

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

FORECASTING LANDSLIDE EVENTS IN EASTERN OKLAHOMA AND  
WESTERN ARKANSAS USING EMPIRICAL METHODS AND  
STATISTICAL MACHINE LEARNING METHODS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

BY

OLUWATOBILOBA F. OYEBANJI

Norman, Oklahoma

2022

FORECASTING LANDSLIDE EVENTS IN EASTERN OKLAHOMA AND  
WESTERN ARKANSAS USING EMPIRICAL METHODS AND  
STATISTICAL MACHINE LEARNING METHODS

A MASTER'S THESIS APPROVED FOR THE  
GALLOGLY COLLEGE OF ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Dean Hougen

Dr. Netra Regmi

Dr. Dimitrios Diochnos



# Acknowledgements

I would like to appreciate Almighty God for the grace and resources to complete this thesis.

My sincere appreciation to my committee chair, Prof. Dean Frederick Hougen for his time, counsel, constructive criticism and support. His style of mentorship is rare and one to emulate.

My sincere gratitude goes to Dr. Netra Regmi. The resources and research ideas would not have been possible without him. I appreciate him for being there always to give research insights that have helped with data collection and analysis. His feedback were very helpful for the successful completion of this thesis report.

I also had the pleasure of working with and learning from Dr. Dimitrios Diochnos, a wonderful lecturer and a personable mentor. Always ready to listen and give his sound, and theoretical insights to any research problem.

My sincere thanks to my go-to guy, Benjamin Allen, always ready to help with thesis review, corrections, and recommendations.

In the midst of academic life demands and pressure, I have been blessed to get the needed attention and care from special friends and families. My family back at home have been the pillar. My gratitude to Aline and Alex Hurlimann for being an integral part of my graduate life journey. Alison Baker, a lovely friend

always willing to reach out and support. My Eternally grateful to Joe Sher who played an immense role in my success story at OU. Lastly, my Oklahoma family Jenny, Evie, Peyton, and Brent Barbour....thank you so much.

# Abstract

Understanding the trend of landslide occurrence in eastern Oklahoma and western Arkansas is crucial to the human and social development of the region. Studies suggest rainfall is one of the major landslide triggering mechanisms in the area. However, the association of landslides and rainfall is yet to be fully understood. Here, I analyze rainfall patterns at numerous landslide locations to better characterize the major rainfall conditions that can trigger landslides in the study area. The rainfall-landslide association was determined based on empirical threshold analysis and statistical machine learning approaches. The empirical threshold curve implemented aligns with similar curves developed for regions around the globe.

The developed empirical threshold captures 95 % of the landslide events above the threshold but fails to discriminate over half of non-landslide rainfall events from the landslide rainfall events. The empirical threshold trend also indicates that a small rainfall intensity can trigger landslides in the region.

Five machine learning approaches including K-nearest neighbor, logistic regression, random forest, gradient boost classifier, and voting ensemble classifier were used to determine important rainfall conditions that can trigger landslides. Rainfall features such as rainfall intensity, cumulative rainfall and antecedent rainfall conditions were integrated as input features for the models.

The ensemble classifier model correctly predicted all landslide events in the test dataset with a recall of 100%. The two-week antecedent rainfall data is ranked as the most important feature for forecasting landslide occurrence. Results from this study can be used in developing a landslide early warning and forecasting system as well as in making decisions for hillslope and natural resources management.

# Contents

<b>Acknowledgements</b>	<b>iv</b>
<b>Abstract</b>	<b>vi</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Literature Review</b>	<b>4</b>
<b>3 Study Area</b>	<b>7</b>
3.1 Geology and Geomorphology . . . . .	8
3.2 Climate and Vegetation . . . . .	10
3.3 Landslide . . . . .	13
<b>4 Materials and Methods</b>	<b>14</b>
4.1 Materials . . . . .	14
4.1.1 Data Sources . . . . .	14
4.1.2 Antecedent rainfall data . . . . .	17
4.1.3 Cumulative rainfall . . . . .	18
4.1.4 Rainfall Intensity . . . . .	19
4.2 Methods . . . . .	21
4.2.1 Empirical Rainfall Threshold . . . . .	21
4.2.2 Exploratory Data Analysis . . . . .	22
4.2.3 Statistical machine Learning methods . . . . .	31
4.2.4 K-Nearest Neighbors (KNN) Classifier . . . . .	39
4.2.5 Logistic Regression Classifier . . . . .	41
4.2.6 Random forest (RF) Classifier . . . . .	43
4.2.7 Gradient Boost Classifier . . . . .	45
4.2.8 Model Optimization: Ensemble Voting Classifier . . . . .	47

<b>5</b>	<b>Results</b>	<b>48</b>
5.1	Empirical Rainfall Threshold . . . . .	50
5.2	Statistical Machine Learning Models . . . . .	53
5.3	Discussion . . . . .	68
<b>6</b>	<b>Conclusions</b>	<b>70</b>
<b>7</b>	<b>Future Work</b>	<b>71</b>
	<b>References</b>	<b>72</b>
<b>A</b>	<b>Appendix</b>	<b>77</b>

# List of Figures

1.1	A photograph of a mountainous terrain, identifying exposed earth at the landslide site at Sugarloaf Mountain, Arkansas. (Modified after Regmi and Walter, 2020)	3
3.1	A map showing the distribution of landslides in the Ozark Mountains and the Ouachita Mountains region in eastern Oklahoma and western Arkansas. The red bounding polygon is the study area. The landslide points were recorded by the Oklahoma Geological Survey (Regmi and Walter, 2020)	8
3.2	A map showing parts of Oklahoma, Arkansas, Missouri, and Kansas accommodating the extent of the Ozark province and the Ouachita Province and some landmarks around them (Modified from Foti and Bukenhofer, 1999)	10
3.3	A map showing the mean annual precipitation (inches) in Oklahoma based on 1971 – 2000 precipitation data (modified after Johnson and Luza, 2008)	12
4.1	A Schematic illustrating steps involved in creating non-landslide rainfall events. For daily rainfall datasets (above), rainfall events are assessed daily from the profile. For the hourly rainfall dataset (below), rainfall events are assessed daily from the profile	17
4.2	A plot of empirical threshold profiles listed in Table 4.3. Each profile is labeled in terms of the area, extent, and landslide type in the legend (Modified from Guzzetti et al., 2007)	24
4.3	A histogram of the distribution of the target class. The landslide is highly imbalanced	25
4.4	The distribution of the independent features of the dataset before any pre-processing steps.	27
4.5	The distribution of the independent features of the dataset after scaling and normalizing the dataset.	28
4.6	A display of the correlation metrics between the rainfall intensity and Antecedent Rainfall Index features.	30

4.7	A schematic showing the iteration process starts from the top to bottom during training as each fold is used for validation from left to right for k-Fold cross-validation. . . . .	32
4.8	A typical AUC–ROC profile showing the range of decision thresholds for the model. . . . .	35
4.9	A schematic plot of a sigmoid function for the logistic regression method. . . . .	42
5.1	A rainfall intensity-duration scatterplot used to build an empirical threshold curve using the hourly dataset. . . . .	51
5.2	A rainfall intensity-duration scatterplot used to build an empirical threshold curve using the daily dataset. . . . .	51
5.3	A plot showing the new threshold curve for the study area overlain on the thresholds developed for all the landslide regions mentioned in table 4.3. The dataset used is the hourly dataset. . . . .	52
5.4	A plot showing the new threshold curve for the study area overlain on the thresholds developed for all the landslide regions mentioned in table 4.3. The dataset used is the daily dataset. . . . .	52
5.5	A plot showing the rank of the features’ importance in descending order. The top 5 features are selected. . . . .	53
5.6	A plot of the PR curve showing the performance of the KNN model in terms of possible range of precision and recall it can achieve. The optimal KNN would perform at recall of 42.85% and precision of 75%. The default model achieved an optimal performance at recall of 42.85% and precision of 75%. . . . .	56
5.7	A Plot showing the ROC-AUC curves of the KNN model. The AUC for the training is larger than AUC of testing. Model overfitting is inferred. . . . .	57
5.8	A confusion matrix showing the performance of the KNN model on the test data. . . . .	57
5.9	A plot of the PR curve showing the performance of the LR model in terms of possible range of precision and recall it can achieve. The optimal LR would have a recall of 58% and a precision of 70%. The default LR model did not achieve an optimal performance at recall of 14.285% and precision of 50%. . . . .	58
5.10	A Plot showing the ROC-AUC curves of the LR model. The AUC for the training is larger than AUC of testing. Both training and testing have relatively the same threshold profile and same AUC. . . . .	59
5.11	A confusion matrix showing the performance of the LR model on the test data. . . . .	60

5.12	A plot of the PR curve showing the performance of the RF model in terms of possible range of precision and recall it can achieve. The default RF model did not achieve an optimal performance at recall of 28.57% and precision of 66.66%. . . . .	61
5.13	A Plot showing the ROC-AUC curves of the LR model. The AUC for the training is larger than AUC of testing. Both training and testing have relatively the same threshold profile and same AUC. . . . .	62
5.14	A confusion matrix showing the performance of the RF model on the test data. . . . .	63
5.15	A Plot showing the ROC-AUC curves of the GBC model. The AUC for the training is larger than AUC of testing. Model overfitting is inferred. . . . .	64
5.16	A plot of the PR curve showing the performance of the GBC model in terms of possible range of precision and recall it can achieve. The default model did not achieve an optimal performance as the recall is at 57.14% and the precision is 80%. . . . .	65
5.17	A confusion matrix showing the performance of the GBC model on the test data. . . . .	65
5.18	A plot of the PR curve showing the performance of the Voting classifier model in terms of possible range of precision and recall it can achieve. The model is optimized for recall performance, and it achieved an optimal recall of 100% and precision of 70%. . . . .	66
5.19	A Plot showing the ROC-AUC curves of the Ensemble voting classifier model. The AUC for the training is approximately the same as the AUC of testing. . . . .	67
5.20	A confusion matrix showing performance of the Voting Classifier model on the test data. . . . .	68

# List of Tables

4.1	Rainfall characteristics associated with 32 landslide events recorded for this study. 1_D_R represents one day antecedent rain, 1_W_R represents 1 week antecedent rain. . . . .	20
4.2	A list of thresholds generated from across the world identified as local, regional, and global (Modified from Guzzetti et al., 2007) . . . . .	23
4.3	Confusion Matrix for binary classification . . . . .	38
5.1	Statistics of all event features (landslide and non-landslide events in mm . . . . .	48
5.2	Statistics of all event features for landslide events in mm. . . . .	49
A.1	Landslide events sourced from news. . . . .	78

# Chapter 1

## Introduction

*Landslide events* are generally a natural phenomenon that occur when a body of rock or earth debris slides down a slope under the effect of gravity (Highland & Bobrowsky, 2008) (USGS, 2004). Synonymous to and interchanged with slope failure and mass movement, landslides are mainly acted upon by external factors such as earthquakes, rainfall, volcanic eruptions, flooding, prolonged intense precipitation and human influences in addition to the intrinsic geological and geomorphologic causes such as substrate type, slope gradient, and hillslope relief (Densmore et al., 1997). The human and socio-economic impacts of landslides can be devastating (Petley, 2012). The consequences result in catastrophes ranging from loss of lives to destruction of infrastructure. Hence, there is need to incorporate landslide studies into regions with growing infrastructure and population especially regions dominated by steep and unstable terrain (Highland et al., 2008).

The majority of known natural landslide events are mostly triggered by heavy *precipitation* (Larsen & Simon, 1993; Jakob & Weatherly, 2003) and earthquakes (Rodriguez et al., 1999). Precipitation increases pore water pressure which initi-

ates the needed disturbance for gravity to displace a body of the earth from its in-situ state. Understanding the relationship between landslides and rainfall characteristics is thus important to develop a landslide forecasting system (NASA Earth Observatory,2016) <https://earthobservatory.nasa.gov/images/92018/predicting-landslide-hazards-in-near-real-time>).

Landslides are one of the major geohazards in the Ozark and Ouachita Mountains of eastern Oklahoma and western Arkansas and have the potential to cause social and economic losses each year. For example, a large landslide that occurred in the Ouachita Mountains near Red Oak severely damaged Oklahoma State Highway 82 and a forest road (Cerato et al., 2014). The goal of this study is to determine the rainfall characteristics that trigger landslides in eastern Oklahoma through the help of empirical hydroclimatic threshold methods and machine learning methods. Objectives that were met to achieve this goal include:

1. map spatiotemporal characteristics of landslides based on remote sensing products,
2. determine rainfall threshold that trigger landslides based on the empirical approach of computing rainfall intensity-duration threshold, and
3. determine important rainfall conditions for triggering landslides using machine learning approaches.

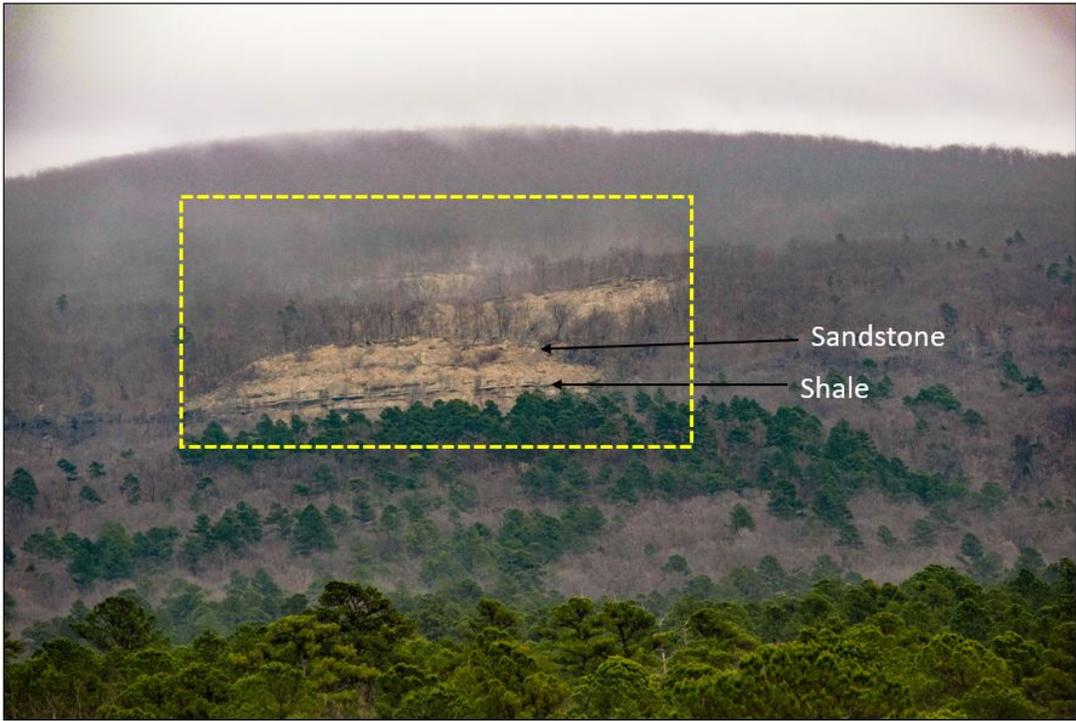


Figure 1.1: A photograph of a mountainous terrain, identifying exposed earth at the landslide site at Sugarloaf Mountain, Arkansas. (Modified after Regmi and Walter, 2020)

# Chapter 2

## Literature Review

An Empirical rainfall threshold are lower bounded lines constructed from rainfall datasets that initiate landslide events (Giannecchini, 2005). Coupled with non-landslide related rainfall datasets, empirical thresholds are used as separators for landslide and non-landslide rainfall.

Campbell (1975) was one of the pioneer researchers in the studies of rainfall threshold analysis for landslide initiation. Campbell illustrated how antecedent precipitation (which is related to the magnitude of prior precipitation) is a function of soil moisture and how it triggers landslide events. (Taylor, 1975) analyzed the effect of rainfall intensity and antecedent rainfall on the occurrence of landslides along slopes in San Francisco, California and computed the threshold curves for the study area. Caine (1980) studied 73 natural shallow landslides (approximately 3m deep) triggered by rainfall in Sierra Norte de Puebla, Mexico. With the help of local rainfall records, Caine (1980) devised the first upper rainfall threshold for landslide initiation (equation 2.1);

$$I = 14.82D^{0.39} \tag{2.1}$$

where  $I$  is the mean rainfall intensity (mm/hour), and  $D$  is the rainfall duration (hours).

Over the decades, with the increase in research interest and improvement in data acquisition techniques such as the Shuttle Radar Topography Mission (SRTM) system, the study of natural landslides has evolved. In an effort to improve landslide early warning systems, different methodologies have been proposed for computing rainfall thresholds across a range of scales. For example, empirical thresholds peculiar to climatic and geographic conditions (Guzzetti et al., 2020). (Komac, 2005) computed rainfall thresholds for different lithological units on the entire Slovenia territory by applying the statistical chi-square method. To account for precipitation effects from excess discharge from other neighboring regions, (Jakob & Weatherly, 2003) incorporated streamflow data with precipitation data to develop a discriminant functions analysis model that scores the chance of landslides and non-landslides.

Recently machine learning methods have been used to develop a landslide warning system. These approaches have the capability to integrate various environmental covariates (landslide conditioning factors) with precipitation data to build a more robust landslide early warning system. Chen et al. (2018) used 95 landslide locations with 14 landslide conditioning factors as input features to train Bayes' net model, radial basis function, logistic model tree, and, random forest model for the landslide susceptibility model in Chongren County, China. To understand the effect of static environmental covariates such as slope, elevation, and curvatures on landslide initiation, Jacquemart and Tiampo (2021) developed a system that assessed landslide susceptibility using InSAR interferometric coherence data and Normalised Difference Vegetation Index (NDVI).

The role of antecedent rainfall on soil moisture prior to landslide events has

been studied and identified as a critical factor for landslide occurrence. With the high water-retentive nature of the lithology in the mountainous region of Puerto Rico, antecedent rainfall events play a huge factor in landslide occurrence (Larsen & Simon, 1993). Recently, Jordanova et al. (2020) identified rainfall as the most prevalent mechanism for slope failure in Slovenia and developed a probabilistic empirical threshold scheme. Jordanova classified the area of study based on two classes, mean annual rainfall and lithological units.

In the study area, (Regmi & Walter, 2020) provides a background study of landslide dynamics. With 185 shallow landslides mapped across approximately 150 km<sup>2</sup> area of Cavanal Hill and Sugarloaf Mountains. They investigated the nature of soil-mantled hillslopes, landslide frequencies, magnitudes, and their approximate ages using historical aerial photographs. Their study shows that the timing of most of the mapped landslide occurrences correspond to the period of extreme and high-frequency precipitation that occurred between May 2015 to December 2015. Their study did not establish a relationship between the landslide events and the rainfall datasets. This study is an expansion of their findings and to establish a relationship between precipitation data and landslide occurrence.

