aharonov, E., and Dugan, B. (2014). Controls on the size and geometry of landslides: Insights from discrete element numerical simulations. *Geomorphology*, 220:104–113. 6.4

Kenigsberg, A. R., Rivière, J., Marone, C., and Saffer, D. M. (2020). Evolution of elastic and mechanical properties during fault shear: The roles of clay content, fabric development, and porosity. *Journal of Geophysical Research: Solid Earth*, 125(3):e2019JB018612. 6.3.1

Kirschbaum, D. B., Adler, R., Hong, Y., Hill, S., and Lerner-Lam, A. (2010). A global landslide catalog for hazard applications: method, results, and limitations. *Natural Hazards*, 52:561–575. 6.2

Kjekstad, O. and Highland, L. (2009). Economic and social impacts of landslides. In *Landslides–disaster risk reduction*, pages 573–587. Springer. 1.3

Klose, M., Maurischat, P., and Damm, B. (2016). Landslide impacts in germany: A historical and socioeconomic perspective. *Landslides*, 13(1):183–199. 6.2

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25. 4.4.3

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*, volume 26. Springer. 2.4, 3.2, 5.1

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer. 5.1, 6.5.2

Kursa, M. B. (2014). Robustness of random forest-based gene selection methods. *BMC bioinformatics*, 15(1):8. 2.4, 5.1, 6.5.2

LaHusen, S. R., Duvall, A. R., Booth, A. M., Grant, A., Mishkin, B. A., Montgomery, D. R., Struble, W., Roering, J. J., and Wartman, J. (2020). Rainfall triggers more deep-seated landslides than cascadia earthquakes in the oregon coast range, usa. *Science Advances*, 6(38):eaba6790. 6.3.5

Lee, C.-F., Huang, W.-K., Chang, Y.-L., Chi, S.-Y., and Liao, W.-C. (2018). Regional landslide susceptibility assessment using multi-stage remote sensing data along the coastal range highway in northeastern taiwan. *Geomorphology*, 300:113–127. 1.4

Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., and Chen, M. (2014). Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, pages 844–848. IEEE. 4.4.3, 4.4.3

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22. 2.4, 5.1

Lima, P., Steger, S., Glade, T., and Mergili, M. (2023). Conventional data-driven landslide susceptibility models may only tell us half of the story: Potential underestimation of landslide impact areas depending on the modeling design. *Geomorphology*, page 108638. 6.2, 6.4

Loche, M., Alvioli, M., Marchesini, I., Bakka, H., and Lombardo, L. (2022). Landslide susceptibility maps of italy: Lesson learnt from dealing with multiple landslide types and the uneven spatial distribution of the national inventory. *Earth-Science Reviews*, page 104125. 4.7

Lombardo, L., Bakka, H., Tanyas, H., van Westen, C., Mai, P. M., and Huser, R. (2019). Geo-

statistical modeling to capture seismic-shaking patterns from earthquake-induced landslides. *Journal of Geophysical Research: Earth Surface*, 124(7):1958–1980. 4.2

Lombardo, L. and Mai, P. M. (2018). Presenting logistic regression-based landslide susceptibility results. *Engineering geology*, 244:14–24. 6.2

Lombardo, L., Opitz, T., Ardizzone, F., Guzzetti, F., and Huser, R. (2020). Space-time landslide predictive modelling. *Earth-science reviews*, 209:103318. 4.2

Lombardo, L. and Tanyas, H. (2021). From scenario-based seismic hazard to scenario-based landslide hazard: fast-forwarding to the future via statistical simulations. *Stochastic Environmental Research and Risk Assessment*, pages 1–14. 4.2

Lombardo, L., Tanyas, H., Huser, R., Guzzetti, F., and Castro-Camilo, D. (2021). Landslide size matters: A new data-driven, spatial prototype. *Engineering geology*, 293:106288. 4.2

Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. *Scientific reports*, 3(1):1–8. 6.2

Luo, Y. (2023). Sensing the shape of functional proteins with topology. *Nature Computational Science*, pages 1–2. 6.2

Malamud, B. D., Turcotte, D. L., Guzzetti, F., and Reichenbach, P. (2004). Landslide inventories and their statistical properties. *Earth Surface Processes and Landforms*, 29(6):687–711. 4.2

Marc, O., Stumpf, A., Malet, J. P., Gosset, M., Uchida, T., and Chiang, S. (2018). Initial insights from a global database of rainfall-induced landslide inventories: the weak influence of slope and strong influence of total storm rainfall. *Earth Surface Dynamics*, 4(6):903–922. 2.2

Marin, R. J., García, E. F., and Aristizábal, E. (2020). Effect of basin morphometric parameters on physically-based rainfall thresholds for shallow landslides. *Engineering Geology*, 278:105855. 4.2

