Landslide Susceptibility Prediction Model using Random Forest for Kalutara District, Sri Lanka

Authors:

L.C. Liyanage 15020381

S.T. Palliyaguru 15020452

O.S. Weerakoon 15020843

Supervisor:

Dr. G.D.S.P. Wimalaratne

This dissertation is submitted to the University of Colombo School of Computing in partial fulfillment of the requirements for the Degree of Bachelor of Science Honours in Information Systems

University of Colombo School of Computing

35, Reid Avenue, Colombo 07

Sri Lanka

February 2020

Declaration

I, L.C. Liyanage(15020381), hereby certify that this dissertation entitled Landslide Susceptibility Prediction Model using Random Forest for Kalutara District, Sri Lanka is entirely my own work and it has never been submitted nor is currently been submitted for any other degree.

Signature:

Date: February 20, 2020

I, S.T.Palliyaguru(15020452), hereby certify that this dissertation entitled Landslide Susceptibility Prediction Model using Random Forest for Kalutara District, Sri Lanka is entirely my own work and it has never been submitted nor is currently been submitted for any other degree.

Signature:

Date: February 20, 2020

I, O.S.Weerakoon(15020843), hereby certify that this dissertation entitled Landslide Susceptibility Prediction Model using Random Forest for Kalutara District, Sri Lanka is entirely my own work and it has never been submitted nor is currently been submitted for any other degree.

Signature:

Date: February 20, 2020

I, G.D.S.P. Wimalaratne, certify that I supervised this dissertation entitled: Landslide Susceptibility Prediction Model using Random Forest for Kalutara District, Sri Lanka conducted by L.C. Liyanage, S.T.Palliyaguru and O.S.Weerakoon in partial fulfillment of the requirements for the degree of Bachelor of Science Honours in Information Systems.

Signature:

Date: February 20, 2020

Acknowledgement

We would like to take this opportunity to express our gratitude to the people who have lent their support all along this dissertation and thus enabled us to concentrate on our work.
First and foremost, we would like to extend our special thanks to our research supervisor, Dr. G.D.S.P. Wimalaratne for his continuous guidance, support and insightful comments which motivated us to improve and move our research further. We are grateful for our lecturer-in-charge, Dr. Thushani Weerasinghe for constantly providing the guidance and assistance to carry out this research successfully through lectures and meetings.
We would also like to express our sincere gratitude to Ms. Hasali Hemasinghe, Research Scientist at National Building Research Organization(NBRO) and Meteorologists at Department of Meteorology for providing us with the data required for this research.
Also, we would like to thank all our colleagues and friends for all their help, support, interest and valuable advice and input. Finally, we would like to thank all others whose names are not listed particularly but have given their support in many ways and encouraged me to make this research a success.

Abstract

Landslides are one of the most recurrent and prominent natural disasters in Sri Lanka. An area of nearly 20,000 sq. km encompassing 10 districts is prone to landslides. According to statistics provided by the National Building Research Organization landslides have destroyed over 800 lives in Sri Lanka over the last decade. In 2017 Kalutara district reported the maximum number of deaths of 101 due to landslides. Owing to haphazard, unplanned land use, inappropriate construction methods, wanton human intervention and other other geological and morphological causes, the trend of landslide occurrence will continue in the next eras. Therefore prediction of landslide susceptibility is indispensable for disaster management and ensure sustainability of developments.

The main focus of this study was to investigate the applicability of 12 landslide conditioning factors including slope, aspect, hydrology, Stream Power Index(SPI), Topographic Wetness Index(TWI), Sediment Transport Index(STI), geology, land form, land use, soil type, soil thickness and rainfall in the prediction of landslide susceptibility in Kalutara district using Random Forest machine learning algorithm. In order to achieve this a Geographical Information System(GIS) was used to manipulate and analyze the spatial data while the implementation of the prediction model was carried out using python.

A pilot study was carried out to analyze the correlation between the landslide conditioning factors and landslide occurrence and to select the most appropriate set of conditioning factors for the prediction. A landslide inventory of 84 landslides occurrences in Kalutara district, was utilized along with randomly generated 84 non-landslide locations from the landslide-free area of Kalutara district. Random Forest (RF), a non-parametric supervised classification algorithm was employed to construct the prediction model. The efficiency of the Random Forest model was evaluated using Receiver Operating Characteristic(ROC), accuracy, sensitivity and specificity. The results indicated 76.92% specificity, 84.00% specificity, and accuracy of 80.39%. The area under the ROC curve demonstrated 79.46% of predictive capability for the model.