# Chapter 3

## Study Area

The study area is located around the Ouachita Mountains and Ozark Mountains in eastern Oklahoma and western Arkansas (Figure 3.1).

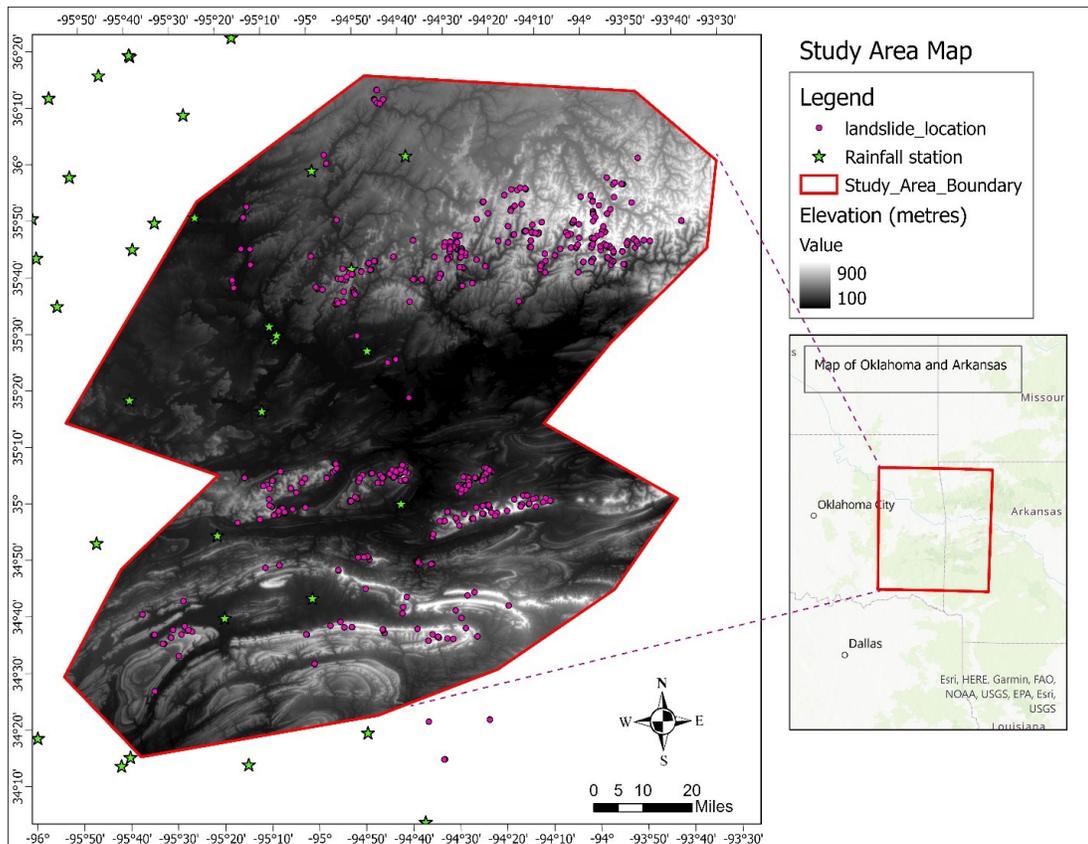


Figure 3.1: A map showing the distribution of landslides in the Ozark Mountains and the Ouachita Mountains region in eastern Oklahoma and western Arkansas. The red bounding polygon is the study area. The landslide points were recorded by the Oklahoma Geological Survey (Regmi and Walter, 2020)

### 3.1 Geology and Geomorphology

The Ozark and Ouachita Mountains in eastern Oklahoma and western Arkansas are underlain by sedimentary rocks, including Quaternary alluvium to Devonian-Silurian marine sandstone, shale, and Quaternary deposits (Tyrl et al., 2007; Stoesser, 2003). The main lithologies found on outcrops are thinly bedded to medium-grained sandstones and thick-bedded and weathered fossiliferous shales to deformed shales. (Regmi & Walter, 2020). These rocks are highly fractured

and weathered. Given its diverse geomorphologic structures, the Ozark uplift is comprised of mountains, domes, highlands, and plateaus (Foti & Bukenhofer, 1999). The Arkansas River valley is a 40-mile-wide trough that separates the Ozark Mountains in the north and the Ouachita Mountains in the south. The Ouachita Mountain region is a mountainous, geologically complex area located between the West Gulf Coastal Plain region to the south and the Arkansas Valley region to the north.

The orogeny of Ouachita was developed during the Pennsylvanian period resulting in folding deformations of the strata. The Ozark Mountains were developed as an asymmetrical uplift during the orogeny making the strata relatively horizontal (Adamski & NWQAP, 1995). Some mountains around this region like St. Francois Mountains had originally existed as a result of volcanic intrusion almost 1.5 billion years ago before the orogeny. Since its formation, the Ozark has been frequently eroded by streams and valleys, therefore, creating several isolated steep-sided plateaus.

The region consists of several plateaus: The Boston Mountains, Springfield Plateau, and Salem Plateau. It begins in Oklahoma and ends in the central part of Arkansas. River gorges abound and some are up to 1,500 feet deep, created through the years by streams currents, cutting through the region.

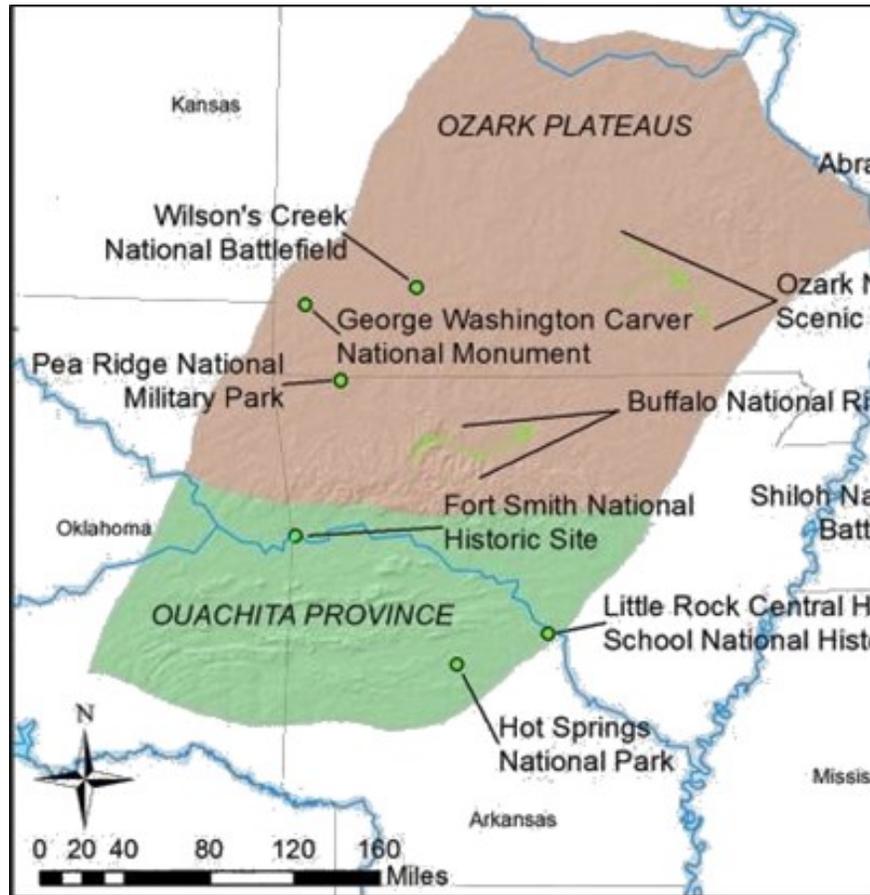


Figure 3.2: A map showing parts of Oklahoma, Arkansas, Missouri, and Kansas accommodating the extent of the Ozark province and the Ouachita Province and some landmarks around them (Modified from Foti and Bunkenhofer, 1999)

### 3.2 Climate and Vegetation

In terms of climatic conditions, Oklahoma’s seasonal and geographical variation is controlled by temperature and precipitation. Relative to western Oklahoma, the moisture from the Gulf of Mexico strongly affects the climatic conditions of eastern Oklahoma (Tyrl et al., 2007). This also dominates the regional climatic condition of Western Arkansas. The huge variation in climatic conditions ranges from 20 inches of annual precipitation in northwestern Oklahoma to 52 inches of

annual rainfall in southeast Oklahoma from 1971 to 2000 (Figure 3.3). Maximum precipitation is recorded during summer (June - August) and minimum precipitation is recorded during winter (December, January, and February) (Johnson and Luza, 2008) The vegetation primarily consists of various Oak species and short-leaf pine. Oak-pine forest is dominant in upland slopes and blackjack forest is dominated in lowlands. Other vegetation includes blubberies, flowering dogwood, hophorn beam, sugar maple, and redbud service berry (Johnson, 2008).

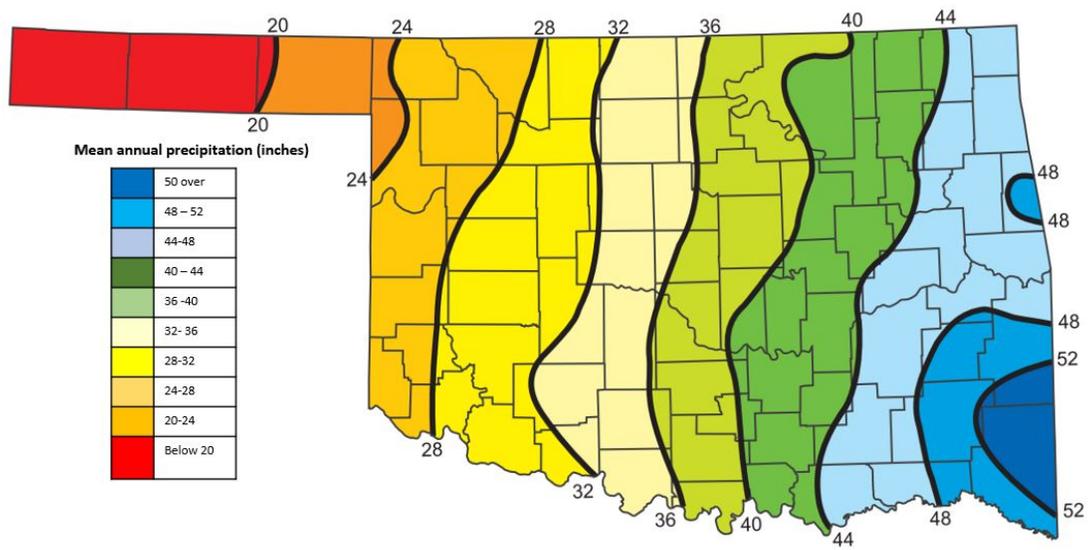


Figure 3.3: A map showing the mean annual precipitation (inches) in Oklahoma based on 1971 – 2000 precipitation data (modified after Johnson and Luza, 2008)

### **3.3 Landslide**

The study area shows clear evidence of landslides including old deep-seated landslides, recent shallow landslides, and soil creep (Regmi & Walter, 2020; He et al., 2014). For example, (Regmi & Walter, 2020) mapped hundreds of shallow landslides (depth of slip surface less than the tree root depth) including soil creep, debris flow, debris slide, soil slide, rock slide, and rock fall across the study area. They suggest most of these landslides are small in size and could have been triggered by intense precipitation (i.e., record precipitation of May 2015) alone or in conjunction with Oklahoma earthquakes.

# Chapter 4

## Materials and Methods

### 4.1 Materials

The majority of the dataset used is hourly and daily time series rainfall datasets associated with both landslide and non-landslide events. The datasets are used to develop empirical threshold models and machine learning models.

#### 4.1.1 Data Sources

We used the landslide inventory of landslides developed by the Oklahoma Geological Survey (Figure 2.1). The inventory consists of over 700 landslide occurrences. The timing of 9 of the landslide events was sourced from local news while 23 were sourced from time series Landsat 7 ETM+, Landsat 8, and Planet imagery. The timing of landslide occurrence in terms of hourly precision is mostly unknown. Hence, the time of the peak rainfall within the day landslide occurred is estimated as the landslide time. Hence, there is a margin of error for analysis using hourly rainfall data. In summary, the sources for landslide information are;

1. Landslide events from the local news: The most accurate landslide infor-

mation in terms of time, magnitude, and scale are obtained from local news. The event is captured in news reports at an hourly approximation.

2. Rainfall data was sourced from Mesonet, and National Oceanic and Atmospheric Administration, NOAA . Over 144 stations provide rainfall information across the state of Oklahoma and western Arkansas

3. Mapping landslides events and timing was conducted using Landsat imagery from the United States Geological Survey (USGS). Regular field trips to the study area were conducted to validate landslides mapped based on remote sensing data as well as geologic details like soil types and geomorphology.

### **Data Collection**

Two datasets (hourly and daily) comprising a total of 214 events (32 landslides and 184 non-landslide), the date of the rainfall event, and coordinates of the rainfall stations were prepared. Sample datapoint location and sampling frequency are crucial to the success of this research, hence a need to evaluate which dataset will be effective for the research. Rain-gauge stations in Oklahoma and Arkansas are spread evenly and they capture daily and hourly rainfall information. Table 1 in the appendix shows an example of raw daily and hourly rainfall data obtained from the Mesonet database.

### **Estimation of rainfall at Datapoint**

A rainfall station is not always stationed at landslide locations. Computing the amount of precipitation at a specific landslide location requires a form of interpolation from neighboring rainfall stations. Inverse Distance Weighting (IDW) interpolation is a deterministic and mathematical way of estimating rainfall values at unknown data points. The IDW technique assumes that the rainfall station

with rainfall information is uniform and closely located to the unknown point (i.e., the landslide location) (Lu & Wong, 2008). The rainfall value  $Z_p$  at location  $p$  is computed by summing rainfall of the rainfall value  $Z_i$  at  $i$  stations located at distance  $d_i$  from the landslide location divided by the inverse of their respective distance,  $d_i$  (equation 4.1). The IDW estimated rainfall value,  $Z_p$  is the linear combination of the weight which is the inverse of the distance from station  $i$ , and the observed rainfall value at station  $i$ .

$$Z_p = \frac{\sum_{i=1}^n \frac{z_i}{d_i}}{\sum_{i=1}^n \frac{1}{d_i}} \quad (4.1)$$

### Creating Non-landslide event datasets

Creating datasets of rainfall events that did not result in landslides is crucial to the development of the rainfall threshold. A logical procedure with sets of rules was adopted in the choice of location and time of selecting the rainfall event that did not trigger landslides. At each landslide location, prior records of non-landslide rainfall events were selected if the event and its antecedent rainfall did not trigger any landslide. The absence of a landslide in the given location and its surrounding is also confirmed using the Landsat imagery. Locations of the landslide event are used to calibrate the creation of non-landslide datasets. A schematic of the procedure is shown in Figure 4.1. The figure shows the sequence for creating non-landslide observations and is articulated below:

- Pick a landslide event, its location, and time.
- Plot rainfall events at the location for the period (90 days) before the landslide event.

- At the same location, use LANDSAT and Planet imagery to check for any landslide event.
- If no landslide is recorded, iterate at an interval of 90 days backward.
- At each iteration, pick a peak rainfall event or prolonged rainfall period.
- Compute intensity and antecedent rainfall at the selected location.



Figure 4.1: A Schematic illustrating steps involved in creating non-landslide rainfall events. For daily rainfall datasets (above), rainfall events are assessed daily from the profile. For the hourly rainfall dataset (below), rainfall events are assessed daily from the profile

#### 4.1.2 Antecedent rainfall data

These are soil moisture conditions created by cumulative rainfall over a specified time before a rainfall event that can trigger a landslide. Studies have shown that they are important phenomena for landslide-induced rainfalls (Campbell, 1975; Glade, Crozier, & Smith, 2000). Due to the orographic effect, landslide-prone mountainous terrain in the islands of Puerto Rico experienced trade-winds that delivered a relatively continuous flow of moisture-filled air, coupled with

low intensity, frequent precipitation (Larsen & Simon, 1993). The Antecedent Precipitation Index (API) is the summation daily estimate of the catchment wetness calculated from the rainfall that occurred over the preceding days. A decay,  $k$  parameter (lies within 0.85 - 0.98) which is a function of the type of the topsoil moisture and rainfall-runoff model (Fedora & Beschta, 1989) is multiplied by each prior day's cumulative rainfall (Heggen, 2001). Antecedent precipitation Index is defined in equation 4.2 as:

$$API_t = k \cdot API_{t-\Delta t} - P_{\Delta t} \quad (4.2)$$

where  $API_t$  is the antecedent precipitation index (mm) at time  $t$ ,  $P_{\Delta t}$  is the cumulative precipitation during the period from  $t - \Delta t$  to  $t$ ,  $K$  is the decay parameter related to soil moisture.

The following 10 antecedent precipitation index features were computed, one day rain (1\_D\_R), two days rain (2\_D\_R), three days (3\_D\_R), four days rain (4\_D\_R), five days rain (5\_D\_R), six days (6\_D\_R), one week rain (1\_W\_R), two weeks(2\_W\_R), three weeks (3\_W\_R), and four weeks (4\_W\_R).

### 4.1.3 Cumulative rainfall

This is the total amount of precipitation (mm) per episode of the rainfall event. For landslide events, it is the total precipitation measured from the start of a rainfall event to the time of landslide occurrence for landslide event. For non-landslide events, it is the total precipitation measured from the start of the rainfall event to the end of the rainfall event.

#### **4.1.4 Rainfall Intensity**

Rainfall intensity is the rate of precipitation recorded over a period commonly measured in mm per hour or inches per hour. Depending on the frequency of rainfall observation, the intensity can be computed as hourly rainfall intensity or daily rainfall intensity per event. The period of precipitation for the hourly dataset is the duration (in hours) of non-zero rainfall values during events. The period of precipitation for the daily dataset is the duration (in days) of non-zero rainfall values recorded during events. Hence, these two datasets produce differing threshold curves.

