LANDSLIDE ANALYSIS AND PREDICTION USING MACHINE LEARNING

NURFARAHIN BINTI AMIR HAMZAH

UNIVERSITI TEKNOLOGI MALAYSIA

PSZ 19:16 (Pind. 1/23)

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name	:	NURFARAHIN BINTI AMIR HAMZAH			
Student's Matric No.	:	MSC241020	Academic Session	:	2024/2025
Date of Birth	:	16/03/1999	UTM Email	:	nurfarahin99@graduate.utm.my
Choose an item. Title	:	LANDSLIDE ANALYSIS AND PREDICTION USING MACHINE LEARNING			

I declare that this thesis is classified as:

☒	**OPEN ACCESS**	I agree that my report to be published as a hard copy or made available through online open access.
□	**RESTRICTED**	Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*
□	**CONFIDENTIAL**	Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the Choose an item. belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this Choose an item. for academic exchange.

Signature of Student:

Signature :

Full Name
Date :

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I	Full Name of Supervisor II
Date :	Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

This letter should be written by a supervisor and addressed to Perpustakaan UTM. A copy of this letter should be attached to the thesis.

Date:
Librarian
Jabatan Perpustakaan UTM,
Universiti Teknologi Malaysia,
Johor Bahru, Johor

Sir,

CLASSIFICATION OF THESIS AS RESTRICTED/CONFIDENTIAL
TITLE: LANDSLIDE ANALYSIS AND PREDICTION USING MACHINE LEARNING
AUTHOR'S FULL NAME:NURFARAHIN BINTI AMIR HAMZAH

Please be informed that the above-mentioned thesis titled **_________** should be classified as RESTRICTED/CONFIDENTIAL for a period of three (3) years from the date of this letter. The reasons for this classification are
(i)
(ii)
(iii)

Thank you.

Yours sincerely,

SIGNATURE:
NAME:NURFARAHIN BINTI AMIR HAMZAH
ADDRESS OF SUPERVISOR:

"I hereby declare that I have read this proposal and in my
opinion this proposal is sufficient in term of scope and quality for the
award of the degree of Master of Data Science"

Signature : ________________________________

Name of Supervisor I :

Date :

Signature : ________________________________

Name of Supervisor II :

Date :

Signature : ________________________________

Name of Supervisor III :

Date :

Declaration of Cooperation

This is to confirm that this research has been conducted through a collaboration Click or tap here to enter text. and Click or tap here to enter text.

Certified by:

Signature :
Name :
Position :
Official Stamp
Date

* This section is to be filled up for theses with industrial collaboration

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar :

Nama dan Alamat Pemeriksa Dalam :

Nama Penyelia Lain (jika ada) :

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

LANDSLIDE ANALYSIS AND PREDICTION USING MACHINE LEARNING

NURFARAHIN BINTI AMIR HAMZAH

A proposal submitted in fulfilment of the
requirements for the award of the degree of
Master of Data Science

Choose an item.
Faculty of Computing
Universiti Teknologi Malaysia

JANUARY 2025

DECLARATION

I declare that this proposal entitled “LANDSLIDE ANALYSIS AND PREDICTION USING MACHINE LEARNING ” is the result of my own research except as cited in the references. The proposal has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : ....
Name :
Date : 17 JANUARY 2025

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Professor Madya Dr. Haza Nuzly for encouragement, guidance, critics and friendship.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my Master study. Librarians at UTM, Cardiff University of Wales and the National University of Singapore also deserve special thanks for their assistance in supplying the relevant literatures.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

This research is focus on investigating global and Malaysia specific landslide factors that can be used to learn and predict landslides using Machine Learning. These include Random Forest for evaluating landslide risk and history, Linear Regression for the slope magnitude analysis, and Time series analysis for the scale of landslides. Supported by the Global Landslide Catalog data, this work focuses on both data pre-processing, feature exploration, and feature creation to enhance the model's predictive power. The findings should improve disaster risk management frameworks, as the model can provide recommendations for specific landslide prevention and management methods.

ABSTRAK

Penyelidikan ini memberi tumpuan penyiasatan faktor tanah runtuh global dan Malaysia yang boleh digunakan untuk mempelajari dan meramal tanah runtuh menggunakan pembelajaran mesin. Model seperto Random Forest digunakan untuk menilai risiko dan sejarah tanah runtuh, Regresi Linear untuk analisis magnitud cerun, dan Analisis siri masa untuk skala tanah runtuh. Disokong oleh data Katalog Tanah Runtuh Global, kajian ini memfokuskan pada kedua-dua pra-pemprosesan data, penerokaan ciri dan penciptaan ciri untuk meningkatkan kuasa ramalan model. Penemuan ini seharusnya menambah baik rangka kerja pengurusan risiko bencana, kerana model itu boleh memberikan cadangan untuk kaedah pencegahan dan pengurusan tanah runtuh yang khusus.

TABLE OF CONTENTS

TITLE **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Network
GA	-	Genetic Algorithm
PSO	-	Particle Swarm Optimization
MTS	-	Mahalanobis Taguchi System
MD	-	Mahalanobis Distance
TM	-	Taguchi Method
UTM	-	Universiti Teknologi Malaysia
XML	-	Extensible Markup Language
ANN	-	Artificial Neural Network
GA	-	Genetic Algorithm
PSO	-	Particle Swarm Optimization

LIST OF SYMBOLS

δ	-	Minimal error
D, d	-	Diameter
F	-	Force
v	-	Velocity
p	-	Pressure
I	-	Moment of Inersia
r	-	Radius
Re	-	Reynold Number

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

1.2 Problem Background

Landslides have been continually ranked as one of the most destructive natural disasters across the globe in terms of lives lost and infrastructural, economic losses, and impacts on environment. These movements are induced through the environmental factors for instance: rainfall intensity, slope gradient and stability of the surrounding soil which combine with the human activities: deforestation and urbanization (Petley, 2012). Generally, due to global climate change that has made extreme events like rainfall and storms frequent and prolonged, cases of landslides in many regions of the world have also been made frequent and severe (Huggel et al., 2012). Such disasters not only affect people's living but also mean extremely high cost in terms of economy, especially in the Canary Islands, Asian countries or African which hardly can afford the increase in infrastructure or the setting up of relevant mechanism.

Malaysia is famous for the tropical climate and mostly consists of steep lands; also the speed of urbanization increases the vulnerability of landslides. Monsoon rains occur annually and some of them surpass critical levels, thus causing landslides in both upland rural tracts and developed or urbanised zones (Latif et al., 2018). Population growth has also intensified these risks as natural water flow is disrupted, vegetation cover depleted and undermined soil structures due to population growth and urbanization. Prominent tragic incidents, including the 1993 Highland Towers failure and other current landslides in the vicinity of cities, further prove the need for better risk estimation and management in understanding environment and applying it to the Malaysian condition (Shahrin et al., 2021).

In the world, major improvements in technology and data access have led to the development of new approaches for landslide prediction based on machine learning, which can analyse various inputs and determine the main environmental criteria for landslides. However, these models are best developed in certain locations and might not consider the distressed environment of locations such as Malaysia. Also, issues such as variability, collection, and uniformity of data from the global dataset and the local dataset limit the direct use and comparison analysis (Froude Petley, 2018). An understanding of the circumstances that cause such cases is equally crucial in disseminating global knowledge into actionable competence for the endangered territory of Malaysia.

This study fills this gap by separately determining and comparing the environmental and geospatial factors controlling landslides at both global and local scales using the ML method. In the attempt to reach these goals, the research aims to analyze the variability and patterns of susceptibilities, temporal frequencies and severities of events within the framework of the Global Landslide Catalog. This method of comparison offers a ground from which specific prediction models and measures relevant to the conditions of Malaysia can be derived as well as enhance strategies of disaster risk reduction and mitigation.

1.3 Problem Statement

In the rapid development of technology, more effective analysis and prediction are needed to identify the areas with a high risk of landslides to make careful preparations and planning for effective mitigation measurements. By introducing machine learning in creating analysis and prediction models for landslide risk assessment, this tool can enhance the precision of landslides prediction and provide the accurate decision-making for mitigations planning and land-use management with cost effective and saving time. The problem questions for this study are:

a) How global landslide and Malaysian related based on environmental and geospatial analysis of factors.

b) How can machine learning models, such as Random Forest, Time Series, and Regression, be utilized to predict landslide susceptibility, temporal patterns, and event severity in both global and localized contexts?

c) What are the differences in landslide risk patterns between global trends and Malaysia-specific cases, and how can these insights contribute to developing tailored prediction frameworks and mitigation strategies?

1.4 Research Goal

The objective of this research is to gain deep insight of environmental and geospatial aspects of landslides and thereby apply machine learning strategies to compare global data with specific to Malaysia cases. The study is aim to produce strong landslide susceptibility, temporal and severity models to enhance the assessment of disaster risks and therefore come up with prevention mechanisms which are universal as well as those of specific areas of the world.

1.4.1 Research Objectives

The objectives of the research are :

(a) To study and identify the environmental and geospatial factors of the landslide hazard in Malaysia and global and the relationship between the data.

(b) To build and apply machine learning models, including Random Forest, Time Series, and Regression techniques, to predicting and analysing the landslide temporal trends, and event severity using global and Malaysia data.

(c) To generate an idea and informed decision-making concerning disaster risk management and risk reduction based on environmental and geospatial characteristics of Malaysia in relation to global standards.

1.5 Research Goal

The objective of this research is to gain deep insight of environmental and geospatial aspects of landslides and thereby apply machine learning strategies to compare global data with specific to Malaysia cases. The study is aimed to produce strong landslide susceptibility, temporal and severity models to enhance the assessment of disaster risks and therefore come up with prevention mechanisms which are universal as well as those of specific areas of the world.

1.6 Scope of Study

This research focuses on the landslide analysis for global and Malaysia. The machine learning models are applied to improve the spatial analysis and prediction performance of landslide risk and get the visualization of landslide prone areas, the impacts and mitigation to reduce the landslide occurrence in future.

1.7 Significance of Study

This study is important for disaster risk management which enables a prediction of areas prone to landslides. Machine learning helps geologists analyse the big and huge data and find out the factors which play a significant role for landslide occurrence and can also develop models with high degree of accuracy. These models assist in partitioning areas by the vulnerability thus facilitating planning and resource deployment for mitigation measures of disasters where it can save peoples live and minimized the losses. The analysis enhances understanding and management of spatial resources within cities, thus enhancing the sustainable development of cities. Moreover, it promotes and builds the capacity for multidisciplinary engagement and commits to improving community disaster capacities, responses and protection of the environment and decisions.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Landslides categorized as one of the geological hazards. Landslide are classified as geological processes whereby rocks, surfaced and debris move down slopes along a natural incline under the sole influence of gravity (Thirugnanam et al., 2020). They are triggered by natural factors and / or human activities can happen in mountainous regions, hills and flat area such as heavy rainfall, earthquakes, soil moisture condition, loss vegetation, geological factors, volcanic activities and freeze-thaw cycle (Acharya et al., 2022).

Landslides develop in a variety of materials: soil, debris, rock and organics and at various speed ranging from mm yearly to tens of meters per second and in different ways such as topple, fall, flow, slide spread. A landslide can be in any of several stages including relic, dormant, or active which migrate along flat or curved surfaces, can be shallow or deep or they can be retrogressive, progressive, advancing, or enlarge and catastrophic (Geertsema et al., 2020)

Even though there are few instances of major earthquake in Malaysia, cases of landslide and floods are recurring natural disasters which are due to over flooding caused by long periods of rainfall. In Malaysia for instance, the number of cases of landslides increases during the monsoon periods. Majority of the said happenings happened in areas such as Hulu Kelang, Cameron Highlands as well as Genting Highlands (Rahman & Mapjabil, 2017).

2.1.1 Types of Landslides

Landslides can be categorized into five types which are a) rockfall, b) topple, c) slides (rotational landslide and translational landslide), d) lateral spreading and e) flows such in Table 1 below (Kazmi et al. 2016 & Varnes 1978).

i. **Rock falls:** Rockfall is defined as a movement of soil, rock or both downhill on a steep incline with little or no shear movement other than falling, bouncing or sliding. They are normally developed on steep or vertical hazel ridges of river channels or coastal cliffs with falling soil fragments, which may break up on impact and runner until they encounter level base.

ii. **Topple:** Causes of toppling may include gravitational forces on the side of an uphill material, water or ice behind the crack, vibrations, undercutting or differential weathering, controlled excavation of slopes or stream erosion. When it happens swiftly that is with high velocity it can prove very destructive.

iii. **Slides:** Slides can be divided into two types which are:

- **Rotational landslides** prevail in homogeneous materials for example in fill areas and are often alleged by duration rain, rapid snowmelt, or by incoming or fluctuating ground water level. Others causes include general abrasion due to rise and fall in water levels, reservoir encroachment and earth movements.
- **Translational landslides** is the movement is in an outward or downward path across a gently inclined surface with little change in leaning. Usually, they occur in areas with weak foundation, gentle rocks, or both, and are invariably sited at faults, joints, bedding planes, or rock/soil contact planes.
- **Lateral spreads** refer to prostrate movement of stiff ground that has ruptured crosswise on week ground such as liquefied clay, without shifting a sharp divide. Such movements can lead to considerable destruction of buildings and constructions; the rate of these movements depends on the water content in the soil.
- **Debris flows** are most frequently observed in steep terrains as gullies, canyons and sparsely vegetated areas, on slopes that was subjected to wildfires or logging, and on volcanic readily eroded soils. These flows occur when rapid rainfall or snowmelt occur and move at high velocities (up to

35mp/h) transporting soil, rocks and other sediments down slopes in fan-shaped deposists at the base of slopes and have been known to cause a significant amount of damage to structures and infrastructures owing to their high velocity, instability and capacity to transport large loads.

In the region humid tropical climate like in Malaysia, landslide type which often occurs on natural slopes is the type of slip or slide. Due to the content high water and often associated with rain heavy, then slip failure or sliding often followed by flow (Jamaluddin et al., 2020). The presents of landslides can be identified using the geomorphology features of land such in **Figure 2.1**. The features consist of crown, main scarp, crown cracks, zone of depletion and zone of accumulation (Varnes 1978).

Table 2.1 Summary of types of landslides by Kazmi et. al 2016

Movement type	**Material classification**		
	Bed rock	Engineering soils	
Fall	Rock fall	Debris fall	Earth fall
Topples	Rock topple	Debris topple	Earth topples
Slides Rotational Translational	Rock slide	Debris slide	Earth slide
Lateral spreading	Rock spread	Debris spread	Earth spread
Flows	Rock flow	Debris flow	Earth flow

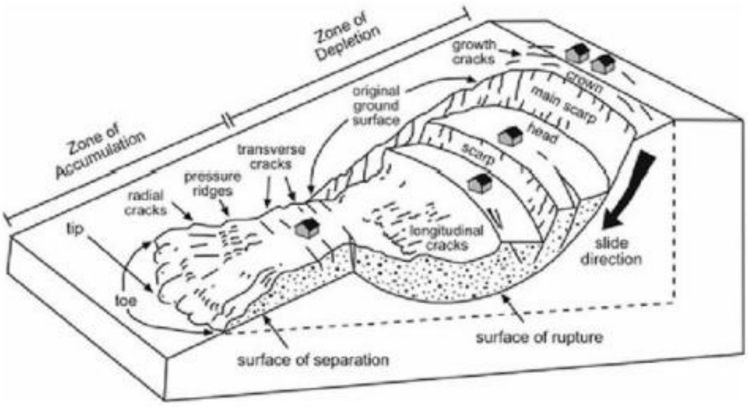


Figure 2.1.1 Illustration of geomorphological features of landslides

2.1.2 Types of Landslides

Landslides are natural and anthropogenic events on a global scale, of which rainfall is one of the leading causes (Qi et al., 2024). Heavy or continuous rain causes pore pressure buildup, the dissipation of soil strength, and slope failure, areas for which are common in the tropics and subtropics (Parizia et al., 2024). Earthquakes are another major cause whereby ground motions bring about landslides specially in the seismically active areas (Dou et al., 2024). Also, volcanic eruptions cause landslides; for instance, the areas in the Pacific Ring of Fire, pyroclastic and ash fall affect soil stability.(Melosh et al., 2024). Climate change conditions these natural activators and enhances the severity and frequency of extreme conditions that necessitate landslides (Abdulahi & Egli, 2024).

Secondly, human activities including mining and quarrying and other unfavorable land use practices also cause landslides (Qi et al., 2024). For instance, mining exercises in South American slopes have precipitated instability of slopes, which has led to numerous landslides (Parizia et al., 2024). Landslide has similarly been linked to uncontrolled logging in south east Asia since the vegetation causes the tree to peel deeper and weakens the cohesion of the soil (Dou et al 2024). Such activities give emphasis on the requirement for specific and stringent environmental policies and efficient methods of sustainable development globally (Melosh et al., 2024).

Malaysia has tropical monsoonal climate, thus; the major cause of landslides in this country is attributed to the heavy rainfall (Ya'acob et al 2024). Long duration of rainfall whether from Northeast Monsoon leads to the occurrence of heavy rain that when soils becomes saturated, the shear strength of the soil decreases and leads to slope failures (Daud et al., 2024). The Malaysia climate defined by many hours of rainfalls increases the frequency of shallow landslides due to the geographical structure characterized by hilly terrain and steep slopes (Qi et al., 2024). Some of the examples include the genting and cameron highland where there have been a lot of landslide common in areas that the soil is instable due to rain water (Ya'acob et al., 2024).

Landslide disasters in Malaysia are worsened by human interventions in various sectors including; deforestation, and land development (Parizia et al., 2024). Conversion of vegetation cover for cultivation purposes and other human activities leads to the modification of water drainage and slope stability (Dou et al., 2024). For instance, Infrastructure development for road construction in hilly areas would also disrupt the soil structure that leads to more slopes to fail during intense rainfall (Daud et al., 2024). An uncontrolled development especially in the growing world's urban zones has resulted to several fatal mishaps such as the Highland Towers landslides in the year 1993 (Ya'acob et al., 2024).

Furthermore, the combined consequences of geological and hydrological factors increase the potential of landslides (Daud et al. 2024). The physical characteristics of soils in the country include weathered granite and sedimentary soils, high groundwater and rainy conditions that triggers slope instability during rainfall (Ya'acob et al., 2024). Hence steep slopes such as those established by the Titiwangsa Range are more susceptible to such influences because besides having inherently weak structures to support them, element increase the intensity of the failures through anthropogenic factors (Parizia et al., 2024). These landslides in these regions have associated impacts including blocking of rivers leading to floods downstream (Qi et al., 2024).

To reduce the effects of landslides, Malaysia and the rest of the world is using the following measures, early warning systems and community-based disaster risk reduction (Abdulahi & Egli, 2024). In Malaysia, the integration of satellite-based rainfall monitoring system with real time alert has enhanced its preparedness against the rainfall-triggered landslides (Daud et al., 2024). In general, technologies like

LiDAR and remote sensing have been deployed to identify landslides and forecast future landslides in high-risk areas more effectively around the world (Dou et al., 2024). Intensification and proper land-use planning, executing reforestation schemes are necessary prerequisites to decreasing anthropogenic disturbances on slope stability (Qi et al., 2024).

2.1.3 Global Landslide History

In order to reduce the effects of landslides, Malaysia and the rest of the world is using the following measures, early warning systems and community based disaster risk reduction (Abdulahi & Egli, 2024). In Malaysia, the integration of satellite-based rainfall monitoring system with real time alert has enhanced its preparedness against the rainfall-triggered landslides (Daud et al., 2024). In general, technologies like LiDAR and remote sensing have been deployed to identify landslides and forecast future landslides in high risk areas more effectively around the world (Dou et al., 2024). With reference to land use planning, it is critical for the diminishment of certain anthropogenic influences on slope stability, with attaches to elastic reforestation projects (Qi et al., 2024).

These geographic mass movements recur through human history influenced by climatic, geologic, and socioanthropic conditions (Parizia et al., 2024). Civilians of the ancient Indus Valley, China and the Mediterranean region, for example, noted landslides resulting from seismic activity and intense rainfall (Dou et al. 2024). Hence in Middle Ages important events that occurred for instances the 1618 Goldau landslide disaster in Swiss involved horrifying disasters that claimed many lives and also property (Abdulahi & Egli, 2024). New data from archives show that landslides played a crucial role not only in the formation of global terrains but in the determination of human dwellings since the population steered clear of dangerous locations (Melosh et al., 2024) .

Over the current centuries the rate of urbanization and cutting down trees has also increased which has further enhance the problem of landslides all the globe (Qi et al., 2024). For instance, the 1999 Vargas disaster in Venezuela demonstrated the cost of eventual growth and densification without control in the event of extremely heavy precipitation; around 20,000 people died, and thousands more became homeless

(Collalti, 2024). Similarly, Kedarnath disaster of 2013 of India elaborated the poor management of development activities, and impact of extreme weather to show the susceptibility of mountain region (Rybalchenko, 2024). Several investigations have established that landslides are catalyzed and influenced by human activities such as mining, construction of roads and deforestation (Dou et al., 2024).

Landslide dynamics have been eased with the assistance of developed monitoring and remote sensing techniques (Fan et al., 2024). The two major landslides of 2017 Xinmo in China and the most recent 2020 Chamoli in India were analyzed using satellite imagery, drone data, and geophysical surveys (Melosh et al., 2024). These have provided additional information on the triggers and causes of landslides and the secondary effects they cause, which were useful in threat management and prevention (Parizia et al., 2024).

Climate change has been causing more frequent incidents of landslides, which is a major global concern stated by Qi et al., (2024). Global warming and climate changes heighten the chances a slide due to heat and unpredictable rainfall increases landslide triggers in areas with vulnerable geological structures due to weather changes (Rybalchenko, 2024). For example, landslide that occurred in July 2024 in Ethiopia due to enormous rainfall demonstrate the need to incorporate climate resilience into disaster risk management (Abdulahi & Egli, 2024). In this continuing threat, future require the application of advanced technologies, the integration of a community practice model, and effective policies (Fan et al., 2024).

These phenomena are still a significant global issue because the occurrence of landslides is increasing because of climate change (Qi et al., 2024). Global warming also exposes the area to climate volatility that raises the likelihood of landslides on sensitive landform location as indicated in 2024; Rybalchenko. For instance, the Ethiopian case of landslides in the event of July 2024 floods resulting from intense rainfall also points towards climate vulnerability mainstreaming into the disaster risk. That modern threat calls for future contingencies for managing it that include the combination of the new technologies, community programs or polices, and sound institutions (Fan et al., 2024).