Keywords: Landslide Susceptibility, Machine Learning, GIS, Random Forest

Contents

Declaration i

Acknowledgements iii

Abstract iv

Contents vii

List of Figures ix

List of Tables x

List of Code Listings xi

Acronyms xi

1 Introduction 1

1.1 Problem Statement 2

1.2 Motivation 3

1.3 Research Questions 3

1.4 Aims and Objectives 4

1.4.1 Aim 4

1.4.2 Objectives 4

1.5 Significance of the Study 5

1.6 Outline of Research Methodology 6

1.7 Scope 7

1.8 Delimitations 7

1.9 Contribution 8

1.10 Structure of the Thesis 8

2 Background **9**

2.1 What is a Landslide? . . . 9

2.2 Impact of Landslides in Sri Lanka . . . 11

2.3 Contributing Factors of Landslides in Sri Lanka . . . 13

2.4 Landslide Evaluation Tools and Technologies . . . 15

2.4.1 Mapping . . . 15

2.4.2 Remote Sensing . . . 16

2.4.3 Real-time Monitoring . . . 17

2.5 Landslide Susceptibility Zonation Maps . . . 17

2.6 Review of Similar Research . . . 19

2.7 Summary . . . 30

3 Methodology and Design **31**

3.1 Overview of the Methodology . . . 31

3.1.1 Landslide Inventory Mapping and Preparation of Training and Test Data Sets . . . 32

3.1.2 Preparation of Landslide Conditioning Factor Maps . . . 32

3.1.3 Correlation Analysis between Landslides and Conditioning Factors . . . 34

3.1.4 Selection of Conditioning Factors . . . 34

3.1.5 Construction of Landslide Susceptibility Prediction Model . . . 35

3.1.6 Evaluation of the Performance of the Model . . . 35

3.2 Design Assumptions . . . 37

3.3 Random Forest Algorithm . . . 37

3.3.1 Mathematical Foundation . . . 37

3.3.2 Bagging . . . 38

3.3.3 Random Feature Selection . . . 39

3.4 Summary . . . 41

4 Implementation **42**

4.1 QGIS . . . 42

4.2 Landslide Inventory . . . 43

4.3 Thematic Maps of Conditioning Factors . . . 44

4.3.1 Raster Maps . . . 44

4.3.2 Vector Maps . . . 48

4.4 Frequency Ratio and Information Gain Ratio Calculation 49
4.5 Feature Pool 50
4.6 Prediction Model 55
4.6.1 Preprocessing 55
4.6.2 Training and Testing 56
4.6.3 Model Assessment 57
4.7 Summary 58

5 Evaluation and Results **59**
5.1 Evaluation of the Thematic Maps 59
5.2 Pilot Study 70
5.2.1 Correlation Analysis 71
5.2.2 Attribute Relevance Analysis 74
5.3 Results of Landslide Susceptibility Prediction Model 75
5.4 Summary 81

6 Conclusion **82**
6.1 Future Work 83

Appendices **91**

A Prediction Model **92**

List of Figures

2.1 Classification of Causes of Landslides [20] 9
2.2 Landslide Research Approaches 11
2.3 Taxonomy of Landslide Susceptibility Mapping Methods 20

3.1 Process Flow of the Research 31
3.2 Boostrap Aggregation 39
3.3 Random Feature Selection in Random Forest 40

4.1 Landslide Inventory 44
4.2 Digital Elevation Model(DEM) 45
4.3 Slope, Aspect, SPI, TWI, and STI Thematic Maps 46
4.4 Rainfall Thematic Map 47
4.5 (a)Geology, (b)Hydrology, (c)Land Form, (d)Land Use, and (e)Soil Type and Thickness Thematic Maps 49
4.6 Attribute Table - Geology 51
4.7 Attribute Table - Slope 51
4.8 Attribute Table - Hydrology 52
4.9 Field Calculator for Slope 54
4.10 Final Dataset 55

5.1 Geology Thematic Map 60
5.2 Land Form Thematic Map 61
5.3 Land Use Thematic Map 62
5.4 Hydrology Thematic Map 63
5.5 Soil Type and Thickness Thematic Map 64
5.6 Slope Thematic Map 65
5.7 Aspect Thematic Map 66
5.8 SPI Thematic Map 67