Table 4.1: Rainfall characteristics associated with 32 landslide events recorded for this study. 1\_D\_R represents one day antecedent rain, 1\_W\_R represents 1 week antecedent rain.

date	Cumm	Dur(hr)	Inten	1_D_R	2_D_R	3_D_R	4_D_R	5_D_R	6_D_R	1_W_R	2_W_R	3_W_R	4_W_R
5/11/2015	198.91	168	1.184	2.881	5.187	6.417	7.536	7.76	7.831	7.831	8.62	9.425	13.189
5/30/2015	259.95	432	0.602	0.636	1.369	1.481	1.488	2.9	5.683	6.391	9.855	15.03	17.898
6/18/2015	136.46	144	0.948	2.22	3.09	3.939	4.015	5.373	5.373	5.373	5.373	6.87	12.901
6/18/2015	161.35	144	1.12	2.708	3.475	4.632	4.777	6.352	6.352	6.352	6.354	7.727	13.109
6/18/2015	161.36	144	1.121	2.71	3.476	4.633	4.778	6.353	6.353	6.353	6.355	7.727	13.109
7/9/2015	60.904	96	0.634	1.022	2.387	2.398	2.398	2.5	3.851	3.977	4.616	6.223	9.971
7/9/2015	60.892	96	0.634	1.022	2.387	2.397	2.397	2.499	3.849	3.976	4.615	6.221	9.967
1/21/2016	4.726	72	0.066	0.157	0.186	0.186	0.186	0.192	0.192	0.192	0.824	0.829	10.717
1/21/2016	4.739	72	0.066	0.157	0.187	0.187	0.187	0.193	0.193	0.193	0.825	0.83	10.72
7/4/2016	43.382	48	0.904	1.708	1.708	2.218	2.218	2.218	2.218	2.226	2.246	2.356	2.801
7/4/2016	43.458	48	0.905	1.711	1.711	2.172	2.172	2.172	2.172	2.181	2.203	2.309	2.804
9/19/2016	62.645	144	0.435	0.587	0.914	1.895	1.935	2.466	2.466	2.468	2.473	2.878	4.179
9/19/2016	62.65	144	0.435	0.586	0.914	1.894	1.934	2.467	2.467	2.468	2.473	2.881	4.193
9/19/2016	62.64	144	0.435	0.586	0.912	1.893	1.933	2.466	2.466	2.467	2.472	2.878	4.187
5/23/2017	107.56	168	0.64	0.039	0.041	2.41	3.415	4.144	4.235	4.235	5.22	5.561	8.758
5/23/2017	107.93	168	0.642	0.038	0.041	2.427	3.425	4.158	4.249	4.249	5.226	5.568	8.749
5/23/2017	108.04	168	0.643	0.038	0.041	2.431	3.428	4.162	4.253	4.253	5.227	5.571	8.746
4/16/2015	74.006	96	0.771	0.031	0.222	2.543	2.543	2.543	2.543	2.551	2.817	3.095	3.967
5/7/2015	42.871	48	0.893	1.688	1.688	1.688	1.688	1.688	1.688	1.688	2.264	3.869	5.639
5/20/2015	147.71	192	0.769	4.897	4.904	5.525	5.538	5.557	5.56	5.815	12.82	13.01	14.791
5/24/2015	120.55	72	1.674	4.579	4.746	4.746	5.043	6.429	6.429	7.833	9.727	16.17	17.191
5/30/2015	296.88	432	0.687	0.612	1.518	1.621	1.622	3.151	5.734	6.624	11.34	17.32	19.835
5/30/2015	301.45	432	0.698	0.615	1.537	1.636	1.637	3.176	5.753	6.65	11.53	17.44	19.936
6/13/2015	28.05	24	1.169	1.104	1.104	1.104	1.104	1.104	1.104	1.104	1.106	8.895	12.981
6/13/2015	36.372	24	1.516	1.432	1.432	1.432	1.435	1.435	1.435	1.435	1.447	7.93	10.29
6/17/2015	287.63	120	2.397	7.396	9.047	9.597	11.324	11.32	11.32	11.32	11.32	14.77	20.094
12/23/2015	1.71	24	0.071	0.067	0.067	0.068	0.068	0.069	0.078	0.078	3.92	3.929	8.421
12/27/2015	235.44	48	4.905	9.269	9.269	9.269	9.421	9.421	9.421	9.426	11.71	13.14	13.527
5/23/2017	107.68	168	0.641	0.039	0.041	2.416	3.418	4.149	4.239	4.239	5.222	5.563	8.754
5/23/2017	53.361	96	0.556	0.226	0.23	2.101	2.618	3.574	3.732	3.732	4.318	4.684	8.311
5/23/2017	54.159	96	0.564	0.235	0.24	2.132	2.647	3.647	3.805	3.805	4.397	4.756	8.346
5/23/2017	53.968	96	0.562	0.233	0.238	2.125	2.643	3.634	3.791	3.791	4.382	4.742	8.346
5/23/2017	53.784	96	0.56	0.231	0.235	2.117	2.634	3.613	3.771	3.771	4.36	4.722	8.33
5/23/2017	53.613	96	0.558	0.229	0.233	2.111	2.626	3.596	3.753	3.753	4.341	4.705	8.318

## 4.2 Methods

### 4.2.1 Empirical Rainfall Threshold

A minimum rainfall intensity threshold is the least amount of rainfall amount within a period below which a landslide event does not occur. Hence, a minimum threshold line is derived to distinguish rainfall events that initiate landslides from those that do not initiate landslides. The datapoints are mainly plotted on a cartesian or a logarithmic scale, the rainfall intensity (mm/hour) on the vertical axis and the duration (hours) on the horizontal axis.

Various methods have been applied to determine this threshold line. For Example, (Caine, 1980) proposed and applied a visual inspection methodology whereby the line is fitted optimally by visually fitting the line on the landslide datapoints. Adopting this method is not statistically rigorous and can be prone to overfitting as the line is fitted only to optimize only the landslide events. This method was adopted for the study. Another methodology is by using Bayesian inference (Guzzetti, Peruccacci, Rossi, & Stark, 2007)(Guzzetti, 2007).

#### **Rainfall Intensity threshold Method**

A relationship between the rainfall intensity-duration and cumulative rainfall-duration has been used as an empirical approach to evaluating the pattern of landslide events. A line that separates rainfall events that induced landslides and rainfall events that did not induce landslides can be considered a threshold line. This line or minimum threshold aligns with the minimum rainfall necessary to trigger a landslide over a period (of hours). This line is determined either visually (Caine, 1980) or using Bayesian inference through the power curve equation (Jakob & Weatherly, 2003). The most common variables used to characterize

landslide events are rainfall intensity-duration (ID) and cumulated event rainfall-rainfall duration (ED). Empirical rainfall thresholds created across the world are summarized in table 4.2 and plotted in Figure 4.2

## 4.2.2 Exploratory Data Analysis

Rainfall datasets are preprocessed and prepared as input features for the empirical threshold model analysis and the machine learning models. Some critical procedures are followed to validate and preprocess the dataset for modeling.

### Class Imbalance

It is necessary to capture all the unique instances of rainfall events in the region to train the model on a broad range of events that could occur. Hence, several non-landslide events were acquired which created an unequal proportion of landslide events to non-landslide events (from Figure 4.3). Such a skewed distribution (Figure 4.3) can introduce a bias in the predictive power of the model as the model will poorly predict the minority test data (Brownlee, 2020). Hence, this situation requires us to prioritize the recall metric during training and the evaluation of the machine learning models. A solution to this challenge is to create a synthetic dataset. A popular method is called Synthetic Minority Oversampling Technique (SMOTE) (Das et al., 2020). The SMOTE technique generally identifies the  $k$ -nearest neighbors to the minority sample that will be used to generate sets of synthetic observation neighbors (Chawla et al., 2002). The technique was implemented in this study to generate additional landslide event datasets.

Table 4.2: A list of thresholds generated from across the world identified as local, regional, and global (Modified from Guzzetti et al., 2007)

#	EXTENT	AREA	LANDSLIDE TYPE	Threshold Equation	Range (hours)
1	Global	World	Shallow, Debris	$I = 14.82 * D^{-0.39}$	$0.167 < D < 500$
2	Region	Carinthia and E Tyrol, Austria	Soil Slide	$I = 41.66 * D^{-0.77}$	$1 < D < 1000$
3	Local	Valtellina, Lombardy, N Italy	Soil Slide	$I = 44.668 * D^{-0.78}$	$1 < D < 1000$
4	Local	San Francisco Bay Region, California	Debris	$I = 6.9 + 38 * D^{-1.00}$	$2 < D < 24$
5	Region	Indonesia	Debris	$I = 92.06 - 10.68 * D^{1.0}$	$2 < D < 4$
6	Region	Puerto Rico	Debris	$I = 66.18 * D^{-0.52}$	$0.5 < D < 12$
7	Region	Brazil	Debris	$I = 63.38 - 22.19 * D^{1.0}$	$0.5 < D < 2$
8	Region	China	Debris	$I = 49.11 - 6.81 * D^{1.0}$	$1 < D < 5$
9	Local	Hong Kong	Debris	$I = 41.83 * D^{-0.58}$	$1 < D < 12$
10	Region	Japan	Debris	$I = 39.71 * D^{-0.62}$	$0.5 < D < 12$
11	Local	Mayon, Philippine	Lahar	$I = 27.3 * D^{-0.38}$	$0.167 < D < 3$
12	Region	Switzerland	All types	$I = 32 * D^{-0.70}$	$1 < D < 45$
13	Global	World	Shallow slide	$I = 0.48 + 7.2 * D^{-1.00}$	$0.1 < D < 1000$
14	Local	Moscardo Torrent, NE Italy	All types	$I = 15 * D^{-0.70}$	$1 < D < 30$
15	Region	Eastern Jamaica	Shallow slides	$I = 11.5 * D^{-0.26}$	$1 < D < 150$
16	Region	Shikoku Island, Japan	All types	$I = 1.35 + 55 * D^{-1.0}$	$24 < D < 300$

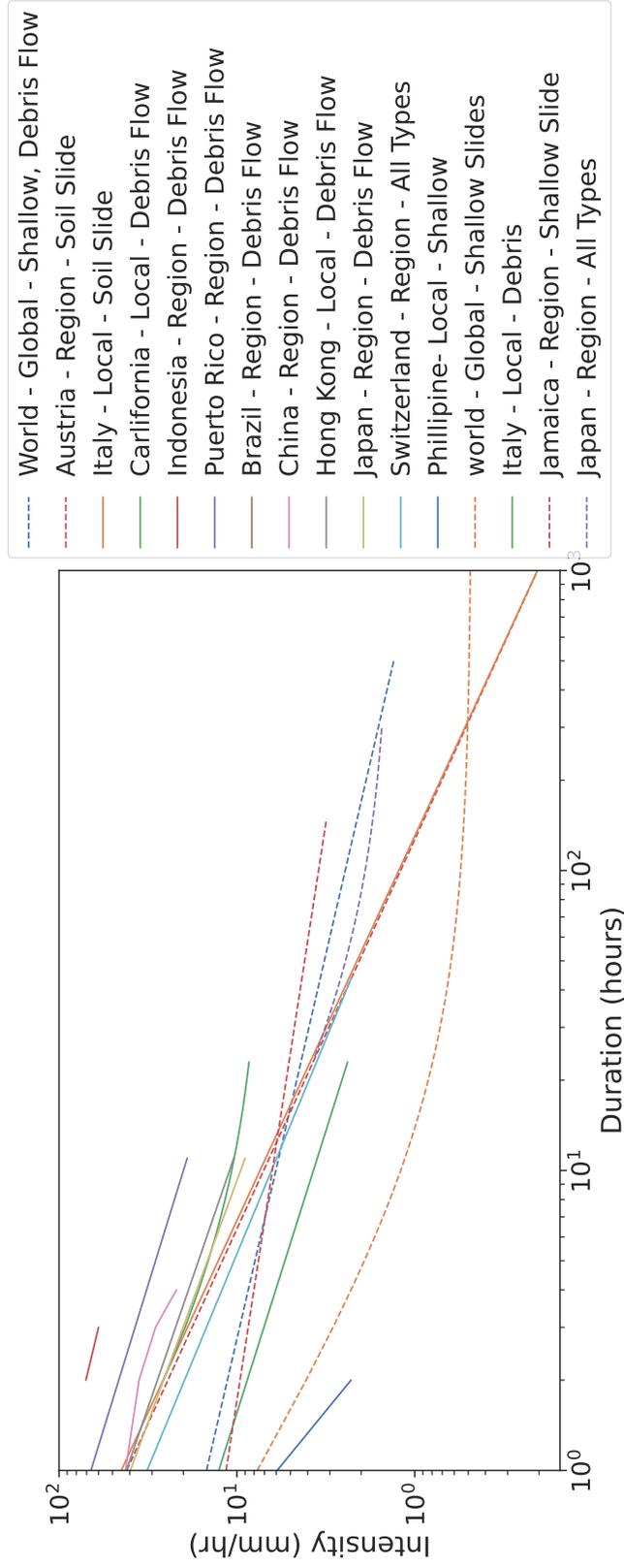


Figure 4.2: A plot of empirical threshold profiles listed in Table 4.3. Each profile is labeled in terms of the area, extent, and landslide type in the legend (Modified from Guzzetti et al., 2007)

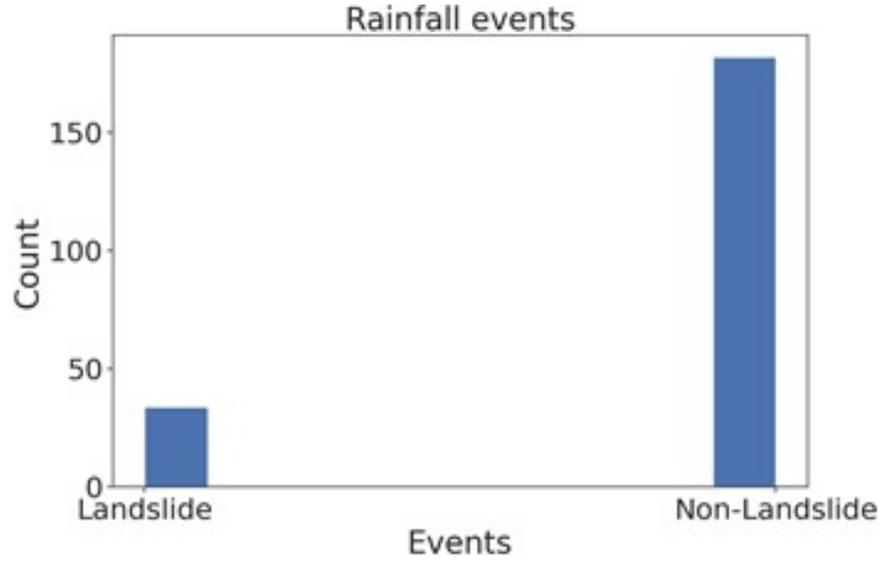


Figure 4.3: A histogram of the distribution of the target class. The landslide is highly imbalanced

### Outlier Analysis

Most statistical techniques are sensitive to outliers. Outliers are values that are abnormally different from other values in a dataset. Finding extreme datapoints that are significantly different from other datapoints is important as they can distort the distribution of univariate or multivariate data. For our study, we implemented the Grubbs test (Urvoy & Autrusseau, 2014) analysis to detect a single outlier for each univariate dataset. By estimating the test statistic  $G$  (in Equation 4.3), Grubbs' test finds extreme deviates from the mean  $\bar{y}$  in the datasets. Here,  $y_t$  is the extreme value while  $s$  is the standard deviation.

$$G = \frac{\max_{i=n} |Y_t - \bar{y}|}{s} \quad (4.3)$$

Given a significance level  $\alpha$  (0.05) for any univariate data, the null hypothesis that there is no outlier in the data is not rejected if the critical value of the test

statistic  $G$  is more than the  $\alpha$  value and will be rejected if the critical value of the test statistics  $G$  is less than the  $\alpha$ .

### **Scaling and Normalization**

Scaling a dataset involves shifting and rescaling the dataset so that values range between 0 and 1. Two popular scaling techniques are min-max scaling and standard scalar. The standard scalar was used in this study. The mean of the scaled dataset is set to 0 with a unit variance ranging from -2.5 to +2.5 (Regmi & Rasmussen, 2018).

A normally distributed dataset is preferred for statistical analysis to reduce the effect of outliers and skewness (Armstrong, 1998; Varouchakis, 2021). Hence, normalization is apt for our dataset given that none of the features in the dataset has a normal distribution. This can also be useful for algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks (Park, Kim, & Lee, 2014). A log transformation of each feature is done using the Yeo-Johnson power transformation (Atkinson, Riani, & Corbellini, 2021). Figure 4.4 and Figure 4.5 show the distribution of the data before and after scaling and normalizing the dataset.

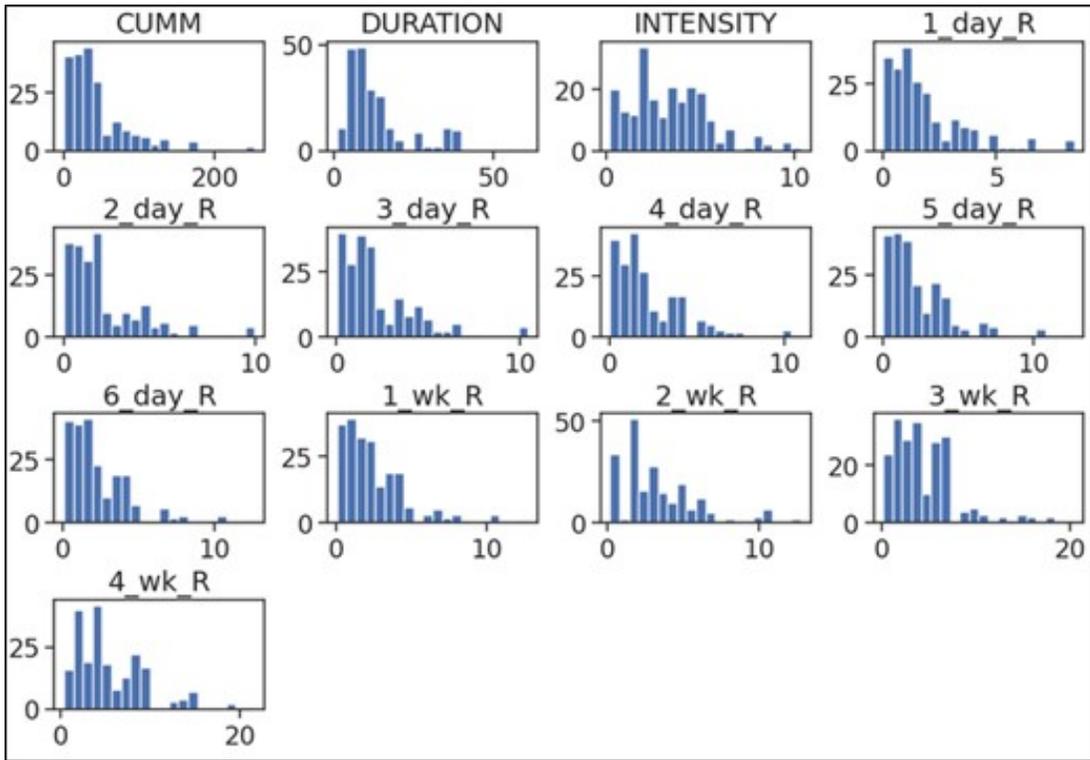


Figure 4.4: The distribution of the independent features of the dataset before any pre-processing steps.