Martha, T. R., Kerle, N., Jetten, V., van Westen, C. J., and Kumar, K. V. (2010). Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods. *Geomorphology*, 116(1-2):24–36. 6.2

Martha, T. R., Roy, P., Jain, N., Khanna, K., Mrinalni, K., Kumar, K. V., and Rao, P. (2021). Geospatial landslide inventory of india—an insight into occurrence and exposure on a national scale. *Landslides*, 18(6):2125–2141. 4.2

Martinotti, M. E., Pisano, L., Marchesini, I., Rossi, M., Peruccacci, S., Brunetti, M. T., Melillo, M., Amoruso, G., Loiacono, P., Vennari, C., et al. (2017). Landslides, floods and sinkholes in a karst environment: the 1–6 september 2014 gargano event, southern italy. *Natural Hazards and Earth System Sciences*, 17(3):467–480. 1.3

Meunier, P., Hovius, N., and Haines, J. (2008). Topographic site effects and the location of earthquake induced landslides. *Earth and Planetary Science Letters*, (275):221–232. 2.2

Milledge, D., Bellugi, D., McKean, J., AL, A. D., and WE.Dietrich (2014). A multidimensional stability model for predicting shallow landslide size and shape across landscapes. *Journal of Geophysical Research: Earth Surface.*, 119(11):2481–2504. 1.7, 2.2, 2.4

Moreno, M., Steger, S., Tanyas, H., and Lombardo, L. (2022). Modeling the size of co-seismic landslides via data-driven models: the kaikōura's example. 4.2

Munch, E. (2017). A user's guide to topological data analysis. *Journal of Learning Analytics*, 4(2):47–61. 4.4.2, 4.4.2, 6.5.1

Nadim, F., Kjekstad, O., Peduzzi, P., Herold, C., and Jaedicke, C. (2006). Global landslide and avalanche hotspots. *Landslides*, 3(2):159–173. 1.4

Nicu, I. C., Lombardo, L., and Rubensdotter, L. (2021). Preliminary assessment of thaw slump hazard to arctic cultural heritage in nordenskiöld land, svalbard. *Landslides*, 18(8):2935–2947. 6.4

Oksanen, J. and Sarjakoski, T. (2005). Error propagation of dem-based surface derivatives. *Computers & Geosciences*, 31(8):1015–1027. 4.7

Okun, O. and Priisalu, H. (2007). Random forest for gene expression based cancer classification: overlooked issues. In *Iberian conference on pattern recognition and image analysis*, pages 483–490. Springer. 5.1, 6.5.2

O.Marc and Hovius, N. (2015). Amalgamation in landslide maps: effects and automatic detection. *Natural Hazards and Earth System Sciences*, 15:723–733. 1.7, 2.2

Osanai, N., Shimizu, T., Kuramoto, K., Kojima, S., and Noro, T. (2010). Japanese early-warning for debris flows and slope failures using rainfall indices with radial basis function network. *Landslides*, 7(3):325–338. 2.2

Öztürk, U. (2018). *Learning more to predict landslides*. PhD thesis, University of Potsdam. 1.1, 1.3, 1.5

Ozturk, U., Bozzolan, E., Holcombe, E. A., Shukla, R., Pianosi, F., and Wagener, T. (2022). How climate change and unplanned urban sprawl bring more landslides. 6.2

Ozturk, U., Pittore, M., Behling, R., Roessner, S., Andreani, L., and Korup, O. (2020). How robust are landslide susceptibility estimates? *Landslides*. 2.2, 4.7

Ozturk, U., Pittore, M., Behling, R., Roessner, S., Andreani, L., and Korup, O. (2021a). How robust are landslide susceptibility estimates? *Landslides*, 18(2):681–695. 1.4

Ozturk, U., Saito, H., Matsushi, Y., Crisologo, I., and Schwanghart, W. (2021b). Can global rainfall estimates (satellite and reanalysis) aid landslide hindcasting? *Landslides*. 4.2

Ozturk, U., Wendi, D., Crisologo, I., Riemer, A., Agarwal, A., Vogel, K., López-Tarazón, J., and Korup, O. (2018). Rare flash floods and debris flows in southern germany. *Science of the total environment*, 626:941–952. 2.2

Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222. 5.1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.,

Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. 2.6

Perkins, S. (2012). Death toll from landslides vastly underestimated. *Nature News*, 8. 6.4

Pittore, M., Ozturk, U., Moldobekov, B., and Saponaro, A. (2018). Emca landslide catalog central asia. v. 1.0. gfz data services. 2.2

Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble machine learning*, pages 307–323. Springer. 5.1