2.1.4 Malaysia Landslide History

Based on Kazmi et al. 2016, the major of landslides occurrences in Malaysia include Highlands Tower collapsed (1993), Bukit Antarabangsa (2008 and 1999), Taman Zooview (2006), Taman Hillview (2002), Puncak Setiawangsa (2012), Penang Hill (2013), Ukay Perdana (2013), Serendah (2016), Gua Tempurung (1996) and Pos Dipang (1996). Highlands Tower collapsed is the most terrible landslides tragedy occur in Malaysia which killed 48 peoples. Most of the landslides caused by heavy rainfalls and human activities (**TABLE 2.2**).

The latest major landslides hazards occurred around 3 am on 16 December 2022 at *Father's Organic Farm,* Batang Kali, Selangor. The landslide occurred along the Jalan Batang kali Genting highlands road which caused 31 dead that consider to be the catastrophic tragedy after Highlands Tower Collapse on 1999 (Harvendhar Singh et al. 2022) caused by the continune rainfall accompanied by unstable weathered granite (Nasir et al., 2024). Landslides areas especially those adjacent to urban areas are on the alert and are occasioned by human factors such as deforestation and rapid urban growth.

East Malaysia that experience high incidence of landslide is the Crocker Range in Sabah because its geological formations are highly weathered with steep slope characterize the area. Heavy monsoon rain caused the 2023 landslide in Penampang and forced several families to flee, which underlined the sensitivity of the area to weather events (Anees et al., 2024). Sarawak's agriculture system has not only increased slope instability but also caused destructive incidents such as the 2020 Baram landslide which affected the connectivity of structures as well as abrupt village accessibility (Umor et al., 2024).

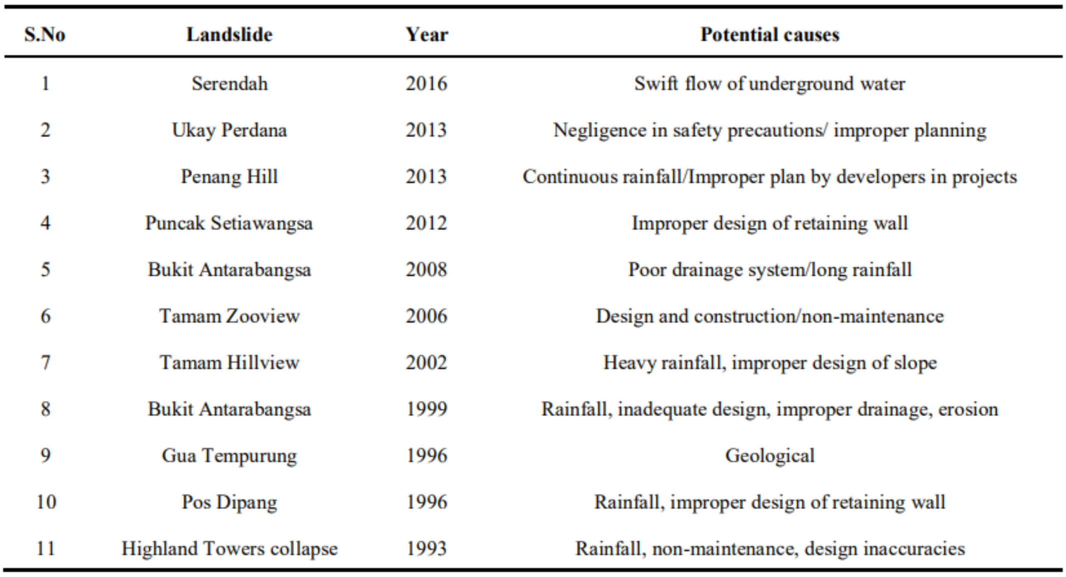
Landslide risks have been escalating due to the enhanced frequency and intensity of rainfall events that have been occasioned by climate change. In the period between 2020 and 2024, more than fifty reported landslides were reported with losses both from the urban and the rural sectors. Some ways are being implemented to have an early warning mechanism and enhance the ways of slope stabilization, though more coherent strategies must be employed to prevent such disasters (Nasir et al., 2024).

Wahab (2024) reported that the Mineral and Geoscience Department listed 38 hotspots area of landslide-prone in Selangor with seven of them classified as critical.

Besides the landslide incident happened in Batang Kali on end of December 2022, there is another landslide in Taman Bukit Permai, Ampang, on March 10 that year, killed four people and destroyed 15 homes and 10 vehicles.

Catastrophic landslides are common in Selangor not only because of natural conditions but also due to various anthropogenic factors like intense urbanization and unlawful uses of land that greatly enhance the susceptibility of such landslides, particularly in heavily developed regions, such as Bukit Antarabangsa (Mafigiri et al., 2022). Thus, for efficient landslides prediction and its control, these factors should become the components of the complex models to avoid the disasters in sensitive territories.

Table 2.2 Major landslides in Malaysia by Kazmi et al. 2016



S.No	Landslide	Year	Potential causes
1	Serendah	2016	Swift flow of underground water
2	Ukay Perdana	2013	Negligence in safety precautions/ improper planning
3	Penang Hill	2013	Continuous rainfall/Improper plan by developers in projects
4	Puncak Setiawangsa	2012	Improper design of retaining wall
5	Bukit Antarabangsa	2008	Poor drainage system/long rainfall
6	Tamam Zooview	2006	Design and construction/non-maintenance
7	Tamam Hillview	2002	Heavy rainfall, improper design of slope
8	Bukit Antarabangsa	1999	Rainfall, inadequate design, improper drainage, erosion
9	Gua Tempurung	1996	Geological
10	Pos Dipang	1996	Rainfall, improper design of retaining wall
11	Highland Towers collapse	1993	Rainfall, non-maintenance, design inaccuracies

2.2 Geology of Malaysia

2.2.1 Malaysia Landslide History

Malaysia is situated in the South-eastern Asia, occupying the North of the equatorial zone, with Thailand in the North, Singapore in the South and the South China Sea Division Malaysia into Peninsular Malaysia and East Malaysia on the Island of

Borneo. For the purposes of this paper, the geology of Peninsular Malaysia can be divided into the Western Central and Eastern Belts as depicted in Figure 2.2.1.

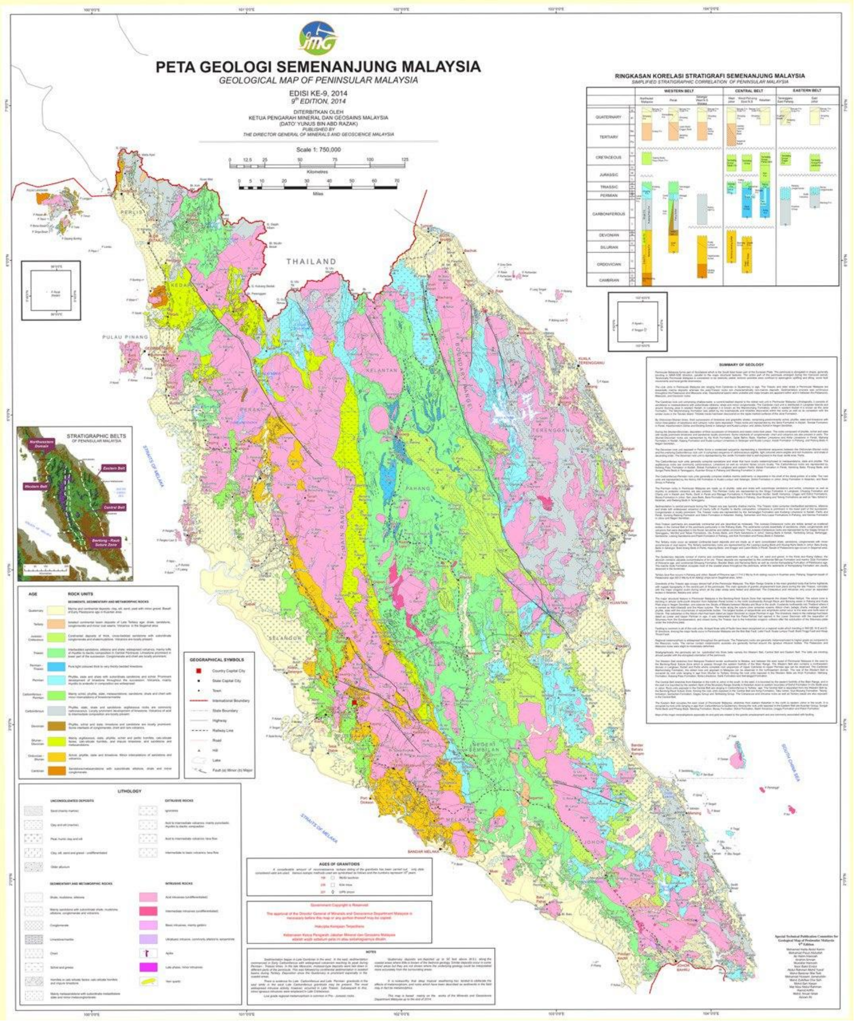


Figure 2.2.1 Geology Map of Peninsular Malaysia

The Western Belt is characterized by Permian-Triassic granites and Paleozoic sedimentary rocks with considerable tin-endowment (Azman, 2000). The central belt cover a wide range of geological units but mainly made up of volcanic and sedimentary rocks such as limestone and shale and is endowed with gold deposits because of its tectonic activity (Makoundi, 2012). Eastern Belt includes young granitic intrusions, basaltic dykes, and marine chemical sediments thus associated with tectonic activity and tin and tungsten deposits (Pour & Hashim, 2015). These belts are generated by the suture of the Sibumasu and Indochina Terranes that characterise the geology of the region.

Northern region of Malaysia, which includes Perlis, Kedah and northern part of Perak has large exposure of paleozoic sedimentary rocks, mainly limestone and shale. Many of these rocks have been highly karsted and created such features as Langkawi and Gunung Keriang. The region also has Mesozoic granites batholiths within Triassic and Jurassic periods, as well as tin bearing alluvial deposits that sustained historic mining (Akingboye et al., 2024). Additionally, the sedimentary rocks in South Malaysia such as Johor and Melaka comprise of Cenozoic age rocks mainly the marine shales and sandstones and these were deposited in a delta fashion during the Miocene age. This area can also be characterized by Quaternary alluvial site, and tuff residual of volcanic origin from the preceding formation (Aminu et al., 2024).

Both Eastern and Western Peninsular Malaysia demonstrate a complex variety of geographical structures. Region IV of Eastern Malaysia comprising of Terengganu, Pahang and Kelantan comprise of sediments from the Mesozoic era, coastal mudstones and inshores, metamorphic landscapes like schists and quartzite. Triassic Main Range granite belt is well identified in this region the sedimentary basins such as the South China Sea are proven to hold significant petroleum resources (Umor et al., 2024). The Sibumasu Terrane extends from western Malaysia, comprising Selangor, Penang, and a part of Perak, consists predominantly of Paleozoic crystalline rocks embraced within schists and gneiss and intruded by Permian-Triassic granites. It also contains Quaternary alluvial plains required for agriculture and urban development as identified by Ogg et al. (2024).

Island of Borneo Malaysia consist Sarawak, Sabah and Labuan make up it seismically active. Crocker Formation of Sabah includes sandstones and shales of Eocene-Miocene age deposited in deep marine, which is followed by tectonic uplifts. Sarawak contains the sedimentary basins with abundant of coals and oils, extensive peat swamp forest area and Baram Delta, a petroleum producing area such Figure 2.2.2.1 . Borneo also has ophiolite complexes – exposures of once-subducting slabs (Syawalia, 2024).

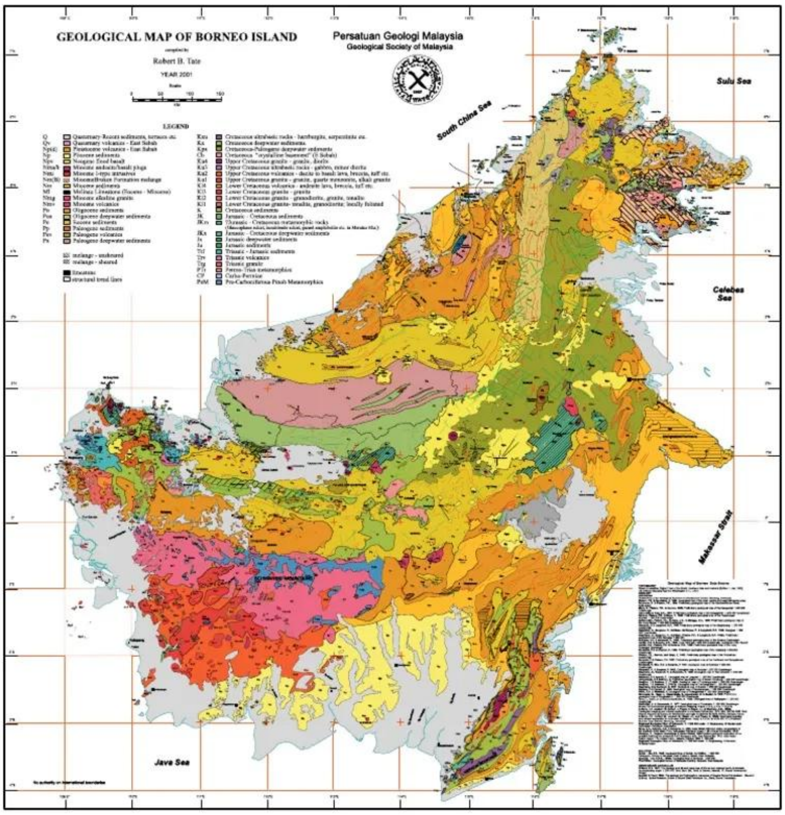

GEOLOGICAL MAP OF BORNEO ISLAND
Robert B. Tate
YEAR 2001
Persatuan Geologi Malaysia
Geological Society of Malaysia
LEGEND
South China Sea
Sulu Sea
Celebes Sea
Makassar Strait
Java Sea

Figure 2.2.2 Geology of Borneo Malaysia

Malaysia is particularly vulnerable to landslides largely due to its geography and topography cause the steep slopes and high intensity rainfall. A major part of the country comprises weathered sedimentary rocks like shale and sandstone that are more susceptible to erosion under high intensity rainfall more so in hilly areas of the central and eastern parts (Nasir et al., 2024). Also, granitic rock formations which are predominant in parts of Peninsular Malaysia are very vulnerable to slope failure because they disintegrate with time to unsound material (Anees et al., 2024). In that case, in Sabah state, the risks are amplified by the regions with tectonic activity and deeply weathered soil, the conditions that are the most favorable for landslides during the monsoon periods (*Nasir et al., 2024*).

2.2.2 Malaysia Climate

Malaysia situated at the geographical coordinates of N 2° 30’ 0” and E 112° 45’ 0”, has tropical climate with near-equatorial temperatures during the year and high percentage of humidity and those includes daily temperatures of between 22°C –33°C and average humidity of about 70%. The average annual precipitation is 2075 mm, this may have wide fluctuations depending in monsoon (Kazmi et al. 2016).

Generally, Malaysia experiences tropical climate, which is characterized by Southwest Monsoon (April – September), and Northeast Monsoon (October – March). These monsoon systems are responsible for the distribution of the country's annual rainfall distribution, with considerable implications on weathering, hydrosystems and structures, ecological systems and social-economic undertakings (Tang, 2019). Malaysia Meteorology Department stated that the Northeast monsoon gives the eastern states of Peninsular Malaysia and Sarawak a heavy shower. During the Southwest Monsoon which is often referred to as the dry season, does not trigger heavy rainfall or powerful wind (Ishak et al. 2021).

For example, the Cameron Highlands and the Ulu Klang regions are most vulnerable for landslides during monsoon seasons owing to the geography and high levels of precipitation. Such areas have recorded several landslides, which have led to loss of lives and properties (Fakaruddin et al., 2019, Mukhlisin et al., 2015, Matori et al., 2012). Landslides have been reported in Ulu Kelang Selangor, majorly during the monsoon season since this leads to fully saturated soil conditions, raise in pore water pressure thus declining slope strength, and is due to steep slopes and increased degree of urbanization. Combined with the reduced vegetation cover this makes the region very susceptible to soil erosion and therefore slope failure during the wet season as pointed out by Tajudin et al., 2021.

2.3 Machine Learning

Machine Learning (ML) has now emerged as a revolutionary method that has been adopted in analysing landslide occurrence for better prognosis, risk evaluation and risk minimisation. Landslides are processes with numerous and diverse causes that may be geological, meteorological, hydrological and anthropogenic in origin. Analysing these disparate variables is a challenging proposition for more conventional approaches to problem-solving. The ML algorithms with the properties such as handling big data, learning non-linear functions and learning and getting better with data are major steps forward in modelling landslide dynamics.

In landslide research, the most comprehensive use of ML to date is in identifying landslides or rather making predictions of them. Classifier models like Support Vector Machine (SVM), Decisions Tree, and Random Forests have been employed to study factors like slope gradient, soil type, precipitation and vegetation cover. For instance, Pradhan et al. (2018) illustrated that SVMs can provide a good way to model the distribution of the possibility of landslide occurrence in tropical areas utilizing geological and climatic data with high predictive precision. These models help decision makers to understand areas that are likely to cause high risk and take precaution measures before such things happen.

Machine learning has also improved the application of remote sensing data in monitoring of landslides. For instance, CNN's and gradient boosting methods analyze satellite imagery or high-altitude UAV imagery, LiDAR data and other photogrammetric data to identify or categorize landslides. One of the recent brilliant works by Huang et al. (2020) applied deep learning convolutional neural network models to classify optical and radar images without much intervention from human experts and found active landslide arrest areas. Such innovation means savings on time and money needed for the traditional on ground surveys while availing near real time information.

Another field in which the importance of ML appears is in the simulation of the factors that cause landslides, such as hydrological and meteorological factors. RNNs and LSTM networks, which enable the latest information at each step to be good for predicting landslides that occur due to rainfall or snowmelt. Zhang, Wang, Wang and Yin (2021) used LSTMs to analyze the relationship between rainfall and landslides with the help of precipitation history and terrain characteristics. They incorporate some form of time dimension which increases the accuracy in early warning systems.

However, the main substantive advantage of using ML is that it allows risk managers to carry out a comprehensive assessment of socio-economic and environmental risks, based not only on predictive indicators, but also with an analysis of significant additional information that meets the same criteria. Ensemble learning approaches and clustering methods, such as k – means, study spatial characteristics and possible effects of landslides on people and buildings. These models help identify which resources are most useful and in what ways can the infrastructure supporting them be made stronger. According to Samia et al., 2019, there have been increased use of ML-based frameworks for risk assessment that has enhanced the readiness in landslide-prone developing regions.

When implemented in landslide research there are some limitations, which include limited data availability, understanding the model's results, and applying the model in various regions with different topographies. Solving these problems needs to involve integration of ML with specific knowledge of the application area as well as enhancements in data acquisition by means of developing analytic capabilities in remote sensing and sensory systems. Explainable AI methods are also utilized in this work to improve the understandability of the model and the level of trust of various stakeholders. Over the years, this field is expected to grow and so is the role ML for incorporation of more sustainable and adaptive forms of approaches for landslide hazard risks.

2.4 Research Gap

Conventional approaches have provided substantial room for improvement in the application of ML algorithms in landslide analysis and prediction, but some challenges still apply. It was noted that the effectiveness of the ML models highly dependent on the availability of the labeled datasets which may be scarce or of poor quality in many regions. This scarcity lowers model performance and applicability, primarily when applied to regions having different geology or climate. Second, the ML models employed can rely solely on the previous information and do not keep track with the new factors like changes in the climate which consequently leads to limited knowledge of the new causes of landslides. There are obvious problems associated with applying complex forms of ML, like black-box algorithms, deep learning in particular, to organizational processes due to their opaque nature, which reigns distrust among stakeholders and does not allow using machine learning-based results to make critical decision.

Other difficulties concern input data, including the exact, high spatial and temporal data that are unavailable in some cases, and a high sensitivity of ML models to biases or errors of inputs. Real-time computation and the constant requirement for maintenance, including data update and model retraining, also form a constraint to the broad application of ML for landslide predictions, more so in developing countries. However, to implement an effective model, researchers in charge of such projects must involve both geoscientists and data scientists so that models can emulate the physical processes leading to a landslide. These limitations indicate that while ML should be employed in conjunction with more conventional approaches and, further development of data resources and improvement of data quality would provide further potential for the utilization of ML in the context of landslide analysis and prediction.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In this research, four machine learning algorithms on Python platform are used to formulate a related theoretical model for the forecast and analysis of landslides in Malaysia. Landslide risk assessment studies of the conventional method involve the use of qualitative assessment or simple statistical assessment that may not be very accurate or practical for large scale assessments. To gain a general insight into global and Malaysia trends, the researchers make use of the Global Landslide Catalog for localization purposes. Landslide data is analyzed using Random Forest, Time Series, and Regression models to determine susceptibility, temporal frequency, and intensity of landslide.

These models offer a complete picture of the underlying factors a primary reason behind feature importance analysis with a temporal perspective. In development of modelling, machine learning such as Random Forest (RF), Linear Regression (LR) and Time Series models are used. These machine learning applied in the analysis of trends in the existing data base and likelihood of occurrence of a landslide in various areas. These models while using ROC curve and AUC, accuracy, precision, F1-score, MSE, RMSE, and MAE, to assess the models guarantee the efficiency and credibility of the data used for model evaluation. This has made the use of GIS in visualization phase where the trained models are incorporated in GIS to generate the landslide susceptibility maps that grouped the areas of landslides into different risks. These maps were then used to disseminate to the various stakeholders involved in decision making and management of the landslides and to monitor the activity in the future. This resulting practical awareness will enhance the effort to support disaster management and minimize risk within the landslide zones of Selangor, Malaysia.

In development of modelling, machine learning such as Random Forest (RF), Linear Regression (LR) and Time Series models are used. This machine learning used to analyze the existing data and find out the possibility of occurrence of a landslide in distinct areas. The models' validity is checked using performance-based measures including Receiver Operating Characteristic (ROC) curve, Area under curve (AUC), accuracy, precision, F1-measure, root mean square error (RMSE), mean absolute error (MAE), and Mean Absolute Percentage Error (MAPE).

3.1.1 Proposed Method

These are multi-faceted disasters involving slippage of large masses of the earth's surface such as slopes, cliffs and other terrains, which pose a great threat to people and property and anywhere in the world, but more especially in regions that are prone to such disasters as the ASEAN country of Malaysia. Conventional approaches to mapping and modeling of landslide susceptibility include the use of GIS information that may be restricted by the type and quality of data, and spatial accuracy. Linear Regression, Time Series Analysis and Random Forest are more numerically oriented; thus, they are reasonably powerful methods for the landslide prediction without involving GIS.

