A MACHINE LEARNING APPROACH FOR LANDSLIDE SUSCEPTIBILITY MODELING FOR RATHNAPURA DISTRICT, SRI LANKA

Madirawalage Anjali Shanika Perera

(189339R)

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2021

A MACHINE LEARNING APPROACH FOR LANDSLIDE SUSCEPTIBILITY MODELING FOR RATHNAPURA DISTRICT, SRI LANKA

Madirawalage Anjali Shanika Perera

(189339R)

Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2021

DECLARATION OF THE CANDIDATE & SUPERVISOR

M.A.S Perera (189339R) hereby declare that this report was carried out by me under the supervision of Dr. Uthayasanker Thayasivam, in the Department of Computer Science and Engineering, University of Moratuwa. It has not been submitted to any other institution or study program by me for any other purpose.

UOM Verified Signature 14/11/2021

....

M.A.S Perera Date

I, Dr. Uthayasanker Thayasivam, certify that the above declaration made by the candidate is true & this report is forwarded for the purpose of evaluation.

UOM Verified Signature 14/11/2021

....

Dr. Uthayasanker Thayasivam Date

Senior Lecturer,

Department of Computer Science and Engineering,

University of Moratuwa,

Katubadda.

ACKNOWLEDGMENT

I would like to express my heartfelt gratitude to my supervisor, Senior Lecture Dr. Uthyasanker Thayasivam, for providing me with excellent guidance, advice, and time from the beginning of this study, as well as for providing me with exceptional experiences during the work, and for his unwavering encouragement and support in various ways.

I would also like to thank, MSc Coordinator, Dr. Charith Chitraranjan for sparing his valuable time and guide me on this work. And I express my gratitude to the entire senior and junior lecturer panel, as well as the faculty and staff of the University of Moratuwa's Faculty of Engineering, for their guidance, supervision, and critical contribution to this study.

Finally, I would like to express my gratitude to my family and friends for their help in making this study a success.

Abstract

In certain areas of the world, landslides are the most common and recurrent natural hazard, resulting in substantial human deaths and property damage. Landslides are extremely common in Sri Lanka, with landslides affecting approximately 30.7%of the country's land area. As the demand for human growth has increased, landslides have become a major problem in Sri Lanka's mountainous regions. As a result, detecting landslide potential associated with terrain data and remote sensing data is crucial for ensuring the long-term viability of projects while minimizing the risk of landslide disasters.

The aim of this study is to develop a susceptibility map for Sri Lanka using a novel data science approach. This study has been used the Rathnapura district in Sri Lanka as the study area. In this study, five ensemble machine learning algorithms: Random Forest, Bagged Decision Tree, AdaBoost, XGBoost, and Gradient Boost were used for landslide prediction and landslide susceptibility map modeling. Using the K-Means clustering algorithm, the class probability values from ensemble-based machine learning algorithms were used to reclassify the study area into susceptibility levels: Extreme Low (EL), Low (L), Moderate (M), High (H), Very High (VH), and Extreme High (EH). In addition, landslide susceptibility maps were generated using the Frequency Ratio technique. The Landslide Susceptibility Index (LSI) was generated using the Frequency Ratio values. The study area was then categorized into six landslide susceptibility classes based on the LSI value: Extreme Low, Low, Moderate, High, Very High, and Very High.

The F-Score, Accuracy, Precision, and Recall values were used to evaluate the landslide prediction results, while the Landslide Density value was used to evaluate the LSMs. Finally, a web application was developed to visualize landslide susceptibility maps, landslide locations, and landslide conditioning factor maps.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

In recent decades, Natural disasters have risen increasingly, with that landslides becoming one of the most damaging natural disasters, causing severe loss of life and property. It is caused with other natural disasters, such as heavy rainfall and earthquakes that make it hard to predict landslides. Heavy rainfall has caused extreme landslides and floods in fifteen districts across Sri Lanka since May 14, 2017. As of June 2017, the Disaster Management Center recorded around 200 deaths, and approximately 700,000 people exaggerated by the landslide and flood situation [1]. Furthermore, the landslide completely destroyed 2,313 homes and slightly damaged another 12,529 homes [1].

1.1 Landslide

Landslides are a common occurrence that have become one of Sri Lanka's most dangerous natural hazards. In the last five years, landslides have directly impacted over 50,000 people, with thousands more at risk today, especially in the districts of Kaluthara, Kegalle, Ratnapura, Badulla, Kandy, Matara, and Galle. The central hills districts of Nuwara Eliya, Badulla, Kandy, Matale, Kegalle, Ratnapura, and Kalutara, as well as the southern hills districts of Matara, Galle, and Hambanthota, are known as highly landslide susceptible areas in Sri Lanka. Landslides are especially dangerous in an area of 20,000 km2 [2].

The term "landslide" encompasses a wide variety of mass movements and processes involving the downward displacement of masses of rock, earth, or debris due to gravity [3]. Soil slips, mudslides, rock falls, rockslides, rock avalanches and rubble flows are just some of the possibilities [4]. Landslides can also be observed in five different terrain types, according to [5]:

- Areas in the uplands that are prone to earthquakes
- Mountains with a lot of relative relief
- Areas with thick deposits of fine-grained materials
- Areas that receive an intensive rain
- Moderate-relief areas with extreme land degradation

Landslides may also be triggered by natural events such as earthquakes, heavy rains, snowmelt, and volcanic eruption. In humid tropical areas, this phase stands out even more. Human disturbances, such as land-use changes, deforestation, excavation, changes in slope profile, or irrigation, cause some destructive landslides, and human activities are increasingly playing a role in the hazard's causation. [6] [7].

Land planning and management, on the other hand, can be useful in reducing the social and economic damages caused by landslides [8]. As a consequence, assessing and forecasting landslide-prone areas is important. Reviewing previous landslide records in conjunction with landslide-related factors may aid in this mission.

As a result, using remote sensing data and other terrain features to identify landslide prone areas are critical for ensuring long-term growth sustainability while reducing the risk of landslide disasters. Landslide susceptibility mapping can be used to determine the risk of a landslide occurring in a specific location. The likelihood of a landslide occurring in a given region is known as landslide susceptibility [9].

1.2 Landslide Susceptibility Mapping

According to [10], The likelihood of a landslide occurring in a given area as a result of local terrain conditions, rather than the probability of occurrence, which is based on the occurrence of triggering factors, is referred to as landslide susceptibility. Landslide assessments are also used to establish land use laws aimed at reducing the number of human losses and property damage [11].

The monitoring and analysis research has been made easier because people are becoming more conscious of the socioeconomic impacts and pressures of urbanization [12]. There are five types of landslide susceptibility mapping methods available [13]:

1) Analysis of landslide inventories
2) Process-based conceptual models
3) Heuristic methods
4) Direct geomorphologic mapping

5) Statistical methods, including technologies like neural networks, fuzzy logic, and expert systems.

Based on a collection of internal causative factors, **landslide susceptibility mapping** aims to classify places where landslides occur across an area.

1.3 Problem Definition

Floods, landslides, cyclones, droughts, wind storms are some of main causes for natural disasters. Nowadays, it is increasing the occurrence of landslide in the hill countries and urbanized areas. So that it is a tremendous risk for existing people and for existing built environment. Mostly, the landslide areas like Nuwara Eliya, Badulla, Hali-ela and Rathnapura districts are highly susceptible for landslides. And it causes massive amount of continuous and severe landslide problems during rainy seasons.

According to surveys, the majority of the landslides occurred in mountainous areas. So, one main reason for landslide is human intervention and human activities. As a result, identifying landslide-prone areas is very important in land use planning and human protection. Accordingly, the Landslide Susceptibility Maps is an essential requirement for several purposes. Since, the occurrence of landslide is a serious problem for development of a country especially for a developing country like Sri Lanka.

1.4 Research Objectives

The main objective of this research was to construct and evaluate landslide susceptibility using the best data science approaches. It was applied to the dataset which was collected from Rathnapura districts in Sri Lanka.

The following goals were chosen for this research study in order to achieve the above main objective.

1. Explain the relationship between landslide frequency and causative factors using the spatial database that has already been developed for the Rathnapura District in Sri Lanka.

2. Using remote sensing and terrain data, modeling a landslide susceptibility map for Sri Lanka.
3. Assist in reducing the risk of landslides and ensuring proper land use planning for Sri Lanka's future development.

CHAPTER 2: LITERATURE REVIEW

This chapter offers a brief overview of the current state of landslide disaster research and forecasting activities both internationally and in Sri Lanka. Furthermore, this chapter thoroughly examines and compares the major techniques utilized for landslide susceptibility mapping. It highly focuses on data science approaches for landslide prediction and landslide susceptibility mapping. Following that, it explains why landslide susceptibility mapping is crucial for Sri Lanka.

2.1 Introduction

Landslide is an extremely unpredictable natural hazard which arises suddenly provoking enormous casualties to the surrounding. If a system exists to reliably predict the likelihood of a landslide, the number of casualties and property damage may be reduced. The landslide, in contrast is based on a huge number of factors, the relationships of which we are unaware. As a consequence, forecasting landslides with both spatial and temporal precision is not a simple or straightforward task.

Landslide susceptibility map (LSM) is a static method which is used to calculate the probability of a landslide at a specific location. Another dynamic approach is a dynamic landslide prediction framework that can produce temporal warnings based on continuously updated data.

Landslides can be influenced by a variety of factors. As a result, defining the relationship between supporting factors will aid in landslide prediction. This chapter focuses on the literature regarding the landslide disaster in Sri Lanka, landslide prediction systems, landslide susceptibility techniques which are in Sri Lanka as well as in worldwide

2.2 Landslide Disaster Risk Reduction in Sri Lanka

Several studies have been performed on the landslide disaster in Sri Lanka. The NBRO (National Building Research Organization) is the leading research and development institute which was established in 1984 in Sri Lanka. It has now developed into a prosperous professional service provider, with experts from different disciplines working to create a disaster-free built environment for Sri Lanka. So NBRO has carried out lots of research studies related to landslide disaster in Sri Lanka.

According to the NBRO, the landslide risk reduction program in Sri Lanka has a cyclic process as shown in figure 2.1. It mainly includes three stages as (1) pre-disaster stage, (2) response stage and (3) post disaster stage. Risk assessment, mitigation and preparedness are the main activities in the pre-disaster stage. In the response stage rescuing and securing the human life and infrastructure is carried out. Eventually, assessment of the damages and restoration of infrastructure are carried out under post disaster stage [14].

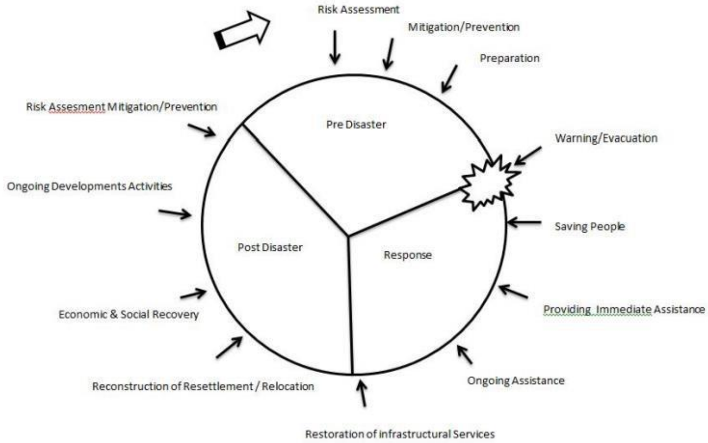


Figure 2.1: Landslide disaster risk reduction process in Sri Lanka

Preparation of landslide hazard zonation maps, building a system of early warning based on automated rain gauges, conducting community awareness and taking structural mitigatory measures are several preparation activities which were carried out in the pre-disaster stage in Sri Lanka. Issuing early warning considering real time rainfall is a kind of prevention activity practiced in Sri Lanka. Eventually, human involvement is highly used for the post disaster stage in rescuing, evacuating, temporary resettling and facilitating activities.

2.3 Landslide prediction systems in Sri Lanka

Recently, Artificial Neural Network (ANN) is widely used in many computer applications. As a result, ANN is used in the prediction of landslides and the application of modern technology to forecast landslides in influential geographical areas. A study [15] was conducted on landslide prediction using Artificial Neural Networks, with an emphasis on the landslides in Sri Lanka's Badulla district. Using existing topographic data such as irrigation, land cover maps, soil, and historical data, a landslide database was developed for this research. Subsequently, this study has identified twelve most prominent factors including external and internal factors which affect for landslide occurrence.

Following that, using ANN, these variables were used to detect the risk of landslides. External factors included data from the metrology department and historical data, while internal factors included current maps [15]. During the training session, this network learns the relationship between landslide factors and their hazard index. They have used three different output classes: landslide occurs, landslide does not occur, and landslide likely to occur. According to the results, the above approach had a high prediction rate of 83% accuracy [15].

The study [16], has developed a landslide prediction model using Decision Tree and Neural network data mining techniques for Badulla and Nuwara Eliya districts in Sri Lanka. According to this study, the Decision Tree model outperformed Neural Network model.

2.4 Landslide Susceptibility Mapping Techniques

A LSM depicts the probability of landslide occurrence in a particular location, ranging from low to high, and is focused primarily on terrain information. There are various factors which are affected for the landslide. These factors are divided into three categories: topographic factors, water-related factors, and human behavior.

It's difficult to express which factors have a greater effect on landslides than others without a thorough review. Recognizing the most important factors and training a model that can establish a connection between them is the most challenging part of creating an LSM. Preferably, an anticipated method should be able to make accurate predictions regardless of the training data set’s nature whether it is biased or balanced. Or else, the model's scalability would be compromised.

2.4.1 Landslide Susceptibility Mapping

The [17] [18] [19] [20] [21] [22] research studies are among methodologies for landslide susceptibility assessment. The current landslide susceptibility mapping methods can be divided into five groups, according to [17] research:

(1) Analysis of landslide inventories
(2) Heuristic methods
(3) Statistical methods like neural networks, fuzzy logic and expert systems
(4) Direct geomorphologic mapping
(5) Process based conceptual models.

Furthermore, several research studies have categorized them into two main types: qualitative (heuristic) approaches and quantitative approaches.

The heuristic approach is based on domain experts' opinions. During this method, experts evaluate the weighting values for landslide causative factors. The probabilistic approach was used to analyze data from remote sensing, observations, and collected data, as well as historical research findings. In the statistical method, combinations of causal factors are calculated statistically, and potential landslides are forecasted using historical evidence.

To investigate landslide vulnerability in Vietnam, researchers [23] used data mining techniques such as support vector machines, decision trees, and nave bayes models. The study used a total of 118 landslide data sets. They've also divided the data into 70% training data and the rest test data at random.

The aim of the [24] was to figure out the most excellent method for landslide susceptibility modeling in data-scare situations. They used 418 landslide inventories and 18 landslide conditioning variables, which they studied. They used SVM, Random Forest, Weight of Evidence, Ridge Regression, and Robust Regression models to construct five susceptibility maps for each experiment. According to the findings, the SVM model had the maximum accuracy (AUC = 0.85). Furthermore, 16 because of the smaller number of sample data used for training, the WoE and RF models were seriously harmed.

For modeling landslide vulnerability in China, the [25] research study also used GIS based ML techniques. For susceptibility modeling, researchers compared four advanced machine learning techniques: Logistic model tree (LMT), Random Forest (RF), Bayes Net (BN), and Radical Basis Function (RBF). For landslide susceptibility modeling, fifteen factors were prepared, and historical data was used to identify 222 landslide locations. Finally, the model was validated with the help of ROC and statistical measurements such as sensitivity, specificity, and accuracy. Among above models, the RF outperforms other methods.

In the [26] research report, various ensemble-based machine learning models such as AdaBoost, LogitBoost, Multiclass Classifier, and Bagging models were used to propose a landslide susceptibility map for South Korea. Figure 2.2 shows how the observed landslide points were fed into the digital elevation model (DEM) data for spatial processing.

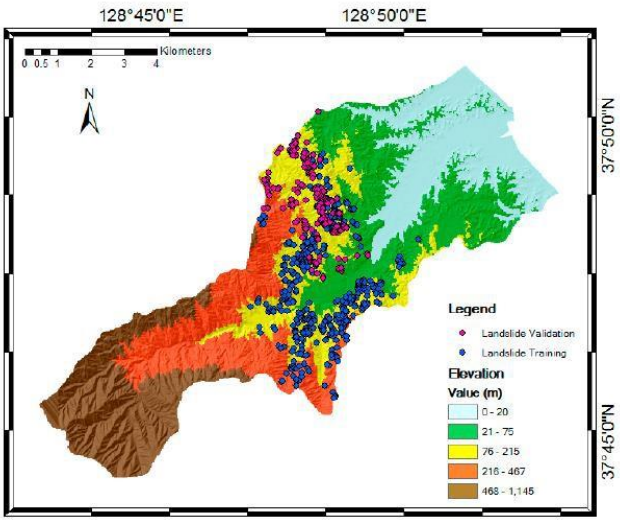


Figure 2.2: Digital elevation model and landslide occurrence in the study area

To create and test the landslide models, the landslide inventory and 20 conditioning factor maps were converted to ASCII format with a 30m cell scale. The positions of all landslides measured with ArcGIS version 10.41 were separated into two groups as 70% for training and 30% for testing. The validation was carried out using the AUC method (Area under Curve). The multiclass classifier system, on the other hand, 17 outperformed the Bagging, LogitBoot, and AdaBoost methods in terms of prediction accuracy

The research [27] used ANN, SVM, and one statistical model, Information Value Model, to evaluate a landslide inventory database of 493 landslides under 16 conditioning factors (IVM). ANN model achieved higher prediction capability than SVM and IVM.

AdaBoost and Alternating Decision Tree were used as an ensemble model for landslide prediction in Malaysia in the study [28]. The model was trained using 152 landslides and 17 different conditioning factors. The AdaBoost model outperforms the ensemble model and the ADTree model in predicting landslide susceptibility, according to the findings.