5.9 TWI Thematic Map 68
5.10 STI Thematic Map 69
5.11 Rainfall Thematic Map 70
5.12 Confusion Matrix 78
5.13 Receiver Operating Characteristic 80
5.14 Landslide susceptibility map -Kalutara District 81

List of Tables

2.1 Deaths Due To Landslides in 2017 [20] 12
2.2 Criteria for Hazard Zonation [20] 18
2.3 Factor Weighing Scale [20] 18
2.4 Rainfall Thresholds [20] 19
2.5 Comparison of Results Obtained in Previous Research 29

3.1 Summary of Research Design 41

4.1 Reclassification of Slope, SPI, TWI, STI, Hydrology and Rainfall 53

5.1 Frequency Ratio values for Conditioning Factors 73
5.2 Information Gain Ratio for Conditioning Factors 75
5.3 Optimal hyper-parameters 78
5.4 Confusion matrix for prediction model 79

List of Code Snippets

4.1 Encode Data 55
4.2 Split Data 56
4.3 RF Classifier Hyper-Parameters 57
4.4 Model Fitting and Prediction 57
4.5 Performance Measures 58
5.1 RF hyper-parameters 77

Acronyms

FNN	Feedforward Neural Network
FR	Frequency Ratio
GDAL	Geospatial Data Abstraction Library
GIS	Geographical Information System
IGR	Information Gain Ratio
LiDAR	Light Detection and Ranging
LSTM	Long-Short Term Memory
MLP-NN	Multi-Layer Perceptron Neural Network
NBRO	National Building Research Organization
OGC	Open Geospatial Consortium
QGIS	Quantum GIS
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SPI	Stream Power Index
STI	Sediment Transport Index
SVM	Support Vector Machine
TWI	Topographic Wetness Index
WMS	Web Map Service
WFS	Web Feature Service
WPS	Web Processing Service

Chapter 1

Introduction

A landslide is defined as the movement of a mass of rock, debris, or earth down a slope [1]. Landslides occur throughout the world, under all climatic conditions and terrains, cost billions in monetary losses, and are responsible for thousands of deaths and injuries each year. Therefore the efforts in the prediction of landslide susceptibility have been a major concern and important aspect in disaster management and ensure the sustainability of developments in countries of the world.

The efficiency of the landslide susceptibility mapping and prediction methods depends on the method employed and the quality of the conditioning factors [2]. The influence of the conditioning factors varies from region to region based on varying environments in the area. Different studies carried out in different parts of the world have employed various types of quantitative and qualitative approaches in the prediction of landslides. Nevertheless, quantitative methods have become very popular in recent years [2].

This study aims to investigate the applicability of 12 conditioning factors including slope, aspect, hydrology, Stream Power Index(SPI), Topographic Wetness Index(TWI), Sediment Transport Index(STI), geology, landform, land use, soil type, soil thickness and rainfall in the prediction of landslide susceptibility in Kalutara district using Random Forest machine learning algorithm. The conditioning factors were selected based on the knowledge acquired through previous literature and consultation of geological scientists at the National Building Research Organization(NBRO). Further analysis of the conditioning factors was carried out using the Frequency Ratio and Information Gain Ratio to select the most appropriate conditioning factors for the study area. Out of three candidate machine learning algorithms identified through the literature, Random Forest was used to implementing the landslide susceptibility prediction model for the Kalutara district.

1.1 Problem Statement

Landslides are considered the third critical natural disaster worldwide [3] causing massive destruction of lives and property. Landslides are among the natural disasters frequently experienced in Sri Lanka. An area of nearly 20,000 sq. km encompassing 10 districts in Sri Lanka is prone to landslides. It is about 30% of Sri Lanka's land area and spread into several districts, namely, Badulla, Nuwara Eliya, Kegalle, Ratnapura, Kandy, Matale, Kaluthara, Mathara, Galle, and Hambantota. Among these 10 districts Kalutara district is among the districts having the highest frequency [4] of landslides. According to the National Building Research Organization (NBRO), Sri Lanka has faced more than 38 [4] landslides in 2017 causing an immense socio-economic impact [5]. In a personal interview with Dr. Gamini Jayatissa, Director General of NBRO, he stated that there have been 101 deaths in Kalutara district in 2017, which is the largest reported number of deaths in that year due to landslides. Considering the wide coverage of landslide damages, planners and decision-makers would need to identify landslide-prone areas to plan mitigation actions. Hence, landslide susceptibility prediction has become indispensable.