Random Forest is one of the most useful machine learning algorithms that are preferred for the assessment of landslides and their likelihood because of its capability to work with reluctant variable relation. Based on simple and easily available environmental factors as rainfall intensity, soil moisture, and topographic attributes, the approach based on Random Forest provides a reliable classification and prediction of landslides susceptibility. Acharya et al. (2023) shown that for the same set of data Random Forest algorithm provided better model than other algorithms when mutual information applied for feature selection has resulted in accuracy and less MSE. Its capability to deal with unbalanced samples, a frequent problem in landslide research, is further expanded by incorporating pre-processing techniques, such as oversampling as highlighted by Song et al. (2023), having revealed better results of recall and AUC in landslide prediction models. Moreover, Sharma & Sandhu (2023) also corroborated its efficiency in computing on voluminous datasets, and therefore it's able to enhance the algorithm's results by adding parameters such as rainfall, geology, and cover type that all are important in the evaluation of a landslide and its prevention methods.

Linear Regression is a widely used statistical and machine learning technique in landslide analysis and prediction due to its simplicity, interpretability, and ability to model relationships between independent variables and the likelihood of landslide events. This approach quantifies the influence of factors like rainfall intensity, soil moisture, and slope on landslide occurrences, providing insights into their individual contributions. Wang et al. (2020) highlighted the effectiveness of Linear Regression in identifying causal relationships in landslide data, enabling data-driven decision-making. Furthermore, Song et al. (2023) demonstrated that preprocessing techniques, such as oversampling and undersampling, improve the performance of Linear Regression on imbalanced landslide datasets, enhancing its predictive accuracy. Sharma and Sandhu (2023) also emphasized its role in analyzing temporal and environmental variables, making it a useful complement to advanced machine learning algorithms for landslide risk assessment and prevention strategies.

Time Series Analysis is a powerful machine learning technique for landslide analysis and prediction, enabling the identification of temporal patterns and trends in landslide occurrences. By analysing data such as rainfall intensity, seasonal variations, and historical landslide events over time, Time Series Analysis provides valuable insights into the temporal dynamics that contribute to landslide susceptibility. Sharma and Sandhu (2023) demonstrated its utility in identifying peak landslide periods, particularly during seasons of high rainfall, aiding in proactive disaster management. Song et al. (2023) emphasized the effectiveness of Time Series Analysis in predicting future landslide occurrences by accounting for recurring environmental factors and anomalies. Acharya et al. (2023) further validated the approach's ability to integrate temporal dependencies, highlighting its role in improving the precision of landslide prediction.

3.2 Researchers Operational Framework

3.2.1 Researchers Design Flowchart

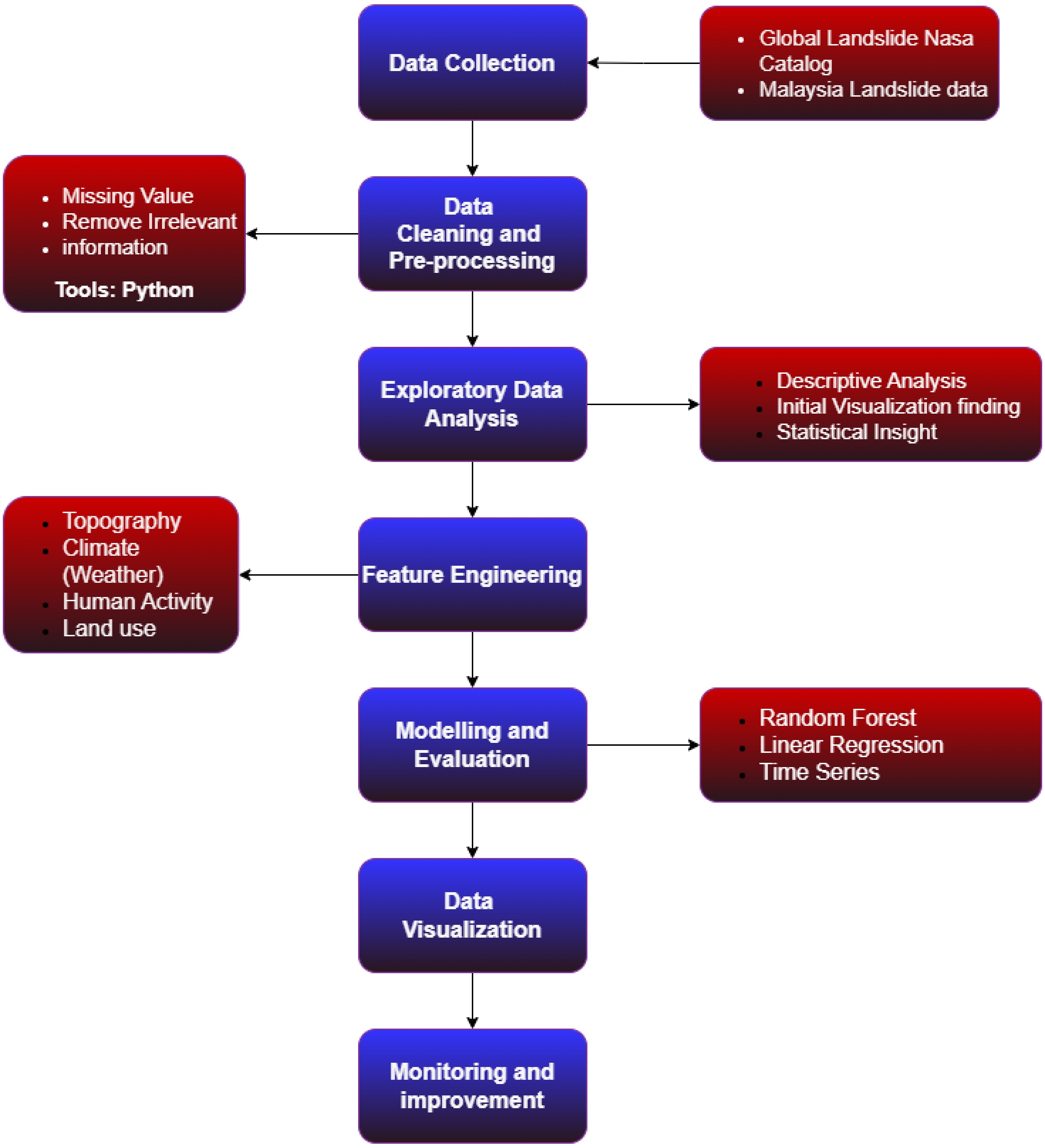


Figure 3.2.1 Flow Chart

3.2.2 Description of Work Breakdown (Step)

i. **Data Collection**

The dataset used in this work has been retrieved from NASA's Global Landslide Catalog, a more general-purpose database intended to aid in the assessment of landslides and their forecast. This catalog aggregates the event data of the landslide incidences that were either reported via the several media platforms, published scientific journals articles, and social platforms. For each landslide event, date and time, geographical coordinates, factors that caused the landslides, sizes and settings, and the related death and injury tolls are included in the dataset set. For this study, the events happening in Malaysia only are of interest and thus the global dataset is filtered by the geographic and administrative criteria. The subset of data collected in the Malaysian region is particularly informative to understand specific facts about local landslides caused by tropical climate conditions and deforesting along with urbanization.

ii. **Data Cleaning and Preparation**

Data cleaning is required when processing such databases as Global Landslide Catalog in the context of landslide analysis. The general data preparation steps include exploring the structure of the data by determiduplicates andlocation, type, and magnitude of the data, dealing with the missing data through imputation or by removing the data as well as removing the duplicates, and finally formatting the data by the conversion of the dates and coordinates to a standard format. Results arising from categorical data should be made accurate, errors in categorization should be deal with appropriately, and any artificially high values should be spotted. There is also a belief that scaling or feature engineering can enhance the accuracy of the analysis. Assuring cleaned data makes for reliability and grouping by the time or space makes identification of patterns possible. High quality cleaned and structured data will help us perform precise landslide risk assessment to assist in the right disaster management.

iii. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is important crucial for prediction and analysis of landslides, that can assists in the identification of patterns that characterize the event as well as nothing abnormal from normal phenomena and to establish relationship between measurable factors such as intensity of rainfall, the soil type, and the slope of the land. Comparing to the features selected by highly dimensional model, EDA helps make significant improvement in the understanding of landslide dynamics, feature selection, and the implementation of ML models through statistical and graphical techniques like scatter plot and heatmap. It helps to solve issues like skewed data and blank data, to improve model accuracy like Random Forest and Gradient Boosting. Research has demonstrated how EDA can assist in the representation of geographical data while preparing possibility maps and forecasting models, imperative for improving and real-time landslide detection with unprecedented environment change.

iv. Feature Engineering

Feature representation is important for landslide analysis and prediction because it provides a meaningful output. Globally relevant features include elevation, slope, rainfall intensity and land cover classification, but in Malaysia, tropical rainfall and deforestation should also be taken into consideration. Occurrence based features such as seasonality and trends, other spatial features such as distance from fault lines, dwelling, and human activities may also be important for classification. Despite its simplicity, Random Forest naturally provides feature importance of selected features and can model high-order interactions since it is capable of handling non-linear characteristics. Linear Regression is sensitive to variables whose relation to the target variable is linear, whereas Time-Series Analysis involves temporal variables to make prediction on trends and seasons. This, additional methods complement each other, and tentative features are incorporated to make landslide prediction and analysis more accurate.

v. Modelling and model evaluation

Random Forest has good performance in dealing with large high dimensional input and works well in capturing non-linearity, therefore is ideal for making global and regional landslides by considering of rainfall, soil and topography of the area. Linear Regression, while easy to understand, interpret and construct, is most suitable in situations where a particular variable depends on one or several others for a continuous variable as in the case of landslide size but less so in situations where there are interactions. An application of Time-Series Analysis will reveal future landslides based on trends and seasonality based on history, which is significant for monsoonal areas such as Malaysia. The integration of this method enhances the overall understanding of landslide risks and in turn assist in risk evaluation and disaster response. This performance of the models depends on accuracy, precision, R-squared, RMSE and such other parameters where Random Forest gives importance to Features Importance, Linear Regression for data fit and Time-series models for forecasting accuracy. Further, adjustment of hyperparameters of these models optimizes them for mature and more accurate predictions. Forecasting accuracy measures. Hyperparameter tuning further optimizes these models for robust predictions.

vi. Data visualization and analysis

Visualization and interpretation plays an important role as part of knowledge distribution on predicted landslide occurrence. A normality check is used in linear regression to assess how relationships are discharged, and a quantile-quantile plot determines the cultural disparity of random forest., a forecasting monster is used to show the accuracy of random forest up to a specified period. Such tools support the detection of main or most hazardous risks and time frames when landslides may happen, useful for preparation for a disaster. The application of models such as linear regression, random forest, and time-series analysis requires incorporation of the models into systems for obtaining real-time or batch predictions. Linear regression shows how big a landslide might be, random forest estimates how likely an area is to have a landslide, and ARIMA is for the actual expectations of future incidents. Real-time, on-going prediction, retraining, and recurrent analysis, which makes automated system reliable in the long run.

vii. **Monitoring and Improvement**

Iterative improvement is about rendering the landslide prediction models better over time by adding new data and changing on aspects of the feature. In the case of linear regression this will entail relooking at assumptions or looking at a more elaborate association. In terms of Random Forest, the accuracy can be tuned by changing the hyperparameters or by adding relevant new data. Dynamic forecasting models can be reconfigured with new features, or they may be extended over longer forecasting horizons. This continuous refinement ensures that models stay accurate and adaptable to environmental changes, enhancing their utility in predicting and managing landslide risks.

3.3 Tools: Phyton

Python has numerous libraries for performing landslides analysis and the prediction based on which techniques: Linear Regression, time series analysis and Random Forest. Scikit-learn and stats models are popular for Linear Regression and has support for cross validation and support for detailed statistical analysis of factors like rainfall and slope. For data pre-processing, the type of data used in Time Series Analysis is furnished by Pandas and to apply models such as ARIMA, Statsmodels is used and for seasonal and weekly holiday forecasting, the tool used is Prophet. For Random Forest, Scikit-learn offers a less complex methodology for modeling in addition to offering hyperparameter tuning tool and feature importance function. Additionally, Imbalanced-learn focuses on tackling data imbalance problems. Matplotlib, Seaborn and Plotly for creating interactive graph, heat maps and scatter plots are also very useful for data interpretation and reportage. These two Python tools combine to improve the prediction of landslides along with disaster management especially in areas of Malaysia.

3.4 Performance Measure

3.4.1 Machine Learning

Linear Regression: Multiple linear regression analysis is used to determine the pattern of association between the landslide probability and potential predictors. In this study, ordinary least absolute deviation regression is employed rather than a single measure since probability of landslide has large variability. The linear function is expressed in equation 3.4.1.

$$Y=a+b1X1+b2X2+...+bnXn \qquad \text{(eq 3.4.1)}$$

where, Y is the dependent variable (landslide probability in our case), X1 represents independent variables (all predictors), a is the constant and b1 is the regression coefficient of the variable X1.

Random Forest Regression: Random Forest Regression is an ensemble model based on decision trees, at the creation of which bootstrap samples of training data are used. Bagging (bootstrap aggregation) works to reduce variance and over training thus improving the model's generalization capabilities. The function is expressed in equation 3.4.2.

$$\hat{y} = \frac{1}{T}\sum_{k=1}^{K} f_t(x) \qquad \text{(eq 3.4.2)}$$

where; K is number of trees, $f_t(x)$ is prediction of the *t*-th tree, and $\hat{y}$ is final prediction. This method effectively balances bias and variance, delivering robust predictions for landslide susceptibility.

Time-series analysis : is a statistical technique used to analyze data points collected over time to identify patterns, trends, and seasonal variations, which can help forecast future values. A widely used model in time-series analysis is ARIMA (AutoRegressive Integrated Moving Average), which combines three components: autoregressive (AR), differencing (I), and moving average (MA). The ARIMA model is represented by the equation:

$$\mathbf{Y_t} = c + \phi_1 Y_{t-1} + Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t \qquad \text{(eq 3.4.3)}$$

where $\mathbf{Y_t}$ is the predicted value at time, c is a constant, ϕ_1 are autoregressive coefficients, ϵ_t is the error term, and p are the order of AR model. This model helps capture the relationships between past values and forecast future events, making it particularly useful for predicting time-dependent phenomena like landslides, where temporal factors such as rainfall or seasonality are crucial for forecasting risks.

3.4.2 Performance Matrix

i. RMSE, MSE, MAE,MAPE

The performance of the models is determined and compared to mean measures including Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). These metrics quantify the differences between predicted and actual values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \text{(eq 3.4.2.1)}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{(eq 3.4.2.2)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(|y_i - \hat{y}_i|)^2 \qquad \text{(eq 3.4.2.3)}$$

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \qquad \text{(eq 3.4.2.4)}$$

Here, y_i denotes the expected value and $\hat{y}_i$ is the predicted value. These metrics provide insights into model accuracy and error distribution.

Where M = mean absolute percentage error, N = number of times the summation iteration happen; A_t = actual value and F_t = forescast value

ii. ROC and AUC

Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) are used to assess the models' classification performance. The AUC quantifies the ability of a model to distinguish between positive (landslide-prone) and negative (non-landslide) areas, with values ranging from 0.5 (random prediction) to 1 (perfect prediction):can be calculated by the integral trapezoidal rule. The equation is written as follows:

AUC (Area Under Curve): Measures classification performance on ROC curve.

$$AUC = \frac{(\sum TP + \sum TN)}{(P + N)} \qquad \text{(eq 3.4.2.5)}$$

iii. *Confusion Matrix*

In this study, the performance of the evaluation model for landslide susceptibility was calculated with the confusion matrix which is a popular method used in binary classification for model assessment. Three statistical measurements including precision, recall, accuracy and F1- Score were used to assess the efficacy of the specific model. It expressed as follows:

- **Precision**: Proportion of correctly identified positive cases.

$$Precision = \frac{TP}{TP + FP} \qquad \text{(eq 3.4.2.5)}$$

- **Recall**: Sensitivity or True Positive Rate.

$$Recall = \frac{TP}{TP + FN} \qquad \text{(eq 3.4.2.6)}$$

- **Accuracy:**

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{(eq 3.4.2.7)}$$

- **F-1 Score :**

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad \text{(eq 3.4.2.8)}$$

where TP is for the correct predicted number of landslide raster, TN is the correct predicted number of non-landslide raster, FP is the number of incorrectly predicted landslide raster and FN is the number of wrong predicted non-landslide raster cells.

In order to calculate these indices, concepts of four kinds of predicted samples for classification learning need to be clarified: Detection accuracy is the ratio of TP+TN to TP+TN+FP+FN and is defined by the four ways the test can be classified: (1) true positive (TP): the patient has the disease and is predicted as positive; (2) false positive (FP): the patient has no disease and is predicted as positive; (3) true negative (TN): the patient has no disease while is predicted as negative; and the prediction disagrees with the actual class. where TP (true) and TN (true negative) denote the correctly classified raster cells, P expresses the total number of landslide raster cells, and N represents the total number of non-landslide raster cells.

Table 3.1 Gantt Chart

Task	Months											
	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept
1. Desk Study	X	X										
2. Proposal Development												
2.1 Research Question and Objective		X	X									
2.2 Literature Review			X									
2.3 Proposal Submission and Approval				X	X							
Project 1 :												
3. Data Collection												
3.1 Global Landslide catalog					X							
3.2 Historical Landslide Data						X						
3.3 Field Surveys							X					
4. Data Processing												
4.1 Data Cleaning							X					
4.2 Exploratory Data Analysis (EDA)							X					
4.3 Feature Selection							X					
5. Model Development												
5.1 Algorithm Selection								X				
5.2 Training Machine Learning Model								X				
5.3 Model Validation and Testing								X	X			
Project 2:												
6. **Results Analysis**												
6.1 Performance Metrics Analysis									X	X		
6.2 Visualization of Results									X	X		
7. Thesis Writing	X	X	X	X	X	X	X	X	X	X		
8. Thesis Presentation and Defense											X	
9. Thesis Submission											X	
10. Thesis Revision and Final Submission												X

CHAPTER 4

PROPOSED WORK

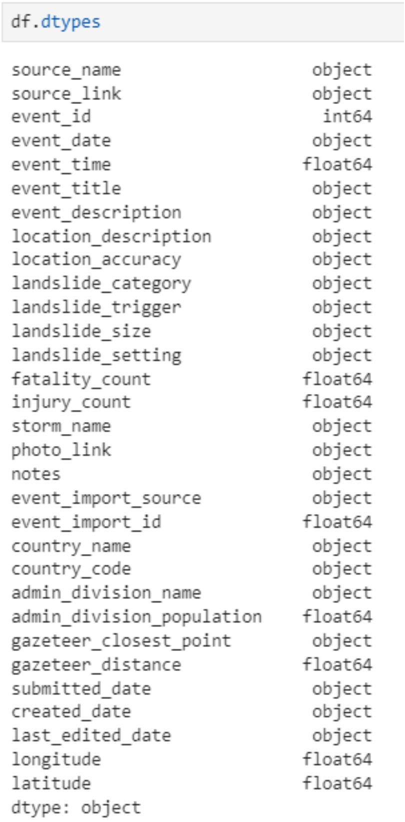
4.1 Introduction

This chapter presents the initial findings from the exploratory data analysis (EDA) conducted on the Global Landslide Dataset. The analysis aims to uncover patterns, relationships, and trends related to landslide occurrences and their triggers, focusing on global patterns while highlighting observations relevant to Malaysia. The findings are presented on statistical summaries, visualizations, and machine learning techniques.