Specially, recently there were several researches carried out on landslide susceptibility mapping in China [29] [30] [31] [32] [33]. For landslide susceptibility mapping in China, the research [29] has used several machine learning algorithms such as Logistic Regression (LR), SVM, Random Forest (RF), Gradient Boosting Machine, and Multilayer Perceptron (MLP) with GIS techniques. This analysis included 16 conditioning factors and 502 landslide points. The RF algorithm was found to outperform the other methods. In the study [30], a neural network model based on semi-supervised learning is proposed. For the comparison, supervised machine learning models such as SVM, Logistic Regression (LR), and Deep Neural Network were used, and the results showed that semi-supervised deep learning outperformed all other models.

For landslide susceptibility mapping, the study [32] used Random Forest (RF) and Frequency Ratio (FR) models. According to the study's findings, the FR model is better suited for landslide susceptibility mapping in their research area. Furthermore, for landslide susceptibility prediction, the study [33] used cluster analysis, probabilistic approaches, and artificial neural networks (ANN). The weight for each factor was calculated using ANN, and the ratings of relative importance of each class belonging to each factor were calculated using the probabilistic method. Finally, the Landslide Susceptibility Index (LSI) value for each pixel was calculated using the ratings and weights. These LSI values were then used to model landslide susceptibility.

The research [34] used Random Forest (RF) and Boosted Regression Tree (BRT) models to evaluate and compare landslide vulnerability in South Korea. 14 landslide conditioning factors were selected using the Information Gain method. According to the findings, the RF model outperforms the BRT model.

2.4.2 Landslide Susceptibility Mapping in Sri Lanka

According to a [35] study, Sri Lanka has recently begun to use the NBRO-developed landslide hazard zonation (LHZ) map, which is focused on hydrological, geomorphologic, and geological factors.

There were few researches have focus on Landslide Susceptibility Mapping. The [36] research study has proposed a landslide susceptibility maps to reduce the damages which occurs from landslides. And this study has compared statistical models, the geographically weighted logistic regression (GWLR), and logistic regression (LR) models. As in other aforementioned studies, this study also has considered nine relevant causative factors and collected the datasets of Badulla district over 1986 to 2014 years. The GWLR model, however, was chosen as the best of all the other methods considered in this analysis. Finally, the ArcGIS standard deviation method was used to identify the probabilities that were obtained for each pixel using models as Low, Medium, and High. And for each of the three groups, a landslide susceptibility map was developed. However, this [36] study has only considered the terrain variables as landslide factors.

The [37] research study used a logistic regression model for LSM modeling in Badulla District. All other information was collected from the NBRO Statistics in Sri Lanka. The likelihood of a landslide occurring using this approach would be equivalent to observed landslides. To calculate the contribution of each factor to the frequency of landslides, the weights of each factor were calculated. After combining coefficients of each, the likelihood of a landslide occurring was calculated, and the LSM was created using the equation 2.1 and 2.2. Here, “u” is assumed to be a linear combination of independent variables, and p is the model's output, which is the likelihood of a landslide. And the coefficients $\beta 1$, $\beta 2$, $\beta 3$ correspond to each component, indicating their contribution to landslide susceptibility.

$$\boxed{\mathbf{P_r = e^u/(1+e^u)}} \tag{2.1}$$

$$\boxed{u = \beta \mathbf{0} + \beta \mathbf{1} x \mathbf{1} + \beta \mathbf{2} x \mathbf{2} + \beta \mathbf{3} x \mathbf{3}} \tag{2.2}$$

The resulted map was categorized into four groups as low, moderate, high and extremely high susceptibility classes. The validation experiment was carried out by comparing the Logistic Regression findings to those from an external dataset. According to the study, approximately 20% of the district is highly vulnerable to landslides, 40% is moderately vulnerable to landslides, and 40% is low susceptible to landslides.

It proposed a generalized ensemble-based machine learning method for susceptibility mapping for the Sri Lankan district of Rathnapura in [38]. Random Forest, XGBoost, and rotation forest were among the machine learning algorithms used. As a consequence of the study's findings, the XGBoost approach outperforms the other approaches. As a result, the Random Forest classifier results were used to build a landslide susceptibility map. There they have created a prediction matrix based on random forest classifier and that prediction matrix was used to create a susceptibility map using ArcGIS software.

Moreover, the study [39], has done a study on landslide susceptibility in Kaluthara district, Sri Lanka with the use of 12 landslide conditioning factors: slope, aspect, geology, hydrology, soil type, soil thickness, land use, landform, sediment transport index, topographic wetness index, stream power index and rainfall. The Random Forest algorithm was used to model landslide susceptibility prediction using the frequency ratio and information gain ratio.

2.5 Summary

As mentioned above, these are some of the data science-based approaches for landslide prediction and susceptibility mapping that have been suggested in Sri Lanka and other countries. In addition, table 2.1 compares among the aforementioned data science-based Landslide Prediction and Landslide Susceptibility Mapping approaches.

According to the table 2.1, most of the researches have used different types of machine learning algorithms and statistical methods for this purpose. But most of the researches didn’t consider more than 3 ensemble machine learning algorithms together for the landslide prediction and for landslide susceptibility mapping. And almost all researches have used ArcGIS software to generate LSMs. In other words, for they didn’t use Machine learning approaches for doing the susceptibility class classifications.

Table 2.1: Comparison between Data Science based Landslide Susceptibility approaches

Landslide Prediction and Landslide Susceptibility Approaches	**Dataset used**	**Algorithms used**	**Accuracy Achieved**
[15]	**Factors including External and Internal factors** • Rainfall • Geology • Aspect • Slope • Curvature • Soil Material • Land use **External Factors** • Data from Metrology department • Historical data **Internal Factors** • Existing maps were used to capture internal factors	• ANN (To recognize the possibility to occur landslides)	• 83% accuracy
[23]	• 118 Landslide data ○ 70% - Training data ○ 30% - Test data • **Factors** ○ Slope Aspect ○ Slope angle ○ Relief amplitude ○ Lithology ○ Distance to faults ○ Distance to roads ○ Land use ○ Rainfalls	• SVM • Decision Tree • Naïve Bayes	• High Accuracy (using SVM)
[24]	• 418 landslide inventories • 18 landslide conditioning factors	• SVM • Random Forest • Robust Regression • Ridge Regression	• AUC = 0.85 (Hight Accuracy using SVM)

		• Weight of Evidence	
[25]	• 15 factors • 222 landslide locations	• Logistic regression • Bayes Net • Radical basis Function classifier • Random Forest	• RF model (Highest sensitivity, specificity and accuracy)
[27]	• 493 landslides • 16 conditioning factors	• SVM • ANN • Informatio n Value Model (IVM).	• ANN (higher prediction capability)
[28]	• 152 landslides • 17 conditioning factors	• AdaBoost and Alternating Decision Tree together as ensemble model	• AdaBoost outperform
[32]	• 987 historical landslides • 5 primary conditioning factors • 22 secondary conditioning factors	• Frequency Ratio (FR) • Random Forest (RF)	• FR is more suitable for landslide susceptibili ty mapping
[34]	• 140 landslide locations • 14 conditioning factors	• Random Forest (RF) • Boosted Regression Tree (BRT)	• RF outperform BRT model

[35]	• 762 landslide records based on reports and aerial photographs • 20 landslide condition factors ○ Forest map ○ Soil map ○ Hydrological factors ○ Topographic factors ○ Geological map	• AdaBoost • LogitBoost • Multiclass Classifier • Bagging Model	
[36]	• 9 causative factors ○ Consider only terrain variables • Data set of Badulla district over 1986 to 2014 years	• Logistic Regression • Geographic ally weighted logistic regression (GWLR)	• 73.5% (For GWLR model) • 64.2% (For Logistic Regressio n Model)
[37]	• Causative factors ○ Land cover (Obtained from geological maps) ○ Slope ○ Aspect ○ Lithology ○ Distance from roads ○ Distance from rivers	• Logistic Regression	• 76% Accuracy
[38]	• 11 causative factors	• Random Forest • XGBoost • Rotation Forest	• Random Forest (Highest accuracy)
[40]	• Identified locations using Aerial photographs and field surveys	• Logistic regression • Frequency ratio •	• 80% (In Logistic Regressio n) • 78.4% (using Frequency Ratio)
[41]	• Satellite images • Historical landslide records	• Hybrid approach of bivariate weights of	• EoE-RF (Highest accuracy)

		evidence (WoE) and Multivariate logistic regression (Woe-LR) • Random Forest (WoE-RF)	
[42]	• 255 landslide polygons • 8 factors ○ Soil ○ Aspect ○ Elevation ○ Land cover ○ Soil type ○ Lithology ○ Distance to fault • Distance to river	• BE-LMtree (Hybridization of Bagging Ensemble and Logistic Mrlr trees) •	• High accuracy (using BE-LMTree model)

CHAPTER 3: STUDY AREA

Sri Lanka is an island in the Indian Ocean and has a total area of 65 610km2, with 64 740km^2 of land and 870km^2 of water. According to estimates, Sri Lanka is one of the world's most disaster-prone nations.

Sri Lanka is divided into 9 provinces by 25 districts. Figure 3.1 shows that the most of landslides tend to occur only in the Southern, Uva, and Central provinces, especially in districts: Rathnapura, Nuwara Eliya, Badulla, and Kegalle, which are Sri Lanka's most landslide-prone districts. Figure 3.2 shows that the Rathnapura district, in comparison to other districts in Sri Lanka, is more vulnerable to landslides. As a result, the focus of this research is on Sri Lanka's Rathnapura districts.

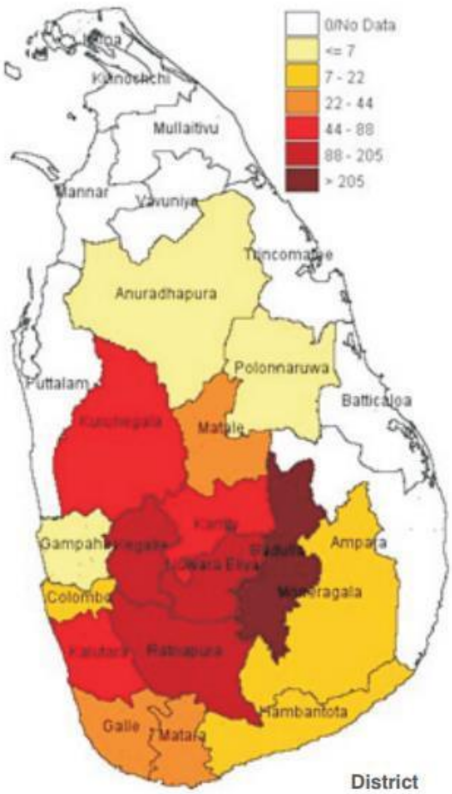


Figure 3.1: Spatial Distribution of Landslides [43]

3.1 Study Area: Rathnapura district

Ratnapura is a district in Sri Lanka's Sabaragamuwa Province. It is situated between the latitudes of 6.53^0 and 6.55^0, and the longitudes of 80.33^0 and 80.48^0, in a mountain range with steep slopes. Even during the driest months, Rathnapura receives a large amount of rainfall. Rathnapura district has an average annual temperature of 27.1°C and an average rainfall of 3679mm [44]. Figure 3.2 depicts the landslide maps of the Rathnapura district, according to Sri Lanka's National Building Research Organization (NBRO).

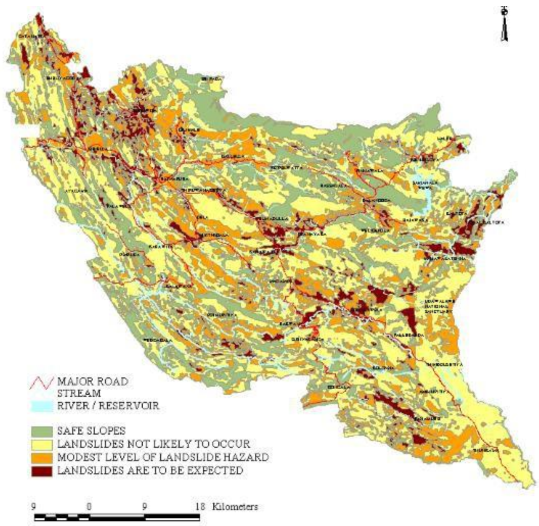


Figure 3.2: Landslide Map in Rathnapura District

Following are the different types of hazards that have been observed in the study area.

- Cutting Failure
- Potential Landslide
- Rockfall + Debris flow
- Old landslide
- Slope failure
- Debris flow
- Landslide cum rack fall
- Rockfall

CHAPTER 4: EXPERIMENTS

This chapter discusses about the dataset and the experiments done on the study.

4.1 Experiments on Dataset

The landslide dataset was collected from different data sources. The majority of the data was gathered by surveying and the development of maps from DEM. All of the surveys were carried out by the NBRO (National Building Research Organization of Sri Lanka), and the results were shown in GIS polygon maps. Overburden, slope, landform, hydrology, land use, and historical landslide locations are among the maps included. The US Geological Survey produced a DEM (Digital Elevation Map). Furthermore, the Vegetation index map was generated using the US Geological Survey's satellite images and the normalized differential vegetation index (NDVI).

By analyzing the variation of slope in the region, the DEM was used to establish the area's stream network. The distance between a cell and the closest stream was then measured. The annual rainfall data map for the Ratnapura area was generated by aggregating precipitation from the Dark Sky API [15] on a regular basis throughout the year.

Then a spatial database has been created using GIS maps for each factor such as Rainfall map, elevation map, aspect map, etc. The maps were then converted to ArcGIS raster maps with a 10m x 10m cell resolution. Each layer reflected a single factor's spatial distribution. To model the study field, each cell value was used to create a matrix, with each column representing a factor. Machine learning models were also trained using this matrix. By definition, the observations were made on a cell-by-cell basis.

4.2 Causative factors of landslide

Identification of the most important landslide conditioning factors is highly important for having better landslide prediction and accurate landslide susceptibility map modeling. Ten landslide conditioning factors were chosen for this analysis

(APPENDIX A) based on the feature selection method and previous studies, including slope, rainfall, land use, distance-to-water-streams, elevation, drainage density, catchment, vegetation, overburden, and landform.

The derived feature maps for the selected study area are shown in APPENDIX A. Table 4.1 shows the origins of the landslide conditioning factors that were used in the previous research [38].

Table 4.1: Landslide conditioning factors with their corresponding sources

Factor	**Source**
Distance to water streams	Derived map from DEM
Annual rainfall	Dark Sky API
Vegetation Index	Derived from Land sat images
Elevation	U.S Geological Survey
Overburden	NBRO Sri Lanka
Aspect	Derived map from DEM
Curvature	Derived map from DEM
Water Catchment area	NBRO Sri Lanka
Drainage Density	NBRO Sri Lanka
Landform	NBRO Sri Lanka
Land use	NBRO Sri Lanka

The following are brief explanations of the collected causative features that were used in this analysis.

4.2.1 Slope

Due to soil instability and stresses, the slope is considered one of the important factors for the occurrence of landslides with sharp slopes. The slope angle was divided into five groups in this analysis, as shown in table 4.2.

4.2.2 Land use

The use of land resources is portrayed on a land-use map. For example, land in rural areas can be used for agriculture, forestry, or water bodies at the same time as land in urban areas is used for industrial purposes or housing. Land use feature was divided into three classes in this study.

4.2.3 Water Catchment

This contains information on water catchment areas. This is the area where the water is collected by the natural landscape. This was divided into four categories, ranging from low to high catchment.

4.2.4 Landform

Landform is about the natural characteristics of the soil surface. This falls under the heading of topographical characteristics. This is divided into four groups, ranging from low to high.

4.2.5 Vegetation

The Normalized Differential Vegetation Index (NDVI) was calculated on land sat images given by the United States Geological Survey. The vegetation index is a metric for assessing the health of live plants in a given region. It also has a major effect on the occurrence of landslides. This study was divided into eight different categories.

4.2.6 Elevation

The elevation distribution of an area is defined by a digital elevation model (DEM). Landslides are more likely to occur at high elevations. The US Geological Survey gathered the elevation data. This has been divided into five sub classes.

4.2.7 Distance to water streams

The likelihood of a landslide was boosted by nearby streams, roads, and other structures. Using ArcGIS, this was divided into eight categories.

4.2.8 Overburden

Overburden [45] refers to the substance that lies over an area that lends itself to economic exploitation, such as the rock, soil, and ecosystem that lies above a coal seam or ore body. This has been divided into 18 categories.

4.2.9 Drainage Density

According to previous research [45] [46], raising the density of the drainage network causes an increase in the number of landslides. There are five distinct classes of drainage density.

4.2.10 Annual Rainfall

The most common cause of landslide occurrence is heavy rainfall. The Rathnapura region's annual rainfall data map was generated by accumulating precipitation from the Dark Sky API [47] on a regular basis throughout the year. There are 13 different categories of annual rainfall.

Table 4.2 represents the classification of the selected conditioning factors for this research. It includes the factors that cause landslides, the number of classes, and the class ranges.