Rana, K., Ozturk, U., and Malik, N. (2021). Landslide geometry reveals its trigger. *Geophysical Research Letters*, 48(4):e2020GL090848. 4.2, 4.4, 4.4.1, 4.7, 5.2.1, 6.3.2, 6.4

Rault, C., Robert, A., Marc, O., Hovius, N., and Meunier, P. (2019). Seismic and geologic controls on spatial clustering of landslides in three large earthquakes. *Earth Surface Dynamics*, 3(7):829–839. 2.2

Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M., and Guzzetti, F. A review of statistically-based landslide susceptibility models. 180:60–91. 6.2

Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M., and Guzzetti, F. (2018). A review of statistically-based landslide susceptibility models. *Earth-science reviews*, 180:60–91. 1.1

Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. (2015). A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4741–4748. 4.4.2, 6.5.1, 7.3

Roberts, N. J. and Evans, S. G. (2013). The gigantic seymareh (saidmarreh) rock avalanche, zagros fold–thrust belt, iran. *Journal of the Geological Society*, 170(4):685–700. 7.2

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67:93–104. 2.4, 5.1

Roelofs, L., Colucci, P., and de Haas, T. (2022). How debris-flow composition affects bed erosion quantity and mechanisms: An experimental assessment. *Earth Surface Processes and Landforms*, 47(8):2151–2169. 7.6

Rossi, M., Guzzetti, F., Reichenbach, P., Mondini, A. C., and Peruccacci, S. (2010). Optimal landslide susceptibility zonation based on multiple forecasts. *Geomorphology*, 114(3):129–142. 6.2

Roy, M.-H. and Larocque, D. (2012). Robustness of random forests for regression. *Journal of Nonparametric Statistics*, 24(4):993–1006. 2.4

Saito, H., Uchiyama, S., Hayakawa, Y. S., and Obanawa, H. (2018). Landslides triggered by an earthquake and heavy rainfalls at aso volcano, japan, detected by uas and sfm-mvs photogrammetry. *Progress in Earth and Planetary Science*, 5(1):15. 2.6, 4.7

Samia, J., Temme, A., Bregt, A., Wallinga, J., Guzzetti, F., Ardizzone, F., and Rossi, M. (2017). Do landslides follow landslides? insights in path dependency from a multi-temporal landslide inventory. *Landslides*, 14(2):547–558. 1.7, 2.2, 2.4, 4.2

Schmitt, R., Tanyas, H., Jessee, A., Zhu, J., Biegel, K., Allstadt, K., Jibson, R., Thompson, E., Westen, C., Sato, H., Wald, D., Godt, J., Gorum, T., Xu, C., Rathje, E., and Knudsen, K. (2018a). An open repository of earthquake-triggered ground-failure inventories. 2.1

Schmitt, R., Tanyas, H., Jessee, N., M.A., Z., Jing, Biegel, K.M., A., K.E., J., R.W., T., E.M., v. W., C.J., S., H.P., W., D.J., G., J.W., G., Tolga, X., Chong, R. E., and K.L., K. (2018b). An open repository of earthquake-triggered ground-failure inventories (ver 2.0, december 2018), u.s. geological survey data release collection. 2.3, 4.3, 4.7

Scott, D. W. (2015a). *Multivariate density estimation: theory, practice, and visualization.* John Wiley & Sons. 2.4

Scott, D. W. (2015b). *Multivariate density estimation: theory, practice, and visualization.* John Wiley & Sons. 3.3, 3.3

Simoes, E. and Araujo, G. Death toll from brazil floods, landslides reaches 57. *Reuters*. 6.2

Slaughter, S., Burns, W., Mickelson, K., Jacobacci, K., Biel, A., and Contreras, T. (2017). Protocol for landslide inventory mapping from lidar data in washington state. *Washington Geological Survey Bulletin*, 82(27):3. 6.6

Stark, C. and Guzzetti, F. (2009). Landslide rupture and the probability distribution of mobilized debris volumes. *Journal of Geophysical Research: Earth Surface*, 114(F2). 6.3.2

Stark, T. D., Baghdady, A. K., Hungr, O., and Aaron, J. (2017). Case study: Oso, washington, landslide of march 22, 2014—material properties and failure mechanism. *Journal of Geotechnical and Geoenvironmental Engineering*, 143(5):05017001. 1.3

Stumpf, A. and Kerle, N. (2011). Object-oriented mapping of landslides using random forests. *Remote sensing of environment*, 115(10):2564–2577. 4.2

Sundriyal, Y., Kumar, V., Chauhan, N., Kaushik, S., Ranjan, R., and Punia, M. K. (2023). Brief communication on the nw himalayan towns; slipping towards potential disaster. *Natural Hazards and Earth System Sciences Discussions*, pages 1–9. 6.4