4.2 Exploratory Data Analysis (EDA)

4.2.1 Overview Dataset

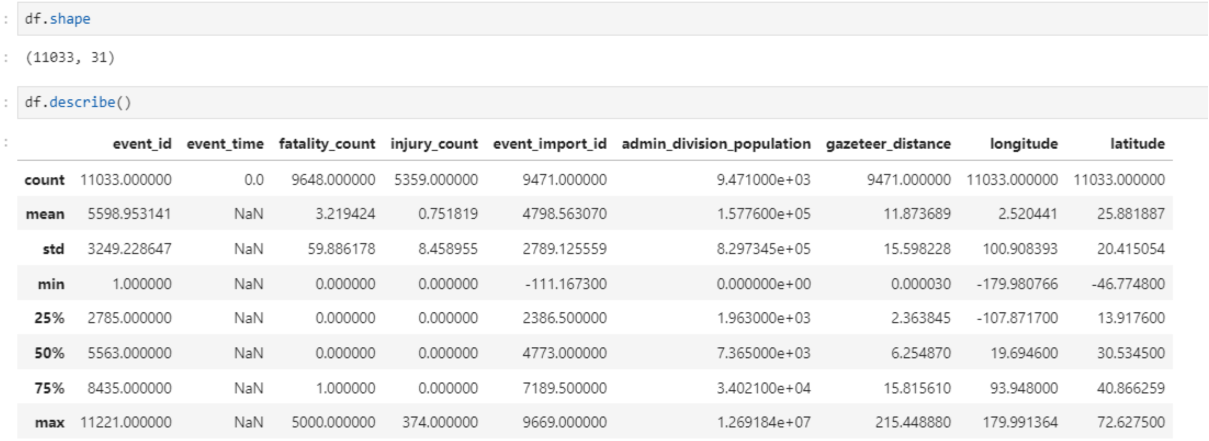
The data set is obtained from landslides site gotten from Nasa datasets webpage, as .csv file with total of 11,033 records and 31 features, which include date the landslide event and geographical location in terms of latitude and longitude on earth as well as factors that caused the event and the effects it had such as number of deaths, injured and others. Types of dataset is illustrated in Figure 4.1.1 while dataset described as Figure 4.1.2 using command `df.describe()`.The columns of dataset classed as information of the landslide event, landslide details and miscellaneous such in Table 4.1. This set of data covers more than one year and gives important statistical and geophysical information relating landslides hazards.



```
df.dtypes
```

```
source_name                    object
source_link                    object
event_id                        int64
event_date                     object
event_time                    float64
event_title                    object
event_description              object
location_description           object
location_accuracy              object
landslide_category             object
landslide_trigger              object
landslide_size                 object
landslide_setting              object
fatality_count                float64
injury_count                  float64
storm_name                     object
photo_link                     object
notes                          object
event_import_source            object
event_import_id               float64
country_name                   object
country_code                   object
admin_division_name            object
admin_division_population     float64
gazeteer_closest_point         object
gazeteer_distance             float64
submitted_date                 object
created_date                   object
last_edited_date               object
longitude                     float64
latitude                      float64
dtype: object
```

Figure 4.2.1 List of data type



```
df.shape
```

```
(11033, 31)
```

```
df.describe()
```

	event_id	event_time	fatality_count	injury_count	event_import_id	admin_division_population	gazeteer_distance	longitude	latitude
count	11033.000000	0.0	9648.000000	5359.000000	9471.000000	9.471000e+03	9471.000000	11033.000000	11033.000000
mean	5598.953141	NaN	3.219424	0.751819	4798.563070	1.577600e+05	11.873689	2.520441	25.881887
std	3249.228647	NaN	59.886178	8.458955	2789.125559	8.297345e+05	15.598228	100.908393	20.415054
min	1.000000	NaN	0.000000	0.000000	-111.167300	0.000000e+00	0.000030	-179.980766	-46.774800
25%	2785.000000	NaN	0.000000	0.000000	2386.500000	1.963000e+03	2.363845	-107.871700	13.917600
50%	5563.000000	NaN	0.000000	0.000000	4773.000000	7.365000e+03	6.254870	19.694600	30.534500
75%	8435.000000	NaN	1.000000	0.000000	7189.500000	3.402100e+04	15.815610	93.948000	40.866259
max	11221.000000	NaN	5000.000000	374.000000	9669.000000	1.269184e+07	215.448880	179.991364	72.627500

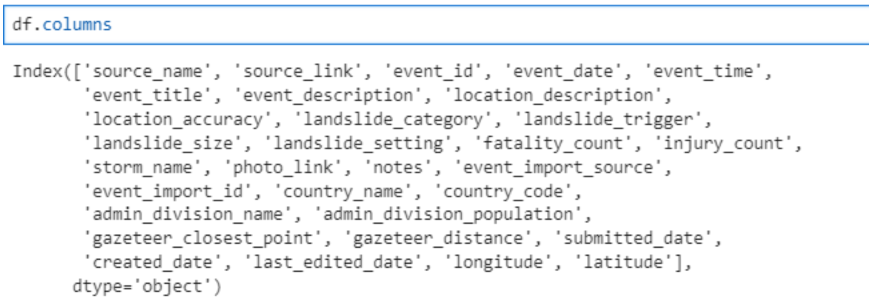
Figure 4.2.2 Data Discription

4.2.2 Data Cleaning and Pre-processing

Data cleaning is important to avoid hopeless inconsistency, error and misleading information in the data that will be analysed into machine learning algorithms. Incomplete and invalid data can lead to lack accuracy, biased models and poor performance affecting the credibility of the decision-making (Batista, Monard & Leite, 2003).

i. Step 1: Check the Column

First, the column checked using `df.columns` command in Python's Pandas library. This command prints the names of columns of a data frame in Pandas which is a two-dimensional data structure. The DataFrame output of `df`, contains several columns relevant that are related to landslide occurrences. This type of columns, for example `‘event_title’`, `‘event_description’`, `‘longitude’`, `and` `‘latitude’`, `‘landslide_category’`, `‘landslide_size’`, `‘fatality_count’`, `‘injury_count’`, `‘source_name’`, `‘submitted_date’`, and others column that stores different types of data. The used of `dtype=object` means that the columns are possibly containing string, integer or date values. The data contained within this DataFrame mostly looks clean and normalized in terms of the landslide event analysis.



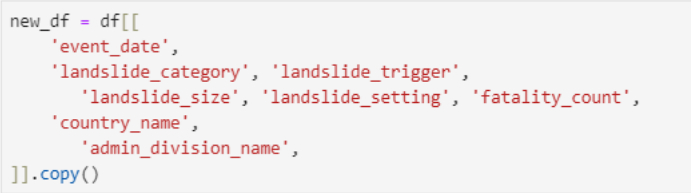
```
df.columns
```

```
Index(['source_name', 'source_link', 'event_id', 'event_date', 'event_time',
       'event_title', 'event_description', 'location_description',
       'location_accuracy', 'landslide_category', 'landslide_trigger',
       'landslide_size', 'landslide_setting', 'fatality_count', 'injury_count',
       'storm_name', 'photo_link', 'notes', 'event_import_source',
       'event_import_id', 'country_name', 'country_code',
       'admin_division_name', 'admin_division_population',
       'gazeteer_closest_point', 'gazeteer_distance', 'submitted_date',
       'created_date', 'last_edited_date', 'longitude', 'latitude'],
      dtype='object')
```

(b)

ii. Step 2: Dropping irrelevant column

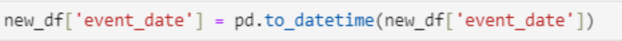
(c) The new DataFrame called `new_df` created. Then, the needed column selected from the original DataFrame `df` and copies them into `new_df`. The selected columns are: `'event_date'`, `'landslide_category'`, `'landslide_trigger'`, `'landslide_size'`, `'landslide_setting'`, `'fatality_count'`, `'country_name'`, and `'admin_division_name'`. The unnecessary columns dropped from the original `df` .



```
new_df = df[[
    'event_date',
    'landslide_category', 'landslide_trigger',
       'landslide_size', 'landslide_setting', 'fatality_count',
    'country_name',
       'admin_division_name',
]].copy()
```

iii. Step 3: Change datetime format

(d) The `'event_date'` column was converted in `new_df` to the datetime data type. The `pd.to_datetime()` function from the Pandas library is used as conversion function that is important when performing date-based analysis and operations.

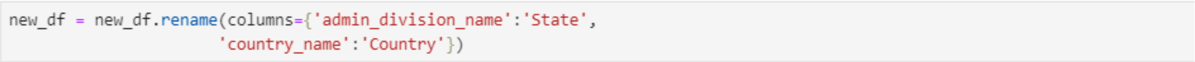


```
new_df['event_date'] = pd.to_datetime(new_df['event_date'])
```

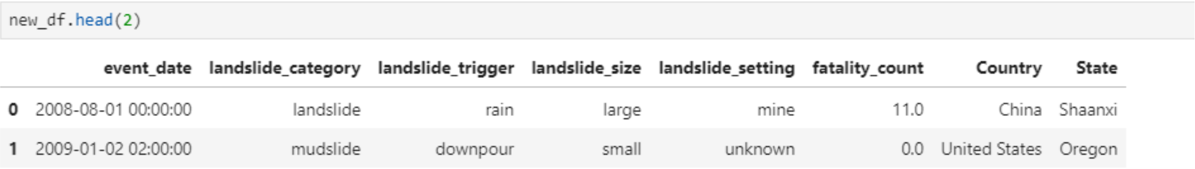
iv. Step 4: Rename Column

(e) This step changed the DataFrame's column names to more understandable. The function `.rename()` method with a dictionary used to change the names of `'admin_division_name'` to `'State'`, and `'country_name'` to `'Country'`. This makes the DataFrame manageable especially to the new user as the new names are more explanatory than the original column names. The `.head(2)` method displayed the the first two rows of the modified DataFrame to visually verify the renaming.

(f)



```
new_df = new_df.rename(columns={'admin_division_name':'State',
                     'country_name':'Country'})
```

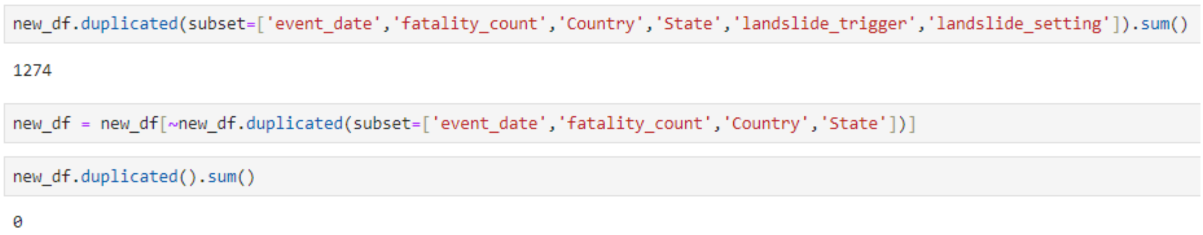


```
new_df.head(2)
```

	event_date	landslide_category	landslide_trigger	landslide_size	landslide_setting	fatality_count	Country	State
0	2008-08-01 00:00:00	landslide	rain	large	mine	11.0	China	Shaanxi
1	2009-01-02 02:00:00	mudslide	downpour	small	unknown	0.0	United States	Oregon

v. Step 5: Dropping all the duplicated rows

(g) Then, then duplicated data eliminated within the dataset by counting the numbers of rows with the same values for the columns named:`'event_date', 'fatality_count', 'Country', 'State',` 'landslide_trigger', and 'landslide_setting'. The result (1274) shows that when using this rather wide list of columns, there will be many duplicates. After that, the code optimizes the identification of duplicates with consideration of only four attributes which are 'event_date', 'fatality_count', 'Country', and 'State'. All the values in these four columns that match each other are labelled as duplicates and automatically deletes them from the DataFrame. The final check proved that the all duplicated rows are excluded from the DataFrame.



```
new_df.duplicated(subset=['event_date','fatality_count','Country','State','landslide_trigger','landslide_setting']).sum()
```

```
1274
```

```
new_df = new_df[~new_df.duplicated(subset=['event_date','fatality_count','Country','State'])]
```

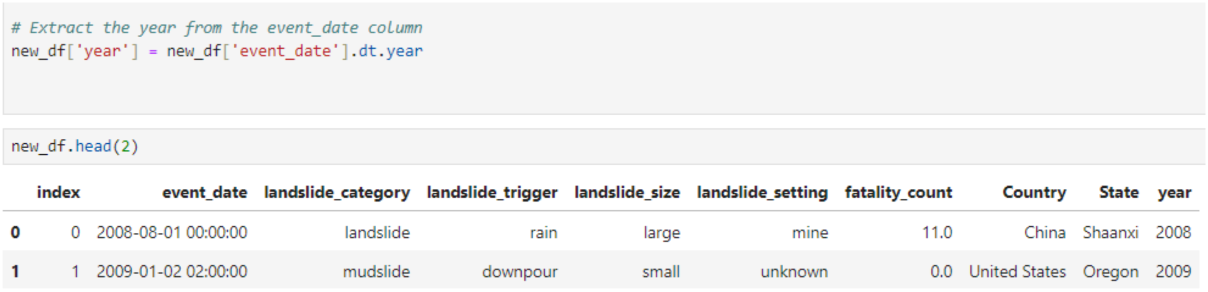
```
new_df.duplicated().sum()
```

```
0
```

vi. Step 6: Add new column ‘year’ in dataframe

This step modifies the structure of DataFrame by creating a new column that hold the year of those occurrences. The year extracted from each 'event_date' using feature selection of pandas called (.dt.year). using feature selection of pandas called .dt[year]. The extracted year information is then allocated to a new column labelled ‘year’, creating a new column with only the year of each landslide event to cater for time-based analysis. The last output demonstrates how the addition of this new column looks like in the DataFrame.

(h) .



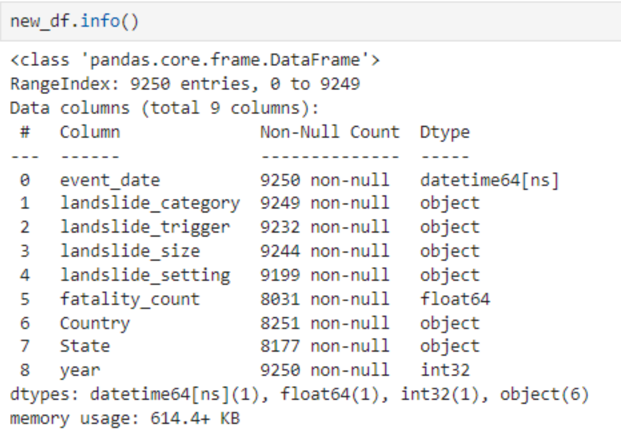
```
# Extract the year from the event_date column
new_df['year'] = new_df['event_date'].dt.year
```

```
new_df.head(2)
```

	index	event_date	landslide_category	landslide_trigger	landslide_size	landslide_setting	fatality_count	Country	State	year
0	0	2008-08-01 00:00:00	landslide	rain	large	mine	11.0	China	Shaanxi	2008
1	1	2009-01-02 02:00:00	mudslide	downpour	small	unknown	0.0	United States	Oregon	2009

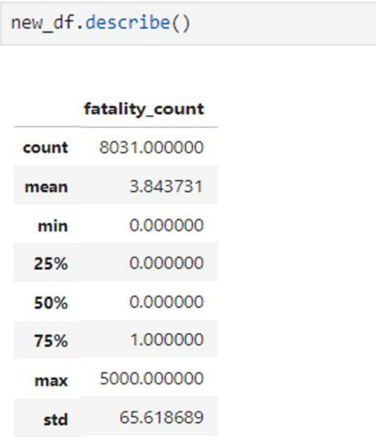
vii. Step 7: Data inspection and summary

(i) The function `.info()` gives brief details about the DataFrame which includes, the number of entries, columns, non-null values, and data types of each column. This reveals that the DataFrame has 9250 cases and 9 columns, with various data types such as `datetime64[ns]` (dates), `float64` (numerical data), i`nt32` (integers), and `object` (strings or mixed data types). The `.describe()` generates statistical information for numerical segments including count, mean, standard deviation, minimum, maximum, and percentile of the `'fatality_count'` column. Finally, .`head(3)` presented the first three rows of the DataFrame, offering a sample view of the data.



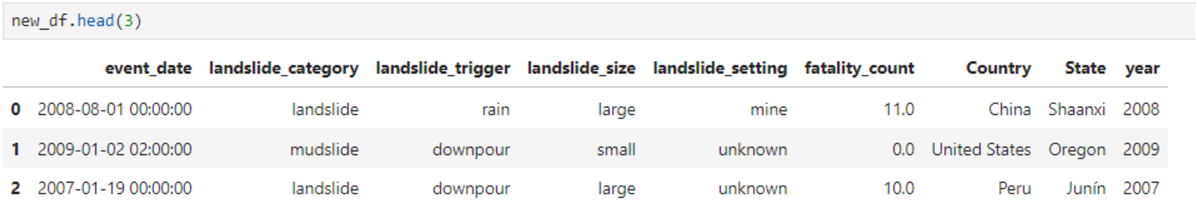
```
new_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9250 entries, 0 to 9249
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   event_date          9250 non-null   datetime64[ns]
 1   landslide_category  9249 non-null   object
 2   landslide_trigger   9232 non-null   object
 3   landslide_size      9244 non-null   object
 4   landslide_setting   9199 non-null   object
 5   fatality_count      8031 non-null   float64
 6   Country             8251 non-null   object
 7   State               8177 non-null   object
 8   year                9250 non-null   int32
dtypes: datetime64[ns](1), float64(1), int32(1), object(6)
memory usage: 614.4+ KB
```



```
new_df.describe()
```

	fatality_count
count	8031.000000
mean	3.843731
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	5000.000000
std	65.618689



```
new_df.head(3)
```

	event_date	landslide_category	landslide_trigger	landslide_size	landslide_setting	fatality_count	Country	State	year
0	2008-08-01 00:00:00	landslide	rain	large	mine	11.0	China	Shaanxi	2008
1	2009-01-02 02:00:00	mudslide	downpour	small	unknown	0.0	United States	Oregon	2009
2	2007-01-19 00:00:00	landslide	downpour	large	unknown	10.0	Peru	Junín	2007

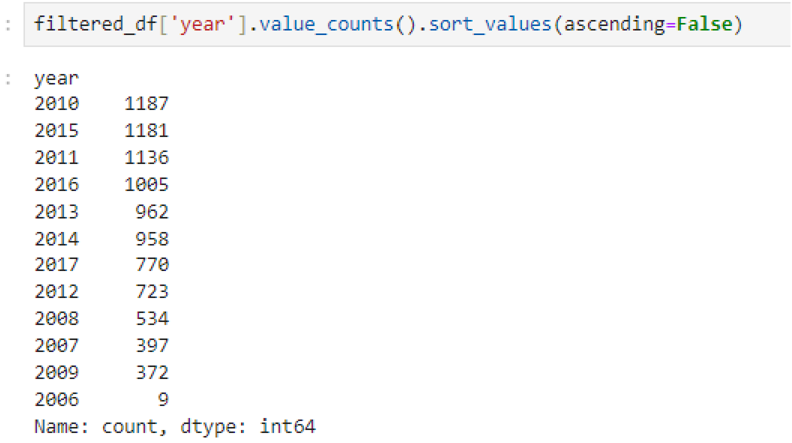
From cleaning and data preparation, it can be summarized that the dataset's columns can be categorized into two types: numerical and categorical. Quantitative data in numerical columns are as follows: `fatality_count` which describe the count of fatalities. Qualitative data is characterised by categorical columns and includes `landslide_category`, `landslide_trigger`, `landslide_size`, `landslide_setting`, `Country`, `State`, and `year`.

These columns hold labels or categories of the attributes of the landslide events. In this dataset, none of the columns is defined as containing mixed data type.

4.3 Descriptive data

4.3.1 Global

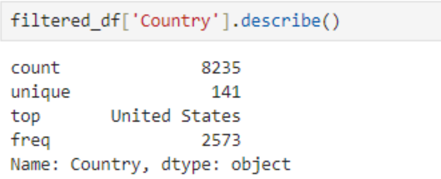
The `(filtered_df['year'])` indicates that the DataFrame was filtered by year from 2006-2017 then counts the data in descreasing order. The results shows that 2010 has the highest number of landslides occurrences with 1,187 cases, followed by 2015 with 1,181 cases, 2011 of 1,136 cases and so on. . On the other hand, the year 2006 had the least number of entries with only 9. As it is shown in the distribution below, this results in a considerably different number of data points for every year with some years represented in the list significantly more than others.



```
: filtered_df['year'].value_counts().sort_values(ascending=False)
```

```
: year
  2010    1187
  2015    1181
  2011    1136
  2016    1005
  2013     962
  2014     958
  2017     770
  2012     723
  2008     534
  2007     397
  2009     372
  2006       9
  Name: count, dtype: int64
```

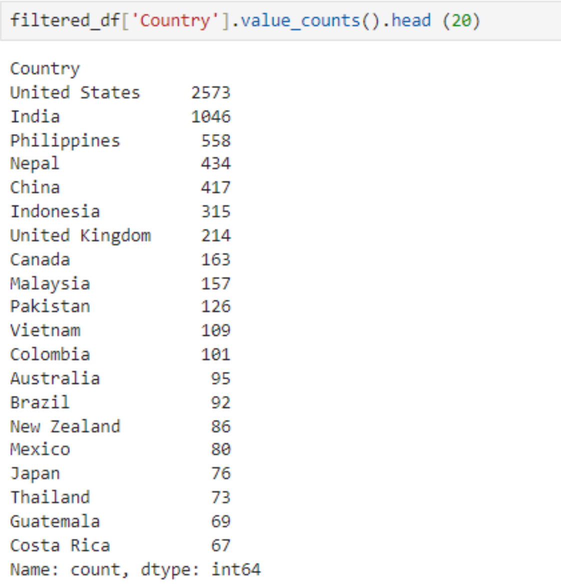
Next, the code of `filtered_df['country']. describe()`provide a total of 8226 entries for 141 countries. The most frequently used country is the "United States" which appears 2566 times, that again shows a highly unbalanced distribution of data based on countries. The data type is specified as an `'object'` means that country names are presented in their string format.



```
filtered_df['Country'].describe()
```

```
count              8235
unique              141
top       United States
freq               2573
Name: Country, dtype: object
```

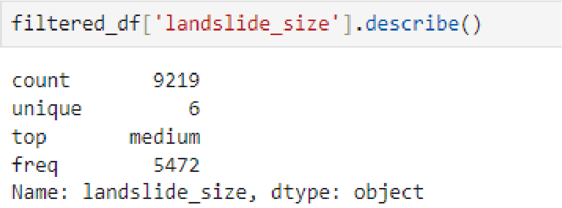
Then, `filtered_df['country'].value_counts().head(20)`, displaying the counts of the most frequently occurring 20 countries. This supports the previous analysis of the distribution where the United States have a significantly higher total number of recorded occurrences, at 2,566 cases followed with India of 1,046 cases and the Philippines with 557 cases. Some of these countries may be at a relatively higher risk, especially by varying climate zone, rugged topography, and potentially large reporting systems in the United States and India. Nepal, China, Indonesia also prominently captured due to high frequency of landslide in hilly tracts, rain, forest removal and earthquake-prone regions. Malaysia also in this category as this country is prone to landslides hazard due to intense rainfall especially during monsoon season.



```
filtered_df['Country'].value_counts().head (20)
```

```
Country
United States     2573
India             1046
Philippines        558
Nepal              434
China              417
Indonesia          315
United Kingdom     214
Canada             163
Malaysia           157
Pakistan           126
Vietnam            109
Colombia           101
Australia           95
Brazil              92
New Zealand         86
Mexico              80
Japan               76
Thailand            73
Guatemala           69
Costa Rica          67
Name: count, dtype: int64
```

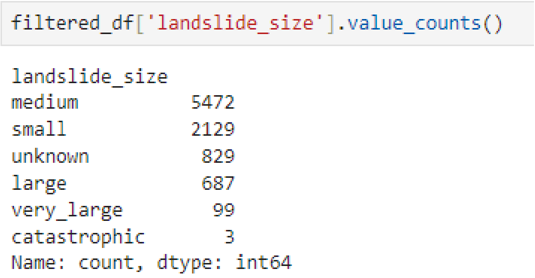
The `filtered_df['lanslide_size'].describe()` function resulting 9219 records and 6 different sizes of the landslide Some of them are presented in the following table In total, ‘medium’ size is observed most often and it is 5472 times. The data type is ‘`object`’ this suggests that sizes could be stored as string and not integers, or Boolean as expected. The fact that it repeats frequently is somewhat misleading: the number of landslides is probably dominated by the ‘medium’ size.



```
filtered_df['landslide_size'].describe()
```

```
count       9219
unique         6
top       medium
freq        5472
Name: landslide_size, dtype: object
```

The `filtered_df['landslide_size'].value_counts()`, gives a detailed frequency count for each landslide size. gives a detailed frequency count for each landslide size. This confirmed that medium landslide is the highest cases of 5472, followed by the small landslides indicated 2129 cases, large landslides of 829 cases, the unknown landslides size of 687, very_large landslide of 99 cases and catastrophic landslide is 3 cases. This shows that the medium sized landslides are more frequent than other sizes in the dataset. The “unknown” specification is an option within the database and the mere occurrence of such a deep impact on the slope structure, indicated by the classification of catastrophic landslides occurring so infrequently clearly shows that further refinement and enhancement of the data could be valuable as well as more research into the nature of infrequent events of landslides.