Table 4.2: Classification of causative factors

Factor	Classification
Annual rainfall	(1) 0.021-0.0213 (2) 0.0213-0.0216 (3) 0.0216-0.0218 (4) 0.0218-0.0221 (5) 0.0221-0.0224 (6) 0.0224-0.0227 (7) 0.0227-0.0230 (8) 0.0230-0.0234 (9) 0.0234-0.0237 (10) 0.0237-0.0240 (11) 0.0240-0.0242 (12) 0.0242-0.0244 (13) 0.0244-0.0246
Land use	(1) 1-12 (2) 12-30 (3) 30-255
Distance to water streams	(1) 0-395.5551 (2) 395.5551-824.0731 (3) 824.0731-1252.5911 (4) 1252.5911-1697.5906 (5) 1697.5906-2142.5901 (6) 2142.5901-2637.0340 (7) 2637.0340-3197.4037 (8) 3197.4037-4202.7730
Elevation	(1) 81-221 (2) 221-301 (3) 301-404 (4) 404-545 (5) 545-65535
Water catchment area	(1) 0-0.07 (2) 0.07-0.2 (3) 0.2-0.5 (4) >0.5
Drainage density	(1) 0.00004-2.3605 (2) 2.3605-5.9012 (3) 5.9012-9.9139 (4) 9.9139-15.8151 (5) 15.8151-60.1916
Vegetation	(1) 0.0188-0.0923 (2) 0.0923-0.1503 (3) 0.1503-0.2063 (4) 0.2063-0.2604 (5) 0.2604-0.3126 (6) 0.3126-0.3590 (7) 0.3590-0.3996 (8) 0.3996-0.5117
Slope	(1) 1 (2) 2 (3) 3 (4) 4 (5) 5-255
Overburden	(1) 1 (2) 2 (3) 3 (4) 4 (5) 5 (6) 6 (7) 7 (8) 8 (9) 9 (10) 10 (11) 11 (12) 12 (13) 13 (14) 14 (15) 15 (16) 16 (17) 17 (18) 18
Landform	(1) 1-7 (2) 7-11 (3) 11-15 (4) 15-255

4.3 Preparation of the training and validation datasets

Landslide dataset contains binary classification where a landslide's pixel is divided into two categories ("landslide" and "non-landslide"). Pixels that are part of a landslide are given the value "1," and those that aren't are given the value "0". The Rathnapura study area consisted of 1,200,000 position points or pixels, 5369 of which were landslide pixels and the rest were non-landslide pixels. However, 70% of the dataset was chosen at random to be used for training, and 30% was chosen at random to be used for validation.

4.4 Experiments on existing methodologies

As mentioned in the literature review section also there are lots of researches were carried out on landslide prediction and landslide susceptibility map modeling. According to the literature most of the researches have used normal machine learning methods such as Linear Regression, SVM, Naïve Bayes and etc. Meanwhile some of the researches have used ensemble machine learning models as Random Forest, XGBoost and etc. Somehow none of them have used more than three ensemble machine learning algorithms together for landslide prediction. So, it one approach that I focused in this research.

When it comes to landslide susceptibility modeling, all most all the researches have used ArcGIS software for landslide susceptibility map modeling while using machine learning models. So that this research tried to focus on using a semi supervised machine learning method for landslide susceptibility map modeling. And since some of the researches have used statistical methods such as Frequency Ratio method for landslide susceptibility map modeling, this research also tried to use Frequency Ratio method newly for Sri Lanka. And it was used for landslide susceptibility map comparison with machine learning model results.

And importantly there is no landslide prediction tool for Sri Lankan authorities to check the severity of landslides. So nicely this research was tried to introduce a user-friendly web application for visualizing landslide data, LSMs and landslide causative factor data.

CHAPTER 5: METHODOLOGY

This study considered establishing a semi-supervised machine learning-based susceptibility map for Sri Lanka. Meanwhile, a traditional statistical Frequency Ratio (FR) method was developed for comparison. Mainly the most important ten landslide conditioning factors were selected for this study. The flow chart which was used in this study is shown in figure 5.1. Moreover, the methodological framework of this study mainly includes the following sub-sections (figure 5.1).

1. Landside causative factor selection.
2. Landslide Prediction using Ensemble Machine Learning methods.
3. Landslide Susceptibility Modeling using semi-supervised machine learning model and using Frequency Ratio model.
4. Landslide prediction evaluation using precision, recall, F-score, and accuracy and Landslide Susceptibility Map Validation using Landslide Density value.
5. Verification and comparison of two Landslide Susceptibility models.

Along with that, the data preparation was handled initially. Then the landslide causative factor selection was carried out. Afterward, the landslide prediction was carried out followed by susceptibility map construction using Ensemble-based Machine Learning methods such as Bagging and Boosting methods. Ensemble-based methods were used for this purpose, as it combines several base models to produce one optimal model. And it will decrease the variance and bias.

Model validation is one of the critical phases of data science-based studies. It's crucial to check whether the model is correctly tailored to the training data without overfitting or under-fitting. According to the prior researches, the dataset has been divided into two parts as 70% and 30% from the whole dataset. 70% of landslides data was used for training purposes and 30% of landslides data was used for validation purposes. And accuracy, f score, precision, and recall will be used as measures for the landslide prediction model validation while Landslide Density value was used for the Landslide Susceptibility Model validation.

Several programming languages and software tools were used in this research. Mainly python language was used for machine learning model building and Frequency Ratio method implementation. In addition to that, ArcGIS 10.4 software tool was used for susceptibility classification in the Frequency Ratio model. Furthermore, Angular, NodeJS, and CesiumJS were used to implement the landslide visualization web application.

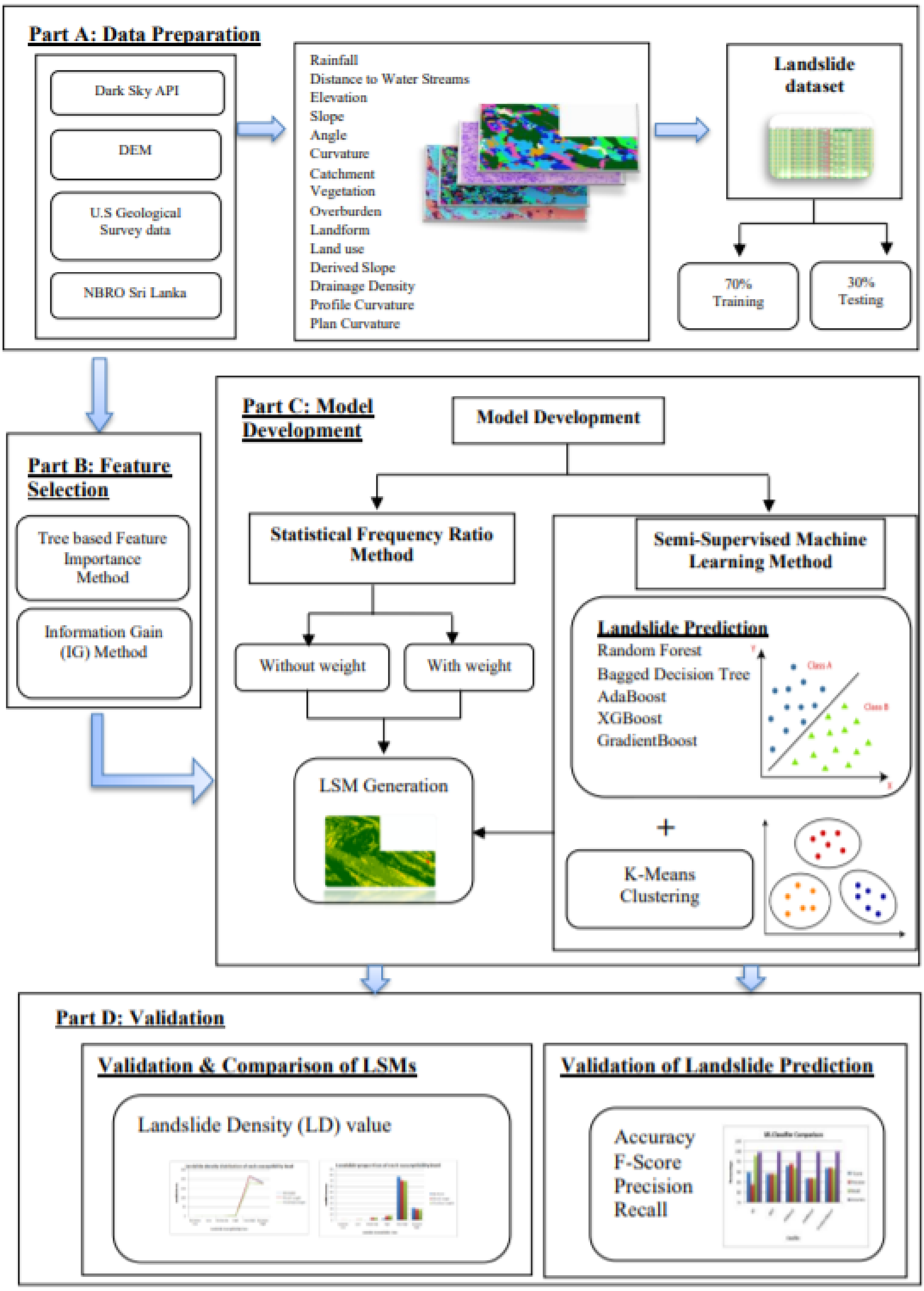


Figure 5.1: Methodological diagram for Landslide Susceptibility map construction

5.1 Landslide Causative factor selection

When using machine learning models to predict landslide susceptibility, the results are heavily reliant on the input features and their accuracy [23]. Selected characteristics may or may not play a role in the occurrence of a landslide. Features with noise, also known as low-quality features, will damage the eligible model's predictive abilities.

To achieve higher precision, features with a high impact on landslide incidence should be identified, and features with a very low impact should be omitted from the training set of features. In this research, several feature important methods such as the tree-based feature importance method, and information gain feature selection methods were used to measure the predictive capability of conditioning factors. These methods are important for identifying the most important factors and improving a machine learning model's overall classification accuracy.

5.1.1 Tree based feature important methods

Both Random Forest Feature Importance method and Extra Trees can be used to estimate the importance of features. However, this “ExtraTreeClassifier” learning method is based on decision trees. This Feature importance values assigns a score to each dataset factor. The factor is more significant or related to the output variable the higher the score. In Python, Tree-based Classifiers have an inbuilt class called Feature Importance. Here the “ExtraTreeClassifier” from the "sklearn" library in python has been used for extracting the most important 10 features out of all the features. The parameters used in this "ExtraTreeClassifier" are listed in Table 5.1.

Table 5.1: ExtraTreeClassifier parameters

Parameter	Parameter description	Parameter value
boostrap	Whether bootstrap samples are used when building trees. If False, the whole dataset is used to build each tree.	false
criterion	The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain.	gini
n_estimators	The number of trees in the forest.	100
class_weight	The “balanced” mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data.	balanced

5.1.2 Information Gain (IG) factor selection method

One of the most effective feature selection algorithms is the information Gain (IG) feature selection process. It splits the dataset according to the value of a random variable and calculates the reduction in entropy or surprise. Here the “feature_selection.mutual_info_classif” from the "sklearn" library in python has been used for extracting the most important 10 features out of all the features. The parameters used in this method are listed in Table 5.2.

Table 5.2: Information gain feature selection classifier parameters

Parameter	Parameter description	Parameter value
x	Feature matrix	Input array
y	Target vector	Output array
discrete_featur es	If bool, then determines whether to consider all features discrete or continuous. If array, then it should be either a boolean mask with shape (n_features,) or array with indices of discrete features. If ‘auto’, it is assigned to False for dense X and to True for sparse X.	auto
n_neighbors	Number of neighbors to use for MI estimation for continuous variables	3

copy	Whether to make a copy of the given data. If set to False, the initial data will be overwritten.	true
random_state	The seed of the pseudo random number generator for adding small noise to continuous variables in order to remove repeated values.	none

5.2 Re-Sampling Techniques for Class Imbalance Issue

In Machine Learning, particularly in classification problems, **class imbalance** is a common problem. There is a class imbalance problem when one class's observation is higher than the other classes' observation. Class imbalance can hamper the accuracy of machine learning models.

If the dataset is unbalanced, predicting the majority class has a reasonable chance of achieving high accuracy. However, it fails to capture the majority class. In other words, accuracy is not an appropriate measure for a study which is having imbalanced datasets. The landslide dataset which is used in this study is highly imbalanced since there are very few landslide locations when compared with the non-landslide locations (Figure 5.2).

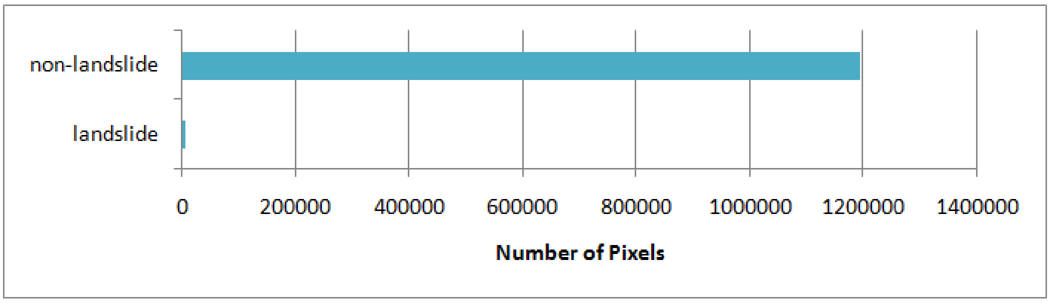


Figure 5.2: Landslide distribution

Therefore, before using the imbalance dataset it is important to take appropriate actions for this class imbalance problem. "Re-sampling" [51] is a commonly used method for working with highly imbalanced datasets. It entails a variety of methods, including eliminating samples from the majority class (under-sampling) and adding more data to the minority class (over-sampling). The following are a couple of these processes.

- Random under-sampling (down-sample the majority class)
- Random over-sampling (up-sample the minority class)
- Synthetic minority oversampling technique (SMOTE)
- Change the performance matrix
- Penalize algorithms (Cost-Sensitive Training)
- Change the algorithm or Use Tree-Based Algorithms

Both “Penalize Algorithms” and “Change the Algorithms methods” techniques were used in this study for re-sampling the landslide. Accordingly, a detailed description of those two techniques is given below.

5.2.1 Penalize Algorithms (Cost-sensitive Training)

For the minority community, this approach raises the cost of classification errors. Penalized-SVM is one of the most widely used algorithms. We may use the class_weight='balanced' claim during the training process to penalize errors in the minority class. If we want to allow probability estimates for the algorithm, we can use the argument probability=True.

5.2.2 Change the Algorithms or use Tree-based Algorithms

While working with Machine Learning related tasks or problems, it is an outstanding way to use a variety of algorithms, most of the time it may be beneficial with class imbalance datasets. Especially, decision tree algorithms frequently work well on imbalanced datasets. Furthermore, modern tree ensembles (Random Forests, Gradient Boosted Trees, and so on) almost always outperform single decision trees and can be used to fix the problem of class imbalance.

So, within this study aforementioned two techniques have been carried out as a solution for the class imbalanced issue in the landslide dataset.

5.3 Landslide Prediction Modeling

One of the primary goals of machine learning is to build a model which represents the incompletely understood or most complex relationships between the target variables and the data [48]. “Supervised” and “unsupervised” machine learning algorithms are the two main types of machine learning algorithms.

The main purpose of supervised algorithms is to construct a machine learning model that connects known inputs and unknown outputs. As a result, the performance values for a new data set can be predicted using previously learned relationships from labeled training data [49]. Classification and regression algorithms are the two most popular types of supervised learning algorithms. The proposed output of the classification method is a semantic class or a name. So, for the problem of identifying potential landslides, the classification will label each pixel of a map as "landslide" or "non-landslide". In contrast, regression aims to predict a continuous variable.

However, to improve performance, some powerful learners have evolved. Those are called “ensemble learning”. Bagging, which reduces uncertainty, boosting, which reduces bias, and stacking, which increases prediction, are the key ensemble learning methods. [50] [51].

In contrast, unsupervised learning methods were used to discover patterns in unlabeled results. Clustering commonly used unsupervised algorithm, which groups data based on similarities. Moreover, another technique is dimensionality reduction and it reduces the variance in the dataset and removes outliers.

Among these techniques, classification and clustering techniques were used for landslide prediction and landslide susceptibility map modeling. Mainly, five ensemble machine learning algorithms which include both bagging and boosting algorithms were used for landslide prediction and landslide susceptibility in this study: Random Forest, Bagged Decision Tree, AdaBoost, XGBoost, and Gradient Boost algorithms. And as a clustering method, K means clustering was used for clustering the resultant classification outputs. Those algorithms are briefly described in the below section.

5.3.1 Random Forest Algorithm

It is an ensemble decision tree classifier which can be used for classification and regression which consisting of large number of individual decision trees [52] [53]. The

RF operates as an ensemble by generating uncorrelated decision trees to produce more accurate prediction results. Each RF tree is generated by re-sampling a random subset of the variable at each [53]. The majority vote of all trees decides the class relationship ("landslide" or "non-landslide") [54].

Furthermore, landslide susceptibility was assessed using "RandomforestClassifer" [55] in "scikit-learn‖ machine learning library" Because of its high performance in landslide susceptibility assessment, the RF algorithm has been widely and effectively used for landslide prediction and landslide susceptibility map modeling all over the world.

5.3.2 Bagged Decision Tree (BDT) Algorithm

Both Random Forest and Bagged Decision Trees [56] are bagging ensemble algorithms. Bagging is a simple ensemble-based algorithm that can be applied to treebased algorithms to increase prediction accuracy. Bagging ensemble methods are useful for algorithms with high variance. One of the best algorithms are the decision tree algorithm. However, BDT is rarely used for landslide prediction.

Usually, bagging ensemble approaches work well for high-variance algorithms, the best of which is the decision tree algorithm. In this analysis, the bagged decision tree ensemble model was developed in the landslide dataset using the sklearn BaggingClassifier function and DecisionTreeClasifier (a classification and regression trees algorithm).

5.3.3 Extreme Gradient Boosting (XGBoost) Algorithm

It is known as one of the best performing supervised learning algorithms and it is an implementation of Gradient Boosting Machines. It can be used to solve both regression and classification problems. Most of the data scientists are using this algorithm as it has higher execution speed when it compares with its core computation. And in the boosting process of XGBoost, it runs cross validation at each iteration.

As, XGBoosts algorithm is having excellent performance, it has become one of the greatest algorithms among all the other algorithms [57]. Moreover, XGBoost algorithm has been used in a few landslide susceptibility-related studies.

However, the most critical part of XGBoost modeling is the basic training phase. XGBoost contains several hyper parameters like learning rate, tree depth and subsample [58]. However, the GridSearchCV function in the "scikit-learn” machine learning library of the Python 3.5 programming language was used to build the XGBoost algorithm for this study. A total of 1000 trees were used, with the learning rate set to 0.1 to prevent over fitting and the tree's maximum depth set to 3.

5.3.4 AdaBoost Algorithm

Freund and Schapire introduced AdaBoost (“Adaptive Booting”) as also an ensemble based algorithm in 1997 [59]. It is a common boosting algorithm. It iteratively creates individual classifiers. Furthermore, AdaBoost selected training samples using an adaptive re-sampling technique. This assigns a weight value for dataset in each iteration and in next iteration it assigns a new weight for misclassified once.