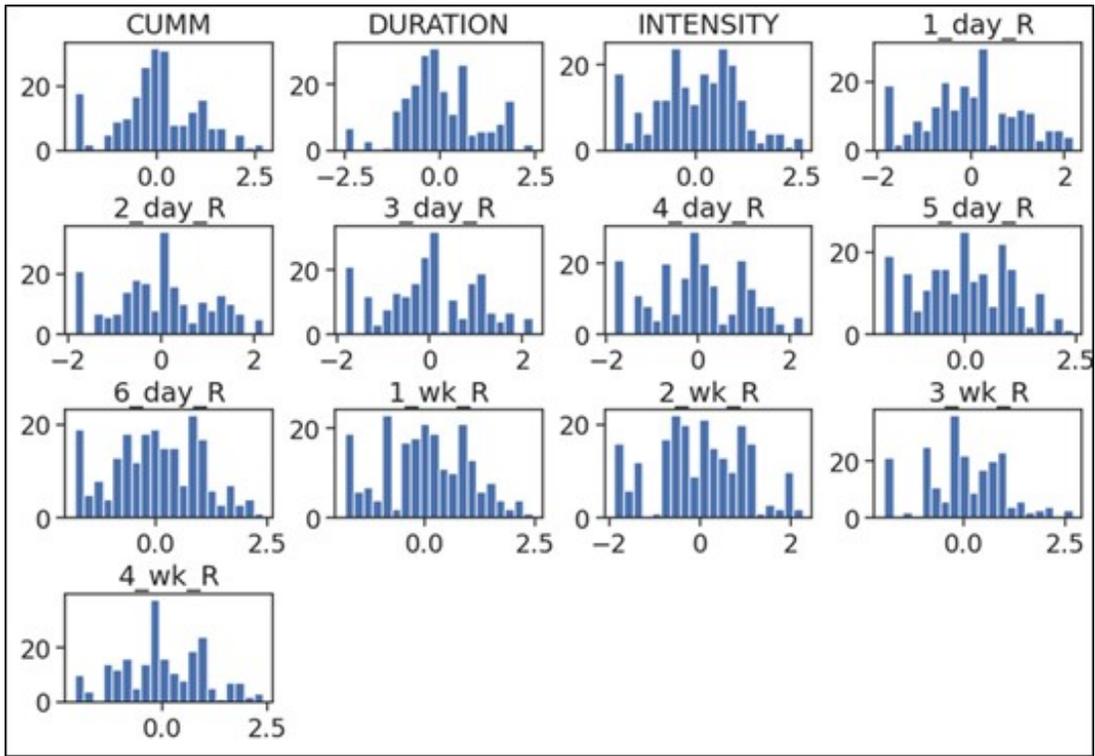


Figure 4.5: The distribution of the independent features of the dataset after scaling and normalizing the dataset.

## **Analysis of Variance Analysis (ANOVA)**

This is a statistical test used to detect whether the difference between independent features is statistically significant. A two-way ANOVA is used to determine the impact of two independent features (for example, rainfall intensity and one-week rainfall) on a dependent feature (like landslide occurrence) and to detect possible interactions between the two independent features.

## **Feature selection and reduction**

In preparing features as model input, it is important to rank important features and avoid redundancy in the datasets (Darst, Malecki, & Engelman, 2018). The Recursive Feature Elimination (RFE) method is used to rank important features (Darst et al., 2018). As a wrapper-based method, it uses a learning algorithm to select subsets of features from all available features recursively. Based on the performance of the desired number of features, features are concurrently added and removed from the subset features used for modeling until an optimal subset of features is reached. Using a greedy optimization technique, the best-performing feature subsets are kept and ranked. RFE allows for the choice of the number of features to keep and thus helps to reduce model complexity.

Pearson's Correlation is a common method used to identify correlated numeric features pair (Bairey Merz et al., 2016). The correlation value is the sum of the product of differences in the feature means divided by the sum by the product of the squared differences from the feature means. With the presence of high correlation pairs in the input features as seen in Figure 4.6, it is imperative to drop less important features that correlate with important ones. Highly correlated features possess the same predictive power and would be redundant, therefore

compromising the performance of the model

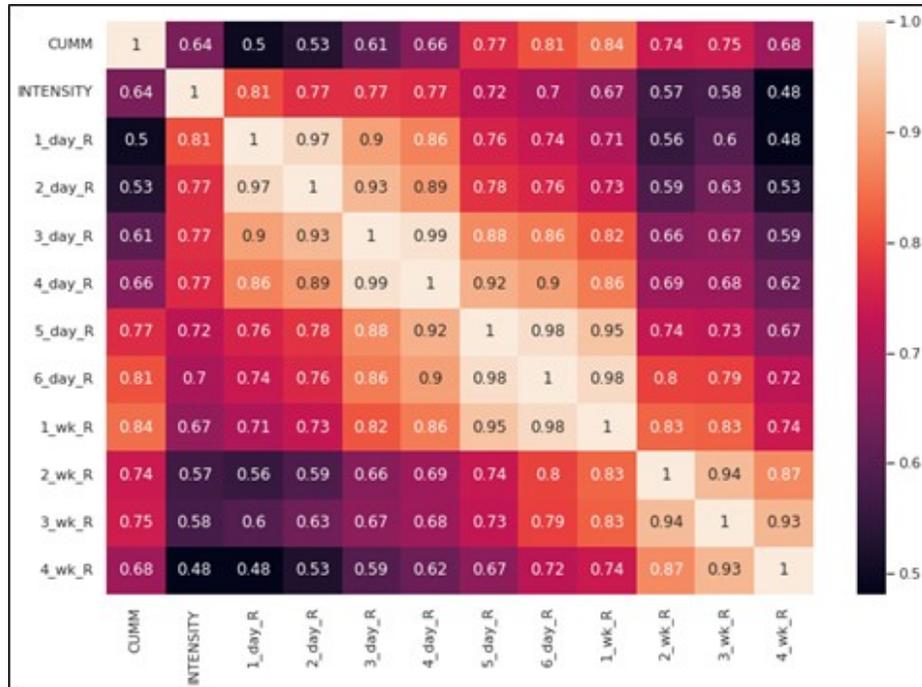


Figure 4.6: A display of the correlation metrics between the rainfall intensity and Antecedent Rainfall Index features.

### 4.2.3 Statistical machine Learning methods

I introduce a statistical machine learning model to find different types of relationships between the dependent variable (like landslide event outcome) and independent variables (like rainfall intensity and antecedent rainfall features). Statistical models also find the significance of each feature relationship (Sugiyama, 2015). By minimizing their respective loss function, machine learning methods learn data patterns that help to improve their predictive power (Jatau, Melnikov, & Yu, 2021) .

Suppose that we observe a quantitative response  $Y$  and  $p$  with different predictors,  $X_1, X_2, \dots, X_p$ . We assume that there is some relationship between  $Y$  and  $X = (X_1, X_2, \dots, X_p)$ , which can be written in the very general form in equation 4.4

$$Y = f(X) + \epsilon \tag{4.4}$$

Here  $f$  is some fixed but unknown function (linear or non-linear) of  $X_1, \dots, X_p$  and  $\epsilon$  is a random error term, which is independent of  $X$  and has an ideal mean value of zero. In this formulation  $f(x)$  is any machine learning method applied to train the model that predicts the binary outcome of a rainfall event,  $Y$ . The input features for the methods that will be explored for the study are 1D continuous numerical datasets. The Python Sci-Kit Learn library provides tools to implement all the methods discussed in the section.

#### Training the model

The prepared dataset  $X$  is taken as input by a training statistical ML model which has intrinsic tunable parameters called hyperparameters (Song et al., 2017). The

model produces a set of parameters  $\phi$ , which defines the model that takes in  $X$  and produces an outcome  $y$  as;  $f(\phi): X \rightarrow y$ . During model training,  $\phi$  is optimized by minimizing the loss function  $\psi$ . Function  $\psi$  penalizes the misclassifications between the actual class  $y$  and the predicted class by the model  $f(\phi)$  and can be expressed as  $(\frac{1}{n} \sum_{i=1}^n \psi(y_i, f_\phi(x_i)))$ .

Cross-validation (CV) is a statistical process of evaluating model training performance. The data is split in a particular pattern to be trained on one part of the dataset and tested on the other. Data splitting is mostly random and trained concurrently with testing. CV averages out overfitted estimations during training and assesses the performance of the method adopted. A frequently used CV technique is the K-fold CV. K-fold CV splits the dataset into  $k$  equal folds (10 folds in this study), then trains the model on 9 folds and uses the last fold to test the model. This process is done in several iterations until all folds are used for testing. The performance metrics of the model are the average performance of all the test folds used. Figure 4.7 shows the schematic steps for K-fold cross-validation.

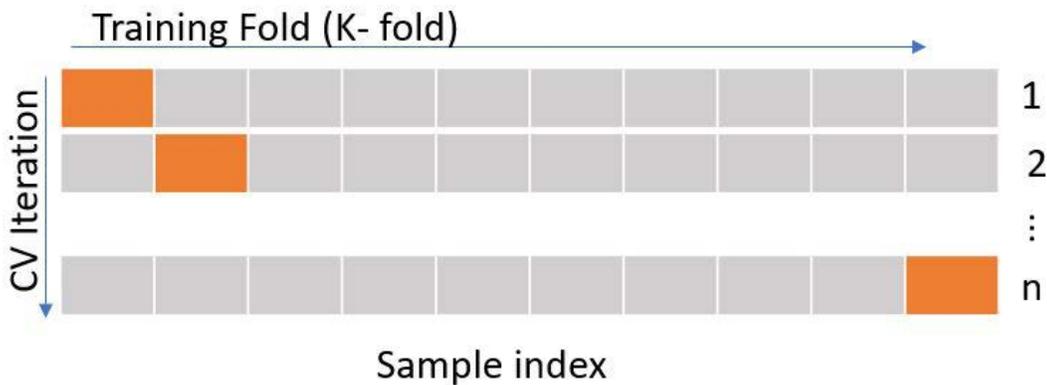


Figure 4.7: A schematic showing the iteration process starts from the top to bottom during training as each fold is used for validation from left to right for k-Fold cross-validation.

## Model performance and evaluation

We explored different types of machine learning approaches. We used parametric, non-parametric, tree-based, and linear regression approaches. To assess the performance of the model using the classification evaluation techniques we used the receiver operating characteristics and the Precision-Recall (PR) Curve.

### Receiver operating characteristics curve

The Receiver Operating Characteristic (ROC) curve defines the performance of a binary classifier system as its discrimination threshold changes (Kuhn & Johnson, 2013). The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a technique that evaluates the model performance given any possible variable thresholds for the model parameters. The ROC curve represents sensitivity as a function of the false positive rate (1-specificity) FPR. It can be generated by plotting sensitivity or true positive rate (TPR) in the  $y$ -axis against the cumulative distribution function of the false positive rate on the  $x$ -axis as seen in Figure 4.8.

The following four (4) quantities are used to compute the AUC-ROC parameters;

1. True Negative: The number of negative outcomes classified accurately.
2. True Positive: The number of positive examples classified accurately.
3. False Positive: The number of actual negative outcomes classified as positive.
4. False Negative: The number of actual positive outcomes classified as negative.

TPR and FPR are defined in equation 4.5 and equation 4.6 as

$$TPR = \frac{TP}{TP + FN} \quad (4.5)$$

and

$$FPR = \frac{FP}{FP + TN} \quad (4.6)$$

where  $TP$  = True positive,  $FP$  = False positive,  $TN$  = True Negative, and  $FN$  = False negative.

It has been widely used as a standard tool for evaluating the general performance of models (Chen et al., 2018). The AUC-ROC is a quantitative measure of the quality of the model, which can be categorized as poor (0.5–0.6), average (0.6 – 0.7), good (0.7 – 0.8) (Hosmer Jr, Lemeshow, & Sturdivant, 2013), very good (0.8 – 0.9), and excellent (0.9 – 1) (Chen et al., 2018). A low AUC value for model training with a low AUC value for model testing value indicates a model that performs poorly. A high AUC value for model training with a low AUC value for model testing value indicates a model that is overfitted. A high AUC value for model training and a high AUC value for model testing value indicates a good model and no overfitting is suspected. ROC curve is useful when the observations are balanced between each class, while the precision-recall curve is applicable for imbalanced datasets. Given the class imbalanced nature of our dataset, generating synthetic data for model training is important, making it possible to assess the model with the ROC curves.

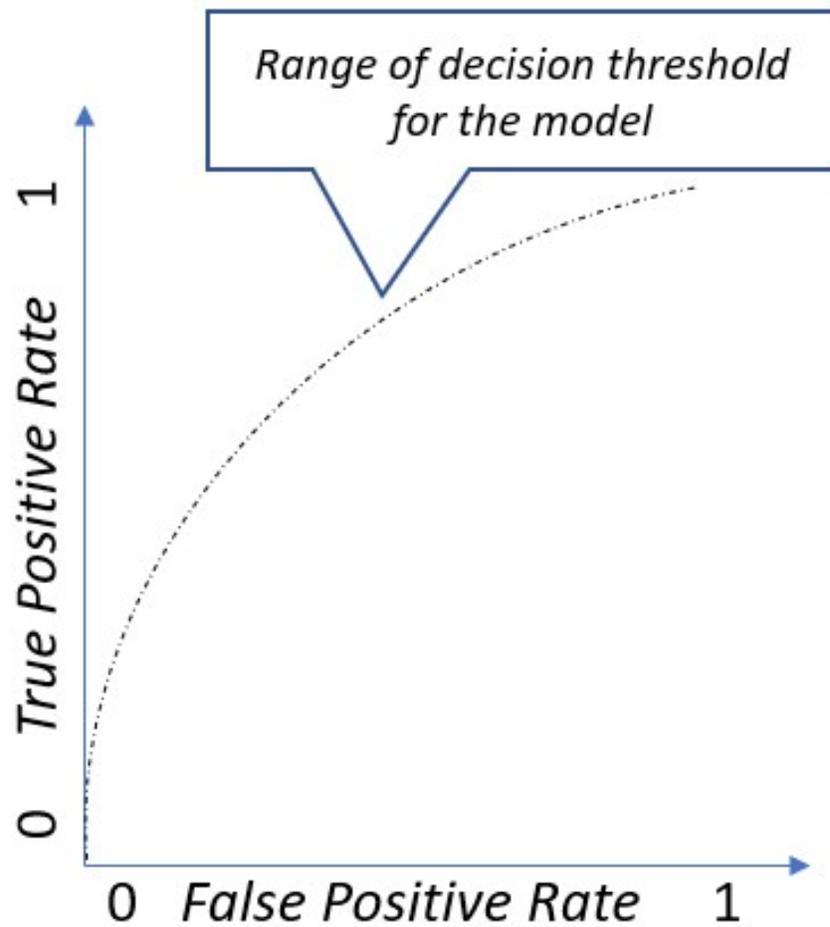


Figure 4.8: A typical AUC–ROC profile showing the range of decision thresholds for the model.

### Precision-Recall (PR) Curve

Another way of diagnosing and fine-tuning model performance is through the Precision-Recall (PR) curve. Through the PR curve, we assess how optimal the model can get, and the trade-offs involved whenever we prioritize any metrics. Given the class imbalanced nature of the dataset reviewing both precision and recall is important as we can choose which metric to prioritize. In the original dataset, there are many non-landslide events and only a few landslide events,

hence the recall metric is important. Prioritizing the recall metric is because the cost of false negatives is more costly than the false positives. The optimal threshold point for any model on the PR curve is the shortest distance from the PR curve to the edge of the perfect model curve.

## Evaluation Metrics

We assessed the performance of the classification models using 4 evaluation metrics: precision, recall, accuracy, and F-1 score.

### Precision

The measure of the correctly predicted positive landslide cases is divided by the predicted positive (landslide event) cases both correct and incorrect (Equation 4.7). With precision, we assess how efficiently our model can be useful to build a landslide warning system. It is defined mathematically in equation 4.7 as

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}. \quad (4.7)$$

### Recall

The measure of the correctly predicted positive landslide cases was divided by all the actual positive landslide cases that were predicted correctly and incorrectly. It is important when the cost of False Negatives is high. Recall helps to assess if the model is predicting landslide events as expected. It is defined mathematically in equation 4.8 as

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}. \quad (4.8)$$

## Accuracy

The measure of all the correctly identified landslide cases and non-landslide cases. This metric is effective if all the target classes are equally distributed. It is defined mathematically in equation 4.9 as

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}. \quad (4.9)$$

## F-1 Score

It measures the harmonic mean of the precision and recall. It helps to measure the nature of the misclassified observations. It gives a better measure of the incorrectly classified cases than the accuracy metric. Whenever the costs of misclassification differ, the F-1 score is a more critical metric than the accuracy metric. It is defined mathematically in equation 4.10 as

$$F - 1\ score = 2 \left( \frac{Recall \times Precision}{Recall + Precision} \right). \quad (4.10)$$

## Confusion Matrix

For binary classification problems, a confusion matrix is an array-like representation of values used to summarize the performance of a supervised classifier model in terms of the count of true positives, false negatives, false positives, and true negatives. Table 4.3 shows the template of a confusion matrix.

Table 4.3: Confusion Matrix for binary classification

		Predicted Class	
		Negative	Positive
True Class	Negative	True Negative (TN)	False Negative (FN)
	Positive	False Positive (FP)	True Positive (TP)

#### 4.2.4 K-Nearest Neighbors (KNN) Classifier

The K nearest neighbor is one of the simplest non-linear supervised learning algorithms. This method is an instance-based method that predicts a new observation based on the closest K neighbor observation from the training dataset (Cover, 1967). KNN method is developed by using individual samples within a predictor space of the training dataset to predict new observations (Kuhn, M. and Johnson, K., 2013). This method first stores all the datasets and then classifies new observations based on similarities within the available categories. The predicted outcome of the new observation is the mean of the response of the K neighbors. The neighborhood of the predictor space is determined mostly by the Minkowski distance defined in equation 4.11 where p is the degree of choice for Minkowski distance measured for k nearest neighbors of the sample majority class of the dataset. The equation is an estimation of the Euclidean geometry when  $p = 2$ . An advantage of the KNN algorithm is that there are no assumptions required about the distribution or the relationship between the dataset features hence it is a non-parametric method. Since the method relies on the distance between observations, scaling the input dataset is important

$$d(\bar{x}, x) = \left( \sum_{j=i}^k (\bar{x}_a - x_{aj})^p \right)^{\frac{1}{p}} \quad (4.11)$$

##### Algorithm implementation

1. Scale and normalize the dataset.
2. Initialize the  $k$  for a chosen number of neighbors.
3. Pick the first  $k$  entries from the sorted collection.