NBRO has prepared landslide zonation maps for the 10 landslide-prone districts employing six terrain causative factors of landslides including slope, geology, hydrology, landform, land use, soil type, and thickness. They have not taken into account other possible conditioning factors such as aspect, topographic wetness index, sediment transport index, rainfall, etc. in the implementation of these maps. Previous studies carried out in the domain shows the importance of considering both the terrain factors and triggering factors in the landslide prediction. Therefore unavailability of a proper prediction mechanism considering both the terrain factors and triggering factors for occurrence in landslides in Sri Lanka has adverse effects on the prevention, mitigation, and preparedness over landslide disasters in Sri Lanka.

With soft computing approaches for landslide susceptibility prediction becoming popular in the world, different models have been implemented using machine learning techniques in different regions of the world showing better accuracy than statistical approaches [6]. When investigating the usage of machine learning approaches to predict landslides in Sri Lanka, it can be identified that Sri Lanka still has not done a sufficient amount of research in the domain. Since the conditioning factors differ from place to place, the applicability of different conditioning factors differs in different regions [2]. Understanding the importance of landslide susceptibility prediction with high accuracy by employing a suitable machine learning algorithm in the Sri Lankan context is of utmost importance.

1.2 Motivation

As discussed above there have been many deaths reported in Kalutara district due to landslides. Therefore this study was focused on implementing a landslide susceptibility model for Kalutara district, Sri Lanka. The lack of studies carried out the landslide domain in Sri Lanka employing both terrain and triggering factors with machine learning techniques was highlighted as a timely concern when going through the previous literature.

When determining the landslide susceptibility, selecting the best set of conditioning factors and identifying the relationship between different conditioning factors and the landslides occurrence is crucial. The main motivation of this study is to investigate the applicability of 12 terrain and triggering conditioning factors identified for the study area and develop a prediction model using Random Forest for efficient identification of landslide susceptible areas in Kalutara District so that it would assist in issuing warnings and minimizing the possible damages to human lives.

1.3 Research Questions

The research was directed to predict landslide susceptibility in Kalutara district in Sri Lanka utilizing suitable landslide conditioning factors. With this intention, the main research question ***how to predict landslide susceptibility using a machine learning employing the data extracted from contour maps, geospatial statistical data, and precipitation data?*** was formulated. Identification of a suitable methodology will assist in the prediction of the occurrence of landslides with high accuracy to help better decision making and thus plan risk mitigation actions in the future. Landslides can be a result of a broad variety of landslide conditioning factors [7]. However, only certain classes of conditioning factors will have a considerable impact on landslide occurrence [8] in a given study area. The selection of the best set of conditioning factors for the study area is identified as crucial [9] as well. Therefore to address these, the sub research questions, ***how to determine the spatial relationship between landslide conditioning factors and landslide occurrence and how to eliminate landslide conditioning factors having low or null predictive capability in the given study area?*** are answered through this research.

Highly developed Geographic Information Systems together with mathematical and machine learning algorithms, have enabled effective landslide modeling [10]. The confusion on

which techniques or models will predict the landslide susceptibility with high accuracy [11] still prevails. A suitable high performing landslide susceptibility model is expected to demonstrate a rise in the prediction accuracy of about 1 or 2% [12] when compared to other models. Thus the sub research question ***what are the machine learning algorithms that can be used to predict landslide susceptibility?*** is answered through this research. The evaluation of the prediction model assists in the identification of the model's efficiency in the prediction of landslide susceptibility. To address this the sub research question ***how to evaluate the proposed approach and assess the accuracy of the proposed model?*** was formulated.

1.4 Aims and Objectives

1.4.1 Aim

To investigate and develop a model to predict landslide susceptibility in Kalutara District using a suitable machine learning technique.

1.4.2 Objectives

The objectives of this study include,

1. Study landslide prediction models implemented in other regions of the world and related literature.
2. Select landslide conditioning factors with the highest predictive capability and correlation to the occurrence of landslides in the given study area
3. Determine the most suitable machine learning techniques to predict the susceptibility of landslides.
4. Implement proof of concept to predict landslide susceptibility in the given study area using the selected advanced machine learning technique.
5. To evaluate the success and failure of the implemented prediction model for landslide occurrence in the given study area.