The weighted average of all classifier base models yielded the final one. The AdaBoost model is developed using the Python programming language in this study. This approach increases efficiency significantly, but it can also lead to over fitting.

In terms of generating classifiers on new data, boosting outperforms bagging. On the other hand, it can sometimes fail in practice by generating a classifier with a lower success rate than an individual classifier created from the same database. According to AdaBoost, the joint classifiers have a major impact on the results.

5.3.5 Gradient Boosted Tree (GBRT) Algorithm

GBRT is another ensemble algorithm that uses a combination of boosting and regression techniques to increase model accuracy and reduce variance. The GBRT algorithm, like the RF method, uses a random subset of variables to identify different decision trees in order to improve prediction accuracy. RF employs the bagging process, which ensures that each event has an equal chance of being chosen in subsequent samples. The GBRT, on the other hand, employs the boosting process, which involves weighting the input data in subsequent trees [60].

In a GBRT algorithm, specially, two parameters: tree complexity and learning rate are needed to be defined. The number of splits in each tree is determined by tree complexity, while the learning rate determines the number of trees to be created. This research used the GBRT technique to equate it to other ensemble algorithms and to model landslide susceptibility.

Here, the Gradient Boosted Tree model was developed using the GradientBoostingClassifier class in sklearn library. There, it used maximum number of estimators as 20, max_features as 2 and max_depth as 2 within this model.

So, the table 5.3 summarizes the details on five different ensemble classifiers. It clearly indicates a description about the classifier, the parameters and the parameter values which were considered within this study.

Table 5.3: Classifier Parameters

Classifier	**Description**	**Parameters**
Random Forest	Using RandomForestClassifier class in sklearn.ensemble library	random_state: 0, n_estimators: 20, class_weight: balanced, boostrap: True, criterion: ‘gini’
Bagged Decision Tree	Using BaggingClassifier class in sklearn.ensemble library	base_estimator: DecisionTreeClassifier (class_weight: ‘balanced’), n_estimators: 150, random_state: 123
XGBoost	Using XGBClassifier from xgboost. Here xgboost.XGBClassifier is a scikit-learn API compatible class for classification.	eta: 0.1, max_depth: 3, random_state: 42, num_parallel_tree: 1000, class_weight: ‘balanced’
AdaBoost	Using AdaBoostClassifier class in sklearn.ensemble library	n_estimators: 30, random_state: 7

GradientBoost	Using GradientBoostingClassifier class in sklearn.ensemble library	n_estimators: 20, max_features: 2, max_depth: 2, random_state: 0

5.4 Landslide Susceptibility Modeling

This section describes the landslide susceptibility modeling which were carried out using a semi supervised machine learning model and using frequency ratio method.

5.4.1 Landslide Susceptibility Modeling using a Semi-supervised machine learning model

Landslide occurrences are a probability, according to LSM. Each cell of the region is classified into several risk levels in LSM modeling. In this study, the initial landslide dataset consists of binary classification data where each cell of the study area was categorized into two categories as "landslide" or non-landslide". But in landslide 39 susceptibility mapping, each cell was classified into six risk levels as Extreme Low, Low, Moderate, High, Very High, and Extreme High. Most researchers have used ArcGIS software to reclassify binary classification data into multi-class susceptibility risk levels.

However, this research focuses on landslide susceptibility map modeling using machine learning models. Essentially, five-set of ensemble machine learning algorithms and K-Means clustering algorithm were used for Landslide Susceptibility Map (LSM) building in this research (Figure 5.1). The LSMs were prepared by generating class probabilities for the "landslide" class using ensemble machine learning algorithms and reclassified the dataset into susceptibility classes: Extreme High, Very High, High, Moderate, Low, and Extreme Low using the K-Means algorithm.

Most of the researchers that are using classification algorithms were tend to do the classifications and getting the output classes using any classification algorithms like Random Forest. It is very common in landslide prediction and landslide susceptibility maps related researches as well. All most all the landslide susceptibility researchers have used machine learning models to predict the output class as Landslide or non-landslide. Then those researchers have done the landslide susceptibility classification using ArcGIS software.

But we can get the probability of getting that class rather than getting only the output class like landslide or non-landslide. In other words, we can get a class probability value for both landslide and non-landslide classes. For example, in the random forest algorithm, calculates the probability of the data point of being class "Landslide" and internally it checks if that probability value is greater than cutoff or less than the cutoff value. If it is greater than the cutoff it will assign the data point to class "Landslide" or else, it will assign to the "non-landslide" class. So here that class probability value which is given for the "landslide" class was considered in this research. Class probability values for the "landslide" class were calculated using ensemble machine learning algorithms: Random Forest, Bagged Decision Tree, AdaBoost, GradientBoost, and XGBoost algorithms. After that, all the class probability values for the "landslide" class were combined for all the pixels in the study area. The class probabilities were taken by using the “predict_proba()” function using the “sklearn” library in python.

Using the K-Means clustering algorithm the resultant class probability values were reclassified into 6 classes: Extreme Low, Low, Moderate, High, Very High, and Extreme High (Figure 5.1).

5.4.1.1 K-Means Algorithm

K-means [61] (Macqueen 1967) utilize an iterative clustering algorithm to approach the best groups of aggregated data. The cluster center is defined as the average value of attributes for all data points in each cluster. The initial step in k-means clustering is to assign the number of clusters and the initial value for each cluster center. Then, assign each data point to a cluster center by the rule that the data point is closer to its grouped center than the other clusters. Finally, the new mean values of all data point in each cluster to obtain a new cluster center. This process iteratively happens until the convergence criteria are satisfied.

So, using this procedure, the class probability values were clustered into six clusters that equal the susceptibility classes: Extreme Low, Low, Moderate, High, Very High, and Extreme High (figure 5.1). Then the landslide susceptibility maps were generated using those clusters using the ArcGIS software tool.

5.4.2 Landslide Susceptibility Modeling using Frequency Ratio (FR) model

Frequency Ratio (FR) is a probability method, for determining the correlation among the landslide causative feature classes and landslide occurrence [62]. The FR can be calculated using equation 5.1 [62] [63] [64].

$$\boxed{FR = \frac{a}{b} = \frac{Nslpix / \Sigma Nslpix}{Ncpix / \Sigma Ncpix}} \qquad (5.1)$$

FR = Frequency Ratio

Nslpix = Number of pixels containing landslide in a class

ΣNslpix = Total number of pixels containing landslide

Ncpix = Total number of pixels in each class in the whole area

ΣNcpix = Total number of pixels in the study area

Initially, each causative factor was categorized into several classes as described in the above section 4.1.1. And FR was calculated for each factor class using equation 5.1 for all the selected causative factors. The Frequency Ratio value of 1 is an average value. And ratio value is greater than one indicates that there is a strong correlation between the causative factor class and the landslide occurrence. On the other hand, a lower ratio value indicates a weak correlation between the causative factor and the landslide occurrence. On the other hand, it means there is a low probability of landslide occurrence when it is having a lower ratio value.

APPENDIX B demonstrates the correlation between the landslide location and landslide conditioning factor classes by showing Frequency values for each factor class. Higher the Frequency Ratio value means a strong correlation between the landslide and the landslide factor classes.

After that, the Landslide Susceptibility Index (LSI) was calculated using the derived frequency ratio values, where the LSI being the sum of each frequency ratio value. Here LSI value was calculated in two ways. As the first method, LSI was calculated using equation 5.2, while the second method is considered a weight factor for each causative factor (equation 5.3). Here the weight factor for each causative factor was the Information Gain value when derived from the feature selection process.

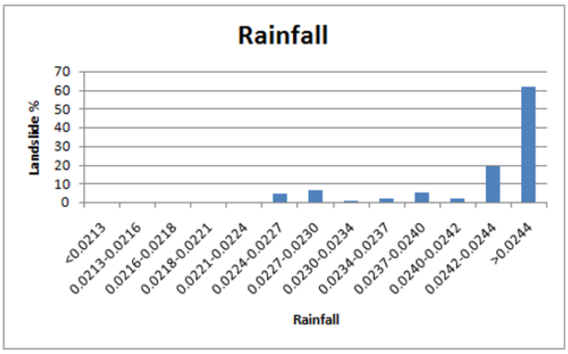
$$\boxed{LSI = Fr_1 + Fr_2 + Fr_3 + \ldots + Fr_n} \tag{5.2}$$

$$\boxed{LSI = Fr_1\, W_1 + Fr_2\, W_2 + Fr_3\, W_3 + \ldots + Fr_n\, W_n} \tag{5.3}$$

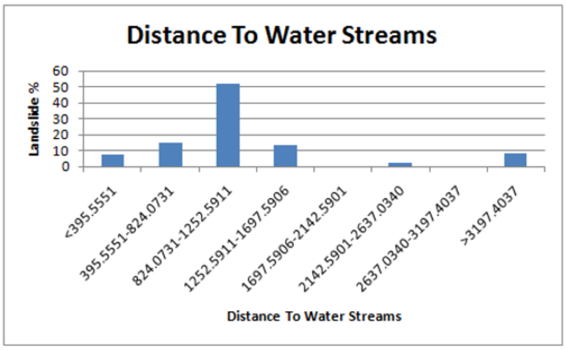
According to equation (5.2) and equation (5.3), Fr is the frequency ratio, n is the number of selected causative factors, LSI is the Landslide Susceptibility Index, and W is the weight factor for each causative factor.

Based on the ArcGIS Natural Breaks (Jenks) classification method, the LSI map was classified into 6 classes as Extreme Low, Low, Moderate, High, Very High, and Extreme High landslide susceptibility classes. So those classified values were used to implement the landslide susceptibility map.

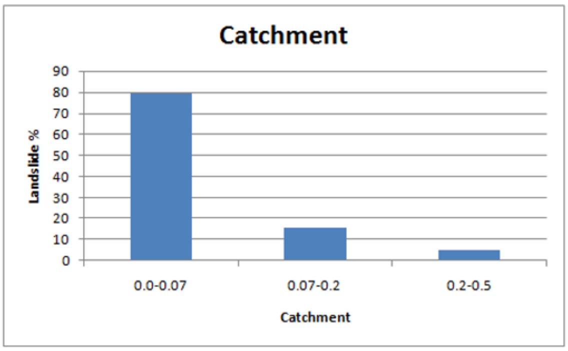
Furthermore, the following figure 5.3 depict the relationship between the landslide and causative factors by showing the landslide percentage value over causative factor categories.


Rainfall
Landslide %
70
60
50
40
30
20
10
0
<0.0213
0.0213-0.0216
0.0216-0.0218
0.0218-0.0221
0.0221-0.0224
0.0224-0.0227
0.0227-0.0230
0.0230-0.0234
0.0234-0.0237
0.0237-0.0240
0.0240-0.0242
0.0242-0.0244
>0.0244
Rainfall

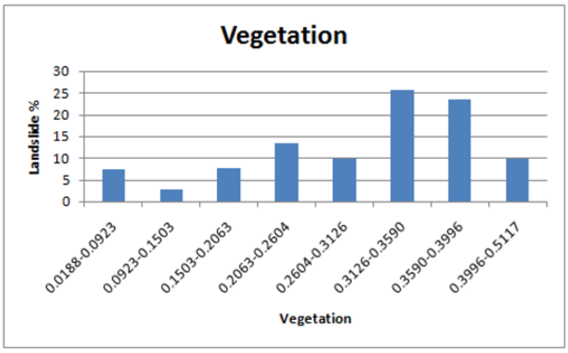
(a)


Distance To Water Streams
Landslide %
60
50
40
30
20
10
0
<395.5551
395.5551-824.0731
824.0731-1252.5911
1252.5911-1697.5906
1697.5906-2142.5901
2142.5901-2637.0340
2637.0340-3197.4037
>3197.4037
Distance To Water Streams

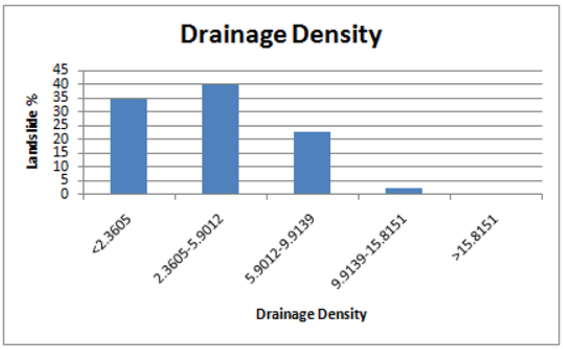
(b)


Catchment
Landslide %
90
80
70
60
50
40
30
20
10
0
0.0-0.07
0.07-0.2
0.2-0.5
Catchment

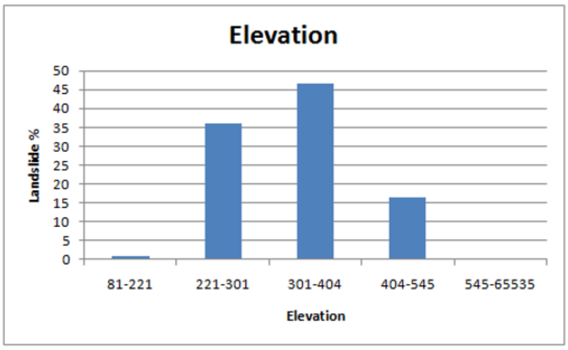
(e)


Vegetation
Landslide %
30
25
20
15
10
5
0
0.0188-0.0923
0.0923-0.1503
0.1503-0.2063
0.2063-0.2604
0.2604-0.3126
0.3126-0.3590
0.3590-0.3996
0.3996-0.5117
Vegetation

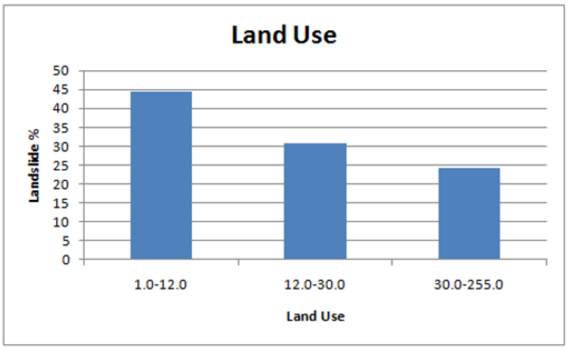
(f)


Drainage Density
Landslide %
45
40
35
30
25
20
15
10
5
0
<2.3605
2.3605-5.9012
5.9012-9.9139
9.9139-15.8151
>15.8151
Drainage Density

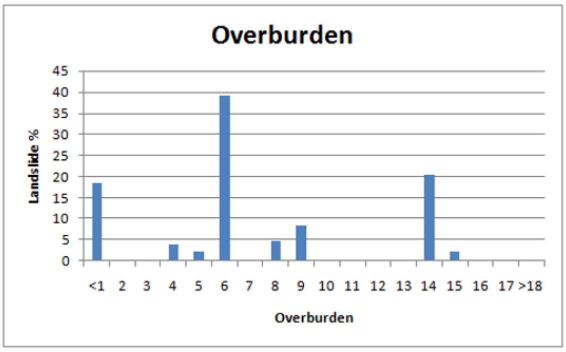
(c)


Elevation
Landslide %
50
45
40
35
30
25
20
15
10
5
0
81-221
221-301
301-404
404-545
545-65535
Elevation

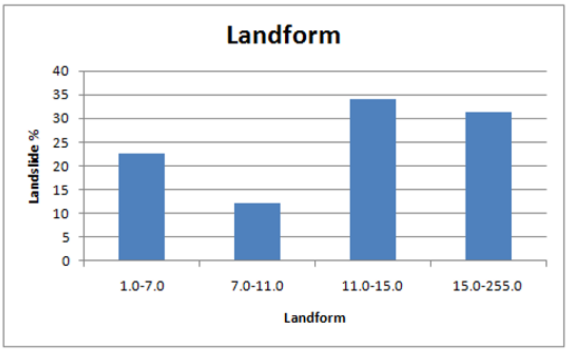
(d)



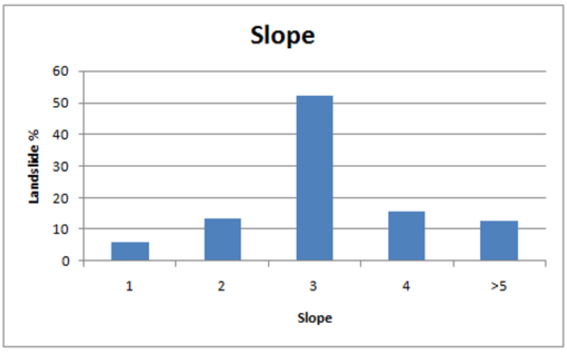
(g)



(h)



(i)



(j)

Figure 5.3: Relationship between landslide and causative factors (a) Rainfall (b) Distance To Water Streams (c) Drainage Density (d) Elevation (e) Catchment (f) Vegetation (g) Land Use (h) Overburden (i) Landform and (j) Slope maps

5.5 Model Validation Techniques

One of the most important tasks in machine learning is model validation. The process of testing the trained model on a test data set is known as model validation. The evaluation dataset is a subset of the same dataset that is used to construct the training set. A trained model's generalization capability is thus given.

The model is tested in this analysis by referencing a recorded landslide and comparing it to the model predictions. The initial strategy was to divide the dataset into two parts: training and testing.

In order to ensure that the model is not overfitted, ten-fold cross-validation was used in this research. And the evaluation was carried out in two steps including binary classification evaluation and multi-class class classification evaluation.

Following are some of the machine learning model evaluation techniques which were used in this study.

5.5.1 Confusion Matrix

In classification problems, the Confusion Matrix is a commonly used metric. It can be used in both binary and multi-class classification problems. A Confusion Matrix (figure 5.4) is a table that is commonly used to describe a classification model's or classifier's performance on a set of test datasets for which the true values are known.