4. Get the labels of the selected  $k$  entries.
5. Calculate the distance between the  $n$  observations feature and the current observation feature.
6. Estimate and return the mode of the  $k$  labels

For KNN model optimization, 3 model hyperparameters can be optimized:

- the number of  $k$  neighbors,
- the choice of distance metric, and
- The weight of the sample observations.

## 4.2.5 Logistic Regression Classifier

Given the linear relationship of the rainfall features with the target class, the logistic regression method is considered as a classifier method in this study. Logistic regression is a method that computes the probability of a discrete outcome given an independent variable (Edgar & Manz, 2017). Like linear regression, it is built on the conditional probability distribution of the predictors given the values of the predictors. When fitting a straight line to a binary outcome, there is a validation that computes and predicts the probability of event  $X$  as  $p(X) < 0$  for some values of  $X$  and  $p(X) > 1$  for others (unless the range of  $X$  is limited). To avoid this problem, we must model  $p(X)$  using a logistic function that gives outputs between non-landslide rain, 0 and non-landslide rain, 1 for all values of  $X$  (equation 4.12 and equation 4.13). The right hand side of equation 4.13 is the Logit or log-odds function. Because the desired outcome is binary, a non-linear log transformation to compute the log odds of the probability that ranges from minus infinity to plus infinity.

$$odds = \frac{p}{1-p} \longrightarrow logit(p) = \ln \frac{p}{1-p} \quad (4.12)$$

$$logit(p) = mx + c = \ln \frac{p}{1-p} \quad (4.13)$$

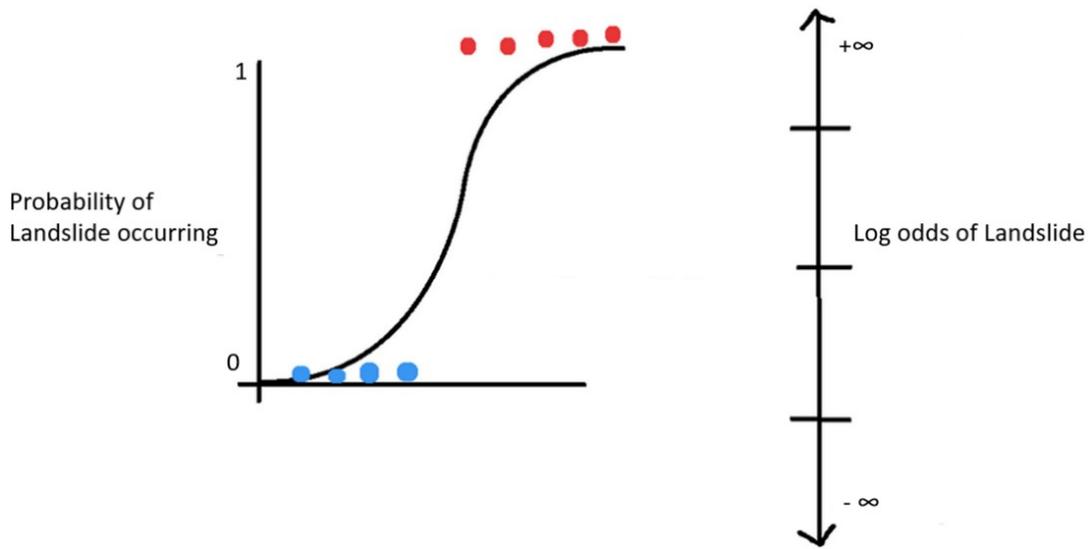


Figure 4.9: A schematic plot of a sigmoid function for the logistic regression method.

Compared to other highlighted methods, the logistic regression method is the most interpretable method as the relationship between predictors and dependent variables is evaluated and the loss function is statistically explained.

The critical hyper-parameters to be optimized for the logistic regression method are:

- regularization penalty: [L1, L2, elasticnet],
- the C parameter to control the strength of penalty: [100, 10, 1, 0.1, 0.01],  
and
- solver: [newton-cg, lbfgs, liblinear].

### 4.2.6 Random forest (RF) Classifier

Random forest is a classification method that uses multiple decision tree predictors whereby each tree depends on the values of randomly chosen vectors that are distributed evenly among all trees in the forest (Herrera, 2019). A decision tree is a set of conditions for optimal classification of outcomes in a hierarchical pattern using leaves and nodes. In standard trees, each node is split using the best split among all variables (landslide conditioning factors). In a random forest, each node is split using the best split among a subset of predictors that are randomly chosen by the node. It has become a popular method for finding hidden patterns within large volumes of data. To determine the best node to split, there are  $n$  variables that can be chosen as random subsets from the training data. One can compute the best node split using Gini criteria (Breiman, 2001). These criteria measure the correlation degree between variables and results. According to the random forest algorithm, the lowest Gini Criteria value is considered to be the best split for each node (Kausar & Majid, 2016). Gini criteria is expressed in equation 4.14 as

$$GiniCriteria = 1 - \sum_j P_j^2. \quad (4.14)$$

$p_j$  represents the distribution of  $j$  class labels at a node. At node  $k$ , the Gini criteria is minimum 0 when all the rainfall events belongs to one unique class. Hence, the feature to split at the node  $k$  will be the feature with the least gini index. A summary of building a random forest model is itemized below;

#### Algorithm Implementation

- input dataset features is sub-sampled with replacement

- use random features for each node partition optimization
- aggregate all decision trees built independently. The probability of the outcome of a rainfall event is the proportion of the tree that classifies the outcome.

For model optimization, the random forest hyperparameters to be tuned are;

1. `n_estimators`: The `n_estimators` hyperparameter is the number of decision trees the method considers as maximum voting inputs and based on the aggregate average of the tree makes a prediction. A higher number of trees reduces the model variance and increases the model performance. A trade-off would increase the computation time.
2. `max_features`: Give a value as the maximum number of features that the algorithm picks to split at the node.
3. `min_sample_leaf`: This determines the minimum number of leafs required to split at an internal node of a decision tree.

## 4.2.7 Gradient Boost Classifier

Gradient Boost Classifier (GBC) is a decision tree ensemble machine learning method that relies on the intuition that the perfect possible next model combines with the previous models to minimize the overall prediction error (Kuhn & Johnson, 2013). The concept of boosting is initiated as a sequence that converts weak learners to strong ones by adding trees to them. A new tree output  $y$  that is scaled down by the learning rate  $v$  is added to a weak learner  $F_0$  for the dataset  $x$  (equation 4.15) as

$$F_1(x) = F_0(x) + vy. \quad (4.15)$$

Generally, the GBC method depends on the information from patterns of model residuals of weak models and improves on subsequent models. At each stage of learning, GBC uses a greedy strategy of selecting the optimal learner which produces an optimal solution that sometimes overfits. There are ways to curtail these drawbacks through optimizations of the method hyperparameters like the shrinkage value.

### Algorithm implementation

1. Initialize the model with a constant value prediction
2. for  $m = 1$  to  $M$  iterations:
  - Calculate the residual  $r_m$  by taking the derivative of the loss function with respect to the previous prediction
  - Train the classification tree with features  $x$  dataset against residual  $r$ .
  - Compute a function  $F_i(x)$  that minimizes the loss function at each terminal node.

3. Update the prediction of the combined model.

The critical hyper-parameters to be optimized for the GBC method are:

- Criterion: This is the loss function chosen to search for the best feature and the threshold required to split the data.
- learning\_rate: This parameter controls the amount of contribution from each new tree. A general range for the learning rate is 0.1 – 0.3.
- max\_depth: The maximum depth of each tree estimator.
- n\_estimators: This is the number of trees to build.
- init: The initial estimator is the  $\log(\text{odds})$  prediction that is converted into a probability by using the logistic function.

### 4.2.8 Model Optimization: Ensemble Voting Classifier

The Voting Classifier model is an estimator that optimizes the performance of all the individual models to build a meta-classifier and cancels out their weaknesses. It introduces the intuition of optimizing the diversity of the individual models to reduce the variance (Lim, 2022). A final voting classifier model is designed based on the use case for the research which is the recall metric.

The recall value of each individual model is used as their weights for the voting classifier. The weights are quantified relative to the magnitude of other individual model weights and can take any numerical value weight (Liu et al., 2020; Brownlee, 2021). The outcome of a voting classifier is the average of the integration of the individual weights and their respective prediction. This classifier is developed in the following three steps:

#### **Algorithm implementation**

1. Build two or more individual predictive models.
2. Train individual predictive models with similar datasets.
3. Predict the result using the average of the individual predictions.

# Chapter 5

## Results

### Summary of Rainfall characteristics

The statistics for all the rainfall events recorded and used for the study are summarized in Table 5.1. The mean, standard deviation, minimum, maximum, 25th, 50th, and 75th percentile are provided.

Table 5.1: Statistics of all event features (landslide and non-landslide events in mm)

	mean	std	min	0.25	0.5	0.75	max
CUMM	44.94	43.34	0.36	17.02	31.32	60.83	253.18
DURATION (hrs)	13.44	10.89	1.00	6.00	9.50	15.00	61.00
INTENSITY	3.37	2.10	0.21	1.92	3.28	4.48	10.37
1_day_R	1.87	1.74	0.01	0.72	1.36	2.41	8.42
2_day_R	2.05	1.98	0.01	0.73	1.52	2.91	9.97
3_day_R	2.20	2.06	0.01	0.82	1.59	3.51	10.43
4_day_R	2.26	2.09	0.01	0.82	1.62	3.51	10.98
5_day_R	2.37	2.19	0.02	0.83	1.72	3.52	12.71
6_day_R	2.41	2.20	0.07	0.86	1.83	3.52	12.71
1_wk_R	2.48	2.21	0.07	0.88	1.95	3.52	12.71
2_wk_R	3.40	2.67	0.17	1.57	2.82	4.96	12.82
3_wk_R	4.59	3.58	0.19	2.02	3.88	6.22	20.42
4_wk_R	5.58	4.12	0.40	2.22	4.20	8.05	21.71

Table 5.2: Statistics of all event features for landslide events in mm.

	<b>mean</b>	<b>std</b>	<b>min</b>	<b>25%</b>	<b>50%</b>	<b>75%</b>	<b>max</b>
<b>CUMM</b>	68.44	54.06	1.74	32.70	60.86	86.03	249.45
<b>INTENSITY</b>	4.53	1.90	0.87	3.48	4.37	4.98	10.36
<b>1_day_R</b>	2.80	1.88	0.07	1.29	2.41	3.98	8.42
<b>2_day_R</b>	3.36	2.28	0.07	1.52	3.39	4.59	9.82
<b>3_day_R</b>	3.61	2.55	0.07	1.52	3.54	4.61	10.43
<b>4_day_R</b>	3.75	2.64	0.07	1.52	3.56	5.41	10.98
<b>5_day_R</b>	4.11	2.88	0.07	1.55	3.63	5.71	12.71
<b>6_day_R</b>	4.17	2.96	0.07	1.56	3.63	5.71	12.71
<b>1_wk_R</b>	4.33	2.98	0.07	2.17	3.64	5.85	12.71
<b>2_wk_R</b>	5.52	3.60	0.25	2.71	4.48	8.42	12.82
<b>3_wk_R</b>	8.15	5.14	0.26	4.46	7.04	9.57	20.42
<b>4_wk_R</b>	9.53	5.29	0.89	6.68	8.32	12.58	21.71

## 5.1 Empirical Rainfall Threshold

Two datasets, the hourly and the daily rainfall dataset were used to develop a regional threshold for the study area. A total of 216 rainfall events are plotted on the rainfall intensity duration curve. A minimum threshold relation was developed for minimum landslide-rainfall event using hourly and daily rainfall records. For the hourly and daily dataset, the optimal threshold fitted to capture the lower boundary of the landslide points and has a power curve equation expressed in equation 5.1 and 5.2 respectively as

$$I = 1.531 + (8.9D^{-0.9064}) \quad (5.1)$$

$$I = 0.01 + (9.13D^{-0.6454}) \quad (5.2)$$

where  $I$  represents the rainfall intensity and  $D$  is the duration in the threshold equation.

The range of duration of the rainfall datapoints for threshold for the hourly data is  $1 \text{ hour} \leq \text{Duration} \leq 64 \text{ hours}$ , while the range of duration of the rainfall datapoints for threshold for the daily data is  $24 \text{ hours} \leq \text{Duration} \leq 432 \text{ hours}$ . The striation seen in the datapoints of the daily rainfall data in Figure 5.2 is a result of the frequency of sample collection (24hrs). The low sampling frequency of data makes it challenging to accurately estimate rainfall intensity. Figure 5.1 shows the moderate estimate of the empirical threshold of the study area.

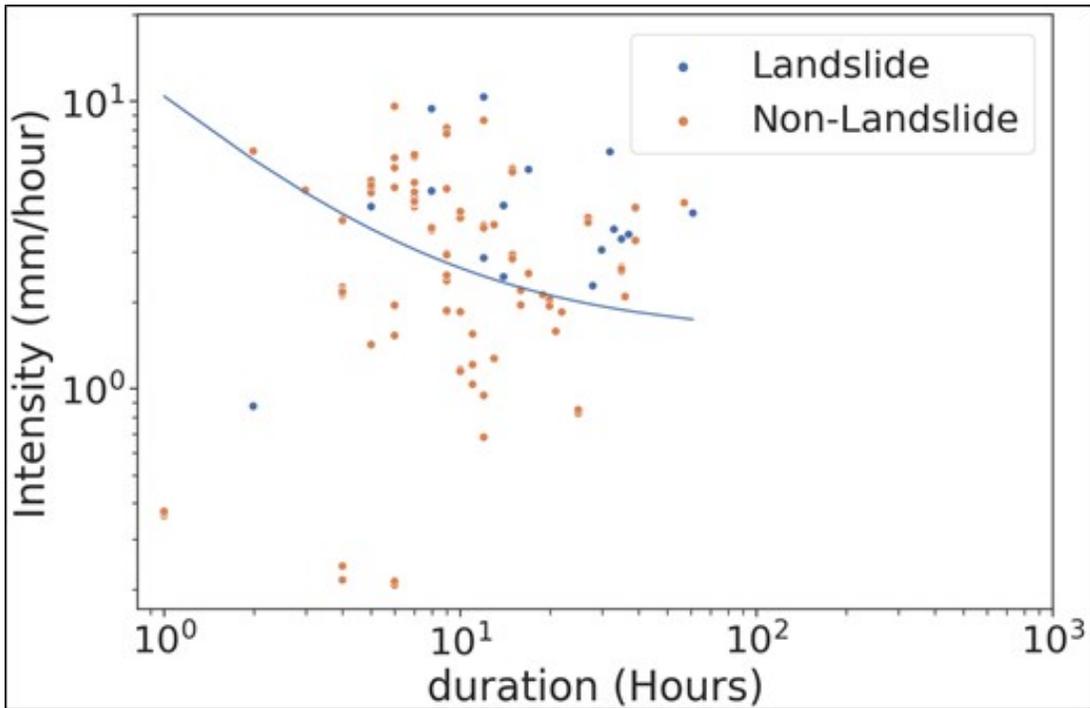


Figure 5.1: A rainfall intensity-duration scatterplot used to build an empirical threshold curve using the hourly dataset.

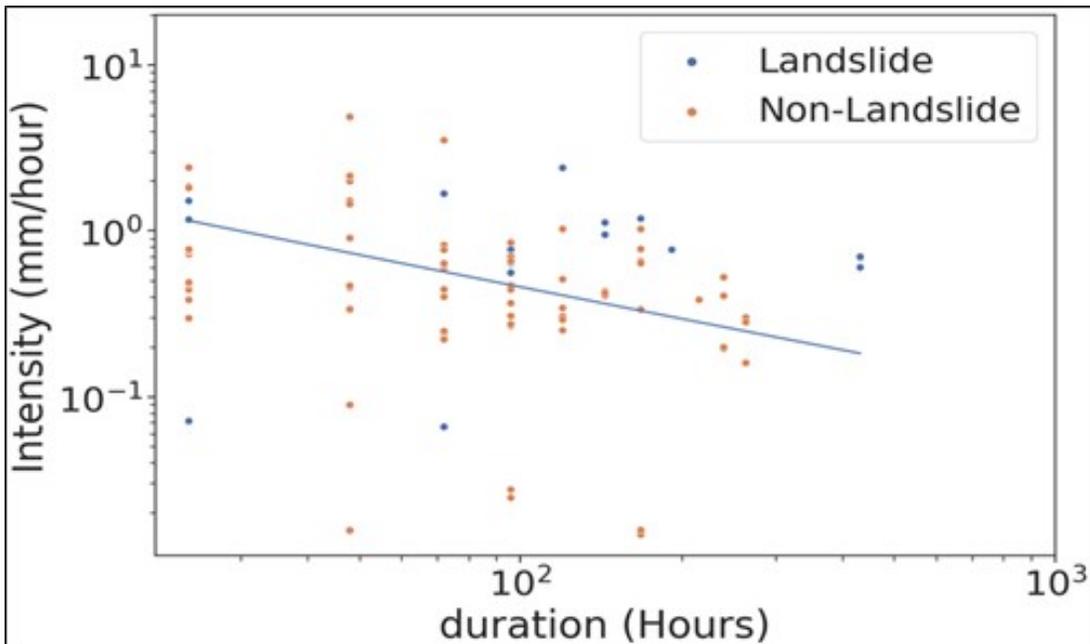


Figure 5.2: A rainfall intensity-duration scatterplot used to build an empirical threshold curve using the daily dataset.