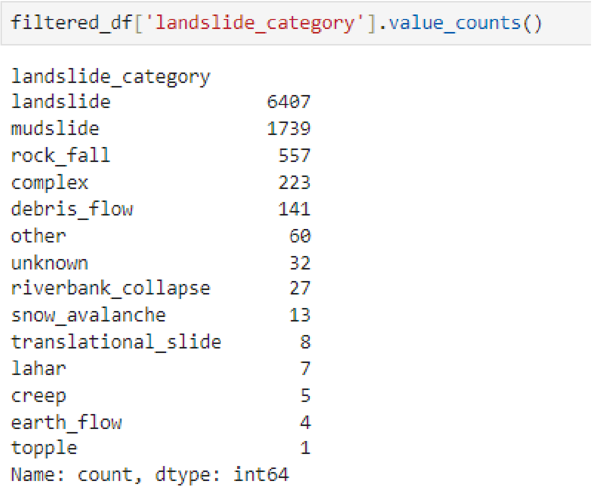


```
filtered_df['landslide_size'].value_counts()
```

```
landslide_size
medium          5472
small           2129
unknown          829
large            687
very_large        99
catastrophic       3
Name: count, dtype: int64
```

The code `filtered_df['landslide_category'].value_counts()` calculates the frequency of each value in the `'landslide_category'` column of the `filtered_df` DataFrame resulting the landslide is the most frequent category with 6,407 cases, followed by mudslide with 1,739 cases and rock fall of 557 cases.

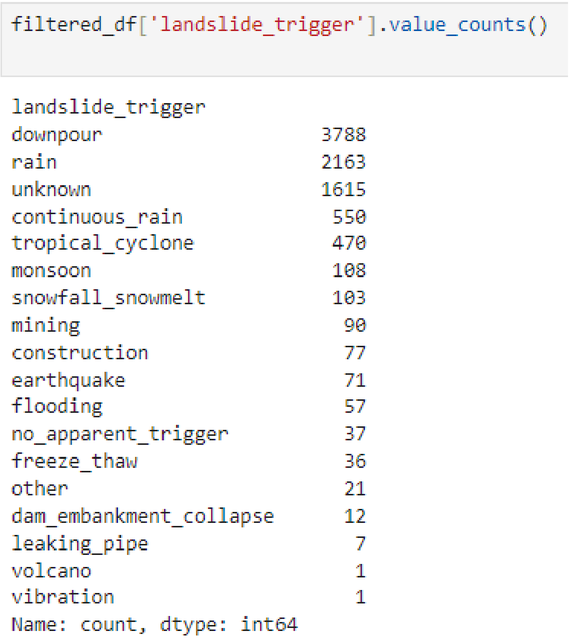
Complex has 223 cases, debris flow of 141 cases, and riverbank collapse, 27 cases. The unusual landslide category such "snow avalanche," "creep," and "earth flow" happened in small numbers. This distribution also demonstrates that few landslide categories namely “landslide” and “mudslide” are more common than the other types in the dataset that covers various types of landslides with a focus on a small number of types.



```
filtered_df['landslide_category'].value_counts()
```

```
landslide_category
landslide              6407
mudslide               1739
rock_fall               557
complex                 223
debris_flow             141
other                    60
unknown                  32
riverbank_collapse       27
snow_avalanche           13
translational_slide       8
lahar                     7
creep                     5
earth_flow                4
topple                    1
Name: count, dtype: int64
```

The snippet, `filtered_df['landslide_trigger'].value_counts(),` counts the occurrences of each unique trigger resulting the ouput of 9,207 landslide hazard with 18 unique triggering factors. Downpour is the most highest trigger factor which is 3,788 cases, followed by rain in 2,163 cases, unknown of 1615 cases, and continuous rain of 550 cases. Tropical_cyclone, monsoon, snowfall_snowmelt, and various human-induced triggers such as construction and mining provide lower value. The ‘unknown’ triggered factor undefined. These findings emphasize the importance of understanding rainfall's role in mitigating landslide risks.



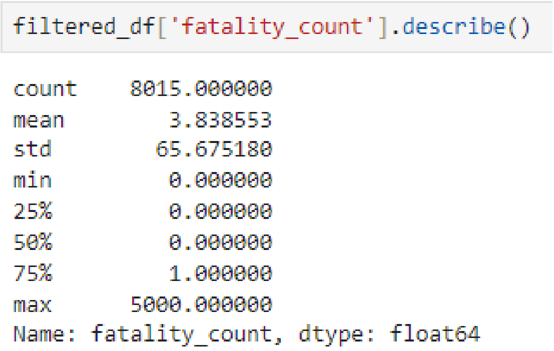
```
filtered_df['landslide_trigger'].value_counts()
```

```
landslide_trigger
downpour                 3788
rain                     2163
unknown                  1615
continuous_rain           550
tropical_cyclone          470
monsoon                   108
snowfall_snowmelt         103
mining                     90
construction               77
earthquake                 71
flooding                   57
no_apparent_trigger        37
freeze_thaw                36
other                      21
dam_embankment_collapse    12
leaking_pipe                7
volcano                     1
vibration                   1
Name: count, dtype: int64
```

The function of `filtered_df. 'fatality_count'.describe()` analyses fatality counts with 8,051 observations and mean 3.84 per landslide. Minimum numbers of people who lost their lives in a landslide event are 0 while the maximum fatalities are 5000. This means that most of the events have few fatalities as evidenced by 25th percentile= 0 and 50th percentile = 0. This implies that majority of the landslide incidences in the dataset did not result to any loss of lives. Positive number of fatalities, 75th percentile value is 1, have indicated that a quarter of the landslide events claimed more than one life.

Distribution of the statistics clearly demonstrates that mean is much higher than median value and the maximum statistic indicated in the table is much higher than other percentiles calculated. This indicates that the mean is skewed by a few occurrences which feature enormously high fatality numbers. This is even made clearer by the large standard deviation, signifying the relative high fluctuation in fatality quantities. It is also observed that most of the events have low fatality numbers (median and quartile 0 and 1 respectively); however, there are a few large values with fatality numbers as high as 500, and hence the mean may not be the most suitable meaningful measure of average fatality numbers as is the median. Another is the data type of fatality counts which is `float 64`, which means that fatality counts are held as floating-point numbers.

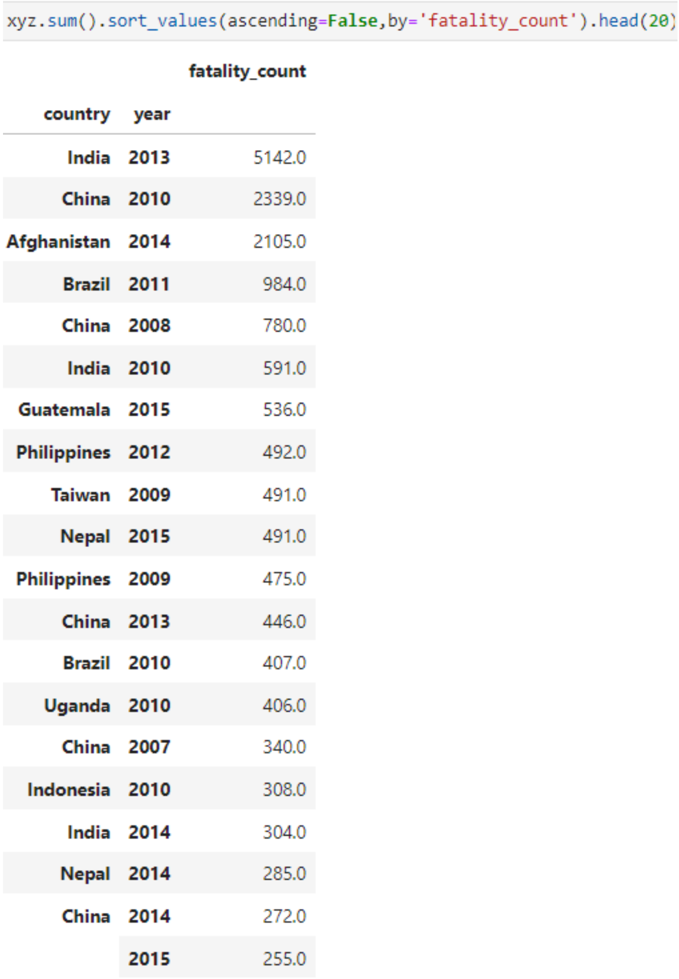
.



```
filtered_df['fatality_count'].describe()
```

```
count    8015.000000
mean        3.838553
std        65.675180
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max      5000.000000
Name: fatality_count, dtype: float64
```

The top 20 country combined with year of the highest landslide fatality counts sorted in descending order suggests that the highest number of fatalities occurred in India on 2013 with 5142 fatalities, followed by China in 2010 of 2339 fatalities and Afghanistan in 2014 of 2105 fatalities recoded. The list includes various countries and years, indicating that high-fatality events were not concentrated in a single location or year. Several countries appear multiple times, suggesting that some countries experience recurring high-fatality landslide events. This information could be valuable for identifying high-risk areas and for developing targeted disaster preparedness and mitigation strategies.

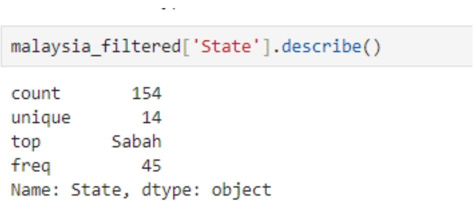


```
xyz.sum().sort_values(ascending=False,by='fatality_count').head(20)
```

		fatality_count
country	**year**	
India	**2013**	5142.0
China	**2010**	2339.0
Afghanistan	**2014**	2105.0
Brazil	**2011**	984.0
China	**2008**	780.0
India	**2010**	591.0
Guatemala	**2015**	536.0
Philippines	**2012**	492.0
Taiwan	**2009**	491.0
Nepal	**2015**	491.0
Philippines	**2009**	475.0
China	**2013**	446.0
Brazil	**2010**	407.0
Uganda	**2010**	406.0
China	**2007**	340.0
Indonesia	**2010**	308.0
India	**2014**	304.0
Nepal	**2014**	285.0
China	**2014**	272.0
	2015	255.0

4.3.2 Malaysia

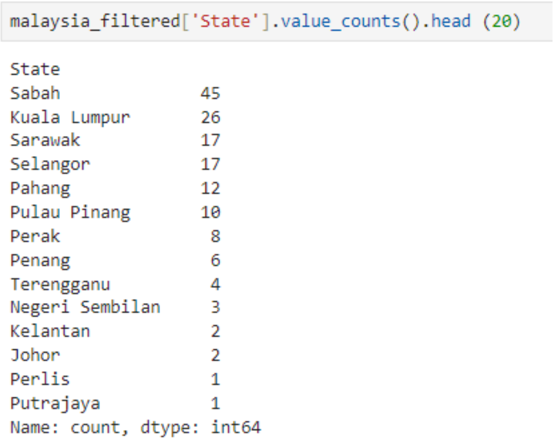
The `malaysia_filtered['state'].describe()` indicate the summary of landslide hazard occurred in Malaysia's state on 2006-2017 including the total count of 154 landslide occurrences and the number of unique states is 14. The highest recorded is Sabah that appearing 45 times and data type confirmation is string. As this overview also demonstrates, the total size of the dataset and the general imbalance of state representations is also rather unequal. The size of this selected sample indicates that only 14 out of 16 states and federal territories are represented raises the possibility that the sample may have been compiled with some limitation or bias, which, in turn, may need to be also considered for subsequent analyses.



```
malaysia_filtered['State'].describe()
```

```
count        154
unique        14
top        Sabah
freq          45
Name: State, dtype: object
```

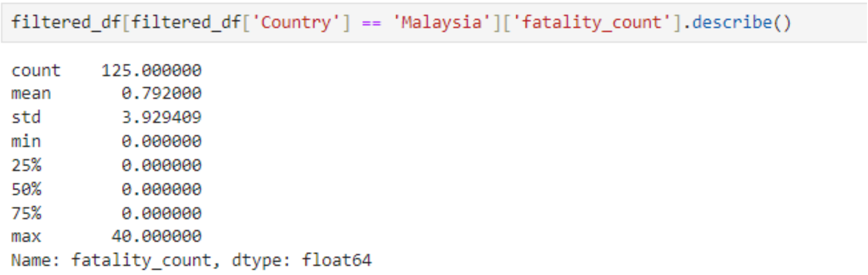
The `malaysia_filtered['State'].value_counts().head(20)`, describe the frequency distribution of the Malaysia’s states. It shows the 20 most frequently landslide occurring states in Malaysia. Sabah experience the highest number of landslide with 45 cases followed by Kuala Lumpur of 26 cases, Sarawak and Selangor both is 17 cases, Pahang 12 cases and Pulau Pinang with 10 cases. Other states shows the value less than 10 cases recorded.



```
malaysia_filtered['State'].value_counts().head (20)
```

```
State
Sabah              45
Kuala Lumpur       26
Sarawak            17
Selangor           17
Pahang             12
Pulau Pinang       10
Perak               8
Penang              6
Terengganu          4
Negeri Sembilan     3
Kelantan            2
Johor               2
Perlis              1
Putrajaya           1
Name: count, dtype: int64
```

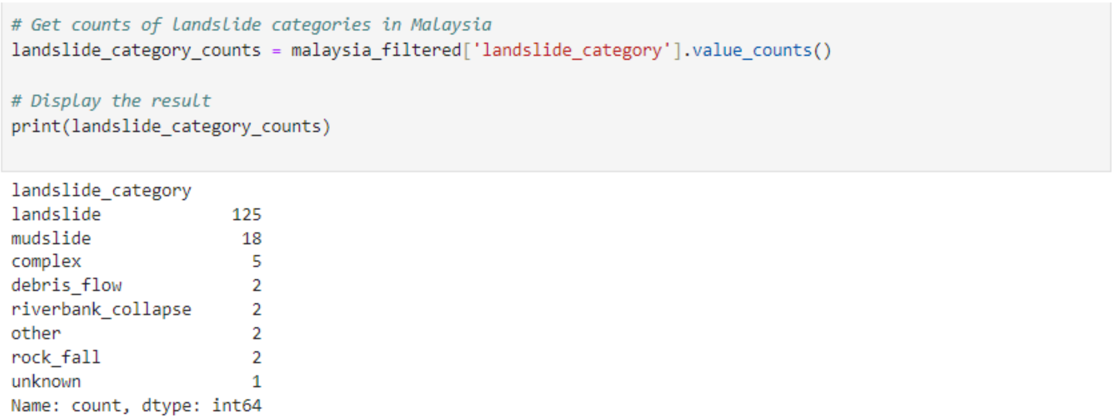
The command of `filtered_df[‘filtered_df’] == ‘Malaysia’]{fality_count’].describe()` calculates the following descriptive measures for ‘`fatality_count’` whereas the analysis is restricted to Malaysia only. The output, it can shows there is 125 observation, the mean of the fatality count is equal to 0.792 and the standard deviation is equal to 3.924. The minimum value is 0, maximum value is 48, which show that majority of the landslides lead to few or no fatalities, while few cause several fatalities. The same idea can be also derived from the position of the median which is 0; In general, the distribution is shifted significantly to the right.



```
filtered_df[filtered_df['Country'] == 'Malaysia']['fatality_count'].describe()
```

```
count    125.000000
mean       0.792000
std        3.929409
min        0.000000
25%        0.000000
50%        0.000000
75%        0.000000
max       40.000000
Name: fatality_count, dtype: float64
```

The code of landslide_category_counts = Malaysia_filtered [‘landslide_category’].value_count()determines the frequency of landslides category in Malaysia. The results show landslide has highest frequency of 125 cases followed by mudslides (18), and complex (5). The rest categories recorded 2 and below where the ‘unknown’ categories need to defined for the improvement and accuracy of the data. It can be concluded that landslide categories dominated by landslide and mudslide due to the topography and climate of Malaysia.

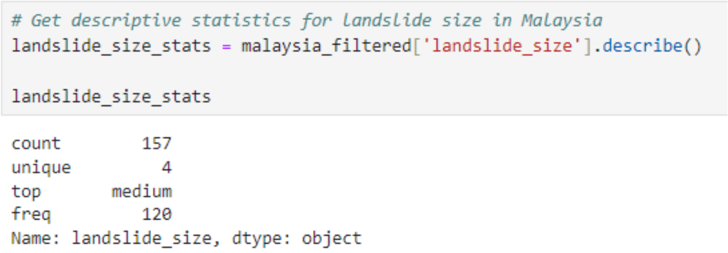


```
# Get counts of landslide categories in Malaysia
landslide_category_counts = malaysia_filtered['landslide_category'].value_counts()

# Display the result
print(landslide_category_counts)
```

```
landslide_category
landslide            125
mudslide              18
complex                5
debris_flow            2
riverbank_collapse     2
other                  2
rock_fall              2
unknown                1
Name: count, dtype: int64
```

The function of landslide_size_stats = malaysia_filtered['landslide_size'].describe() method to create summary statistics for the 'landslide_size' which resulting count of 157 entries, 4 unique size categories, and the top landslides size is medium that occured 120 cases. The data type is as an object (string).

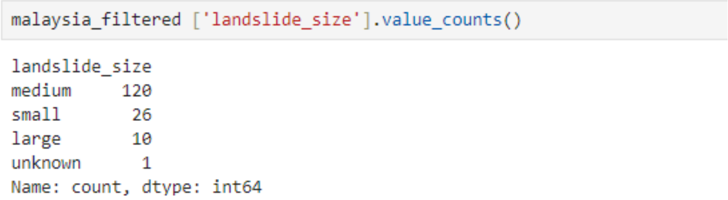


```
# Get descriptive statistics for landslide size in Malaysia
landslide_size_stats = malaysia_filtered['landslide_size'].describe()

landslide_size_stats
```

```
count        157
unique         4
top       medium
freq         120
Name: landslide_size, dtype: object
```

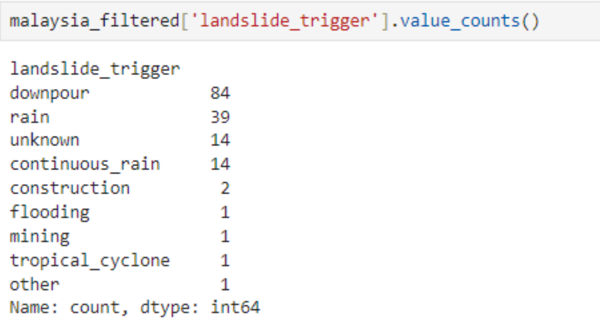
The `malaysia_filtered['landslide_size'].value_counts()` code provides a detailed frequency count for each landslide size category. This confirms that "medium" is indeed the most frequent (120), followed by "small" (26), "large" (10), and "unknown" (1). This analysis reveals a skewed distribution of landslide sizes in the dataset, with a clear dominance of "medium"-sized landslides. The small number of "unknown" entries suggests relatively good data quality regarding landslide size classification.



```
malaysia_filtered ['landslide_size'].value_counts()
```

```
landslide_size
medium     120
small       26
large       10
unknown      1
Name: count, dtype: int64
```

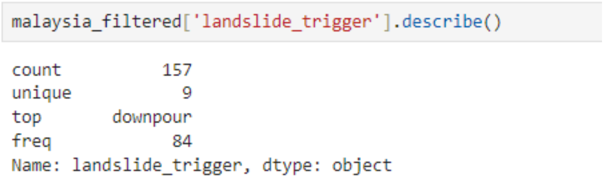
The command, `malaysia_filtered['landslide_trigger'].value_counts()`, displayed the frequency of each landslide trigger. The highest landslide triggered factor is downpour that recorded 84 cases followed by rain of 39 cases, unknown and continuous_rain with both 14 cases. The remaining triggered factor of landslide including human activities such as construction and mining, environment hazard such as flooding, tropical cyclone, and other that rarely occurred which is only once or twice.



```
malaysia_filtered['landslide_trigger'].value_counts()
```

```
landslide_trigger
downpour            84
rain                39
unknown             14
continuous_rain     14
construction         2
flooding             1
mining               1
tropical_cyclone     1
other                1
Name: count, dtype: int64
```

Lastly, the command `malaysia_filtered['landslide_trigger'].describe()` confirmed that 157 total entries with 9 unique triggers, and the highest landslide triggered factor is downpour which happened 84 cases in Malaysia. The data type is strongly defined as the object (variable of the string type, most likely). A clear inference from this analysis is that events associated with rains; downpour, rain, continuous rain, are the principal causes for landslides in this study area in Malaysia. Thus, the study has outlined the factors that define landslide susceptibility in the meteorological factors through the different types of rainfalls. It should be noted that there were many more instances of ‘unknown’ triggers, therefore, these findings imply a requirement for more comprehensive data in following research.



```
malaysia_filtered['landslide_trigger'].describe()
```

```
count          157
unique           9
top       downpour
freq            84
Name: landslide_trigger, dtype: object
```

4.4 Visualization

4.4.1 Global

The line chart on Figure 4.4.1 provides a visual representation of the variations in landslide counts in different years that indicate the trends and anomalies. The increasing trend in landslide cases from 2006 through the mid-2010s indicating there is possibility of high precipitation during that time. However, the dropping in 2012 and 2014 may reflect periods of reduced landslide hazard or inconsistent data collection The results suggest that research into environmental factors such as rainfall distribution, seismicity, and land use may offer a fuller explanation for these trends.