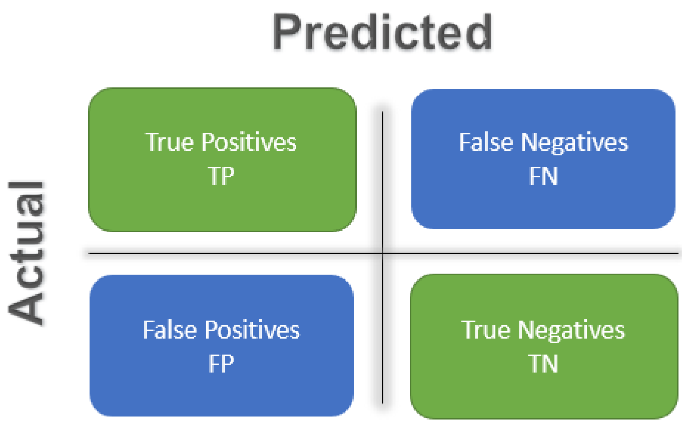


Figure 5.4: Confusion matrix terminology

The confusion matrix mainly represents counts from predicted and actual data values. It contains common terms: True Positive, False Positive, True Negative, and False Negative. Here it explains the above terms.

- **True positives (TP):** Predicted positive and are actually positive.
- **False positives (FP):** Predicted positive and are actually negative.
- **True negatives (TN):** Predicted negative and are actually negative.
- **False negatives (FN):** Predicted negative and are actually positive.

The ***“confusion matrix()”*** function, which is part of the ***“sklearn”*** library, was used to build the confusion matrix in this analysis.

5.5.2 Accuracy

When it comes to classification, accuracy is one of the most widely used metrics. The following formula is used to measure the precision with the use of confusion matrix. However, since this analysis uses an unbalanced dataset, precision is not a suitable criterion to use.

$$\text{Accuracy} = \frac{TN + TP}{TN + FP + FN + TP} \tag{5.4}$$

5.5.3 Precision

The percentage of positive instances out of the total estimated positive instances is known as precision. This is used to determine how accurate the model is when it is correct.

$$\text{Precision} = \frac{TP}{FP + TP} \tag{5.5}$$

5.5.4 Recall

The recall is the proportion of positive instances from total actual positive instances.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.6}$$

5.5.5 F-Score

The harmonic mean of precision and recall is the F-Score. Precision and recall are also included in this. And, higher F-Score is better for a machine learning model. Speciallly, F-Score is very useful matrix for classification problems which involves unbalanced data.

$$\text{F- Score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \tag{5.7}$$

5.5.6 Specificity

The proportion of correctly defined negatives is measured by specificity (True Negative Rate).

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5.8}$$

5.5.7 Sensitivity

The proportion of correctly defined positives is measured by sensitivity (True Positive rate).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.9}$$

5.5.8 Landslide Density Index (LDI)

The “Landslide Density Index” (LDI) [65] [66] is the ratio between the percentage of landslide pixels and the percentage of class pixels in each class in the landslide susceptibility map, which is used to validate the landslide susceptibility index. If the LDI value is increased from low to very high susceptibility class, the susceptibility map is considered correct. The LDI is determined using the formula below. Furthermore, if a higher percentage of landslides occur in high, very high, and extremely high susceptibility levels than in other levels, the suitability of any susceptibility map can be verified [65] [66].

$$\boxed{\text{LDI} = \frac{\text{Percentage of validation landslide pixels}}{\text{Percentage of area pixel}}} \tag{5.10}$$

CHAPTER 6: RESULTS AND EVALUATION

This chapter mainly focuses on showing the results and evaluation on feature selection techniques, landslide prediction method and landslide susceptibility modeling.

6.1 Feature Selection Method Results

Here it depicts the results of the landslide feature selection methods. So, it mainly depicts the results of tree-based feature selection method and information gain method.

6.1.1 Tree-based feature selection method results

The Tree-based Feature Importance Approach was primarily used to measure the relative importance of each landslide conditioning factor. The results of tree-based feature importance method illustrate in figure 6.1. The results indicate that the annual rainfall (0.145) has the highest contribution to landslide susceptibility, followed by distance to streams (0.103), land use (0.103), elevation (0.089), drainage density (0.088), catchment (0.086), vegetation (0.082), overburden (0.079), slope (0.074), landform (0.052), aspect (0.029), derived slope (0.029), profile curvature (0.025), curvature (0.021) and plan curvature (0.014). According to this method almost all the landslide conditioning factors plays a major role in making the study area prone to landslides as all factors got a considerable feature importance value.

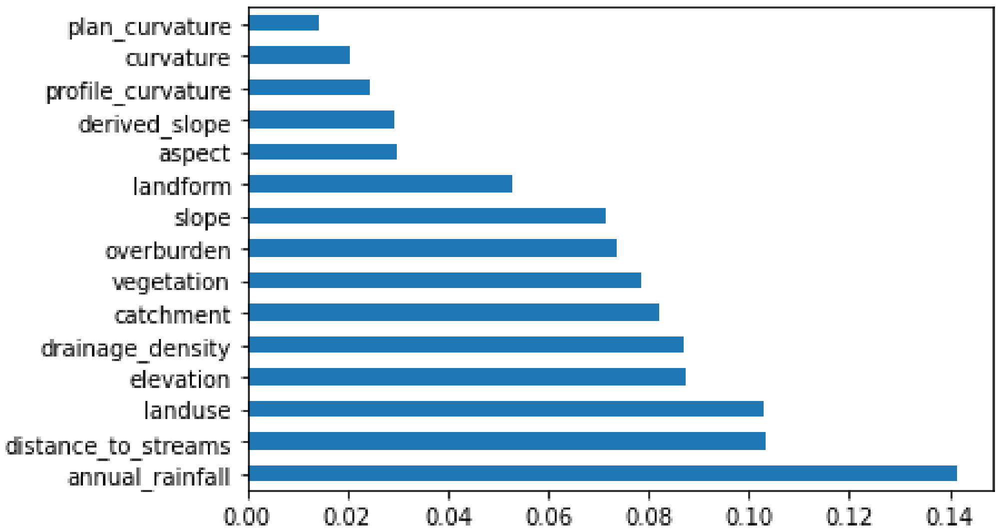


Figure 6.1: Feature Importance of landslide conditioning factors

6.1.2 Information Gain (IG) feature selection method results

As a tree-based feature importance method, Information Gain Feature Selection method was used in order to select the best set of conditioning factors. The figure 6.2 depicts the Information Gain Ratio values calculated for 15 landslide conditioning factors in this study. It observed that the highest information gain feature importance value is for slope (0.090) followed by landform (0.075), overburden (0.038), land use (0.030), vegetation (0.026), drainage density (0.021), catchment (0.020), annual rainfall (0.005), distance to streams (0.004) and elevation (0.003).

According to the IG results, plan curvature, curvature, profile curvature and aspect factors were neglected in this study since there was no considerable feature important scores for those features.

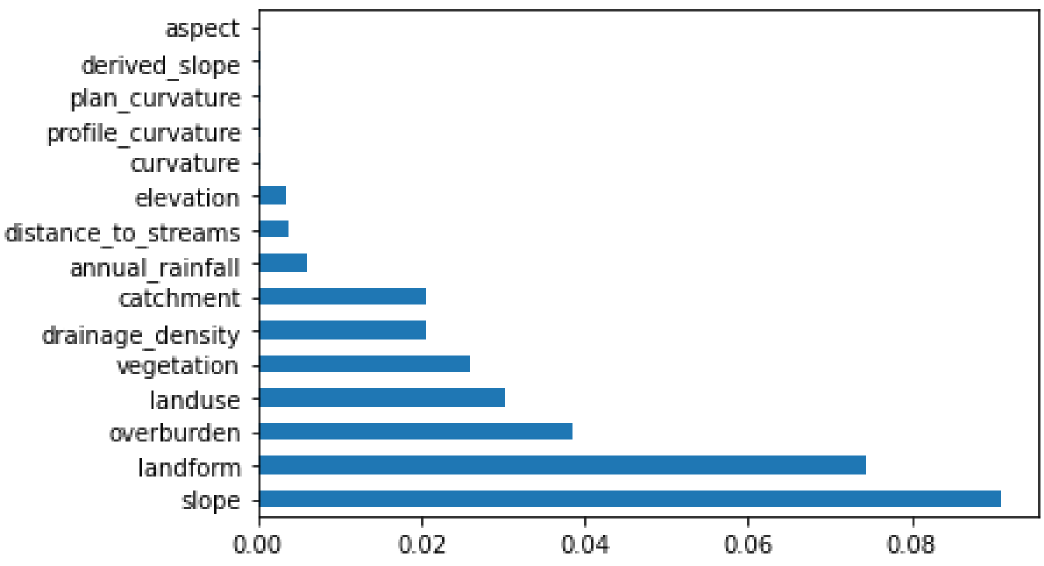


Figure 6.2: Information Gain of landslide conditioning factors

At the end, by considering all the above feature selection methods, especially considering tree-based feature importance method and Information Gain method, ten causative factors were selected out of fifteen set of causative factors. Since, both of those two methods present similar results. When considering the Information Gain results, some of the factors: aspect, derived slope, plan curvature, and curvature can be neglected as they are not giving any considerable IG value. Moreover, same set of factors are having less impact on landslide occurrence according to the tree-based feature important method too. As a result, those mentioned causative factors were neglected and selected the rest of the causative factors: annual rainfall, distance to water streams, elevation, drainage density, catchment, vegetation, overburden, slope, land use and landform for this study.

6.2 Landslide prediction results

With the use of selected causative factors, several ensemble machine learning algorithms: Random Forest, Bagged Decision Tree (BDT), AdaBoost, XGBoost and GradientBoost were built using the training dataset. The results are shown in table 6.1 and figure 6.3. Table 6.1 shows the performance comparison of selected ensemble machine learning algorithms on landslide prediction, while figure 6.3 depicts percentage of actual landslide points predicted from selected ensemble machine learning algorithms.

Table 6.1: Performance comparison of ensemble machine learning algorithms on landslide prediction

Parameters	RF	BDT	XGBoost	AdaBoost	GradientBoost
TP	1288	1433	1490	1409	1478
TN	358354	358209	358287	358172	358263
FP	35	180	102	217	126
FN	323	178	121	202	133
F-Score (%)	90	89	**93**	87	91.9
Precision (%)	84	89	**94**	87	92.1
Recall (%)	**98**	89	92	87	91.7
Sensitivity (%)	79.9	88.9	**92.5**	87.5	91.7
Specificity (%)	99.9	99.9	99.9	99.9	99.9
Accuracy (%)	99.6	99.9	**99.94**	99.8	99.9

However, those results depict that XGBoost has the highest performance in terms of F-Score and Accuracy. They are followed by the GradientBoost, Random Forest, BDT and AdaBoost respectively. In terms of the precision also, XGBoost (94%) had the highest value followed by GradientBoost (92.1%), BDT (89%), RF (84%) and AdaBoost (87%). However, in terms of recall, the highest value has been achieved by RF followed by XGBoost, GradientBoost, BDT and AdaBoost.

Based on these performance matrices results, we conclude that the XGBoost is more accurate than rest of the other machine learning models for landslide predicting in the study area.

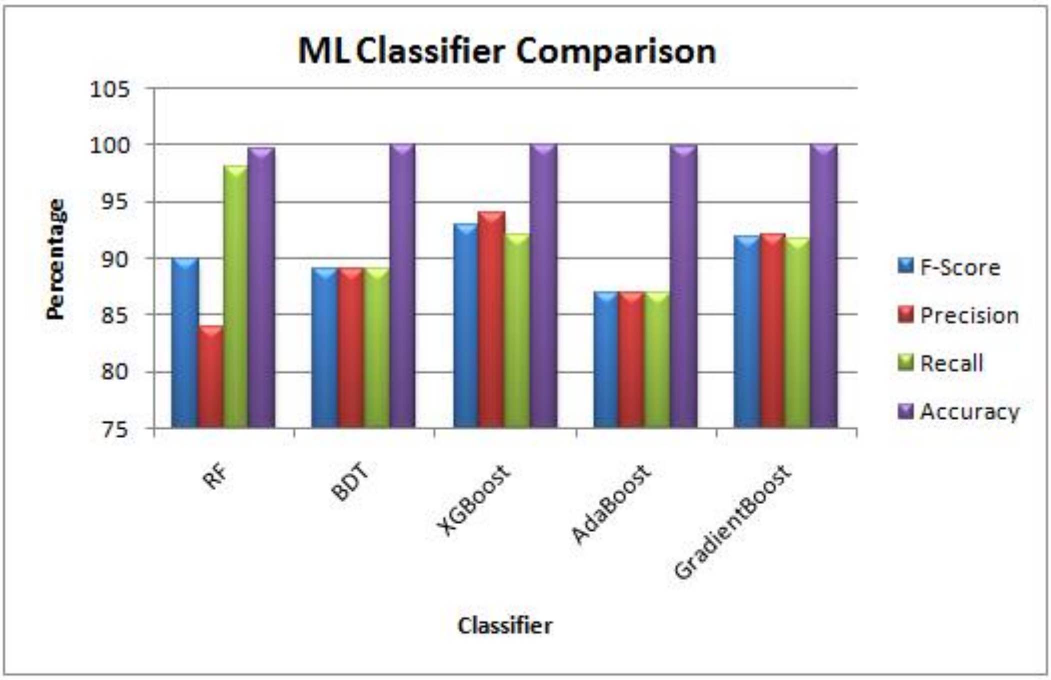


Figure 6.3: Performance comparison of Machine learning classifiers

6.3 Landslide Susceptibility Map (LSM) results

In this section it shows the results and evaluation LSMs which were generated based on semi-supervised machine learning models and on Frequency Ratio method.

6.3.1 LSM by Machine Learning Model Results

As described in the chapter 4, three separated landslide susceptibility maps were prepared for the study area. One LSM was prepared using semi-supervised machine learning model. And rest of the LSMs was prepared using Frequency Ratio method with and without considering weights for each causative factor. The generated landslide susceptibility maps are showing in figure 6.4 and figure 6.5.

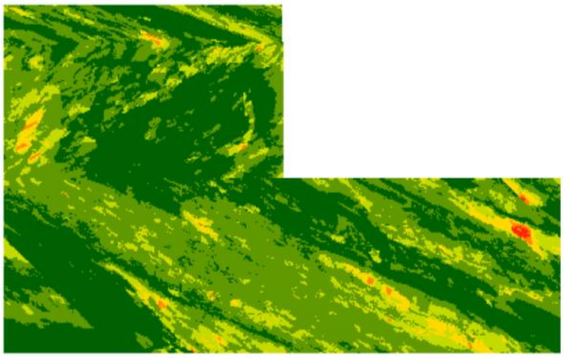
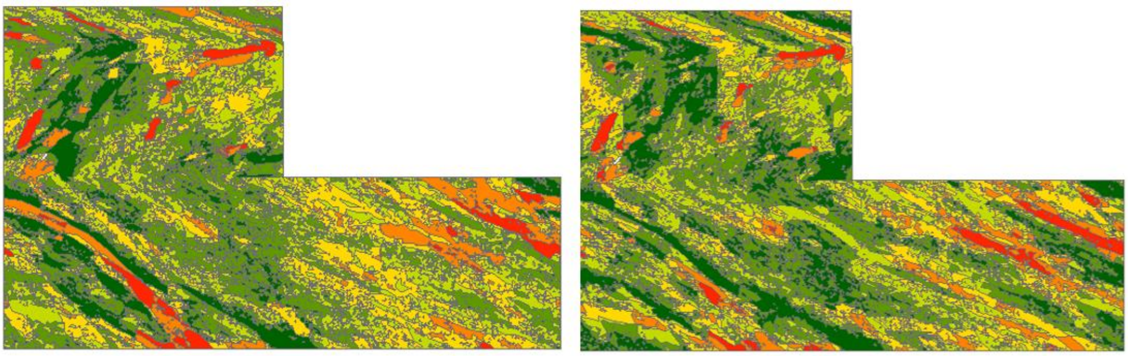


Figure 6.4: Landslide susceptibility maps produced by semi-supervised machine learning model



(a) (b)

Figure 6.5: Landslide susceptibility maps produced by ensemble FR method (a) without considering weights for each causative factor (b) by considering weights for each causative factor

Table 6.2 and figure 6.6 depicts a detail description about landslide susceptibility classes. As presented in table 6.2, the landslide susceptibility map produced by semi-supervised machine learning model, Very High susceptibility covers 76.85% of landslide points, followed by Extreme High (21.16%), High (1.71%), and Moderate (0.28%). In the Frequency Ratio method with considering weight factors for each factor class, Very High susceptibility class covers 70.29% of landslide points, followed by Extreme High (18.77), High (6.95%), Moderate (3.48%), Low (0.35%) and Extreme Low (0.15%). Moreover, in the Frequency Ratio method without considering weight factors for each factor class, Very High susceptibility class covers 69.01% of landslide points, followed by Extreme High (18.53), High (3.22%), Moderate (0.34%), Low (0.01%) and Extreme Low (0.005%).