1.5 Significance of the Study

Many studies concerning the prediction of landslides using machine learning techniques have been carried out in countries such as China, India, Iran, etc. But there is a lack of studies carried out in the prediction of landslides using machine learning techniques in the Sri Lankan context. Among the few types of research carried out in the domain include landslide susceptibility mapping using Logistic Regression model for Badulla District, Sri Lanka [13], predicting landslides in hill country using Decision Trees and Artificial Neural Networks [14] and ensemble approach based on Support Vector Machine (SVM), Naïve Bayes model for landslide prediction in Ratnapura District [15]. These studies possess capabilities to predict landslides incorporating only three or four causative factors of landslides. There are several conditioning factors of landslides including slope, aspect, stream power index, lithology, bedding structure, etc. which have not been considered in these studies. According to several studies [16], [25], [17] done in this domain it can be identified that it is important to select the most appropriate conditioning factors for a study eliminating factors with low or non-predictive capability in predicting landslides in the study area. Elimination of these factors assists in coming up with a model with better predictive capabilities. But it has not been considered in studies done in the Sri Lankan context.

Comparative studies [6], [18] carried out in other regions of the world using machine learning techniques and statistical approaches for prediction of landslides have emphasized that machine learning approaches give better performance than statistical approaches. Benefits of using advanced machine learning approaches such as Artificial Neural Networks, Naive Bayes, Radial Basis Classifier, Random Forest, Decision Trees, etc. are not sufficiently reaped by the minimal set of research carried out in prediction of landslides in Sri Lanka.

The landslide hazard zonation maps prepared by the NBRO also merely focuses on identifying the landslide susceptible zones in different districts using only six terrain conditioning factors and provide warnings to the public based on the rain gauge readings in rainy seasons considering a threshold rainfall value. This methodology does not employ both the terrain and triggering factors in the construction of the susceptibility maps. To provide a more reliable prediction of landslides all the influencing conditioning factors should be considered which is lacking in the current methodology adopted by the NBRO.

1.6 Outline of Research Methodology

Many studies concerning the prediction of landslides using machine learning techniques have been carried out in countries such as China, India, Iran, etc. But there is a lack of studies carried out in the prediction of landslides using machine learning techniques in the Sri Lankan context. Among the few types of research carried out in the domain include landslide susceptibility mapping using Logistic Regression model for Badulla District, Sri Lanka [13], predicting landslides in hill country using Decision Trees and Artificial Neural Networks [14] and ensemble approach based on Support Vector Machine (SVM), Naïve Bayes model for landslide prediction in Ratnapura District [15]. These studies possess capabilities to predict landslides incorporating only three or four causative factors of landslides. There are several conditioning factors of landslides including slope, aspect, stream power index, lithology, bedding structure, etc. which have not been considered in these studies. According to several studies [16], [25], [17] done in this domain it can be identified that it is important to select the most appropriate conditioning factors for a study eliminating factors with low or non-predictive capability in predicting landslides in the study area. Elimination of these factors assists in coming up with a model with better predictive capabilities. But it has not been considered in studies done in the Sri Lankan context.

Comparative studies [6], [18] carried out in other regions of the world using machine learning techniques and statistical approaches for prediction of landslides have emphasized that machine learning approaches give better performance than statistical approaches. Benefits of using advanced machine learning approaches such as Artificial Neural Networks, Naive Bayes, Radial Basis Classifier, Random Forest, Decision Trees, etc. are not sufficiently reaped by the minimal set of research carried out in prediction of landslides in Sri Lanka.

The landslide hazard zonation maps prepared by the NBRO also merely focuses on identifying the landslide susceptible zones in different districts using only six terrain conditioning factors and provide warnings to the public based on the rain gauge readings in rainy seasons considering a threshold rainfall value. This methodology does not employ both the terrain and triggering factors in the construction of the susceptibility maps. To provide a more reliable prediction of landslides all the influencing conditioning factors should be considered which is lacking in the current methodology adopted by the NBRO.

1.7 Scope

The scope of this study include,

1. Investigate the machine learning algorithms that can be used in the prediction of landslides in Kalutara district
2. Identification of the most appropriate conditioning factors to predict landslides in Kalutara district
3. Implement a model to predict landslide susceptibility in Kalutara district using a machine learning algorithm
4. Evaluate the final model with real landslide occurrences in Kalutara district to determine the accuracy

One of the main concerns identified for landslide zonation maps prepared by the NBRO was that they only consider the terrain factors in generating these maps using GIS tools. Therefore this study focuses on incorporating both terrain and triggering factors of landslides in the prediction of the landslide susceptibility using machine learning.