Tacher, L. and Bonnard, C. (2007). Hydromechanical modelling of a large landslide considering climate change conditions. In *Landslides and Climate Change: Challenges and SolutionsProceedings of the International Conference on Landslides and Climate Change. Taylor & Francis, Ventnor*, pages 131–141. 1.3

Tanyaş, H., Görüm, T., Kirschbaum, D., and Lombardo, L. (2022). Could road constructions be more hazardous than an earthquake in terms of mass movement? *Natural hazards*, 112(1):639–663. 4.2

Tanyaş, H., Kirschbaum, D., and Lombardo, L. (2021). Capturing the footprints of ground motion in the spatial distribution of rainfall-induced landslides. *Bulletin of Engineering Geology and the Environment*, 80(6):4323–4345. 6.4

Tanyas, H., van Westen, C. J., Allstadt, K. E., Anna Nowicki Jessee, M., Görüm, T., Jibson, R. W., Godt, J. W., Sato, H. P., Schmitt, R. G., Marc, O., and Hovius, N. (2017). Presentation and analysis of a worldwide database of earthquake-induced landslide inventories. *Journal of Geophysical Research: Earth Surface*, 122(10):1991–2015. 2.2

Tauzin, G., Lupo, U., Tunstall, L., Pérez, J. B., Caorsi, M., Medina-Mardones, A. M., Dassatti, A., and Hess, K. (2021). giotto-tda:: A topological data analysis toolkit for machine learning and data exploration. *J. Mach. Learn. Res.*, 22:39–1. 4.4.2, 6.5.1

Taylor, F., Malamud, B., Witt, A., and Guzzetti, F. (2018a). Landslide shape, ellipticity and length?to?width ratios. *Earth Surface Processes and Landforms*, 43(15):3164–3189. 1.7, 2.2, 2.4, 2.6, 4.2, 4.7

Taylor, F. E., Malamud, B. D., Witt, A., and Guzzetti, F. (2018b). Landslide shape, ellipticity and length-to-width ratios. *Earth Surface Processes and Landforms*, 43(15):3164–3189. 1.1, 1.1, 6.3.2

Tianchi, L., Shumin, W., et al. (1992). *Landslide hazards and their mitigation in China*. Science Press. 1.2

Trigila, A., Iadanza, C., Esposito, C., and Scarascia-Mugnozza, G. (2015). Comparison of logistic regression and random forests techniques for shallow landslide susceptibility assessment in giampilieri (ne sicily, italy). *Geomorphology*, 249:119–136. 1.4

Trigila, A., Iadanza, C., and Spizzichino, D. (2010). Quality assessment of the italian landslide inventory using gis processing. *Landslides*, 7(4):455–470. 6.6

Uchida, T., Nishigugchi, Y., Nakatani, K., Satofuka, Y., Yamakoshi, T., Okamoto, A., and Mizuyama, T. (2013). New numerical simulation procedure for large-scale debris flows (kanako-ls). *International Journal of Erosion Control Engineering*, 6(2):58–67. 2.6

Varnes, D. C. D. (1996). Landslides: investigation and mitigation chapter 3-landslide types and processes. *Transportation research board special report*, 247. 1.1, 1.7, 2.2, 2.6

Varnes, D. J. (1978). Slope movement types and processes. *Special report*, 176:11–33. 1.5.2, 4.2, 4.7, 6.2, 6.4, 6.6

Vega, J. A. and Hidalgo, C. A. (2016). Quantitative risk assessment of landslides triggered by earthquakes and rainfall based on direct costs of urban buildings. *Geomorphology*, 273:217–235. 6.4

von Specht, S., Ozturk, U., Veh, G., Cotton, F., and Korup, O. (2019). Effects of finite source rupture on landslide triggering: the 2016 mw 7.1 kumamoto earthquake. *Solid Earth*, 10(2). 2.2, 2.6

Wang, Z.-Y., Lee, J. H., and Melching, C. S. (2015). Debris flows and landslides. In *River dynamics and integrated river management*, pages 193–264. Springer. 2.6

Watakabe, T. and Matsushi, Y. (2019). Lithological controls on hydrological processes that trigger shallow landslides: Observations from granite and hornfels hillslopes in hiroshima, japan. *CATENA*, 180:55 – 68. 2.6

Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9(4):611–629. 4.4.3

Zangeneh-Nejad, F. and Fleury, R. (2019). Topological analog signal processing. *Nature communications*, 10(1):2058. 6.2

Zhang, C. and Ma, Y. (2012a). *Ensemble machine learning: methods and applications*. Springer. 2.4

Zhang, C. and Ma, Y. (2012b). *Ensemble machine learning: methods and applications*. Springer. 5.1, 6.5.2