Table 6.2: Accuracy statistics for LSM models

Model	Susceptibility Class	No of Class Pixels	No of Landslide Pixels	Class Pixels %	Landslide %	Landslide Density (LD)
Ensemble Machine Learning Model	Extreme High	1401	1136	0.12	21.16	181.24
	Very High	4253	4126	0.35	76.85	216.84
	High	52954	92	4.41	1.71	0.39
	Moderate	299970	15	24.99	0.28	0.01
	Low	756100	0	63.01	0	0
	Extreme Low	85322	0	7.11	0	0
Frequency Ratio method with considering weight for each factor	Extreme High	1273	1008	0.11	18.77	176.98
	Very High	3911	3774	0.33	70.29	215.68
	High	29485	373	2.46	6.95	2.83
	Moderate	130582	187	10.88	3.48	0.32
	Low	518343	19	43.19	0.35	0.01
	Extreme Low	516406	8	43.03	0.15	0.003
Frequency Ratio	Extreme High	1277	995	0.11	18.53	174.15
	Very High	4586	3705	0.38	69.01	180.57

method without considering weight for each factor	High	29928	431	2.49	8.03	3.22
	Moderate	133086	202	11.09	3.76	0.34
	Low	525557	24	43.79	0.45	0.01
	Extreme Low	505368	12	42.13	0.22	0.005

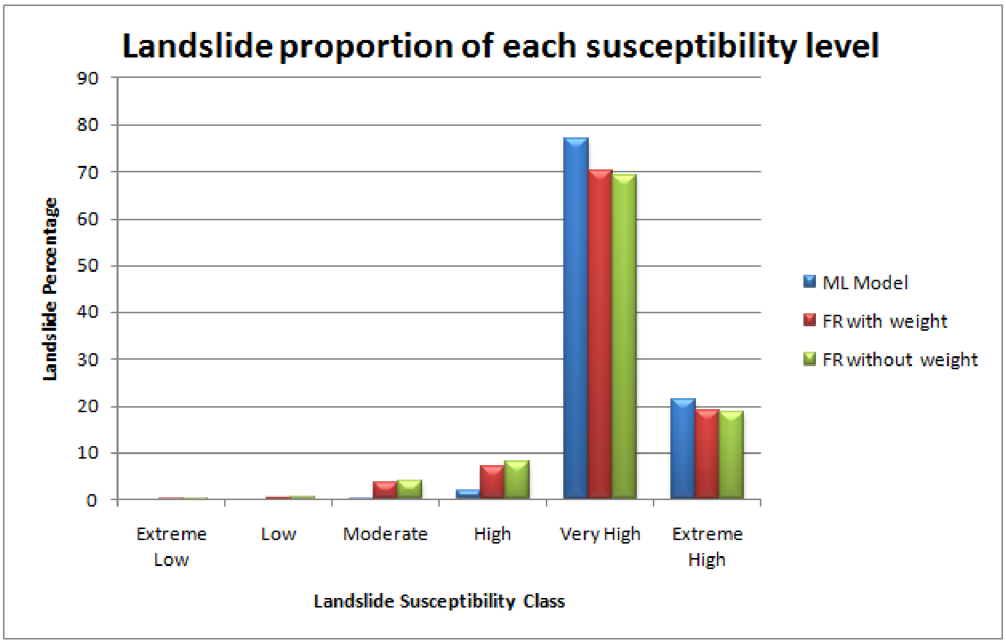


Figure 6.6: Landslide proportion of each susceptibility level

According to the table 6.2, the landslide susceptibility map produced by semi-supervised machine learning model, got high landslide density value 181.24 for Very High susceptibility class, followed by Extreme High (216.84), High (0.39), and Moderate (0.01). In the Frequency Ratio method with considering weight factors for each factor class, Very High susceptibility class got 176.98 values as the high LD value, followed by Extreme High (215.68), high (2.83), moderate (0.32), Low (0.01) and Extreme Low (0.003). Moreover, in the Frequency Ratio method without considering weight factors for each factor class, got high LD value 174.15 for Very High landslide class, followed by Extreme High (180.57), High (3.22), Moderate (0.34), Low (0.01) and Extreme Low (0.005) respectively.

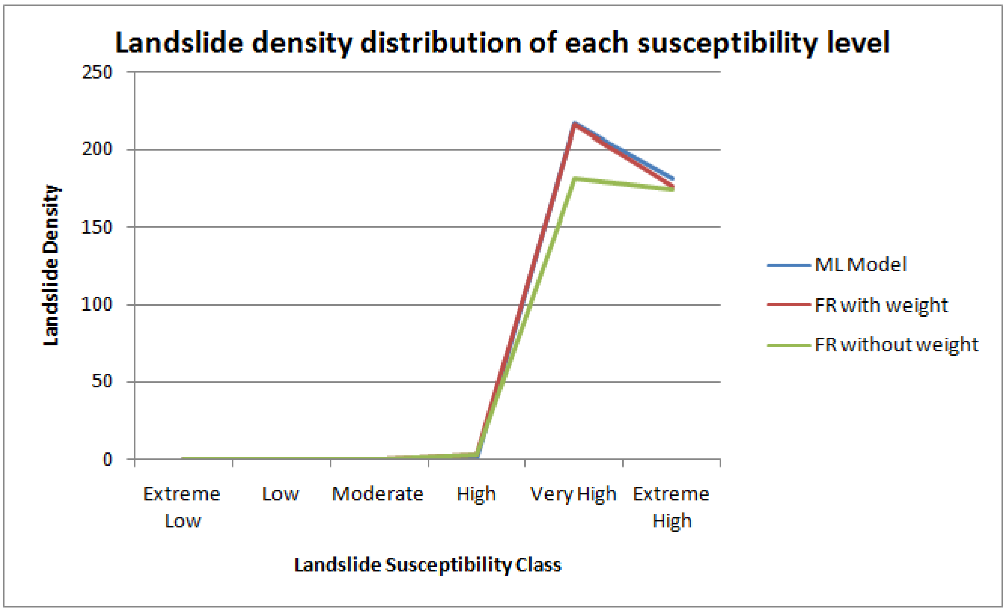


Figure 6.7: Landslide density distribution of each susceptibility level

According to the above LSM results which are shown in figure 6.7, the LD values for the Extreme High susceptibility class are 181.24, 176.98, and 174.15 for the ML dependent model, FR model with factor weight, and FR model without factor weight, respectively. As a result, the ML model outperforms the rest of the models by having the highest LD value.

Furthermore, the LSMs which were generated using machine learning model and frequency ratio models were validated based on the landslide density of each susceptibility class on the LSMs. Generally, for the study area, the value of LD increased gradually, from extreme low to extreme high susceptibility. So according to these results, LD value increased from very extreme to very high classes. But the Extreme high class got less value when compared to the very high class. But those LSMs can be considered as valid LSMs since most of the landslide locations were laid in the Extreme High, Very High and High susceptibility classes in all three LSMs.

CHAPTER 7: DATA VISUALIZATION WEB APPLICATION

Landslide Susceptibility Mapping (LSM) is defined as a useful tool for disaster management and risk mitigation [67]. The LSMs are often useful to decision-makers and authorities in determining the best course of action for mitigating risks. LSM techniques were extensively developed and implemented by several researches using different types of techniques.

As a result, the visualization of the susceptibility map plays an important role in the disaster management phase. Since no landslide prediction tool has been built for Sri Lanka, it is critical to have a landslide susceptibility visualization tool. This study aims to visualize the landslide data and landslide susceptibility data of Rathnapura District, Sri Lanka, as a first step toward creating a visualization tool that covers the entire country.

A prototype web application for Landslide Susceptibility Data visualization was developed as a result of this study, using the derived Landslide Susceptibility Maps, landslide locations, and feature maps. These landslide susceptibility maps, on the other hand, were based on static data. This program can be expanded to become an integrated decision support system that displays dynamic data sets.

This web application gives the capability to view the landslide susceptibility details, landslide data locations and landslide causative factor details on the world map (APPENDIX C).

The programming languages Angular, Typescript, and NodeJS were primarily used in the creation of data visualization web applications. Additionally, the CesiumJS library was incorporated into the application to generate 3D visualization globes and maps. The web application's home page is depicted in APPENDIX C. Using ArcGIS software, the landslide susceptibility information, landslide locations, and causative feature data were formatted in to the geoJson format. The geoJson data was then shown on the cesium globe. Users can select any area from the right-hand tree menu by clicking it.

7.1 Technologies used for web application implementation

This section provides the technologies which were used for data visualization web application development. So it mainly used Angular, NodeJs as programming languages and CesiumJS as a third party library.

7.1.1 Leaflet JavaScript Library

Leaflet [68] is a lightweight and efficient JavaScript library. It's mainly used to build web mapping applications. It works with Angular, NodeJs, HTML5, and CSS3, among other technologies, on both desktop and mobile platforms. It can read data from GeoJSON files or form files to load function data. It has interactive features including scroll zooming, drag panning with inertia, keyboard navigation, and so on.

Leaflet was used to visualize the susceptibility maps and feature maps in this study. However, rendering those maps with high performance was difficult as Leaflet is better suited to small applications that demand low library overhead and high performance.

7.1.2 CesiumJS JavaScript Library with GeoJson files

CesiumJS [69] is an open-source JavaScript library for creating interactive 3D world globes and maps with improved performance, visual quality, and accessibility. CesiumJS is a web framework for developing interactive geospatial data visualization applications.

CesiumJS was used to visualize landslide causative feature maps, landslide susceptibility maps, and landslide data in this study. These maps can be downloaded as vector data (shape files), geoJSON files, or tiff files (raster data). Initially, shape files were used to introduce and evaluate visualization. Shape files are vector files that contain information about geospatial locations and attributes. Polygons, lines, and points are typically used to describe features in shape files. However, since the shape files contain polygons, output is not as good as it could be.

After that, the data was visualized using geoJson. Using ArcGIS program, the shape files were translated to geoJson files. When compared to shape files, geoJson files performed better for landslide data visualization. The screenshots of the web application are depicted in section 6.3.

CHAPTER 8: CONCLUSION AND FUTURE WORKS

This chapter will include some concluding remarks, such as a summary, research Outcome, future work, and discussion.

8.1 Summary

Landslide prediction and susceptibility map modeling are critical for effectively mitigating landslide risks, especially in Sri Lanka. Landslide presents a significant impact at Rathnapura District in Sri Lanka. It has a major impact on human and animal lives, human settlements, agriculture, and the social and economic aspects of rural communities. This study used a Machine Learning model and a Frequency Ratio model for landslide susceptibility modeling to address these concerns.

For this reason, a landslide inventory map with 5369 landslide cells was divided into training and validation datasets with 70% and 30%, respectively. Furthermore, ten landslide causative factors were chosen from fifteen sets of causative factors using the tree- based feature importance method, and information gain method. All fifteen causative factors led to landslide incidence, according to the Information Gain (IG) method, since their IG values were all greater than zero. Slope, annual rainfall, overburden, elevation, drainage density, landform, land use, distance to water streams, catchment, and vegetation are among the selected causative factors. As a result, the selected causative factors were used to conduct landslide prediction and LSM modeling. When compared to other variables, annual rainfall has been found to be closely related to the occurrence of landslides.

For landslide prediction, this study used five different ensemble machine learning approaches. This research made a major contribution by systematically comparing and analyzing five machine learning models for landslide prediction. Since landslide susceptibility mapping is significant, the LSM was carried out using a semi-supervised machine learning algorithm model and using Frequency Ratio method. The landslide data, on the other hand, consists of binary classification data such as "landslide" and "non- landslide," requiring each cell to be categorized into one of six susceptibility classes.

LSMs were generated in the machine learning model using semi-supervised machine learning algorithms, taking into account the class probability values for the “landslide” class. The K-Means clustering algorithm was used to reclassify the class probability values obtained from five ensemble machine learning algorithms into susceptibility classes. Finally, the mean value of each class was used to label the clusters. With the increasing order of class mean values, the classes were classified as Extreme Low, Medium, Moderate, High, Very High, and Extreme High.

The FR process, on the other hand, reclassified selected causative factors into sub groups. After that, a frequency ratio value was determined for each of the causative factor groups. Following that, the appropriate FR value for each cell was used to measure the LSI value for each cell using those frequency ratio values. Finally, these LSI values were reclassified into six landslide vulnerability with the use of ArcGIS Natural-Break’s classification system: Extreme Low, Low, Medium, Moderate, High, Very High, and Extreme High.

The Landslide Density Index (LDI) value was used to validate the output of the LSMs provided by the semi-supervised ML model and the FR model. The results revealed that both models performed exceptionally well. The semi-supervised ML model, on the other hand, had a comparable high success rate and a high LDI value. Furthermore, the landslide susceptibility analysis validation results indicated that the majority of the reported landslide locations were in the Extreme high and high susceptibility classes. And these LSM models appear to be accurate in this research field.

As a result, the LSMs developed in this study could be useful to decision-makers, planners, and engineers in disaster planning in order to reduce economic losses and casualties. So as an outcome of this research, it proposed a pleasant web application to visualize the landslide data, LSMs, and conditioning factor data on it. And it will be highly usable for the Sri Lankan authorities and decision makers.

Python was used for landslide prediction, statistical computation, and data visualization in this study since it is commonly used for machine learning based forecasts, processing, and analysis.

8.2 Future Work

The following areas can be considered for the future work carried out in the field of Landslide Susceptibility Mapping.

- The landslide susceptibility map is a static tool for mapping the likelihood of a landslide occurring at a given site. So that it is important to have a dynamic landslide prediction framework that can produce temporal warnings based on continuously updated data. So, this research can be extended to support a dynamic landslide prediction framework.
- This research has only focused on Rathnapura district in Sri Lanka. So that as a future work this can be carried out for other districts in Sri Lanka.

8.3 Discussion

This main objective of this research was to do the landslide prediction using ensemble-based machine learning methods and to construct and evaluate landslide susceptibility maps using a semi-supervised machine learning method. It was applied to Rathnapura district in Sri Lanka.

As mentioned above as the feature selection methods it has used tree-based feature important method and the information gain method as those two methods are commonly used in landslide susceptibility related researches. Both methods gave similar results, so that it was easy to select most important features out of fifteen factors which were used in this research. Afterward, five ensemble machine learning methods were used for landslide prediction as most of the researches didn't consider more than three ensemble methods together. So, the prediction results have shown that the XGBoost algorithm outperform the other ensemble methods. However, all the methods got comparably better results than using normal machine learning models as SVM, Naïve bayes and etc.

However, the dataset only contained the binary data as "landslide" and "non-landslide". But it is not a good measurement to check the severity of a particular area. So, it was very important to classify the study are into susceptibility classes as Extreme

Low, Low, Moderate, High, Very High and Extreme High. And also, there is no proper mechanism to create LSMs or to check the severity of landslide for Sri Lanka. For that purpose, this research proposed a machine learning method.

And most prominently, this research has used a novel method for landslide susceptibility modeling. Most of the prior researches have created LSMs with the help of ArcGIS software. Somehow this paper presents a new semi-supervised machine learning method for Landslide Susceptibility map modeling. Furthermore, a Frequency Ratio method which is a statistical method was used to create LSMs and for comparing with the machine learning method. So finally, it proved that the semi-supervised machine learning got comparably better results for landslide susceptibility map creation when comparing with FR method.

The generated landslide susceptibility maps within this research can be used by decision makers and by the relevant authorities in Sri Lanka to take appropriate actions in order to mitigate the risks which can be occurred by landslides. Specially these landslide susceptibility maps show the degree of susceptibility of an area to landslide occurrence. So, by considering the generated landslide susceptibility map, the decision makers can identify the high-risk areas in Rathnapura district in Sri Lanka and can take appropriate actions in advance.

Somehow these landslide susceptibility maps are static methods to show the severity of landslide occurrence. On the other hand, there is a dynamic approach to show a dynamic landslide prediction framework that can be produce by using continuously updated data. It will be handled as a future work in this research study.

REFERENCES

[1] National Building Research Organisation. (2019, March) National Building Research Organisation- Ministry of Public Administration & Disaster Management. [Online]. http://www.nbro.gov.lk/index.php?option=com_content&view=article&id=127:damage-assessment-of-recent-landslide-events-in-sri-lanka&catid=2&Itemid=101&lang=en

[2] Nimal Herath, N. Thayalan R.K. Bhandari, "Landslide hazard zonation mapping in Sri Lanka- A Holistic approach, National Building Research Organisation, Colombo, National Symposium on Landslides in Sri Lanka," 1994.

[3] Thomas, and Mike Crozier. Glade, "A Review of Scale Dependency in Landslide Hazard and Risk Analysis," *Landslide Hazard and Risk*, pp. 75-138, January 2005.

[4] Chau, K.T., Y.L. Sze, M.K. Fung, W.Y. Wong, E.L. Fong, and L.C.P. Chan. 2004. "Landslide Hazard Analysis for Hong Kong Using Landslide Inventory and GIS." Computers & Geosciences 30 (4) (May): 429–443.

[5] Jones, D. K. C. 1995. "The Relevance of Landslide Hazard to the International Decade for Natural Disaster Reduction." In Proceedings of Conference on Landslide Hazard Mitigation with Particular Reference to Developing Countries, 19–33. London: Royal Academy of Engineering.

[6] Guzzetti, Fausto, Paola Reichenbach, Mauro Cardinali, Mirco Galli, and Francesca Ardizzone. 2005. "Probabilistic Landslide Hazard Assessment at the Basin Scale." Geomorphology 72 (1-4) (December): 272–299. doi:10.1016/j.geomorph.2005.06.002. http://www.sciencedirect.com/science/article/pii/S0169555X05001911.

[7] Smith, Keith. 2009. "Mass Movement Hazards." In Environmental Hazards-Assessing Risk and Reducing Disaster, 6th ed., 205–226. New York: Routledge.

[8] Soeters, Robert, and C. J. Van Westen. 1996. "Slope Instability: Recognition, Analysis, and Zonation." In Landslide: Investigation and Mitigation, edited by A. Turner and R. Shuster, 129–177. Washington D.C.: National Academy Press.

[9] Saito M., "Forecasting time of slope failure by tertiary creep.," in *Proceedings of the 7th International Conference on Soil Mechanics and Foundation Engineering*, Mexico City, 1969, pp. 2. p. 677–683.

[10] Varnes, 1984. “Landslide Hazard Zonation: A Review of Principles and Practice.” UNESCO. France. http://www.bib.ub.edu/fileadmin/fdocs/landslidehazard.pdf.

[11] Wang, Huabin, Gangjun Liu, Weiya Xu, and Gonghui Wang. 2005. “GIS-Based Landslide Hazard Assessment: An Overview.” Progress in Physical Geography 29 (4): 548–567. doi:10.1191/0309133305pp462ra

[12] Kanungo, D.P., M.K. Arora, S. Sarkar, and R.P. Gupta. 2006. “A Comparative Study of Conventional, ANN Black Box, Fuzzy and Combined Neural and Fuzzy Weighting Procedures for Landslide Susceptibility Zonation in Darjeeling Himalayas.” Engineering Geology 85 (3-4) (June): 347–366. doi:10.1016/j.enggeo.2006.03.004. http://linkinghub.elsevier.com/retrieve/pii/S001379520600113X.