Contour maps, on-field data related to soil type, soil thickness, land use, landform, hydrology were obtained from NBRO while rainfall data were obtained from the Meteorology Department of Sri Lanka.

1.8 Delimitations

The study was focused only on prediction landslides and do not accommodate the prediction of other environmental hazards such as floods and earthquakes etc. Only contour data, geospatial statistical data, and precipitation data were utilized in the implementation of the model.

The model was implemented considering the initiation area of landslide and it did not take the entire displacement area of the landslide into consideration. Prediction of landslide susceptibility was carried out only for Kalutara District, Sri Lanka and was not focused on any other districts.

1.9 Contribution

An existing knowledge gap was filled by this research through providing the knowledge on the applicability of 12 conditioning factors of landslides including slope, aspect, hydrology, Stream Power Index(SPI), Topographic Wetness Index(TWI), Sediment Transport Index(STI), geology, landform, land use, soil type, soil thickness and rainfall for landslide susceptibility prediction in Kalutara district to the research community.

To the best of our knowledge, this is the first landslide susceptibility prediction model implemented for Kalutara District, Sri Lanka employing Random Forest as the machine learning technique and taking a total of 12 landslide conditioning factors into consideration.

Further, one publication has been made to the research community as of the writing of this thesis.

1. "Towards Prediction of Landslide Susceptibility using Random Forest for Kalutara District, Sri Lanka" presented at 2019 IEEE R10 Humanitarian Technology Conference(HTC), Depok, Indonesia (R10'HTC).

The following deliverables have been made to the research community for future research in this area.

1. A landslide inventory containing a detailed register of the distribution and characteristics of past landslides occurred in Kalutara District from 1984 to 2018 with regard to 12 conditioning factors.

2. Proof of concept implementation to predict landslide susceptibility in Kalutara District and an information system to visualize the susceptibility results obtained from the prediction model.

3. Knowledge on the success or failure of the landslide susceptibility prediction model constructed using Random Forest for Kalutara District.

1.10 Structure of the Thesis

The related literature surrounding the problem domain is studied and analyzed in Chapter 2. The design of research architecture and assumptions are included in Chapter 3. Chapter 4 describes the implementation process undertaken in the study and Chapter 5 describes the experimental protocol, experimentation process, and the results gained in the study. Finally, Chapter 6 concludes the research by providing the conclusion and future works.

Chapter 2

Background

2.1 What is a Landslide?

Landslide or mass movement is a phenomenon of denudation process, whereby soil, rock or debris is displaced along the slope by mainly gravitational forces [1]. Research carried out by the National Building Research Organisation(NBRO) indicates that haphazard and unplanned land use, inappropriate construction methods and wanton human intervention have to lead to an increase [19] in landslide susceptibility while other geological and morphological causes also influence the occurrence of landslides. Figure 2.1 illustrates a classification of causes of landslides [20].