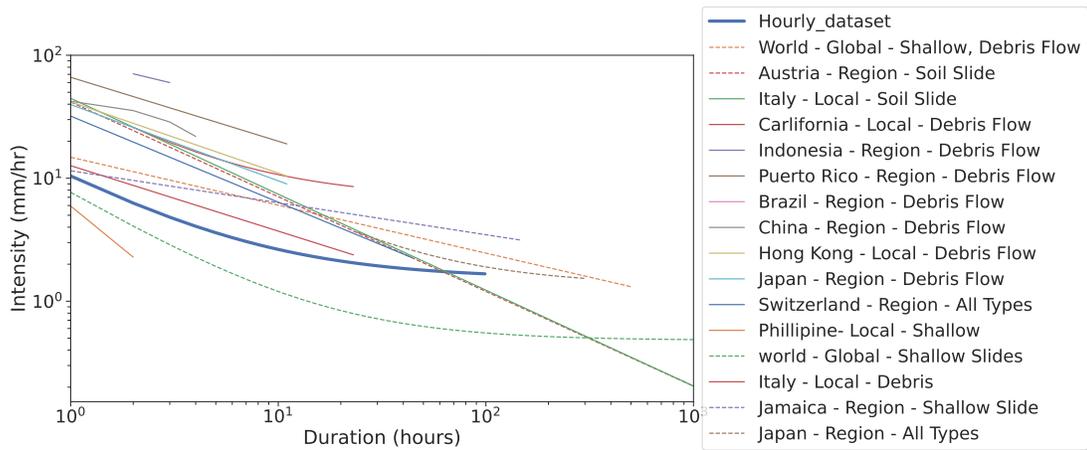


Figure 5.3: A plot showing the new threshold curve for the study area overlain on the thresholds developed for all the landslide regions mentioned in table 4.3. The dataset used is the hourly dataset.

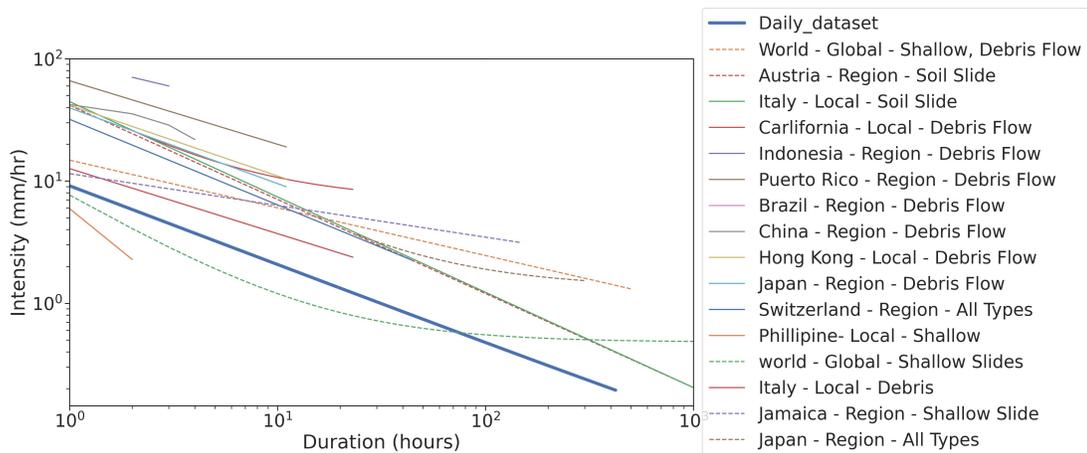


Figure 5.4: A plot showing the new threshold curve for the study area overlain on the thresholds developed for all the landslide regions mentioned in table 4.3. The dataset used is the daily dataset.

## 5.2 Statistical Machine Learning Models

We assess the performance of the models using the confusion matrix and ROC curve. We explore the PR curve and hyperparameters optimization for all the models. From PR curves, events of overfitting during training are checked by plotting both the test and training AUCs together. The individual models, KNN, LR, RF, GBC were trained and tested using the same training and testing dataset.

### Feature importance

From the ranking in Figure 5.5, the top five features, which are two week rain, six day rain, four week rain, rainfall intensity, and two day rain, were used as input features for the machine learning methods. The choice is also validated using the correlated pair information in Figure 4.5 as none of the features had the closest correlated pair in the input features.

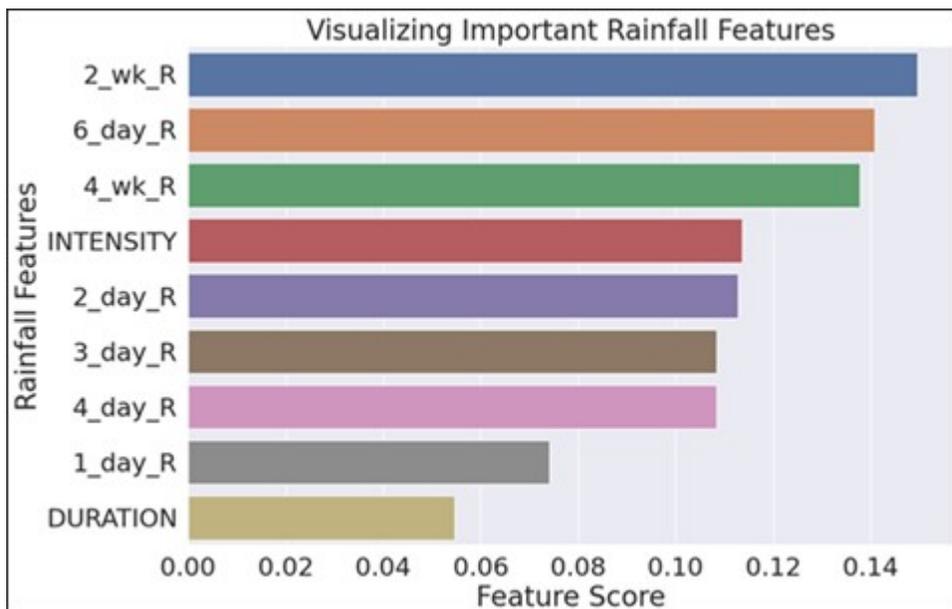


Figure 5.5: A plot showing the rank of the features' importance in descending order. The top 5 features are selected.

## Model optimization

The model optimization approach for this study is hyper-parameter optimization. The hyperparameters of all four models are used to increase their predictive power based on the recall metric. Performance of the models derived based on the use-case of the research application (recall), is used as a weight average for individual models for the voting ensemble classifier model.

## Model Evaluation

From the PR curve, a model's optimal performance is the shortest point along the PR curve closest to the top right corner of the perfect model profile. We adopted a baseline random guess as the least acceptable model performance with a precision of 50%. From the confusion matrix figures, 1 denotes landslide events while 0 denotes non-landslide events.

The default KNN model performance is optimal as a precision of 75% and recall of 42.85% from the confusion matrix (Figure 5.8) is on the same point as the best point on the PR curve (Figure 5.6). The AUC-ROC value of the training and testing differ by a margin of 13%. Hence, model overfitting is suspected during model training.

The default LR model is sub-optimal as the model precision is 50% and recall is 14.285% (Figure 5.11) which are lower than the optimal LR model values of 65% precision and recall of 55% as the point in the PR curve (Figure 5.9). AUC-ROC for training and AUC-ROC for testing differ by approximately 2%. Hence, overfitting is not suspected during training (Figure 5.10).

The default RF model is sub-optimal as the model precision is 66.66% and recall is 28.57% from the confusion matrix (Figure 5.14). The performance metrics

are lower than the optimal RFC model values of 65% precision and 100% recall respectively as the point in the PR curve (Figure 5.12). AUC-ROC for training and AUC-ROC for testing differ by approximately 4%. Hence, overfitting is not suspected during training (Figure 5.13).

The default GBC model's performance on the test dataset is 80% precision and recall is 57.14% from the confusion matrix(Figure 5.17). These performance metrics are lower than the optimal RFC model values of 75% precision and 70% recall, respectively, as the point in the PR curve (Figure 5.16). AUC-ROC for training and AUC-ROC for testing differ by approximately 10%. Hence, overfitting is suspected during training (Figure 5.15).

For weight voting in the ensemble classifier, the weighted average of all four models is given as [0.14, 0.15, 0.45, 0.35] for KNN, LR, RF and GBC respectively. On the same test dataset, the voting ensemble model performed at a recall of 100% and precision of 70% (Figure 5.20)

## Results KNN Classifier

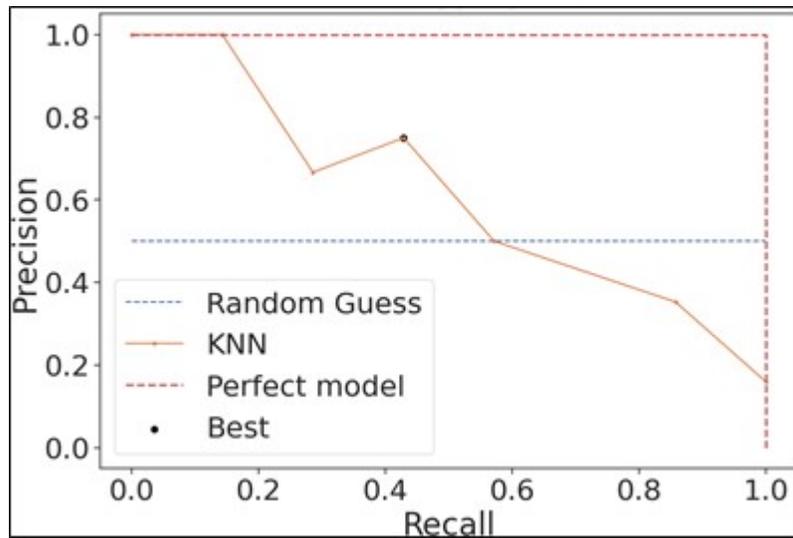


Figure 5.6: A plot of the PR curve showing the performance of the KNN model in terms of possible range of precision and recall it can achieve. The optimal KNN would perform at recall of 42.85% and precision of 75%. The default model achieved an optimal performance at recall of 42.85% and precision of 75%.

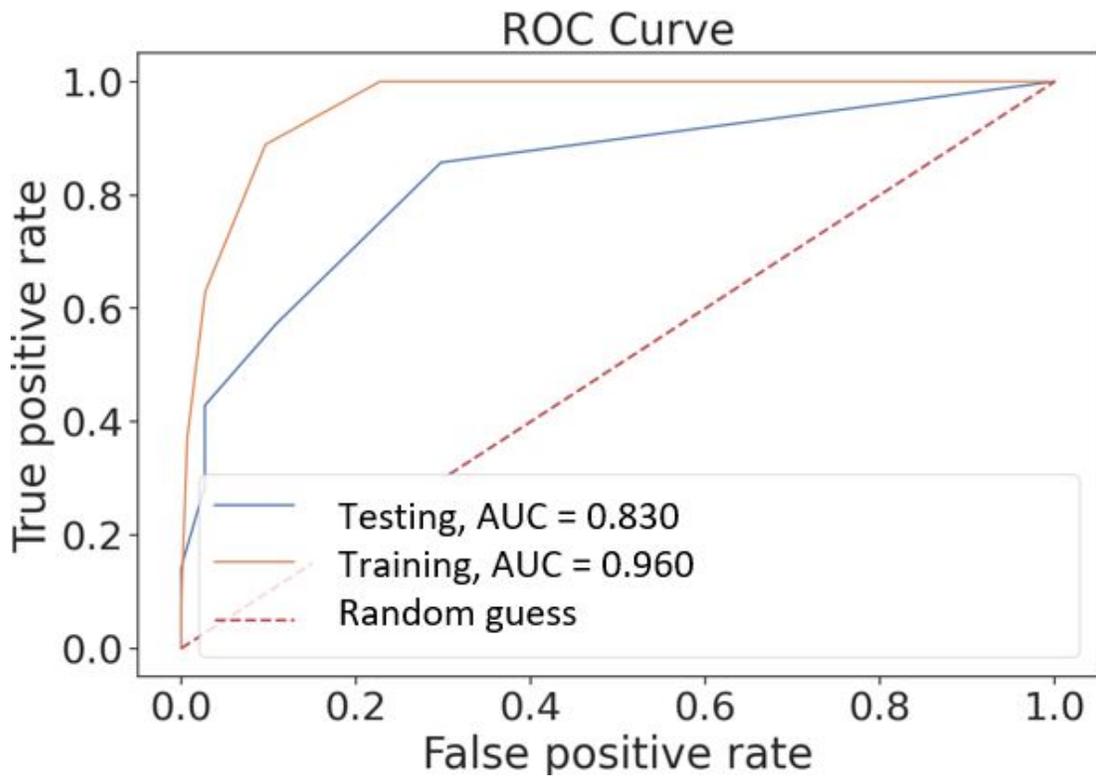


Figure 5.7: A Plot showing the ROC-AUC curves of the KNN model. The AUC for the training is larger than AUC of testing. Model overfitting is inferred.

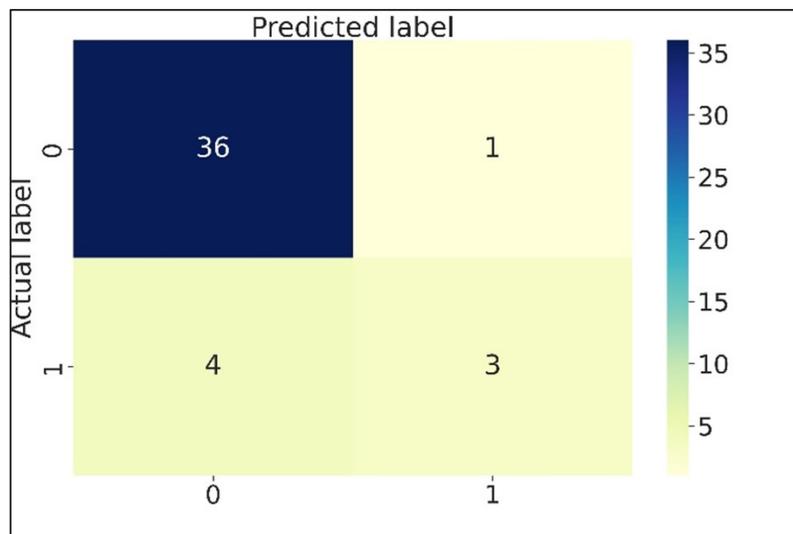


Figure 5.8: A confusion matrix showing the performance of the KNN model on the test data.

## Result Logistic Regression Classifier

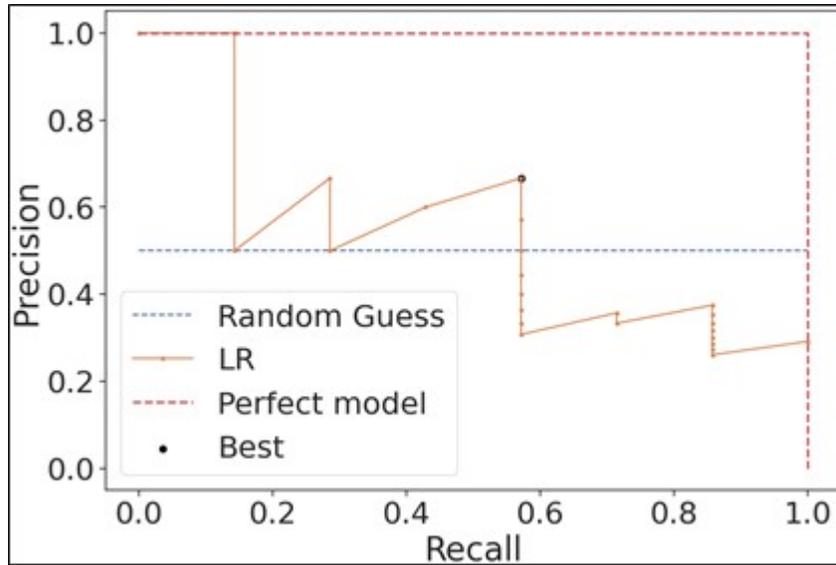


Figure 5.9: A plot of the PR curve showing the performance of the LR model in terms of possible range of precision and recall it can achieve. The optimal LR would have a recall of 58% and a precision of 70%. The default LR model did not achieve an optimal performance at recall of 14.285% and precision of 50%.

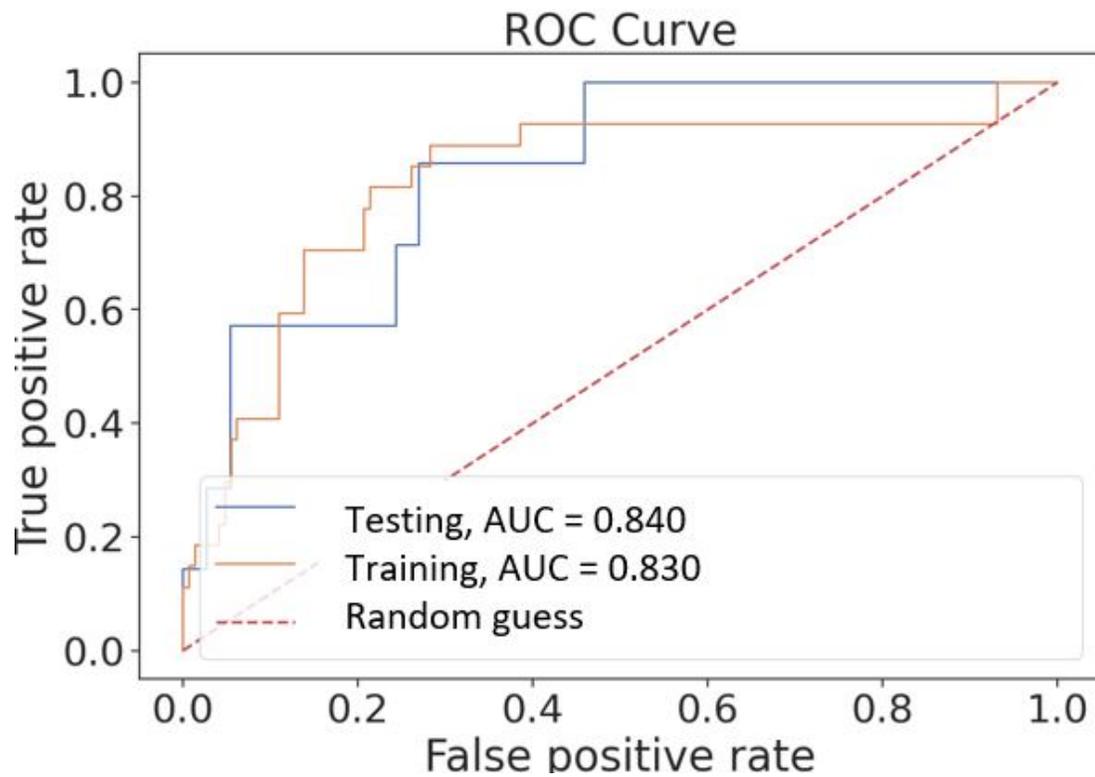


Figure 5.10: A Plot showing the ROC-AUC curves of the LR model. The AUC for the training is larger than AUC of testing. Both training and testing have relatively the same threshold profile and same AUC.