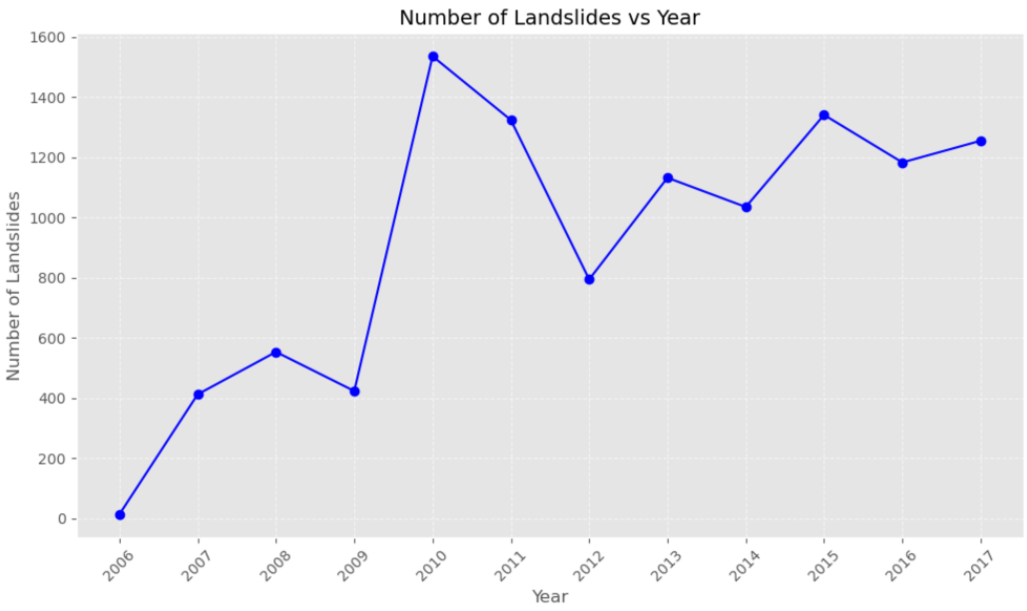


Figure 4.4.1 Line Chart

The bar chart in Figure 4.4.2 show that frequencies of occurrence of landslides drastically vary between the top placed countries and the rest. For instance, Costa Rica and Guatemala suspect much lower incidences of over 70 events possibly because the mentioned countries have smaller land mass or compile inadequate reports. Malaysia also falls on the top 20 of the country which experience most of the landslide’s event. The variability of data implies that geophysical characteristics, environmental conditions, population distribution and networks inf landslide incidences and reporting. More research could be made analysis of the relationship with other factors, such as the intensity of rainfall, stability of the ground and policies in place for disasters and how to avoid them.

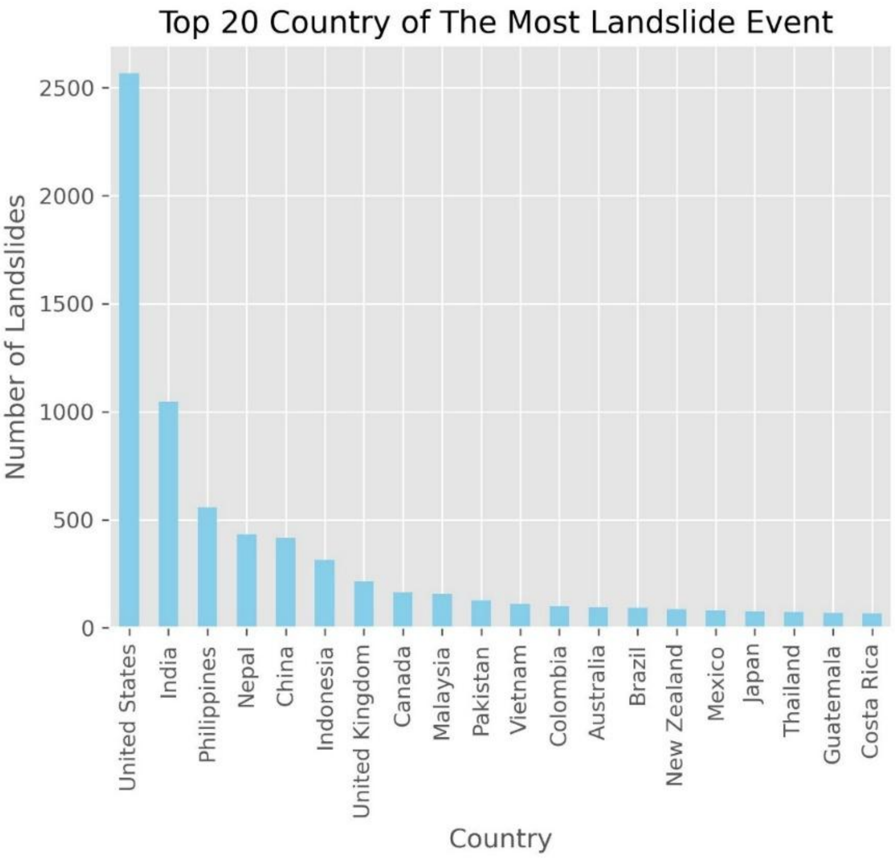


Figure 4.4.2 Bar Chart of Top 20 Countries of Most Landslide Cases

This pie chart in Figure 4.4.3 the proportion of different sizes of landslide. The largest portion, with 59.3% of the overall value, is the “medium” size that remains the most popular. This suggests that most of the landslides have been in this size range because large slumps are rare. “Small” movement type comprises 23.1% of all the movements, while the “large” movement types comprise only 7.5%. There is also 9.0% of ‘unknown’ sized landslides which can mean that there are still some ways to go about sorting out the size classification of some of these landslides. Last of all, “Very Large and Catastrophic” events according to the map only cover 1.1% of the total area. Distribution of the sized landslides fractionally tilts to a higher distribution of ‘medium’ sized landslide occurrences. This tends to imply that the geological conditions per tendency may be better suited for the generation of medium-sized landslides.

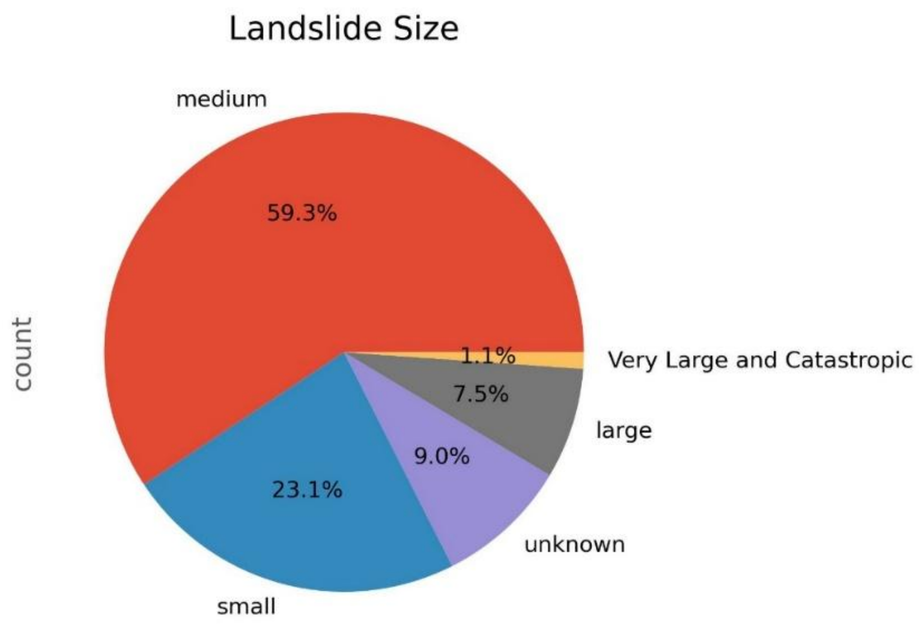


Figure 4.4.3 Pie chart of global landslide size

Landslide categories and percentages are depicted in this pie chart on Figure 4.4.4. The "landslide" is by far the most prominent category, which constituting 69.4% of the total. This indicates that most landslides are bound to fall under this classification to a very large extent. The second most common category is 'mudslide' with overall occurrence of 18.9%. The remaining categories composing the "Other" category include more specific types of mass movements and they include rockfall which has a 3.4%, complex 1.1%, debris flow 1.4%, and the remaining 0.2% making the proportions sum to about 100%. The 'Other' group is the smallest at 1.7 percent, which means that most of the landslides need to be grouped in some of the other described categories. The chart shows that the distribution is highly skewed, and 'landslide' is used considerably oftener than any other category. This implies that geological conditions that trigger, conditions triggering and environmental factors are mostly in favor of this sort of a landslide.

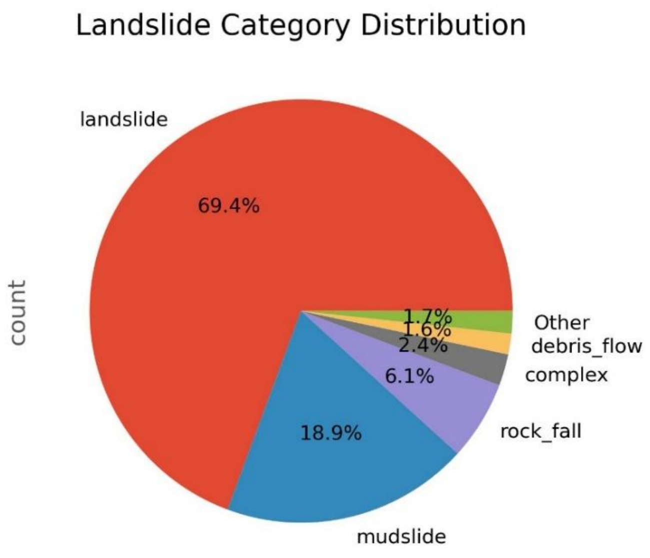


Figure 4.4.4 Pie chart of global landslide category Distribution

Pie chart of Figure 4.4.5 exhibits the distribution of landslide triggering factors. Downpour dominated the pie chart with 41.2 % contribution. This points to the fact that short duration and extreme rainfall produces significant amount of influence on landslides. Rain is the second highest of trigger with 23.5% that indicate the high intensity of the rainfall. "Unknown" recored 17.5%, suggesting that there is limited information available for a significant portion of landslides. Consequently, continuous rain contributes to 6.0 % indicate that heavy rainfall is also has an influence. Tropical cyclone contributed to 5.1% which testimony the role of cyclonic weathers. Other is at 6.7% categorizing all the rest of the triggering factors. From the above chart, it is evident that rainfall, in its different forms; downpour, rain, continuous rain most at times initiate the occurrence of landslides.

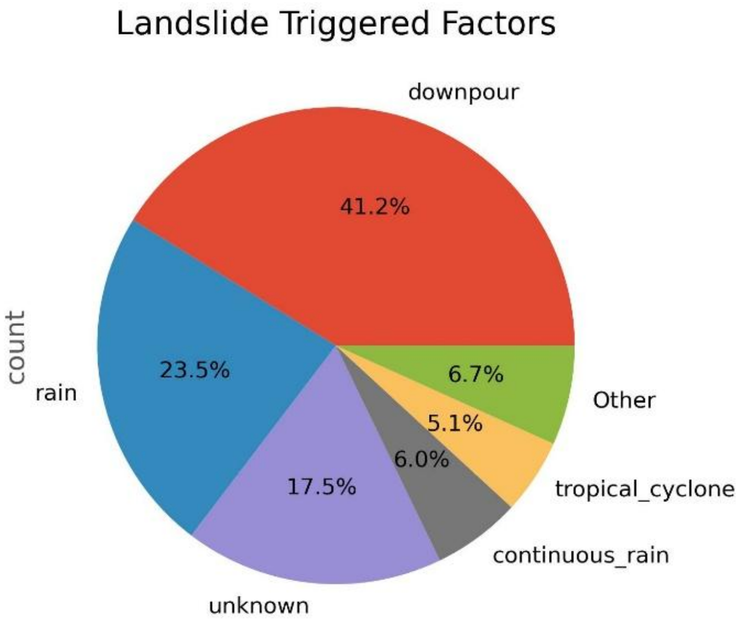


Figure 4.4.5 Pie chart of global landslide triggered factors

The most affected countries by the number of landslide events demostrated in Figure 4.4.6. India and China are at the top of the chart with the highest observed landslides, each over 200. These countries are followed by Indonesia, the Philippines and Nepal which report between 100 and 200 landslide events. Landslide events in United States, Colombia, and Vietnam are relatively moderate in comparison to other countries in terms of shale related landslides while Bangladesh, Malaysia and Italy are of relatively low intensity but they themselves are not negligible in terms of landslides. This visualization shows the extent to which landslides have affected countries geographically and why some countries are more affected than others especially the Asian countries might be due to their ground topography, climates and extreme weathers.

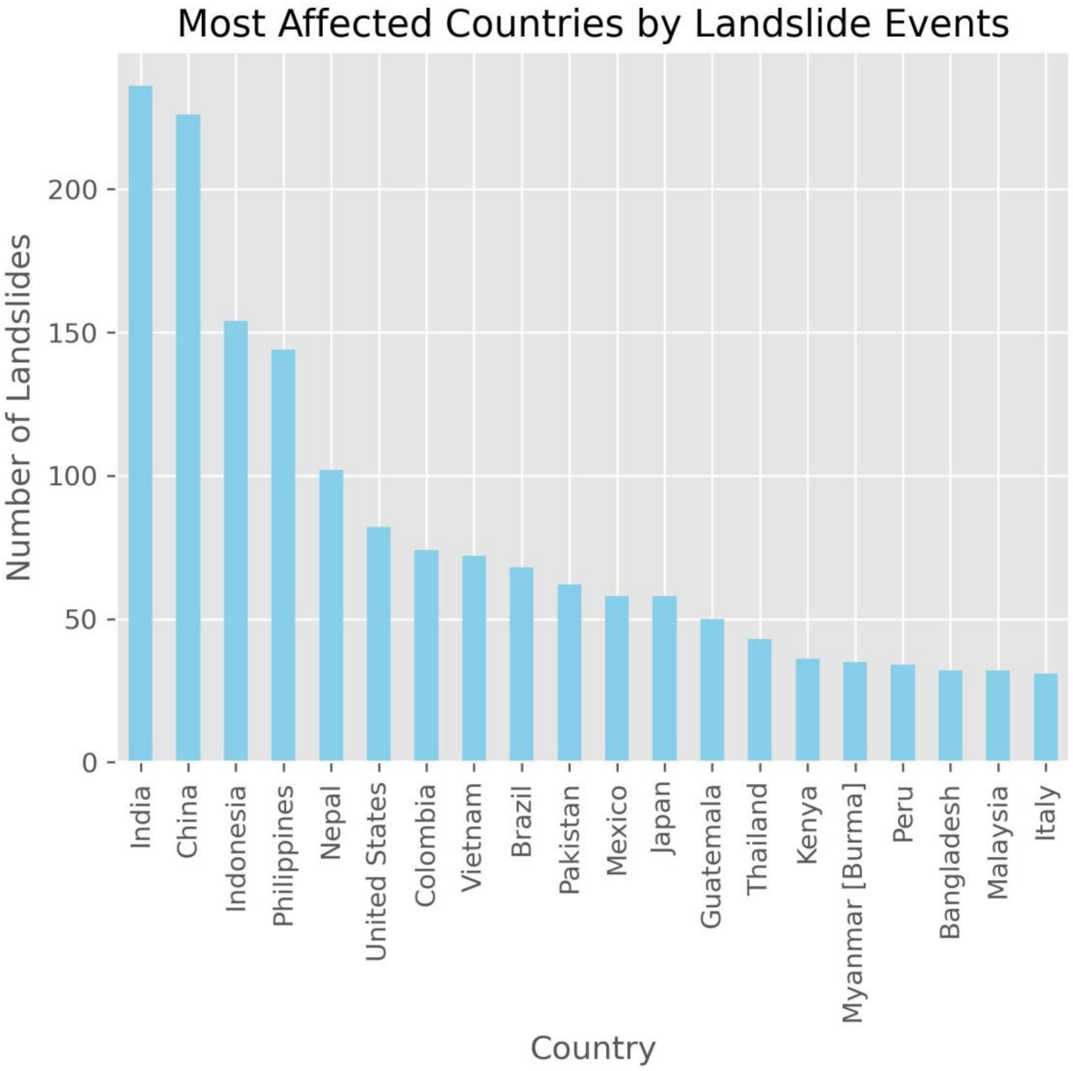


Figure 4.4.6 Bar chart of global most affected countries

4.4.2 Malaysia

The line graph on Figure 4.4.7 shows the Malaysian landslide occurrences between 2006 and 2017. The fluctuating pattern with low number of events seen in 2006 and 2009. There is one intensive wave in 2008 the indicators are reduced and increase in 2011 and 2013 until the highest peak recorded in 2014. The trends sharply drop after 2014 and fluctuates with the least number of events in 2016. The graph indicate that landslide activity in Malaysia during the duration was not consistent with trend of increased and decreased data.

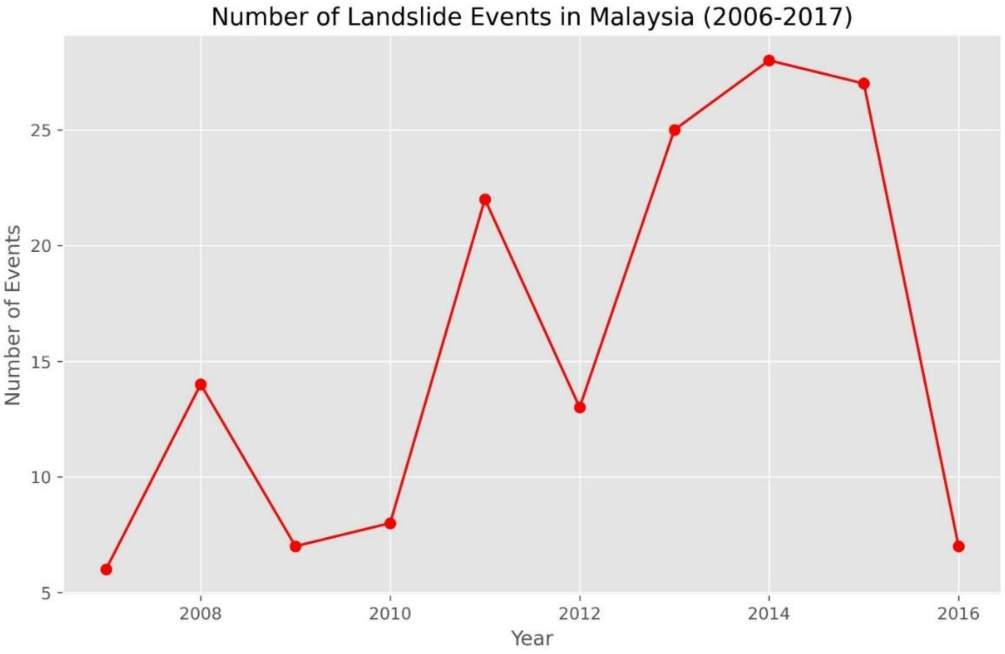


Figure 4.4.7 Line chart of Malaysia Landslide by year

The bar chart in Figure 4.4.9 illustrated the distribution of landslide activity in different states in Malaysia. Sabah recorded the highest number of landslides followed by Kuala Lumpur. The landslides cases for Sarawak, Selangor, Pahang, Pulau Pinang, Perak, Penang, Terengganu, Negeri Sembilan, Kelantan, Johor, Perlis, and Putrajaya decreasing with Putrajaya experiencing the fewest landslides. This shows that there is a distinct regional distribution of the landslide incidence with Sabah state as a clear example. It appears from the data that state peculiarities including topography characteristics, precipitation impacts, and land management all play a very big part in the occurrence of landslides.

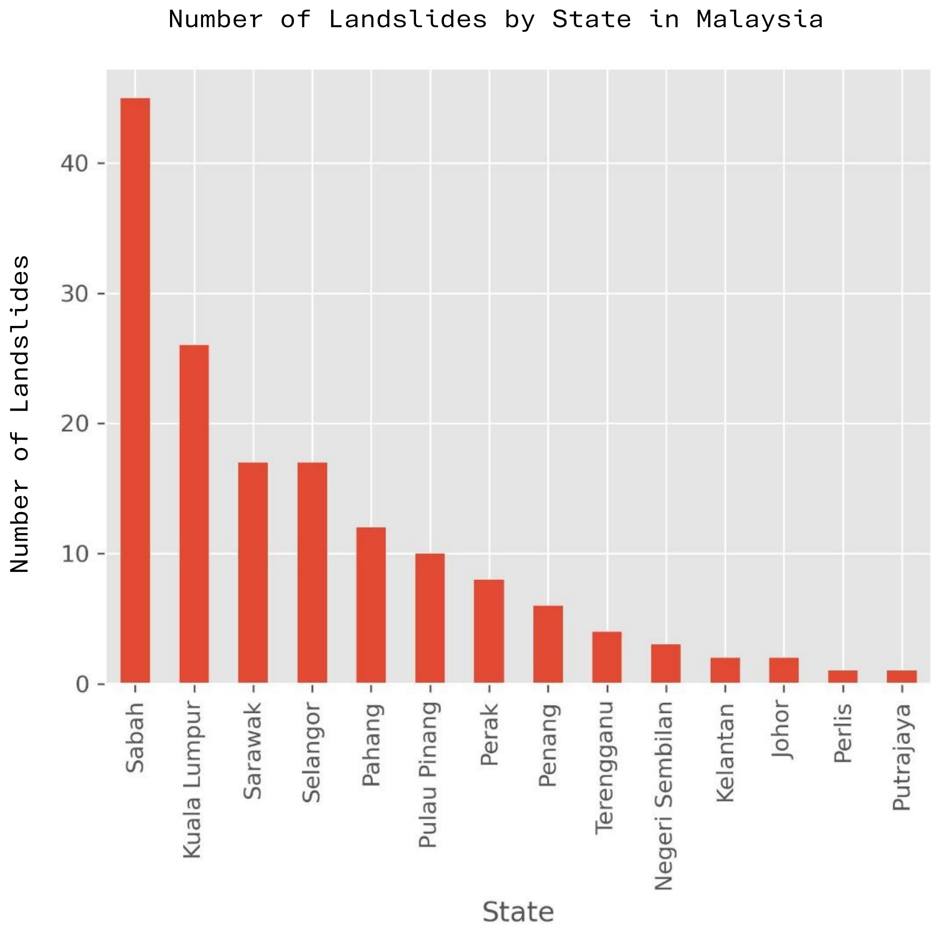


Figure 4.4.8 Bar chart of landslide cases in Malaysia’s state

The bar chart of Figure 4.4.9 shows the top 6 state of number of fatalities cause by landslide activity in Malaysian. Sarawak recorded the highest number of fatalities, followed by Pahang and Selangor. Selangor and Pahang experienced a considerably lower number of fatalities than Sarawak. Kuala Lumpur, Sabah, and Perak reported even fewer landslides, with Perak having the lowest number affected from landslide activity. This visualization highlights a significant disparity in landslide occurrences across these six states, suggesting that geographical factors and other environmental conditions play a crucial role in determining landslide susceptibility.