[13] Guzzetti, Fausto, Paola Reichenbach, Mauro Cardinali, Mirco Galli, and Francesca Ardizzone. 2005. “Probabilistic Landslide Hazard Assessment at the Basin Scale.” Geomorphology 72 (1-4) (December): 272–299 doi:10.1016/j.geomorph.2005.06.002. http://www.sciencedirect.com/science/article/pii/S0169555X05001911.

[14] P. Jayasingha, "Social Geology and Landslide Disaster Risk Reduction in Sri Lanka," Journal of Tropical Forestry and Environment, p. 13, 2016

[15] L.D.C.S. Subhashini, Premaratne (2013), “Landslide Prediction Using Artificial Neural Networks”

[16] Karunanayake, K.B.A.A.M & Wijayanayake, Janaka. (2019). Application of Data Mining Technique to Predict Landslides in Sri Lanka. International Journal of Data Mining & Knowledge Management Process. 9. 53-67. 10.5121/ijdkp.2019.9404.

[17] Guzzetti, Fausto, Alberto Carrara, Mauro Cardinali, and Paola Reichenbach. 1999. “Landslide Hazard Evaluation: A Review of Current Techniques and Their Application in a Multi-Scale Study, Central Italy.” Geomorphology. http://www.sciencedirect.com/science/article/pii/S0169555X99000781.

[18] Dai, Fuchu, and Chack Fan Lee. 2002. “Landslides on Natural Terrain: Physical Characteristics and Susceptibility Mapping in Hong Kong.” Mountain Research and Development 22 (1): 40–47. http://www.bioone.org/doi/abs/10.1659/0276-4741(2002)022%5B0040:LONT%5D2.0.CO%3B2.

[19] Wang, Huabin, Gangjun Liu, Weiya Xu, and Gonghui Wang. 2005. "GIS-Based Landslide Hazard Assessment: An Overview." Progress in Physical Geography 29 (4): 548–567. doi:10.1191/0309133305pp462ra.

[20] Alexander, David E. 2008. "A Brief Survey of GIS in Mass-Movement Studies, with Reflections on Theory and Methods." Geomorphology 94 (3-4) (February): 261–267. doi:10.1016/j.geomorph.2006.09.022.http://linkinghub.elsevier.com/retrieve/pii/S0169555X0700267X.

[21] Corominas, Jordi, and José Moya. 2008. "A Review of Assessing Landslide Frequency for Hazard Zoning Purposes." Engineering Geology 102 (3-4) (December): 193–213. doi:10.1016/j.enggeo.2008.03.018.http://www.sciencedirect.com/science/article/pii/S0013795208001841.

[22] Van Westen, Cees J., Enrique Castellanos, and Sekhar L. Kuriakose. 2008. "Spatial Data for Landslide Susceptibility, Hazard, and Vulnerability Assessment: An Overview." Engineering Geology 102 (3-4) (December): 112–131.doi:10.1016/j.enggeo.2008.03.010.http://linkinghub.elsevier.com/retrieve/pii/S0013795208001786.

[23] Dieu, D. T., Pradhan, B., Lofman, O., &Revhaug, I., (2012). Landslide susceptibility Assessment in Vietnam using Support vector machines, Decision Tree, and Naive Bayes Models. *Mathematical problems in Engineering* **2012**: 1-26. doi: 10.1155/2012/974638.

[24] "Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods," Geomorphology, vol. 303, pp. 284-298, February 2018.

[25] Jianbing Peng, Haoyua nHong, Himan Shahabi, BiswajeetPradhan Wei Chen, "Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China," Science of The Total Environment, vol. 626, pp. 1121-1135, June 2018.

[26] Chang-Wook Lee, Saro Lee Prima Riza Kadavi, "Application of Ensemble-Based Machine Learning Models to Landslide Susceptibility Mapping," vol. 10, no. 8, p. 1252, 2018.

[27] X. Luo, F. Lin, S. Zhu, M. Yu, Z. Zhang, L. Meng and J. Peng, "Minelandslide susceptibility assessment using IVM, ANN and SVM modelsconsidering the contribution of affecting factors", PLOS ONE, vol. 14,no. 4, p. e0215134, 2019.

[28] V.-H. Nhu, A. Mohammadi, H. Shahabi, B. B. Ahmad, N. Al-Ansari, A. Shirzadi, J. J. Clague, A. Jaafari, W. Chen, and H. Nguyen, “Landslide Susceptibility Mapping Using Machine Learning Algorithms and Remote Sensing Data in a Tropical Environment,” International Journal of Environmental Research and Public Health, vol. 17, no. 14, p. 4933, Jul. 2020.

[29] Z. Wang, Q. Liu and Y. Liu, "Mapping Landslide Susceptibility Using Machine Learning Algorithms and GIS: A Case Study in Shexian County, Anhui Province, China", Symmetry, vol. 12, no. 12, p. 1954, 2020. Available: 10.3390/sym12121954.

[30] J. Yao et al., "Assessment of Landslide Susceptibility Combining Deep Learning with Semi-Supervised Learning in Jiaohe County, Jilin Province, China", Applied Sciences, vol. 10, no. 16, p. 5640, 2020. Available: 10.3390/app10165640.

[31] T. Xiao, K. Yin, T. Yao and S. Liu, "Spatial prediction of landslide susceptibility using GIS-based statistical and machine learning models in Wanzhou County, Three Gorges Reservoir, China", Acta Geochimica, vol. 38, no. 5, pp. 654-669, 2019. Available: 10.1007/s11631-019-00341-1.

[32] Y. Wang, D. Sun, H. Wen, H. Zhang and F. Zhang, "Comparison of Random Forest Model and Frequency Ratio Model for Landslide Susceptibility Mapping (LSM) in Yunyang County (Chongqing, China)", International Journal of Environmental Research and Public Health, vol. 17, no. 12, p. 4206, 2020. Available: 10.3390/ijerph17124206.

[33] R. Tang, P. Kulatilake, E. Yan and J. Cai, "Evaluating landslide susceptibility based on cluster analysis, probabilistic methods, and artificial neural networks", Bulletin of Engineering Geology and the Environment, vol. 79, no. 5, pp. 2235-2254, 2020. Available: 10.1007/s10064-019-01684-y.

[34] S. Park and J. Kim, "Landslide Susceptibility Mapping Based on Random Forest and Boosted Regression Tree Models, and a Comparison of Their Performance", Applied Sciences, vol. 9, no. 5, p. 942, 2019. Available: 10.3390/app9050942.

[35] Weerasinghe, Vithanage Primali. (2015). Improving existing landslide hazard zonation map in KMC area, Sri Lanka. SUAN SUNANDHA SCIENCE AND TECHNOLOGY JOURNAL. Volume 02. 24-29.

[36] Landslide susceptibility assessment using statistical models: A case study in Badulla district, Sri Lanka G.J.M.S.R. Jayasinghe1*, P. Wijekoon1 and J. Gunatilake2

[37] R.S.S.Rangali, N.L.Deshapriya, LalSamarakoon HasaliHemasinghe, "Landslide susceptibility mapping using logistic regression model (a case study

in Badulla District, Sri Lanka)," Procedia Engineering, vol. 2012, pp. 1046-1053, 2018.

[38] Bandara, Akila & Hettiarachchi, Yashodha & Hettiarachchi, Kusal & Munasinghe, Sidath & Wijesinghe, Ishara & Thayasivam, Uthayasanker. (2019). A Generalized Ensemble Machine Learning Approach for Landslide Susceptibility Modeling. 10.1007/978-981-13-9364-8_6.

[39] Liyanage, Lasini & Weerakoon, Oshani & Palliyaguru, Sachini & Wimalaratne, Prasad. (2019). Towards Prediction of Landslide Susceptibility using Random Forest for Kalutara District, Sri Lanka. 216-221. 10.1109/R10-HTC47129.2019.9042450.

[40] Lee, S., Digna, G., &Evangelist, (2005). Landslide Susceptibility Mapping Using Probability and Statistical Models in Baguio City, Philippines. In Proceedings of 31st International Symposium on Remote Sensing of Environment, June 20-24, 2005, Saint Petersburg, Russia.

[41] Chen, Wei & Sun, Zenghui & Han, Jichang. (2019). Landslide Susceptibility Modeling Using Integrated Ensemble Weights of Evidence with Logistic Regression and Random Forest Models. Applied Sciences. 9. 171. 10.3390/app9010171.

[42] Truong, Xuan-Luan & Mitamura, Muneki & Kon, Yasuyki & Raghavan, Venkatesh & Yonezaw, Go & Truong, Xuan & Thi Do, Hang & Tien Bui, Dieu & Lee, Saro. (2018). Enhancing Prediction Performance of Landslide Susceptibility Model Using Hybrid Machine Learning Approach of Bagging Ensemble and Logistic Model Tree. Applied Sciences. 8.

[43] S. Herath, Y. (2008). Landslides - Disaster Risk Reduction.

[44] CLIMATE-DATA.ORG. (n.d.). Retrieved January 2019, from https://en.climate-data.org/asia/sri- lanka

[45] “Wikipedia, the free encyclopedia", Wikipedia, 2021. [Online]. Available: https://en.wikipedia.org/wiki. [Accessed: 13- Jan- 2021].

[46] J. Dou et al., "Optimization of Causative Factors for Landslide Susceptibility Evaluation Using Remote Sensing and GIS Data in Parts of Niigata, Japan", PLOS ONE, vol. 10, no. 7, p. e0133262, 2015. Available: 10.1371/journal.pone.0133262 [Accessed 12 November 2020].

[47] Fernandez-Merodo JA, Mulas J, Pastor M, Luzi G, Monserrat O. Herrera G, "A landslide forecasting model using ground based SAR data: the Portalet case study," *Engineering Geology*, vol. 105, no. 3-4, pp. 220-230, May 2009.

[48] Karpatne A, Liess S (2015) A guide to earth science data: summary and research challenges. Comput Sci Eng 17:14–18. https://doi.org/10.1"109/MCSE.2015.127

[49] Bzdok D, Krzywinski M, Altman N (2018) Machine learning: supervised methods. Nat Methods 15:5–6. https://doi.org/10. 1038/nmeth.4551

[50] Dietterich T (2000) Ensemble methods in machine learning. pp 1–15 , Rokach L (2010) Ensemble-based classifiers. Artif Intell Rev 33:1–39. https://doi.org/10.1007/s10462-009-9124-7

[51] Nagpal, A., 2017. Decision Tree Ensembles- Bagging and Boosting. [online] Medium. Available at: <https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9> [Accessed 22 January 2021].

[52] Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[53] Micheletti et al., 2014; Pourghasemi and Kerle, 2016; Rahmati et al., 2018; Zhang et al., 2019

[54] Micheletti, N., Foresti, L., Robert, S. *et al.* Machine Learning Feature Selection Methods for Landslide Susceptibility Mapping. *Math Geosci* **46,** 33–57 (2014). https://doi.org/10.1007/s11004-013-9511-0

[55] Pedre Nagpal, A., 2017. *Decision Tree Ensembles- Bagging and Boosting*. [online] Medium. Available at: <https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9> [Accessed 22 January 2021].

[56] Foley, M., 2021. *8.3 Bagged Trees | My Data Science Notes*. [online] Bookdown.org. Available at: <https://bookdown.org/mpfoley1973/data-sci/bagged-trees.html> [Accessed 15 January 2021].

[57] Chen, T. and Guestrin, C., 2021. *XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [online] Dl.acm.org. Available at: <https://dl.acm.org/doi/10.1145/2939672.2939785> [Accessed 30 January 2021].

[58] C. Python, "XGBoost Parameters | XGBoost Parameter Tuning", *Analytics Vidhya*, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/. [Accessed: 31- Mar- 2021].

[59] Freund, Y.; Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 1997, 55, 119–139. [CrossRef]

[60] "Gradient boosting - Wikipedia", En.wikipedia.org, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Gradient_boosting. [Accessed: 25- Jan- 2021].

[61] I. Dabbura, "K-means clustering: Algorithm, applications, evaluation methods, and drawbacks," Towards Data Science, 17-Sep-2018. [Online]. Available: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a. [Accessed: 04-Apr-2021].

[62] Chen, T. and Guestrin, C., 2021. *XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [online] Dl.acm.org. Available at: <https://dl.acm.org/doi/10.1145/2939672.2939785> [Accessed 30 January 2021].

[63] Wubalem, A., 2020. Modeling of Landslide susceptibility in a part of Abay Basin, northwestern Ethiopia. *Open Geosciences*, 12(1), pp.1440-1467.

[64] Assessment of Landslide Susceptibility using Weight of Evidence and Frequency Ratio Model in Shahpur Valley, Eastern Hindu Kush July 2020 DOI: 10.5194/nhess-2020-167

[65] Mersha, T., Meten, M. GIS-based landslide susceptibility mapping and assessment using bivariate statistical methods in Simada area, northwestern Ethiopia. Geoenviron Disasters 7, 20 (2020). https://doi.org/10.1186/s40677-020-00155-x

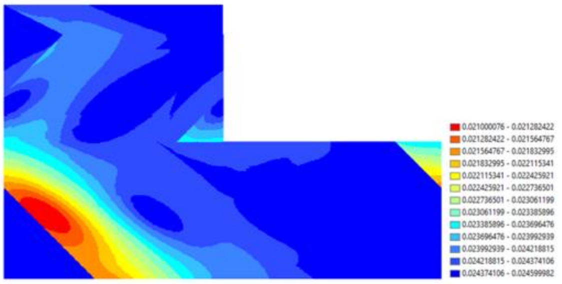
[66] Zhice Fang, Yi Wang, Ling Peng & Haoyuan Hong (2021) A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping, International Journal of Geographical Information Science, 35:2, 321-347, DOI: 10.1080/13658816.2020.1808897

[67] Chen, W. *et al*. Landslide Susceptibility Modeling Based on GIS and Novel Bagging-Based Kernel Logistic Regression. *Appl*. *Sci*, https://doi.org/10.3390/app8122540 (2018)

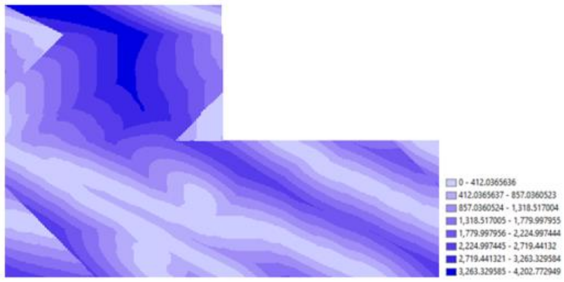
[68] "Leaflet — an open-source JavaScript library for interactive maps", Leafletjs.com, 2021. [Online]. Available: https://leafletjs.com/. [Accessed: 30- Dec-2020].

[69] "CesiumJS", Cesium, 2021. [Online]. Available: https://cesium.com/cesiumjs/. [Accessed: 27- Dec- 2020].

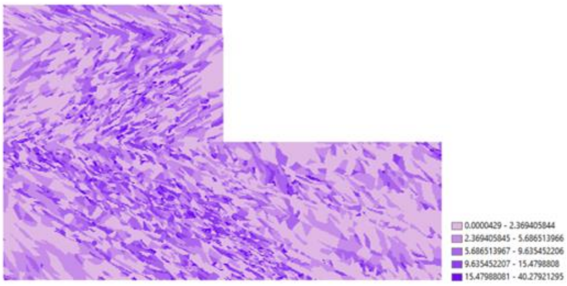
APPENDIX A: GIS FEATURE MAPS



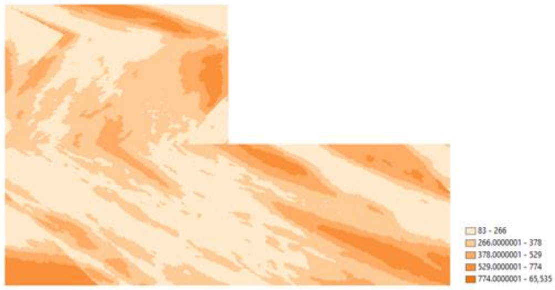
(a)



(b)

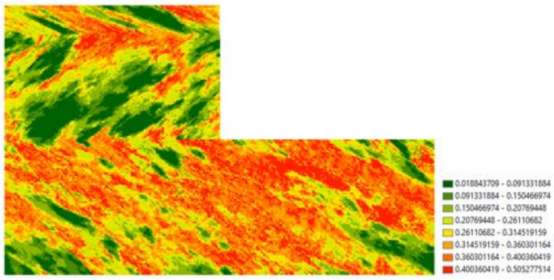


(c)

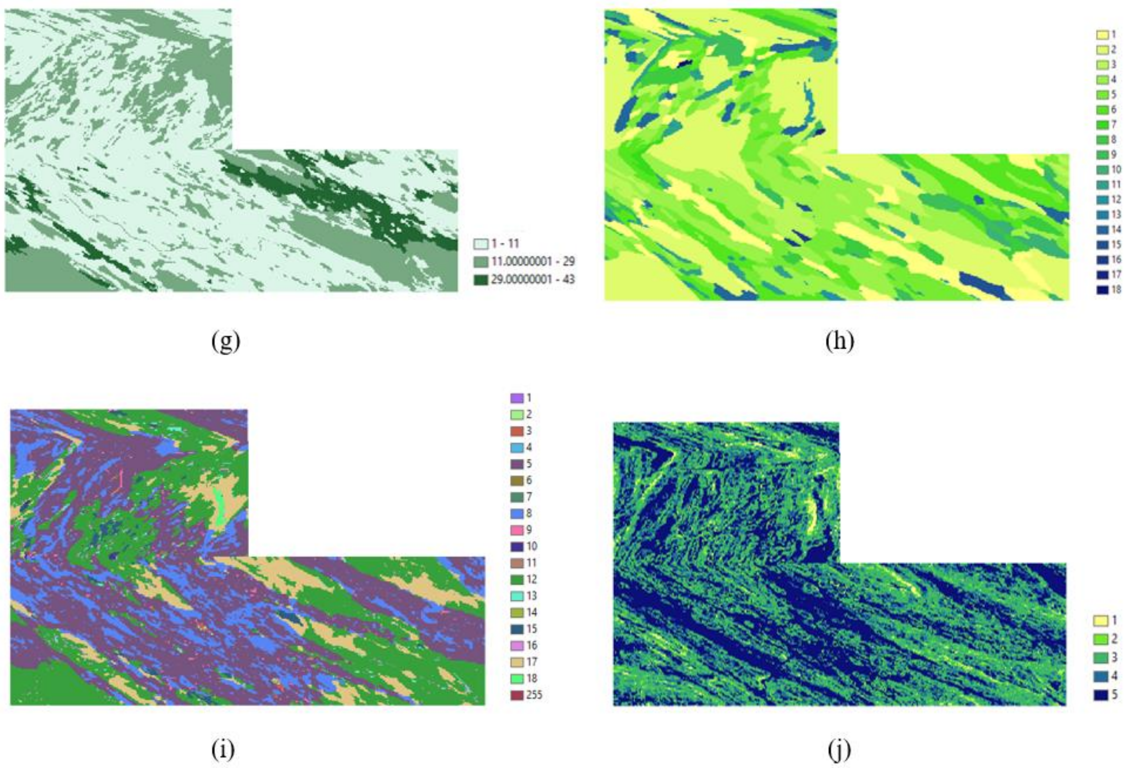


(d)