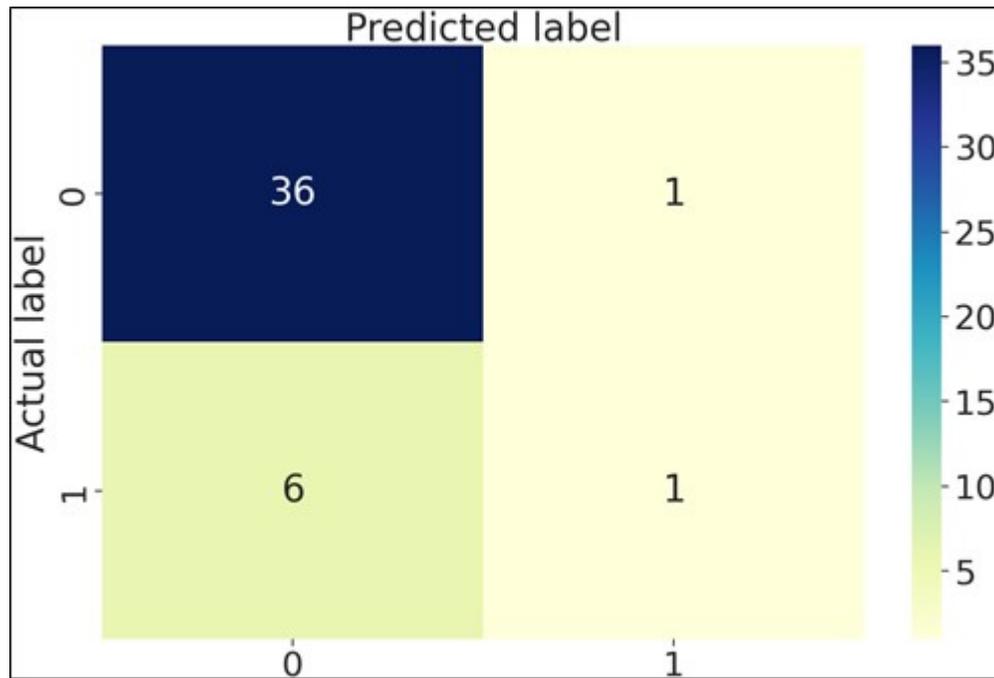


Figure 5.11: A confusion matrix showing the performance of the LR model on the test data.

## Result Random Forest Classifier

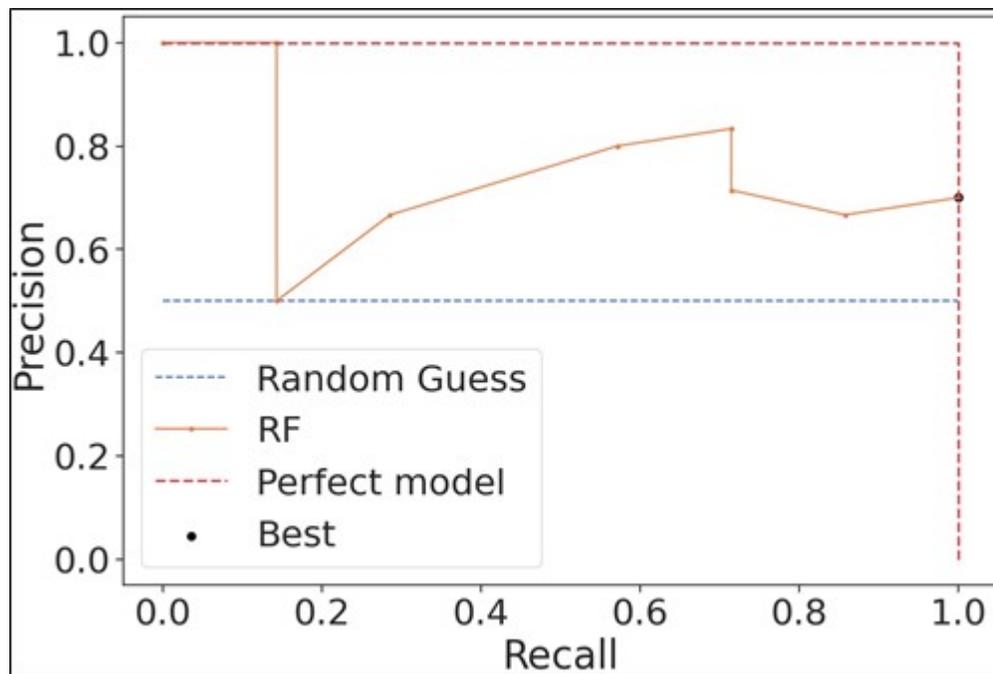


Figure 5.12: A plot of the PR curve showing the performance of the RF model in terms of possible range of precision and recall it can achieve. The default RF model did not achieve an optimal performance at recall of 28.57% and precision of 66.66%.

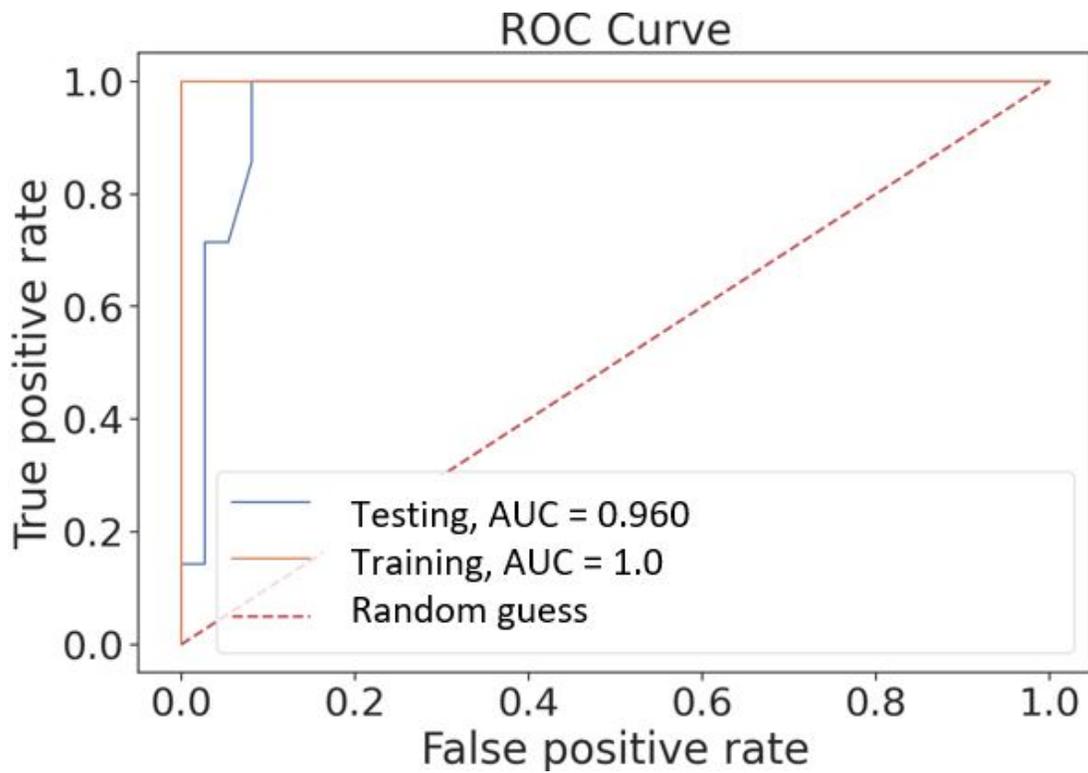


Figure 5.13: A Plot showing the ROC-AUC curves of the LR model. The AUC for the training is larger than AUC of testing. Both training and testing have relatively the same threshold profile and same AUC.

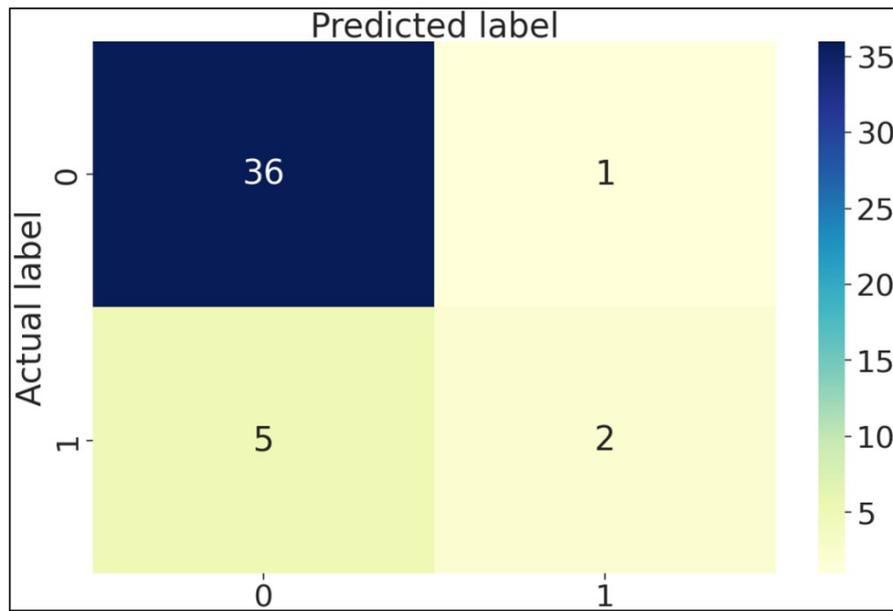


Figure 5.14: A confusion matrix showing the performance of the RF model on the test data.

## Result Gradient Boost Classifier

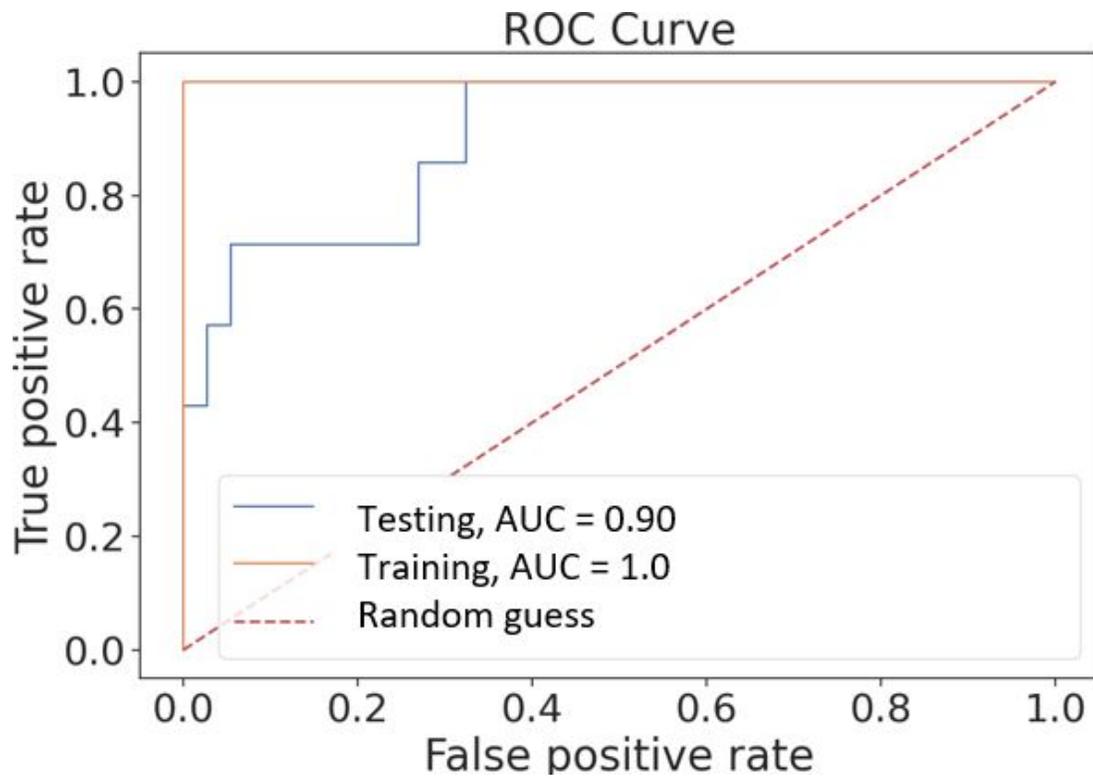


Figure 5.15: A Plot showing the ROC-AUC curves of the GBC model. The AUC for the training is larger than AUC of testing. Model overfitting is inferred.

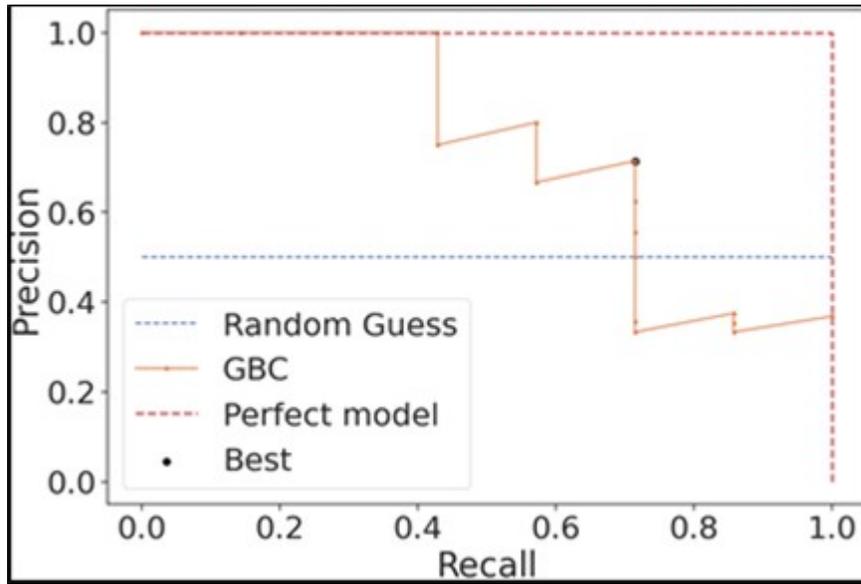


Figure 5.16: A plot of the PR curve showing the performance of the GBC model in terms of possible range of precision and recall it can achieve. The default model did not achieve an optimal performance as the recall is at 57.14% and the precision is 80%.

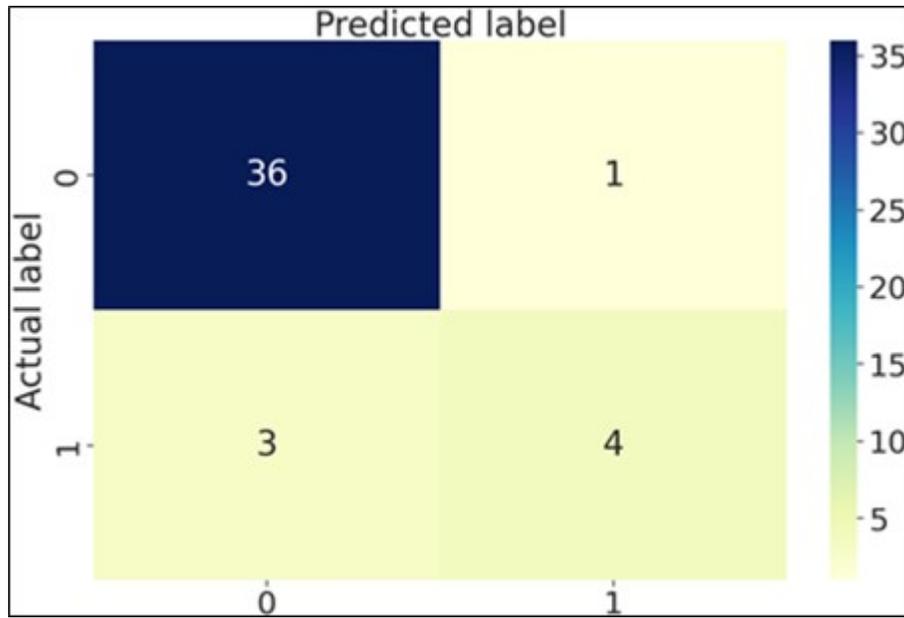


Figure 5.17: A confusion matrix showing the performance of the GBC model on the test data.

## Result Voting Classifier

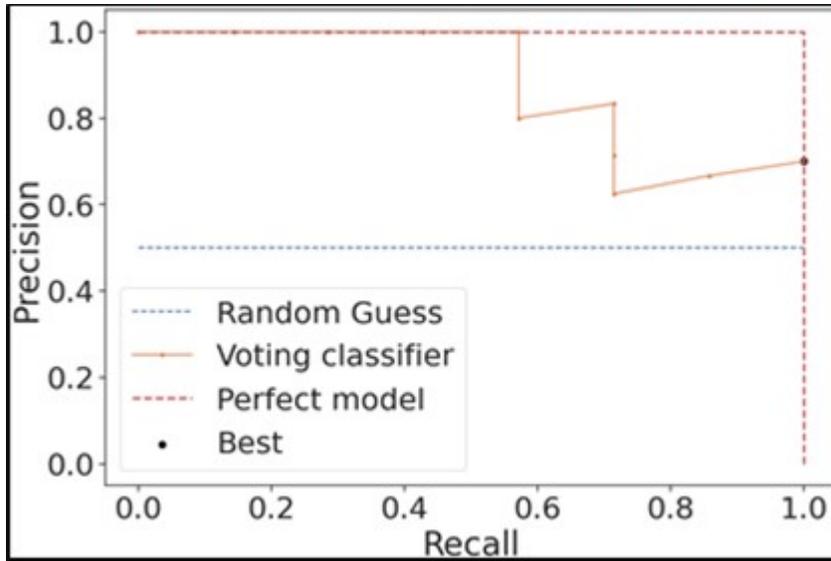


Figure 5.18: A plot of the PR curve showing the performance of the Voting classifier model in terms of possible range of precision and recall it can achieve. The model is optimized for recall performance, and it achieved an optimal recall of 100% and precision of 70%.