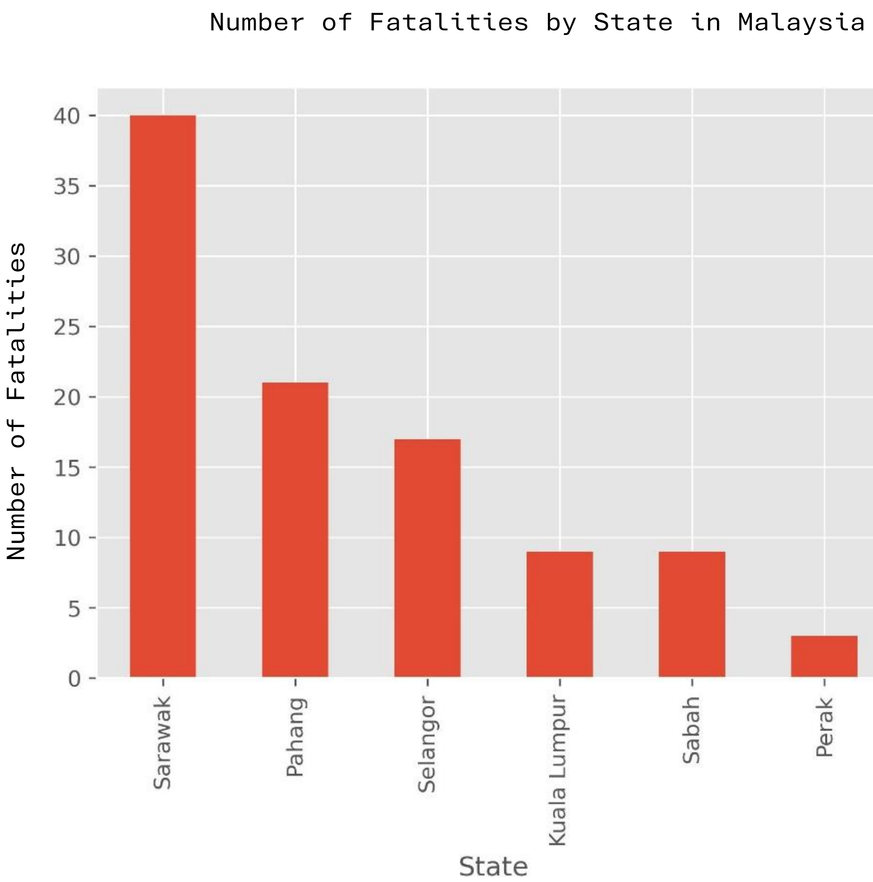


Figure 4.4.9 Bar chart of landslide fatalities in Malaysia’s state

The following pie chart of Figure 4.4.10 shows the Malaysian landslide sizes. Thus, the most unique feature of the database can be referred to the largest group of “medium” sized landslides, which make 76.4% of all cases. This suggests that majority of landslides in Malaysia are of medium size. Small landslides occupy the second biggest share of 16.6% while large landslides represents another 6.4%. However, there is 0.6% landslides data that comes under ‘unknown’ type, even though there still can be gaps with the data on the size of landslides of the overall 5 classification types might be quite representative. The chart clearly shows a skewed distribution, with a strong dominance of medium-sized landslides. This suggests that the geological conditions and triggering mechanisms in Malaysia may be more conducive to the formation of medium-sized landslides

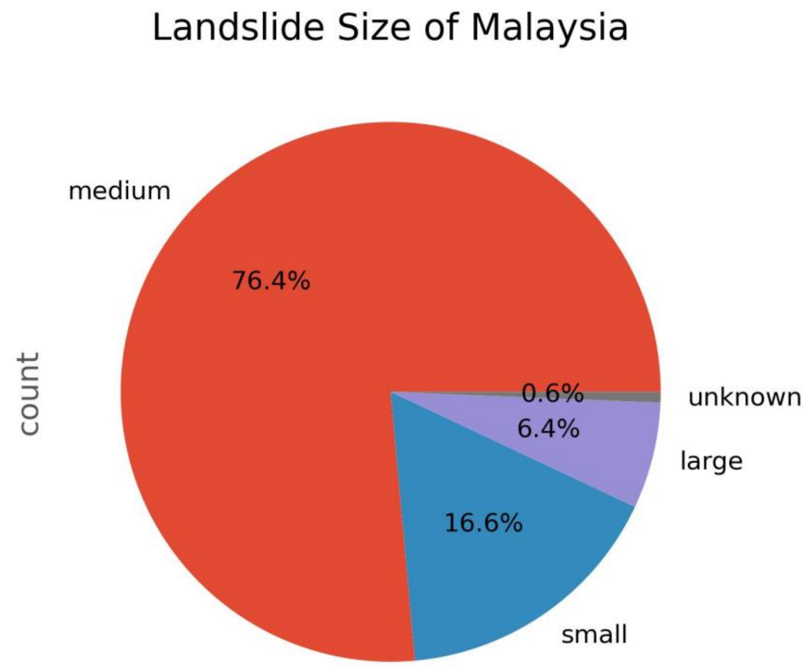


Figure 4.4.10 Pie chart landslide size in Malaysia

Figure 4.4.11 of pie chart illustrates the percentage of Malaysia’s landslide categories. The largest category identified is “landslide” which takes a huge 79.6% of the entire sample. This implies that the greater part of the landslide’s occurrences experienced in this country is classified under this type. The second largest percentage (11.5%) was recorded under the name of “mudslide” which also suggests a great number of landslides are of mudflow type. The other categories include the complex, debris flow, riverbank collapse, other, rock fall, and unknown, which are quite small having an average of less than 3.2%. There is only 0.6% of data classified as “unknown,” which means that landslide categorization is quite accurate. From the chart itself, it can very easily be seen that the distribution is way off on the end with the term ‘landslide’ being significantly more common than all the other terms combined. This leads to the assumption that the geology, factors that initiate, and environment in Malaysia favor this kind of slope failures.

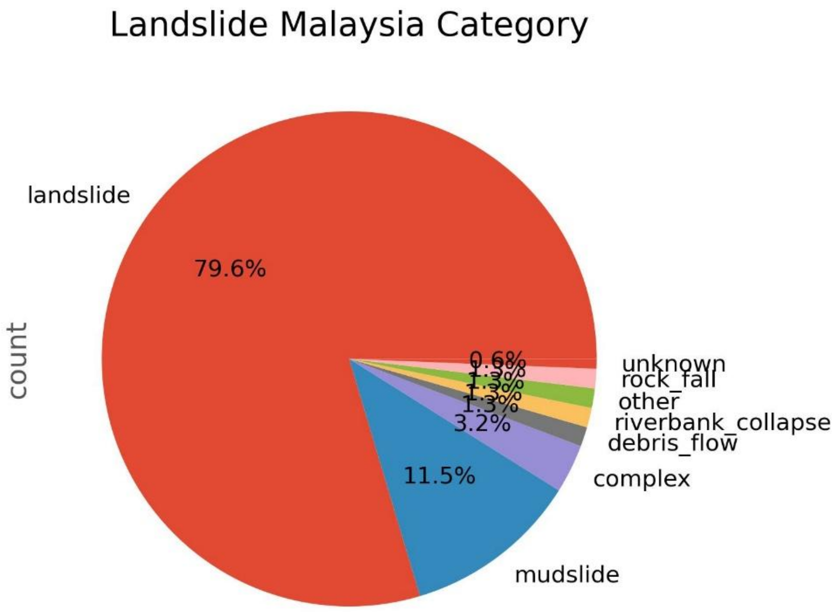


Figure 4.4.11 Pie chart of landslide category in Malaysia

Pie chart of Figure 4.4.12 shows distribution of landslide triggering factors. The largest contribution is toward the category “downpour” that stands 53.5 percent. This has underlined the major role played by heavy convectional rainfall in triggering landslides. The second most used trigger is “Rain” which occurs 24.8% more, which supports the storm hypothesis as the main cause. This shows that heavy rainfall consisting of continuous rain accounts for 8.9 % in overall contribution to landslide frequency. Percentages for “Unknown” also amount to 8.9 % suggesting that there could be episodes where lack of information or its unavailability led to landslides. Lastly, the rest of the triggering factors make up only 3.8% of the entire group which was categorized under ‘Other’. Statistics provided in the chart unambiguously indicate that rainfall in all its manifestations of downpour, rain, and continuous rain is the main cause of landslides. This mean that areas experiencing heavy or persistent rainfall are most likely to be affected by landslides. A large proportion of the events falls under the “unknown” category meant, there could be either gaps in data gathering or data analysis, that should be covered in the subsequent research to enhance the understanding of the landslides’ triggering mechanisms.

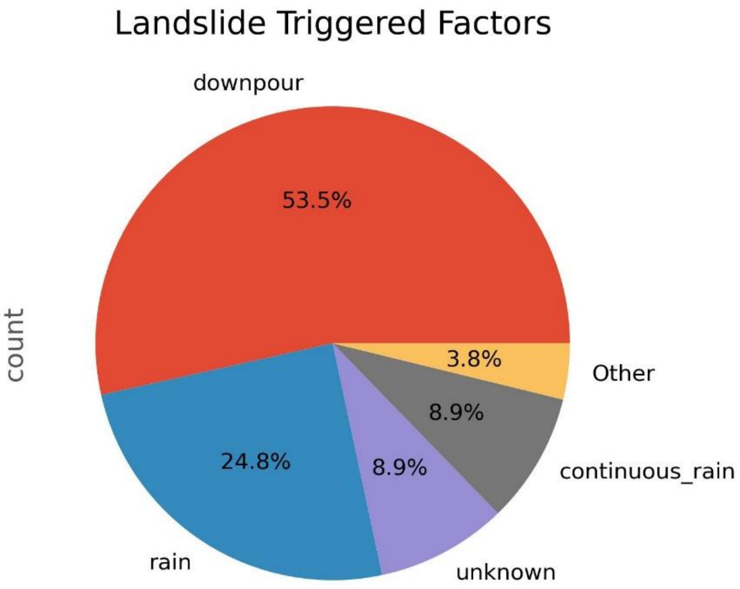


Figure 4.4.12 Pie chart of landslide triggered factor in Malaysia

4.5 Statistical Insights

4.5.1 Global Scatterplot Analysis

The scatterplot of Figure 4.5.1 indicate the number of fatalities caused by landslides from 2006 to 2017, categorized by landslide size. When it comes to year-to-year variability in the fatal accident rate, there is no trend: casualties tend to fluctuate erratically. 2010 and 2014 show a higher number of fatalities where it can not be realted with the number of deaths. The size of the landslide affect on fatalities where larger landslides correlated with higher fatality counts, even though no clear relationship based on he scatter in the data.

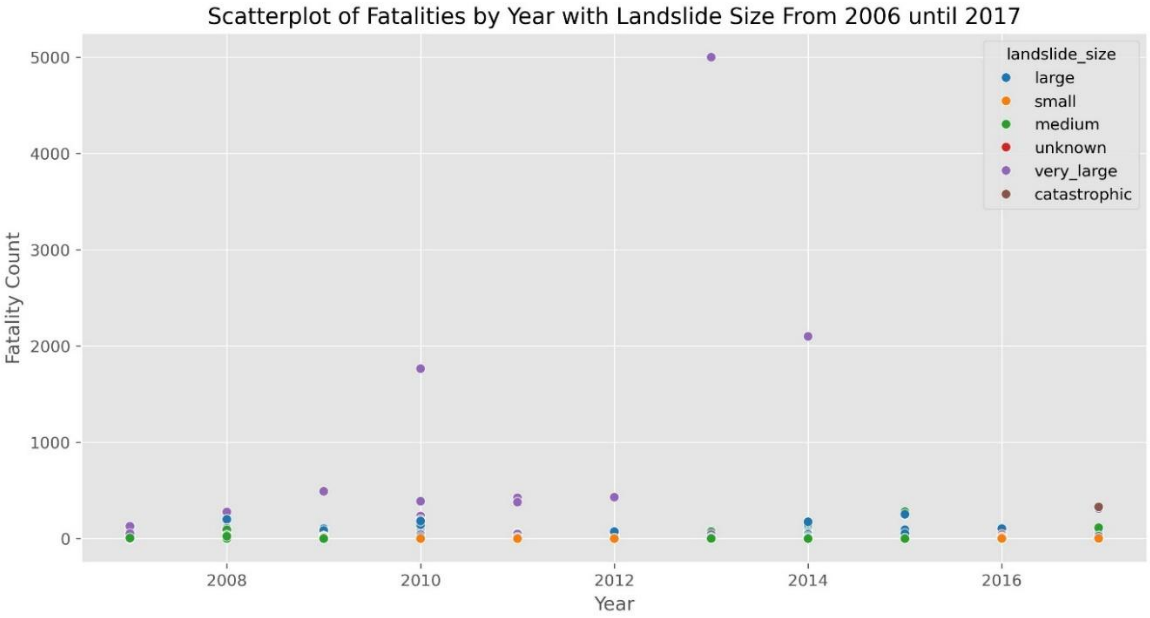


Figure 4.5.1 Global scatterplot of landslide events

4.5.2 Malaysia Scatterplot Analysis

The scatter plot of Figure 4.5.2 demostrate the number of fatalities due to landslides incident in Malaysia between 2006 and 2017, categorized by size (medium, small, large, size unknown) and type of landslides (landslide only, mudslide, complex, debris flows, riverbank failure, other, rock fall, type unknown). The data resulting the inconsistent fatalities year-to-year, with no specific trend. Even though medium-sized landslides seem to entail a higher number of deaths in some years, there are certain limitations in analyzing the results, due to a high percentage of unknown categories both by size and by type. This variety of landslide types however influence fatalities raise the proposition that there are multivariate aspects about these occurrences. Consequently, the variability in their data over the years and across categories do not allow for assessment of relative risk of the various types and sizes of landslides that occurred in this region during the specified period.

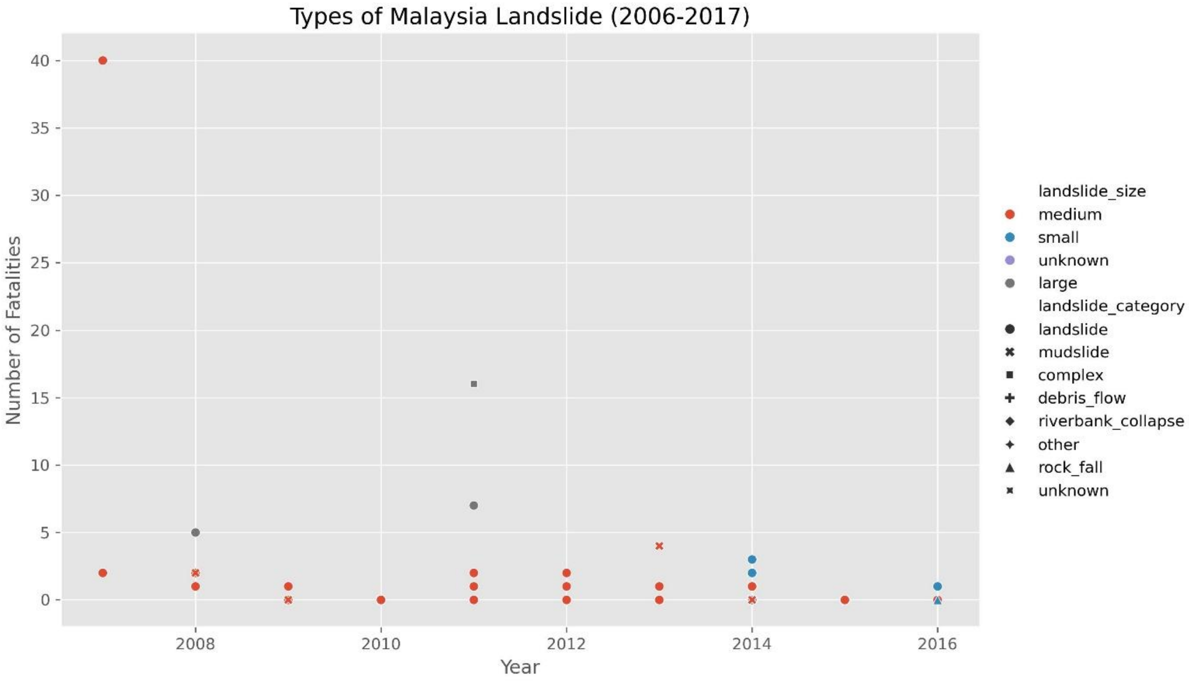


Figure 4.5.2 Malaysia scatterplot of landslide events

4.5.3 Malaysia Boxplot Analysis

The boxplot of Figure 4.5.3 focuses on the impact of the number of landslide fatalities and the different landslide categories in Malaysia. The y axis gives different types of landslides namely: ‘landslide,’ ‘mudslide,’ ‘complex,’ etc while the x axis represents the number of fatalities. The number of landslides events can be counted from each circle distinctive to a particular event. It is shown in the plot that majority of the landslide classes are depicted to have low values of fatalities less than 5. The first one ‘’landslide’’, indicates the highest number of deaths, several cases with more than 0 and one with approximately 40 meaning it was a severe case. Other categories indicate cases of instances or no deaths at all. Interestingly, the plot shows that no direct relationship can be established between some of the specific types of landslides and the high fatality rates, except for the general classification where the ‘landslide’ type was ranked highest in terms of variability of fatalities.

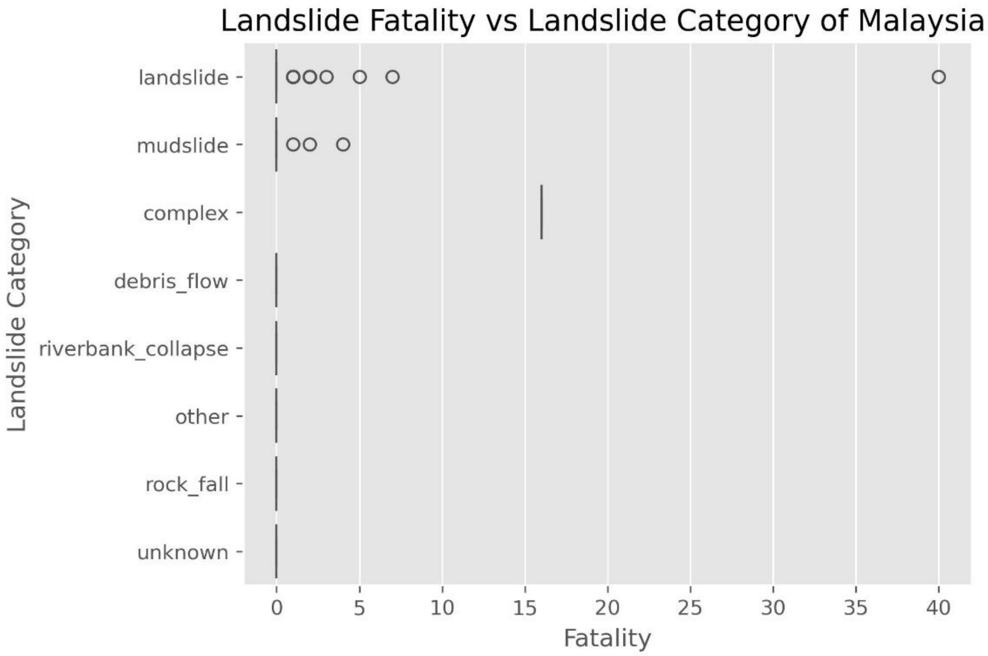


Figure 4.5.3 Malaysia boxplot of landslide category vs fatality

Figure 4.5.4 shows the deaths resulting from landslides regarding landslide size, categorized as “medium”, “small”, “unspecified,” and “large”. The horizontal axis is the number of deaths, while the vertical axis is categorised by size of the landslide. The rectangle denotes the interquartile range (IQR) of the data and a line within this rectangle represents the median of the data set. The whiskers go up to the level of the highest and lowest data points up to 1.5 the IQR. Any values beyond this range are given as outliers. The plot also shows that large landslides have an average of around 4 fatalities per event, and this value has a relatively little variability. Moderate and small-scale landslides have more variability, showing that most of the events only cause a small number of deaths but with some values showing a few events with much higher fatality figures. There are virtually no data points for size of ‘unknown’ landslides for ‘small’ and ‘medium’ sized landslides, the plot shows that the variability in the number of fatalities is high though it does show that there is broadly a gradient of increasing fatalities as the size of the landslide increases.

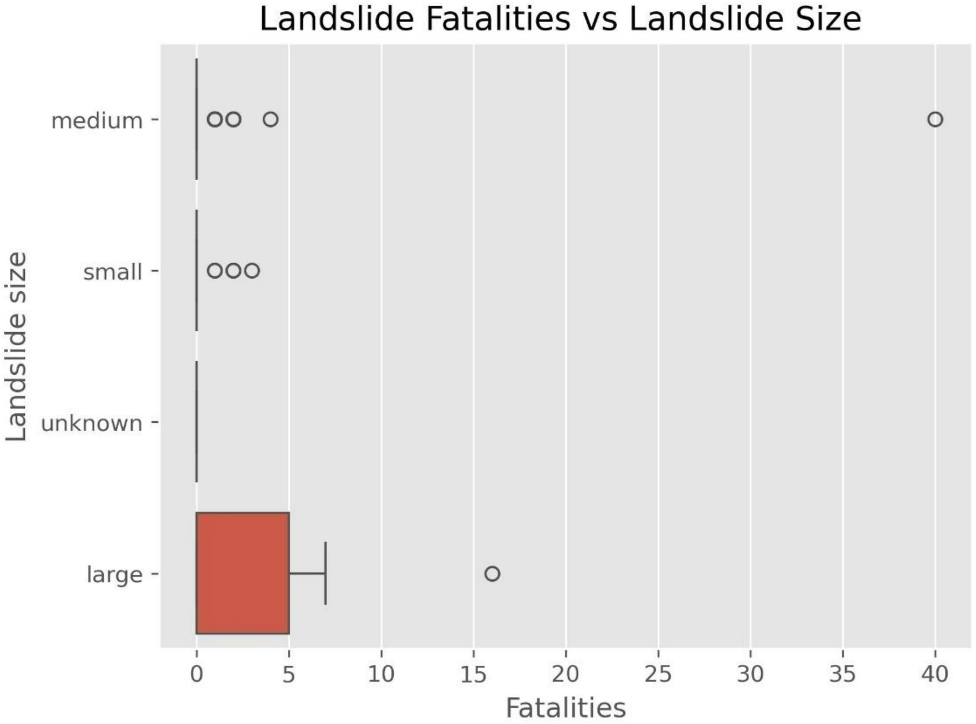


Figure 4.5.4 Malaysia boxplot of landslide size vs fatalities

4.5.4 Malaysia Heatmap Analysis

Heatmap of Figure 4.5.5 shows the numbers of large, medium, small, and unknown landslide occurrences of the following years, 2007 to 2016. The intensity of the color is symbolic of the number of landslides recorded, the darker the color the higher the number. As seen from the heatmap, “medium” sized landslides dominate almost all the years since they record higher counts than all the other sizes. There are 100 to 200 small landslides every year while few huge landslides occur, and its type is also rare. Some years have higher overall frequencies than others in each size category there is some year-to-year variation in counts of each size category. For example, 2013 is the year with many “medium” sized landslides reported. There seems to be, based on the heatmap, a possible temporal distribution of the landslides as well as the distribution of size of the landslides which could be further explored using the time series data or complemented with data such as rainfall data or seismic data to explain the occupying variations. This means that there is a good level of classification accuracy in the dataset since the number of “unknown” landslides is comparatively small.