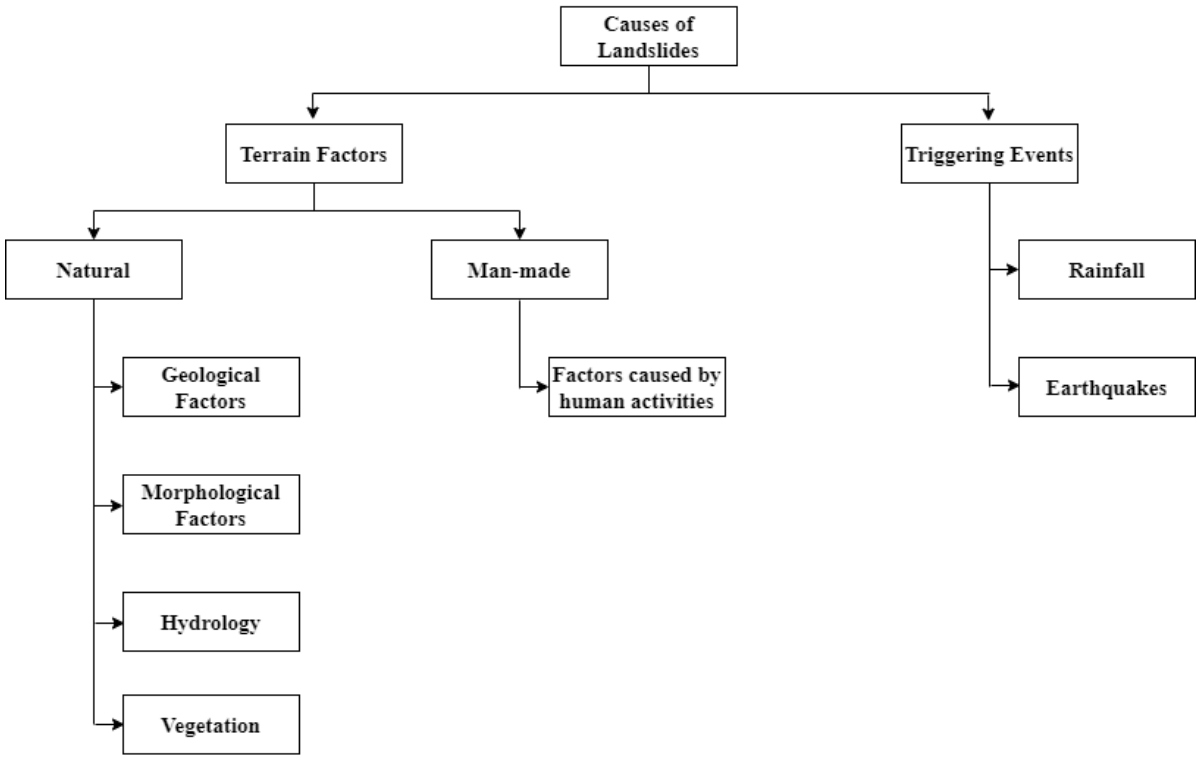


Figure 2.1: Classification of Causes of Landslides [20]

According to the U.S. Geological Survey(USGS), a landslide encompasses five modes of slope movement [1] including falls, topples, slides, spreads, and flows. A fall begins with the detachment of soil or rock, or both, from a steep slope along a surface on which little or no shear displacement has occurred. A topple is recognized as the forward rotation out of a slope of a mass of soil or rock around a point or axis below the center of gravity of the displaced mass. A slide is a down-slope movement of a soil or rock mass occurring on surfaces of rupture or relatively thin zones of intense shear strain. Spread is an extension of cohesive soil or rock mass combined with the general subsidence of the fractured mass of cohesive material into the softer underlying material. A flow is a spatially continuous movement in which the surfaces of shear are short-lived, closely spaced, and usually not preserved.

A landslide can be a movement of either a sloping mass or the crest or the foot of a hill or even the cut surface of a slope. Similarly, the material that flows down can also vary according to circumstances. It could be a sliding huge soil mass at one time or a giant mudslide at another. It may also be an instance of a falling mixture of rock and soil down a slope. At times, it is possible for a large boulder resting unstably on higher ground to fall down a slope. As such, a landslide can mean differently depending on the circumstances and conditions.

Out of the different categories of landslides such as Rock Slides, Earth Flows, Debris Slides, Debris Flows, and Rock Falls; Debris Flows and Rock Falls occur in Sri Lanka [21]. Rock Falls occur when rock material on a higher elevation falls freely as fragments, splinters, etc and Debris Flows occur on escarpments with a very rapid downward flow of muddy water and soil, stone, as well as clay and gravel. Landslides can occur almost anywhere on the land from sloping terrain, valleys to even plains - even the seabed can be subjected. However, it is usually believed that they commonly occur on hill slopes at an inclination ranging from 15° to 45° [19] to the horizontal. Although landslides are common at inclinations below 45°, occurrences at inclinations above 45° are seldom for the obvious reason that soil layers will not accumulate on such surfaces for sliding at such angles. Rockfalls may occur in such areas instead, but it has been observed that the only terrain unduly tampered is subjected to such landslides [19].