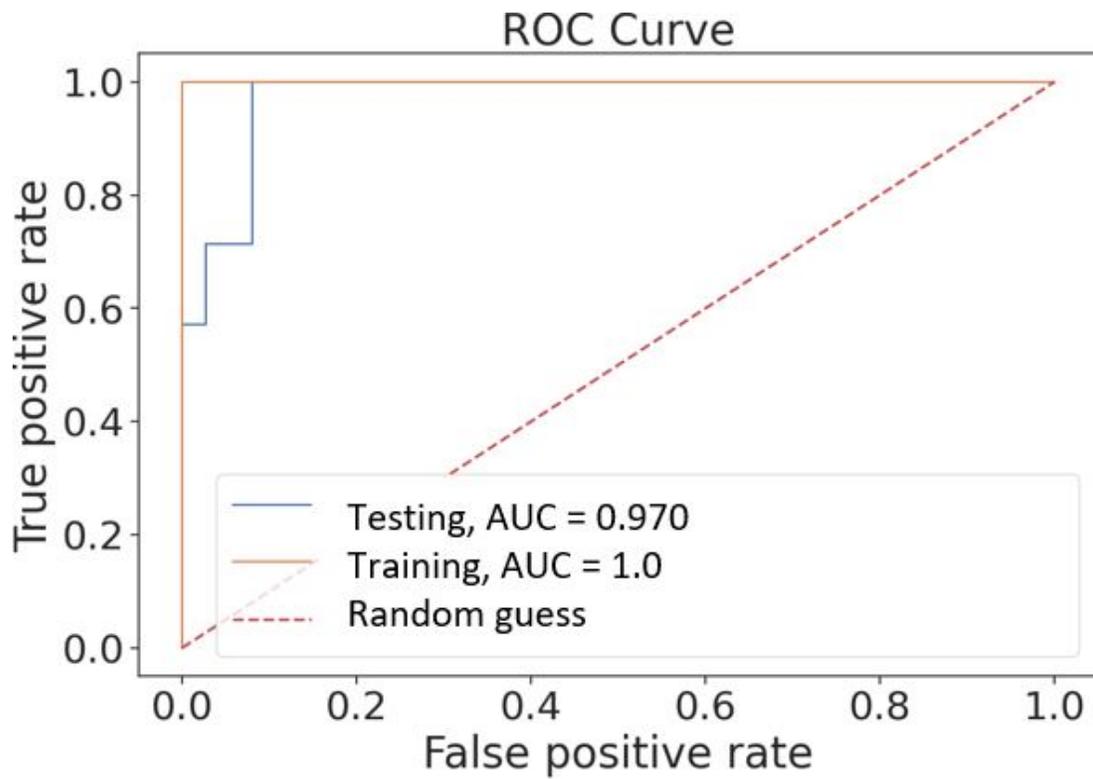


Figure 5.19: A Plot showing the ROC-AUC curves of the Ensemble voting classifier model. The AUC for the training is approximately the same as the AUC of testing.

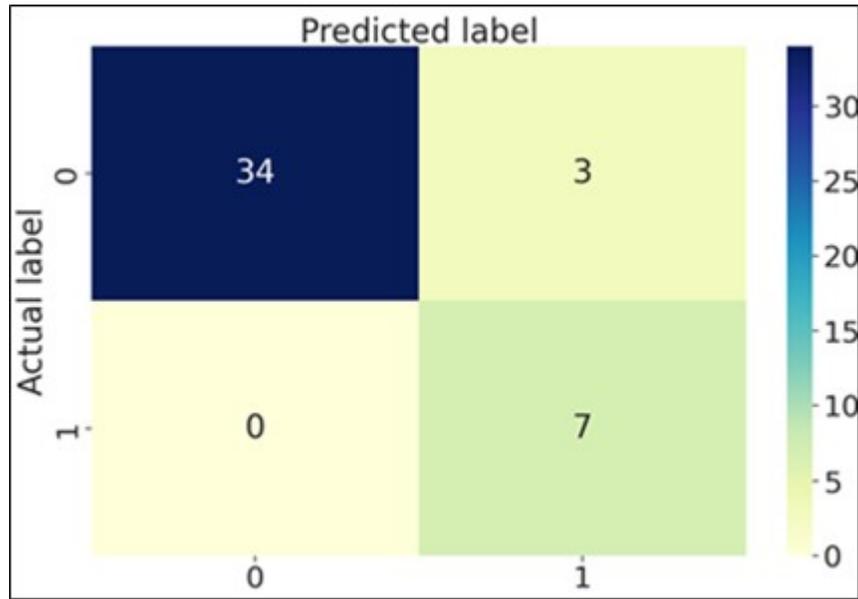


Figure 5.20: A confusion matrix showing performance of the Voting Classifier model on the test data.

### 5.3 Discussion

Using only rainfall intensity-duration relationship, the empirical threshold analysis barely distinguished landslide events from non-landslide events. We can infer from the threshold curves that the average rainfall intensity required to initiate landslide event in the study area is relatively lower than majority of the rainfall intensity thresholds obtained for various locations around the world.

Integrating antecedent features with other rainfall features shows a distinct difference in model performance. The ML models performed better than empirical methods in the forecast of landslide events (Figure 5.6 to Figure 5.20). With the ranking of feature importance in Figure 5.5, three antecedent rainfall features have been identified to be more related to landslide initiation than rainfall intensity.

With high and frequent antecedent rainfall events, landslide initiation occur-

ring with little rainfall intensity is possible. Machine learning models have better results and several parameters to tweak to achieve efficient performance. With an ensemble of all the models, we achieved 100% recall.

# Chapter 6

## Conclusions

The empirical rainfall intensity-duration threshold developed for this study suggests that the landslides in the region are initiated with minimal rainfall intensity values. In addition, comparison of the curve with similar curves obtained for different regions across the world indicates alignment with that trend.

This study implemented different ML approaches to determine major rainfall conditions that can trigger landslides in the study area. From the ML model analysis, the two-week antecedent rainfall events play a crucial role in triggering landslides. Results suggests there are more than one feature that affect landslide initiation. Compared to the individual machine learning models' performance, there is marginal improvement in the ensemble classifier model as the recall and the accuracy are improved in the final model used. In order to suit the desired project's use case, the hyperparameters of the ensemble classifier model is optimized in favor of the recall metrics to obtain a recall of 100%.

Both the empirical threshold model and machine learning models have been able to establish statistical relationships between landslide events and precipitation data (at prior events and at current time).

# Chapter 7

## Future Work

There are models that can be expended to accommodate various complexities in mapping and forecasting landslides. To develop a regional landslide hazard emergency management system, a more robust landslide forecast warning system that considers spatial variations of other environmental factors is crucial. These environmental covariates include geomorphology, land-use development, elevation, distance to stream, composition of soil cover, aspect ratio of slope, drainage pattern, type of vegetation, and slope geometry. These provide extensive data about the terrains that strongly connect to landslide occurrences.

Upon incorporating other static environmental covariates, we can consider other dynamic environmental covariates like earthquake vibrations. A fitting machine learning method to explore for this challenge is time series analysis and recurrent neural network modifications like long-short term memory (LSTM).

# References

- Adamski, J., & NWQAP. (1995). *Environmental and hydrologic setting of the ozark plateaus study unit, arkansas, kansas, missouri, and oklahoma*. National Water-Quality Assessment Program. Retrieved from <https://books.google.com/books?id=pGnuAAAAMAAJ>
- Armstrong, M. (1998). *Basic linear geostatistics*. Springer Science & Business Media.
- Atkinson, A. B., Riani, M., & Corbellini, A. (2021). The boxâcox transformation: Review and extensions. *Statistical Science*.
- Bairey Merz, C. N., Handberg, E. M., Shufelt, C. L., Mehta, P. K., Minissian, M. B., Wei, J., ... others (2016). A randomized, placebo-controlled trial of late na current inhibition (ranolazine) in coronary microvascular dysfunction (cmd): impact on angina and myocardial perfusion reserve. *European heart journal*, 37(19), 1504–1513.
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3), 199–231.
- Brownlee, J. (2020). A gentle introduction to imbalanced classification. *Machine Learning Mastery*(2). Retrieved from <https://machinelearningmastery.com/what-is-imbalanced-classification/>
- Brownlee, J. (2021). How to develop a weighted average ensemble with python. *Machine Learning Mastery*(2). Retrieved from <https://machinelearningmastery.com/weighted-average-ensemble-with-python/>
- Caine, N. (1980). The rainfall intensity: Duration control of shallow landslides and debris flows. *Geografiska Annaler. Series A, Physical Geography*, 62(1/2), 23–27. Retrieved 2022-09-14, from <http://www.jstor.org/stable/520449>
- Campbell, R. (1975). *Soil slips, debris flows, and rainstorms in the santa monica mountains and vicinity, southern california*. U.S. Government Printing Office. Retrieved from <https://books.google.com/books?id=>

decps6y5fGQC

- Cerato, Yang, H., Xiaodi, Y., Xiaogang, H., & Wassim, T. (2014, 1). Real-time monitoring of slope stability in southeastern oklahoma. *Meteorology and Atmospheric Physics*, 98, 91. Retrieved from [http://192018.eos-intl.net/elibsql16\\_L92018\\_Documents/FHWA-0K-14-06%202241%20Cerato.pdf](http://192018.eos-intl.net/elibsql16_L92018_Documents/FHWA-0K-14-06%202241%20Cerato.pdf)
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321–357.
- Chen, W., Zhang, S., Li, R., & Shahabi, H. (2018). Performance evaluation of the gis-based data mining techniques of best-first decision tree, random forest, and naïve bayes tree for landslide susceptibility modeling. *Science of The Total Environment*, 644, 1006-1018. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0048969718324653> doi: <https://doi.org/10.1016/j.scitotenv.2018.06.389>
- Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC genetics*, 19(1), 1–6.
- Das, R., Biswas, S. K., Devi, D., & Sarma, B. (2020). An oversampling technique by integrating reverse nearest neighbor in smote: Reverse-smote. In *2020 international conference on smart electronics and communication (icosec)* (p. 1239-1244). doi: 10.1109/ICOSEC49089.2020.9215387
- Densmore, A. L., Anderson, R. S., McAdoo, B. G., & Ellis, M. A. (1997). Hillslope evolution by bedrock landslides. *Science*, 275, 369 - 372.
- Edgar, T., & Manz, D. (2017). *Research methods for cyber security*.
- Fedora, M., & Beschta, R. (1989). Storm runoff simulation using an antecedent precipitation index (api) model. *Journal of Hydrology*, 112(1), 121-133. Retrieved from <https://www.sciencedirect.com/science/article/pii/0022169489901844> doi: [https://doi.org/10.1016/0022-1694\(89\)90184-4](https://doi.org/10.1016/0022-1694(89)90184-4)
- Foti, & Bukenhofer. (1999). Ouachita and ozark plateaus provinces. *National Park Service, U.S. Department of the Interior*. Retrieved from <https://www.nps.gov/articles/ouachitaandozarkplateaus.htm>
- Giannecchini, R. (2005). Rainfall triggering soil slips in the southern apuan alps (tuscany, italy). *Advances in Geosciences*, 2, 21–24.
- Glade, T., Crozier, M., & Smith, P. (2000). Applying probability determination to refine landslide-triggering rainfall thresholds using an empirical antecedent daily rainfall model. *Pure and Applied Geophysics*, 157(6), 1059–1079.

- Guzzetti, F., Gariano, S. L., Peruccacci, S., Brunetti, M. T., Marchesini, I., Rossi, M., & Melillo, M. (2020). Geographical landslide early warning systems. *Earth-Science Reviews*, *200*, 102973. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0012825219304635> doi: <https://doi.org/10.1016/j.earscirev.2019.102973>
- Guzzetti, F., Peruccacci, S., Rossi, M., & Stark, C. (2007, 12). Rainfall thresholds for the initiation of landslides in central and southern europe. *Meteorology and Atmospheric Physics*, *98*, 239-267. doi: 10.1007/s00703-007-0262-7
- He, X., Hong, Y., Yu, X., Cerato, A. B., Zhang, X., & Komac, M. (2014). Landslides susceptibility mapping in oklahoma state using gis-based weighted linear combination method. In *Landslide science for a safer geoenvironment* (pp. 371–377). Springer.
- Herrera, K. T. V. F., V.M. (2019). Random forest implementation and optimization for big data analytics on lexisnexisâs high performance computing cluster platform. *Journal of Big Data*, *6*(68). Retrieved from <https://doi.org/10.1186/s40537-019-0232-1>
- Highland, L., & Bobrowsky, P. T. (2008). The landslide handbook: a guide to understanding landslides (p. 129). *Reston: US Geological Survey*.
- Highland, L., Bobrowsky, P. T., et al. (2008). *The landslide handbook: a guide to understanding landslides*. US Geological Survey Reston.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.
- Jacquemart, M., & Tiampo, K. (2021). Leveraging time series analysis of radar coherence and normalized difference vegetation index ratios to characterize pre-failure activity of the mud creek landslide, california. *Natural Hazards and Earth System Sciences*, *21*(2), 629–642. Retrieved from <https://nhess.copernicus.org/articles/21/629/2021/> doi: 10.5194/nhess-21-629-2021
- Jakob, M., & Weatherly, H. (2003). A hydroclimatic threshold for landslide initiation on the north shore mountains of vancouver, british columbia. *Geomorphology*, *54*(3-4), 137–156.
- Jatau, P., Melnikov, V., & Yu, T.-Y. (2021). A machine learning approach for classifying bird and insect radar echoes with s-band polarimetric weather radar. *Journal of Atmospheric and Oceanic Technology*, *38*(10), 1797 - 1812. Retrieved from <https://journals.ametsoc.org/view/journals/atot/38/10/JTECH-D-20-0180.1.xml> doi: 10.1175/JTECH-D-20-0180.1
- Johnson, L., Kenneth Sutherland. (2008, 09). Earth sciences and mineral re-

sources of oklahoma. *Oklahoma department of libraries*.

- Jordanova, G., Gariano, S. L., Melillo, M., Peruccacci, S., Brunetti, M. T., & Jemec Aulflič, M. (2020). Determination of empirical rainfall thresholds for shallow landslides in slovenia using an automatic tool. *Water*, *12*(5), 1449.
- Kausar, N., & Majid, A. (2016). Random forest-based scheme using feature and decision levels information for multi-focus image fusion. *Pattern Analysis and Applications*, *19*(1), 221–236.
- Komac, M. (2005). Rainstorms as a landslide-triggering factor in slovenia. *Geologija*, *48*(2), 263–279.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.
- Larsen, M. C., & Simon, A. (1993). A rainfall intensity-duration threshold for landslides in a humid-tropical environment, puerto rico. *Geografiska Annaler. Series A. Physical Geography*, 13–23.
- Lim, Y. (2022). Ensemble averaging â improve machine learning performance by voting. *towards data science*. Retrieved from <https://towardsdatascience.com/ensemble-averaging-improve-machine-learning-performance-by-voting-246106c753ee>
- Liu, L., Zhang, X., Liu, Y., Zhu, W., & Zhao, B. (2020). An ensemble of multiple boosting methods based on classifier-specific soft voting for intelligent vehicle crash injury severity prediction. In *2020 ieee sixth international conference on big data computing service and applications (bigdataservice)* (p. 17-24). doi: 10.1109/BigDataService49289.2020.00011
- Lu, G. Y., & Wong, D. W. (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Computers & geosciences*, *34*(9), 1044–1055.
- Park, H., Kim, N., & Lee, J. (2014). Parametric models and non-parametric machine learning models for predicting option prices: Empirical comparison study over kospi 200 index options. *Expert Systems with Applications*, *41*(11), 5227-5237. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0957417414000566> doi: <https://doi.org/10.1016/j.eswa.2014.01.032>
- Petley, D. (2012, 10). Global patterns of loss of life from landslides. *Geology*, *40*(10), 927-930. Retrieved from <https://doi.org/10.1130/G33217.1> doi: 10.1130/G33217.1
- Regmi, N. R., & Rasmussen, C. (2018). Predictive mapping of soil-landscape relationships in the arid southwest united states. *CATENA*, *165*, 473-486. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0341816218300766> doi: <https://doi.org/10.1016/j.catena.2018.02.031>

- Regmi, N. R., & Walter, J. I. (2020). Detailed mapping of shallow landslides in eastern oklahoma and western arkansas and potential triggering by oklahoma earthquakes. *Geomorphology*, *366*, 106806. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0169555X19301965> (The Binghamton Geomorphology Symposium: 50 years of Enhancing Geomorphology) doi: <https://doi.org/10.1016/j.geomorph.2019.05.026>
- Rodriguez, C., Bommer, J., & Chandler, R. (1999). Earthquake-induced landslides: 1980–1997. *Soil Dynamics and Earthquake Engineering*, *18*(5), 325–346.
- Song, C., Ristenpart, T., & Shmatikov, V. (2017). Machine learning models that remember too much. In *Proceedings of the 2017 acm sigsac conference on computer and communications security* (p. 587â601). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/3133956.3134077> doi: 10.1145/3133956.3134077
- Stoeser, D. B. (2003). A digital geologic map database for the state of oklahoma. *US Geological Survey Open-File Report*, *3*, 247.
- Sugiyama, M. (2015). *Introduction to statistical machine learning*. Elsevier Science. Retrieved from <https://books.google.com/books?id=4YsdCAAQBAJ>
- Taylor, N. R. M., Fred A. (1975). *Distribution and cost of landslides that have damaged manmade structures during the rainy season of 1972-1973 in the san francisco bay region, california* (Tech. Rep.). US Geological Survey.
- Tyrl, J., Bidwell, T., Masters, R., Elmore, R., & Weir, J. (2007, 09). Oklahoma's native vegetation types. *Oklahoma Cooperative Extension Service Publication No. E-993*.
- Urvoy, M., & Autrusseau, F. (2014). Application of grubbs' test for outliers to the detection of watermarks. In *Ih & mmsec '14*.
- Varouchakis, E. A. (2021). Gaussian transformation methods for spatial data. *Geosciences*, *11*(5). Retrieved from <https://www.mdpi.com/2076-3263/11/5/196> doi: 10.3390/geosciences11050196

**Appendix A**

**Appendix**

Table A.1: Landslide events sourced from news.

<b>Event Date</b>	<b>location (Lat and Long)</b>	<b>Type of Landslide</b>	<b>Trigger Mechanism</b>	<b>link</b>
4/10/2008	<b>36.367848, -95.772719</b>			USGS
5/9/2015	<b>35.50, -93.68</b>	Landslide	Rainfall	5newsonline
5/22/2015	<b>34.52, -95.29</b>	Landslide	Flooding	new9
5/28/2015	<b>35.66, -95.29</b>	Mudslide	Rainfall	MuskogeePhoenix
6/16/2015	<b>35.71, -93.80</b>	hill slide	Rainfall	5newsonline
6/17/2015	<b>35.110, -95.646</b>	mudslide		mcalesternews.com
6/18/2015	<b>34.44, -97.13</b>	Landslide	Rainfall	Kltv
12/26/2015	<b>36.03, -94.92</b>	Rockfall	Rainfall	ODOT
12/29/2015	<b>35.87, -93.91</b>	Mudslide	Rainfall	USGS