(e)



(f)



Landslide impact factors. (a) Rainfall (b) Distance to water streams (c) Drainage density (d) Elevation (e) Catchment (f) Vegetation (g) Land use (h) Overburden (i) Landform (j) Slope

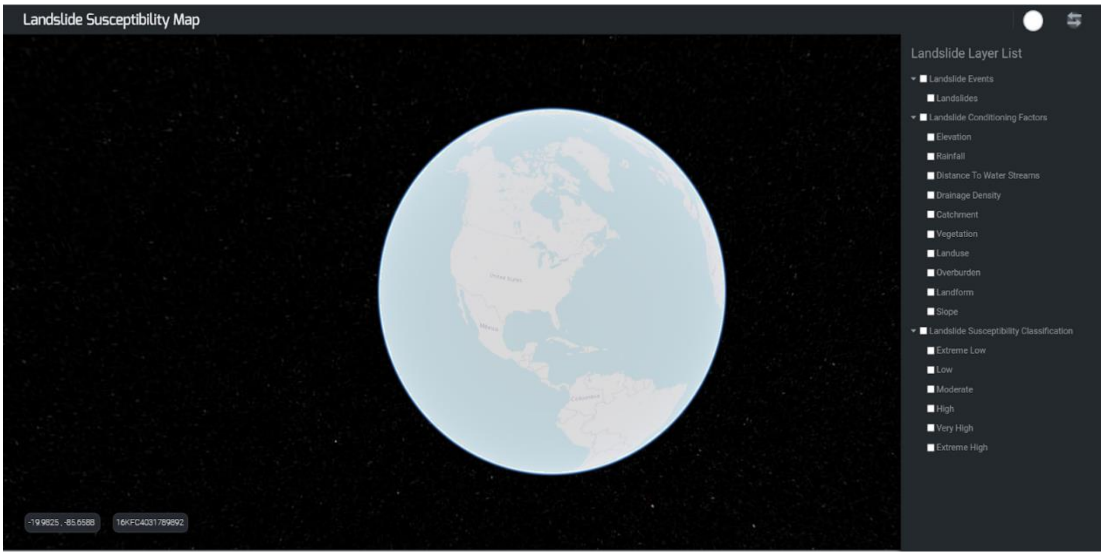
APPENDIX B: THE CORRELATION BETWEEN THE LANDSLIDE LOCATION AND LANDSLIDE CONDITIONING FACTOR CLASSES

Factors	Class	Nslpix	a = Nslpix/ΣNslpix	Ncpix	b = Ncpix/ΣNcpix	FR=a/b
Rainfall	0.021-0.0213	0	0	1268	0.0108	0.0
	0.0213-0.0216	0	0	1247	0.0106	0.0
	0.0216-0.0218	0	0	20431	0.0170	0.0
	0.0218-0.0221	0	0	15609	0.0130	0.0
	0.0221-0.0224	0	0	15059	0.0125	0.0
	0.0224-0.0227	232	0.0432	15570	0.0130	3.3303
	0.0227-0.0230	330	0.0615	18452	0.0154	3.9972
	0.0230-0.0234	27	0.0050	22879	0.0191	0.2638
	0.0234-0.0237	95	0.0177	29611	0.0247	0.7170
	0.0237-0.0240	255	0.0475	39069	0.0326	1.4588
	0.0240-0.0242	87	0.0162	116874	0.0974	0.4664
	0.0242-0.0244	1037	0.1931	273432	0.2279	0.8477
	0.0244-0.0246	3306	0.6158	607299	0.5061	1.2167
Distance to streams	0-395.5551	402	0.0749	231575	0.1930	0.3880
	395.5551-824.0731	807	0.1503	211507	0.1763	0.8528
	824.0731-1252.5911	2802	0.5219	174722	0.1456	3.5843
	1252.5911-1697.5906	747	0.1391	170107	0.1418	0.9815
	1697.5906-2142.5901	16	0.0030	161185	0.1343	0.0222
	2142.5901-2637.0340	143	0.0266	126272	0.1052	0.2531

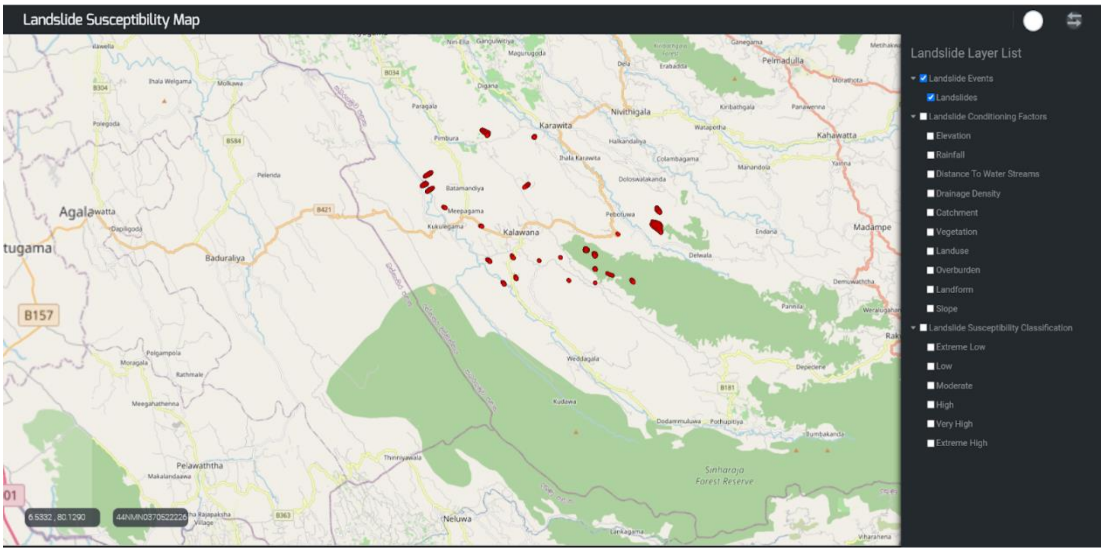
	2637.0340-3197.4037	0	0	76022	0.0634	0
	3197.4037-4202.7730	452	0.0842	48610	0.0405	2.0782
Drainage density	0.00004-2.3605	1867	0.3477	531951	0.4433	0.7844
	2.3605-5.9012	2125	0.3958	343938	0.2866	1.3809
	5.9012-9.9139	1219	0.2270	214087	0.1784	1.2726
	9.9139-15.8151	123	0.0229	92259	0.0769	0.2980
	15.8151-60.1916	35	0.0065	17294	0.0144	0.4523
Elevation	81-221	43	0.0080	193566	0.1613	0.0497
	221-301	1934	0.3602	499080	0.4159	0.8661
	301-404	2503	0.4662	277685	0.2314	2.0146
	404-545	889	0.1656	151877	0.12656	1.3083
	545-65535	0	0	77792	0.0648	0
Catchment	0.0-0.07	4258	0.7931	928272	0.7736	1.0252
	0.07-0.2	847	0.1578	250962	0.2091	0.7543
	0.2-0.5	264	0.0492	20766	0.0173	2.8414
Vegetation	0.0188-0.0923	388	0.0723	107368	0.0895	0.8077
	0.0923-0.1503	152	0.0283	100668	0.0839	0.3375
	0.1503-0.2063	416	0.0775	124095	0.1034	0.7493
	0.2063-0.2604	772	0.1345	156252	0.1302	1.0328
	0.2604-0.3126	526	0.0980	128227	0.1069	0.9168
	0.3126-0.3590	1379	0.2568	190197	0.1585	1.6205
	0.3590-0.3996	1266	0.2358	251400	0.2095	1.1255
	0.3996-0.5117	520	0.0969	141793	0.1182	0.8197
	1-12	2416	0.4450	720076	0.6001	0.7499

Land Use	12-30	1657	0.3086	391677	0.3264	0.9455
	30-255	1296	0.2414	88247	0.0735	3.2824
Overburden	1	992	0.184764	110075	0.091729	2.014238
	2	4	0.000745	359981	0.299984	0.002484
	3	0	0	48972	0.040810	0
	4	213	0.039672	242875	0.202396	0.196013
	5	118	0.021978	134587	0.112156	0.195960
	6	2108	0.392624	110532	0.092110	4.262559
	7	0	0	4062	0.003385	0
	8	260	0.048426	45075	0.037562	1.289215
	9	449	0.083628	34453	0.028711	2.912776
	10	4	0.000745	24226	0.020188	0.036903
	11	7	0.001304	32305	0.026921	0.048430
	12	6	0.001118	10251	0.008542	0.130820
	13	0	0	239	0.000199	0
	14	1091	0.203204	31524	0.026270	7.735195
	15	117	0.021792	7877	0.006564	3.319807
	16	0	0	1693	0.001411	0
	17	0	0	533	0.000444	0
	18	0	0	740	0.000617	0
Landform	1-7	1214	0.226113	502064	0.418387	0.540440
	7-11	652	0.121438	184735	0.153946	0.788835
	11-15	1826	0.340101	370436	0.308697	1.101731
	15-255	1677	0.312349	142765	0.118971	2.625422
Slope	1	330	0.061464	45603	0.038003	1.617366
	2	730	0.135966	97124	0.080937	1.679903
	3	2797	0.520954	430886	0.359072	1.450835
	4	836	0.155709	201595	0.167996	0.926860
	5-255	676	0.125908	424792	0.353993	0.355679

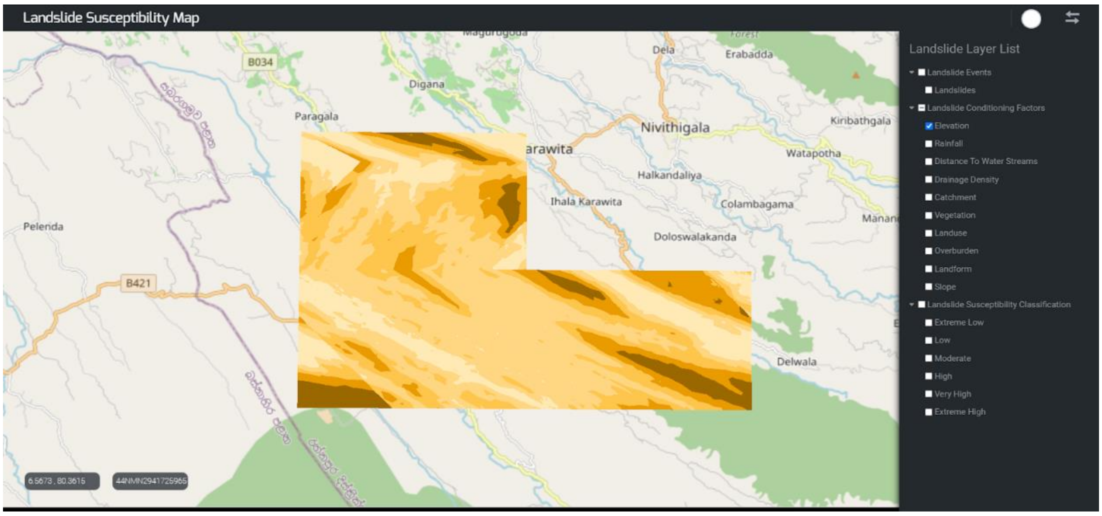
APPENDIX C: SCREENSHOTS OF THE APPLICATION



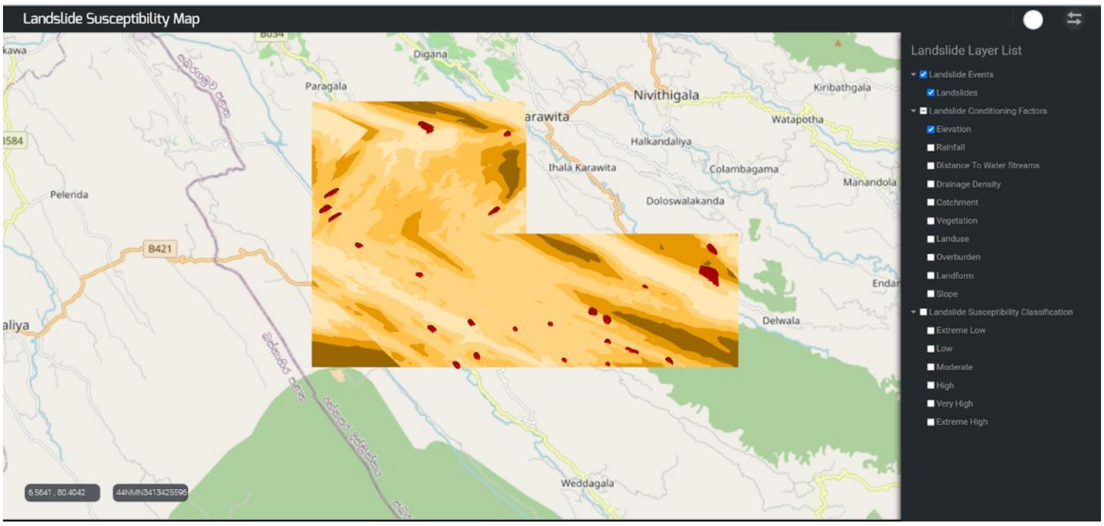
Home page of web application



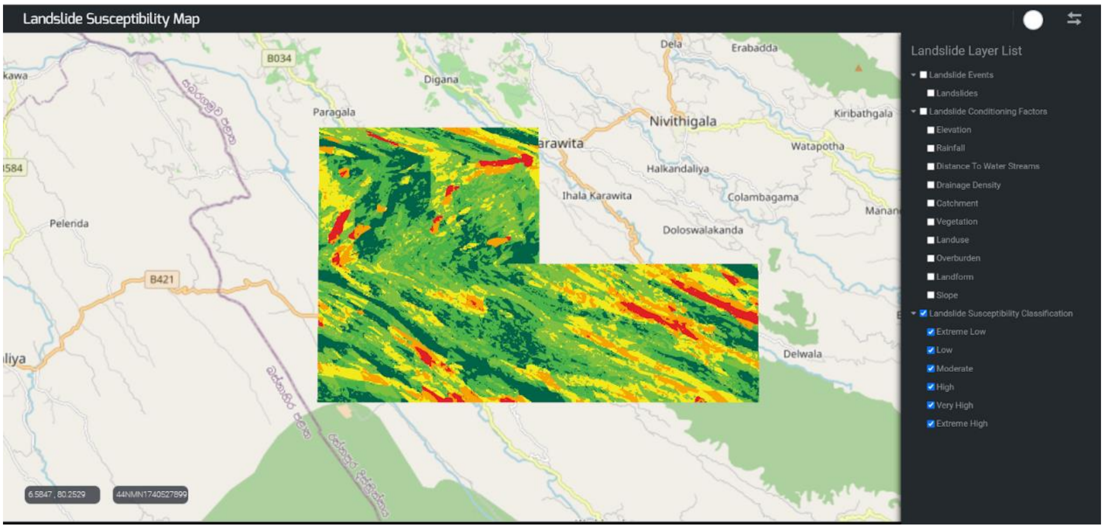
Visualizing landslide locations on the globe



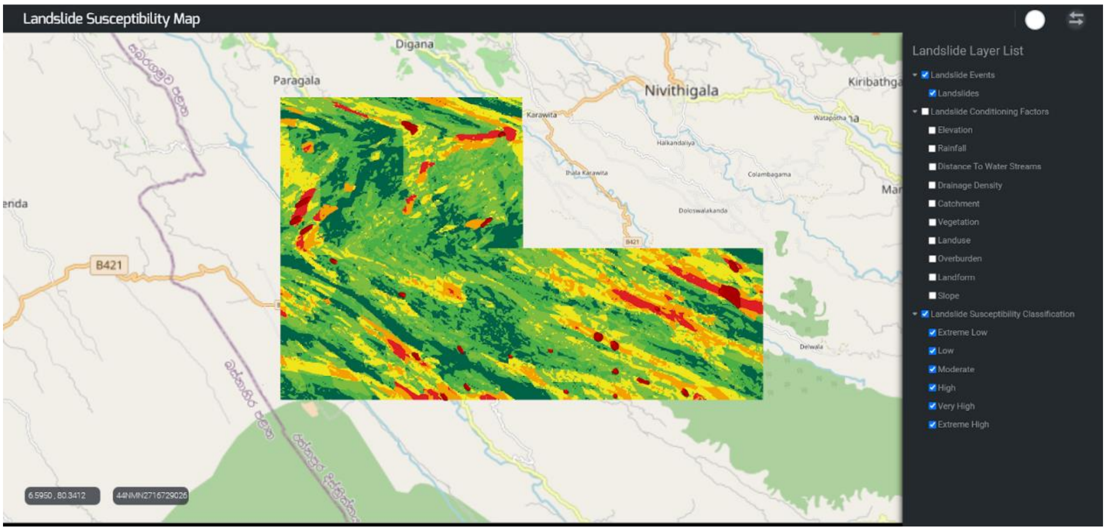
Visualizing Elevation feature map



Visualizing elevation map with the landslide locations



Visualizing landslide susceptibility map on the application



Visualizing LSM and the landslide locations