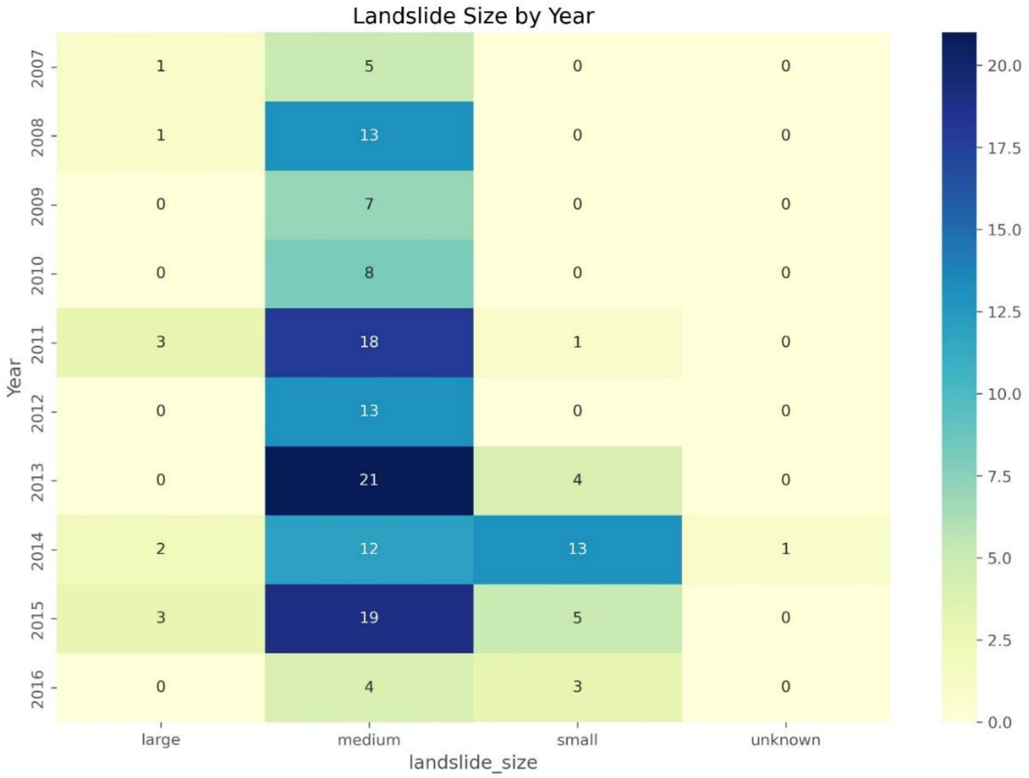


Figure 4.5.5 Malaysia heatmap of landslide size and year

Figure 4.5.6 displayed the Malaysia's landslide types typing form 2007 to 2016, where rows represent the year and columns represent the assorted types of the landslides: "complex", "debris_flow", "landslide" and others. Here, colour darkness was used to represent the number of landslides: darker shades of blue suggesting higher frequencies. The 'landslide' category records the highest by years result in most of the year indicating that it is the most frequent type. It is also observed that some of them 'Mudslide' also occurs with relatively high number of years some of other categories 'complex', 'debris_flow', 'riverbank_collapse', 'rock_fall' and 'unknown' are much less than the above categorized years. If we compare the counts of some of these categories from one year to the other there are fluctuations that might be attributed to the number of environmental factors such as seasonal rainfall. The heatmap explains that, apart from landslide, other types of landslides also happen but on a less frequent basis and the frequency of all kinds of landslides has a certain temporal inclination. More complex quantitive analysis involving factors that may be obtained externally to the dependent analytics, such as rainfall or geological data, might give deeper understanding into the fundamental drivers of these fluctuations. The low values obtained for ‘’unknown’’ implies that there is relatively high degree of accuracy for classifying landslides using the data set.

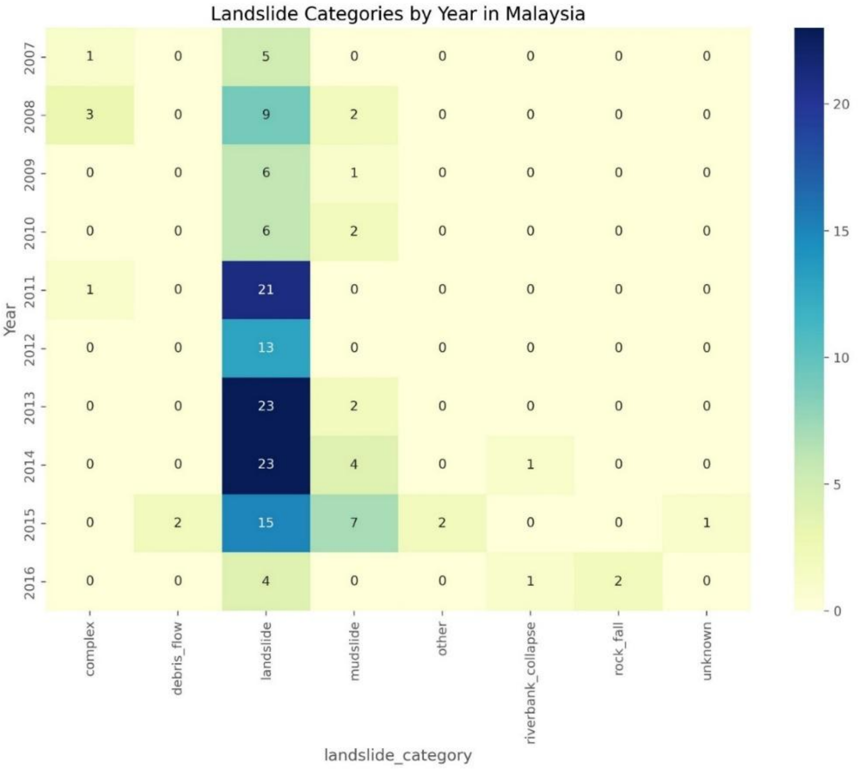


Figure 4.5.6 Malaysia heatmap of landslide category and year

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Research Outcomes

From the initial finding of EDA analysis Global Landslide Dataset, they are some conclusions can be made. Global landslides are mainly caused by rainfall, such as downpours and continuous rain, which dominated the occurrences of landslide cases. Medium-sized. The number of fatalities shows highly skewed distribution, where most cases associated with low fatalities, but catastrophic fatality is rare that influence the extremely to the overall impact. United States, India and China has highest number of cases and combined with environmental factors that influence this number. Rates of occurrence were fluctuant across the years with a high incidence in 2010 and this may be attributed to climatic fluctuations or reporting bias.

Focusing on Malaysia, rainfall are the most frequent type of trigger for the landslides and downpour is common. Sabah records the most frequent landslides event highlighting its vulnerability because of regional climate and terrain characteristics. Hence, there are many medium-sized landslides and low fatalities affected from the landslide events. Seasonal nature of the landslide and in particular the fluctuating trends of occurrence in the different states and years indicate some cases causes such as rainfall, urbanization, and deforestation. In conclusion, further analysis in crucial to give importance view of specific initial measures to minimize rainfall induced landslides and improvement in the data collection process to minimize the ‘unknown’ or undefined landslide activities to improve the understanding and policy making.

5.2 Future Works

The analysis will continue with future engineering and next method to achieve a higher level of insights and key actionable results. Advanced Machine Learning methods like Random Forest, Linear Regression and Time Series will be used for Forecasts of Susceptible Areas by analyzing few parameters such as the Intensity of Rainfall, Geographical Features and Land-Use Maps. They can approximate next landslides and put into term risks that are present in various areas. The time series analysis, using methods such as ARIMA or Long Short-Term Memory (LSTM) networks, will be applied to identify seasonality and periodicity of the landslide event, as well as provide alerts when there is high risk period of the event.

Furthermore, the integration of outside data such as current rainfall data, real-time seismographic data and the changes in the cover of the land is important in the development of a complete risk management plan. Other analysis tools that, that are not used in this study like spatial analysis tools that are useful for extending the study of landslide occurrences geographically. In addition, increasing form insights by dealing with missing values, increasing the level of data details, and unifying the reporting procedures will contribute to higher reliability and evidentiality. Last, the production of web-based interactive dashboard and tools will be applied to support the policy makers as well as the disaster management authorities where the gap between data analysis and application can be sealed.

REFERENCES

Abdulahi, M. M., & Egli, P. E. (2024). Landslides triggered by the July 21–22, 2024, heavy rainfall in the Gofa Zone, Southern Ethiopia. *Landslides*, *22*(1), 267–270. https://doi.org/10.1007/s10346-024-02397-4

Acharya, T., Shah, S., & Sharma, S. (2023). Enhancing landslide susceptibility mapping using machine learning: A case study of predictive models. *International Journal of Environmental Sciences*.

Ahn, S. A., Lee, J. H., & Park, H. J. (2023). Assessment of landslide susceptibility in Jecheon using deep learning based on exploratory data analysis. *The Journal of Engineering Geology*.

Akingboye, A. S., Bery, A. A., Aminu, M. B., Dick, M. D., Bala, G. A., & Ale, T. O. (2024). Surface–subsurface characterization via interfaced geophysical–geotechnical and optimized regression modeling. *Modeling Earth Systems and Environment*, *10*(4), 5121–5143. https://doi.org/10.1007/s40808-024-02054-8

Anees, M. T., Bakar, A. F. B. A., & Khan, M. M. A. (2024). Regional Rainfall-Induced Landslide Risk Assessment Using Susceptibility Mapping and Unexpected High-Intensity Rainfall. *BIO Web of Conferences*, *131*, 04020. https://doi.org/10.1051/bioconf/202413104020

Cui, Y., Cheng, D., Choi, C. E., Jin, W., Lei, Y., & Kargel, J. S. (2019). The cost of rapid and haphazard urbanization: lessons learned from the Freetown landslide disaster. *Landslides*, *16*(6), 1167–1176. https://doi.org/10.1007/s10346-019-01167-x

Daud, N. N. N., Mulwarman, M. I., & Goh, S. C. (2024). Slope Remediation Using Hydroseeding Technique: Signal Grass. *7th International Workshop Proceedings*.

Dharmasaputro, A. A., Fauzan, N. M., Kallista, M., Wibawa, Ig. P. D., & Kusuma, P. D. (2022). *Handling Missing and Imbalanced Data to Improve Generalization Performance of Machine Learning Classifier*. https://doi.org/10.1109/ismode53584.2022.9743022

Dou, J., Fu, X., & Pradhan, B. (2024). Water-Related Natural Disasters in Mountainous Area, Volume II. *Springer*.

Endut, Z., Ng, T. H., Abdul Aziz, J. H., & Teh, G. H. (2015). Structural analysis and vein episode of the Penjom Gold Deposit, Malaysia: Implications for gold mineralisation and tectonic history in the Central Belt of Malaysia. *Ore Geology Reviews*, *69*, 157–173. https://doi.org/10.1016/j.oregeorev.2015.02.012

Fakaruddin, F. J., Yip, W. S., Diong, J. Y., Dindang, A., K. Chang, N., & Abdullah, M. H. (2019). Occurrence of meridional and easterly surges and their impact on Malaysian rainfall during the northeast monsoon: a climatology study. *Meteorological Applications*, *27*(1). https://doi.org/10.1002/met.1836

Fan, Q., Zhang, S., Niu, Y., Si, J., Li, X., Wu, W., Zeng, X., & Jiang, J. (2024). Formative Period Tracing and Driving Factors Analysis of the Lashagou Landslide Group in Jishishan County, China. *Remote Sensing*, *16*(10), 1739. https://doi.org/10.3390/rs16101739

Gariano, S. L., & Guzzetti, F. (2016). Landslides in a changing climate. *Earth-Science Reviews*, *162*(1), 227–252. https://doi.org/10.1016/j.earscirev.2016.08.011

Ghani, A. A. (2000). The Western Belt granite of Peninsular Malaysia: some emergent problems on granite classification and its implication. *Geosciences Journal*, *4*(4), 283–293. https://doi.org/10.1007/bf02914037

Haque, U., Blum, P., da Silva, P. F., Andersen, P., Pilz, J., Chalov, S. R., Malet, J.-P., Auflič, M. J., Andres, N., Poyiadji, E., Lamas, P. C., Zhang, W., Peshevski, I., Pétursson, H. G., Kurt, T., Dobrev, N., García-Davalillo, J. C., Halkia, M., Ferri, S., & Gaprindashvili, G. (2016). Fatal landslides in Europe. *Landslides*, *13*(6), 1545–1554. https://doi.org/10.1007/s10346-016-0689-3

Highland, L. M., Bobrowsky, P., Geological Survey of Canada, & United States Geological Survey. (2008). *The Landslide Handbook- A Guide to Understanding Landslides* (p. 1325). US Geological Survey. https://pubs.usgs.gov/circ/1325/pdf/C1325_508.pdf

Huang, R., Li, Q., & Wang, J. (2020). Landslide detection and monitoring using deep learning methods. *Remote Sensing of Environment*, 240, 111697.

Hutchison, C. S., & Tan, D. N. K. (2009). *Geology of Peninsular Malaysia* (1st ed., pp. 73–79). Universiti Malaya and Geological Society of Malaysia.

Ismail, N. E. H., Taib, S. H., & Abas, F. A. M. (2019). Slope monitoring: an application of time-lapse electrical resistivity imaging method in Bukit Antarabangsa, Kuala Lumpur. *Environmental Earth Sciences*, *78*(1). https://doi.org/10.1007/s12665-018-8019-9

Jiang, X., Smith, R., & Lee, A. (2021). *Landslide Prediction Using Machine Learning Models: A Case Study*. Journal of Geosciences and Risk Management, 12(3), 45-59.

Kuradusenge, M., Kumaran, S., & Zennaro, M. (2020). Rainfall-Induced Landslide Prediction Using Machine Learning Models: The Case of Ngororero District, Rwanda. *International Journal of Environmental Research and Public Health*, *17*(11), 4147. https://doi.org/10.3390/ijerph17114147

Lim, C.-S., Jamaluddin, T. A., & Komoo, I. (2019). Human-induced landslides at Bukit Antarabangsa, Hulu Kelang, Selangor. *Bulletin of the Geological Society of Malaysia*, *67*, 9–20. https://doi.org/10.7186/bgsm67201902

Mafigiri, A., Faisal Abdul Khanan, M., Che Din, A. H., & Abdul Rahman, M. Z. (2022). Assessing the Influence of Anthropogenic Causal Factors on Landslide Susceptibility in Bukit Antarabangsa, Selangor. *International Journal of Built Environment and Sustainability*, *10*(1), 43–60. https://doi.org/10.11113/ijbes.v10.n1.1051

Mafigiri, A., Khanan, A., Hassan, A., & Rahman, A. (2022). Assessing the Influence of Anthropogenic Causal Factors on Landslide Susceptibility in Bukit Antarabangsa, Selangor. *International Journal of Built Environment and Sustainability*, *10*(1), 43–60. https://doi.org/10.11113/ijbes.v10.n1.1051

Majid, N. A., Taha, N. R., & Selamat, S. N. (2020). Historical landslide events in Malaysia 1993-2019. *Indian Journal of Science and Technology*, *13*(33), 3387–3399. https://doi.org/10.17485/ijst/v13i33.884

Makoundi, C. (2012). Geology, geochemistry, and metallogenesis of selected sediment-hosted gold deposits in the Central Belt, Peninsular Malaysia. *University of Tasmania*. https://doi.org/10.25959/23206031.v1

Matori, A. N., Basith, A., & Harahap, I. S. H. (2012). Study of regional monsoonal effects on landslide hazard zonation in Cameron Highlands, Malaysia. *Arabian Journal of Geosciences*, *5*(5), 1069–1084. https://doi.org/10.1007/s12517-011-0309-4

Melosh, B. L., Bodtker, J. W., & Valin, Z. C. (2024). Geologic map and structure sections along the southern part of the Bartlett Springs Fault Zone and adjacent areas from Cache Creek to Lake Berryessa, northern Coast Ranges, California. *Scientific Investigations Map*. https://doi.org/10.3133/sim3514

Mukhlisin, M., Matlan, S. J., Ahlan, M. J., & Taha, M. R. (2015). Analysis of Rainfall Effect to Slope Stability in Ulu Klang, Malaysia. *Jurnal Teknologi*, *72*(3). https://doi.org/10.11113/jt.v72.4005

Nasir, N. F., Omar, H. A., Rosly, M. H., & Mohamad, H. M. (2024). Review of Quantifying Slope Stability and Assessing Landslide Susceptibility in Sabah, Malaysia. *International Journal of Business and Technology Management*, *6*(3). https://doi.org/10.55057/ijbtm.2024.6.3.15

Nor Diana, M. I., Muhamad, N., Taha, M. R., Osman, A., & Alam, Md. M. (2021). Social Vulnerability Assessment for Landslide Hazards in Malaysia: A Systematic Review Study. *Land*, *10*(3), 315. https://doi.org/10.3390/land10030315

Parizia, F., Roberti, G., Clague, J. J., Alberto, W., Giardino, M., Ward, B., & Perotti, L. (2024). Landslide Deposit Erosion and Reworking Documented by Geomatic Surveys at Mount Meager, BC, Canada. *Remote Sensing*, *16*(9), 1599–1599. https://doi.org/10.3390/rs16091599

Petley, D. (2012). Global patterns of loss of life from landslides. *Geology*, *40*(10), 927–930. https://doi.org/10.1130/g33217.1

Pradhan, B., & Lee, S. (2018). Landslide susceptibility assessment and prediction using machine learning algorithms. *Geosciences*, 12(5), 303-317.

Qi, T., Meng, X., & Zhao, Y. (2024). Landslide Susceptibility Assessment in Active Tectonic Areas Using Machine Learning Algorithms. *Remote Sensing*, *16*(15), 2724–2724. https://doi.org/10.3390/rs16152724

Rosly, M. H., Mohamad, H. M., Nurmin Bolong, & Herayani, S. (2022). An Overview: Relationship of Geological Condition and Rainfall with Landslide Events at East Malaysia. *Trends in Sciences*, *19*(8), 3464–3464. https://doi.org/10.48048/tis.2022.3464

Rybalchenko, S. V. (2024). Geospatial conditions of landslide formation on the territory of Sakhalin Island. *BIO Web of Conferences*, *93*, 04001–04001. https://doi.org/10.1051/bioconf/20249304001

Samia, J., & Murphy, K. (2019). Machine learning for risk assessment in landslide-prone regions. *Natural Hazards Review*, 20(3), 100185.

Selamat, S. N., Majid, N. A., Taha, M. R., & Osman, A. (2022). Landslide Susceptibility Model Using Artificial Neural Network (ANN) Approach in Langat River Basin, Selangor, Malaysia. *Land*, *11*(6), 833. https://doi.org/10.3390/land11060833

Sharma, R., & Sandhu, K. (2023). Exploring temporal and environmental predictors of landslides through advanced machine learning models. *Natural Hazards and Earth System Sciences*.

Singh, U., Nandan, R., & Tiwari, A. (2024). Recent Trends and Techniques in Landslide Hazard Assessment. *Qeios*. https://doi.org/10.32388/lbyeqn

Song, Y., Wang, F., & Liu, Z. (2023). Improving machine learning performance in landslide susceptibility mapping through data balancing techniques. *Applied Geosciences and Remote Sensing Journal*.

Soon, L. W. (2023, February 2). Beef Up Protection Of Environmentally Sensitive Areas To Prevent Deadly Landslides – Experts. *Bernama*. https://www.bernama.com/en/bfokus/news.php?environment&id=2161129

Sulaiman, N., Robin, M. F. A., Muhammad Jamil, R., Sulaiman, N., Udin, W. S., Shafiee, N. S., & Sulaiman, F. R. (2024). Geology and Landslide Susceptibility Using GIS at Kampung Belahat, Jeli, Kelantan. *BIO Web of Conferences*, *131*, 04009. https://doi.org/10.1051/bioconf/202413104009

Tajudin, N., Ya'acob, N., Ali, D. M., & Adnan, N. A. (2021). Soil moisture index estimation from landsat 8 images for prediction and monitoring landslide occurrences in Ulu Kelang, Selangor, Malaysia. *International Journal of Power Electronics and Drive Systems/International Journal of Electrical and Computer Engineering*, *11*(3), 2101–2101. https://doi.org/10.11591/ijece.v11i3.pp2101-2108

Tang, K. H. D. (2019). Climate change in Malaysia: Trends, contributors, impacts, mitigation and adaptations. *Science of the Total Environment*, *650*(2), 1858–1871. https://doi.org/10.1016/j.scitotenv.2018.09.316

Umor, M. R., Ghani, A. A., Leman, S., Ali, C. A., & Sian, C. (2018, December 15). *Geochemistry and Petrology of Klang Gate Quartz Dyke in Gombak, Selangor and Its Bearings on Tectonic Evolution of Peninsular Malaysia*. Konferens Tapak Warisan Selangor Ke Arah Pengiktirafan Dunia. https://doi.org/10.13140/RG.2.2.27312.61446

Umor, M. R., Kayode, J. S., Rafek, A. G., Arifin, M. H., & Ghazali, M. A. (2024). A novel subsurface slopes hazardous mapping with engineering geologic and geophysical characterizations. *Heliyon*, *10*(10), e31308–e31308. https://doi.org/10.1016/j.heliyon.2024.e31308

Wahab, F. (2024, October 11). *Selangor gears up for rescue ops*. The Star. https://www.thestar.com.my/metro/metro-news/2024/10/12/selangor-gears-up-for-rescue-ops

Wang, S., Li, Y., & Huang, J. (2020). Machine learning approaches for landslide susceptibility prediction: A comprehensive review. *Geotechnical Engineering Review*

Ya'acob, N., Basah, N., Lelono, D. L., Tajudin, N., Yusof, A. L., Kassim, M., & Naim, N. F. (2024). Web-Based Landslide Early Warning System using GPM Precipitation Estimates. *Journal of Advanced Research in Applied Sciences and Engineering Technology*. https://doi.org/10.37934/araset.XX.X.172184

Zhang, X., & Lu, C. (2021). Rainfall-induced landslide prediction using LSTM networks. *Environmental Modelling & Software*, 134, 104866.

Zhang, Y., Meng, X., Jordan, C., Novellino, A., Dijkstra, T., & Chen, G. (2018). Investigating slow-moving landslides in the Zhouqu region of China using InSAR time series. *Landslide*, *15*(7), 1299–1315. https://doi.org/10.1007/s10346-018-0954-8