The methods and techniques that can be employed in the landslide research vary depending on the use of simple expert knowledge to sophisticated mathematical procedures. These techniques can be divided into as physically based and statistics-based correlation analysis [9]. An outline of the physically-based methods and statistical methods [22], [23] are given in figure 2.2

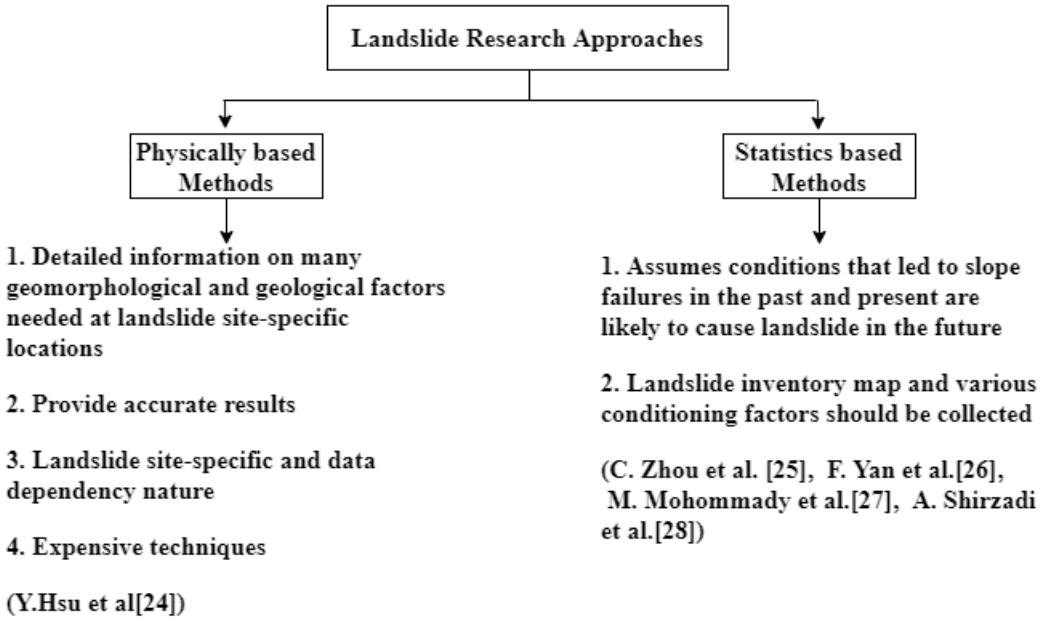


Figure 2.2: Landslide Research Approaches

2.2 Impact of Landslides in Sri Lanka

Approximately 30% of Sri Lanka's land area has been identified susceptible to landslide occurrence including districts, namely, Badulla, Nuwara Eliya, Kegalle, Ratnapura, Kandy, Matale, Kalutara, Matara, Galle, and Hambantota [21]. As per statistics provided by the NBRO, more than 400 lives have been lost due to landslides during the past three years in Sri Lanka. In 2016 and 2017, 151 [21] and 230 [21] lives were lost respectively with an estimated 10,000 families being displaced. In 2018 due to cutting failures, six lives were lost. Sri Lanka was affected by 38 severe landslides [21] in 2017 while the spread of disaster situation was confined only to 15 districts compared to the 24 districts in the previous year. Table 2.1 gives the number of deaths in six of the affected districts in 2017.

Affected Districts	Number of Deaths
Kalutara	101
Ratnapura	79
Matara	11
Hambantota	11
Galle	09
Kegalle	04

Source: NBRO

Table 2.1: Deaths Due To Landslides in 2017 [20]

When the above statistics were analyzed, it was seen that Kalutara district has reported the largest number of deaths in 2017 due to landslides. Communities attribute many landslides in Kalutara District to the uprooting of rubber plantations to provide the raw materials for Medium-Density Fibreboard(MDF) manufacturing and conversion of forest land for tea cultivation in other areas.

Apart from the damage to life and property, several infrastructural, as well as economically important facilities, have also been affected due to these landslides, especially water distribution pipes, hydroelectricity generating centers, and communication systems [19]. At times, social interests such as education and health services are also severely disrupted [19]. According to statistics provided by the NBRO it was evident that the education sector suffered damages and losses in Kalutara, Ratnapura, Galle, Matara, and Hambantota districts. Moreover, frequent landslides have threatened the destruction to the environment, including the flora and fauna of the areas concerned. Such damage caused to the environment, at times is irreversible and therefore cannot be estimated and perhaps will never be known.

Due to deforestation and other human-related activities [29] the trend of landslide occurrence seem to continue to next era as well. Therefore, considering the wide coverage of landslide damages, planners and decision-makers need to identify landslide-prone areas to plan mitigation actions. Landslide susceptibility prediction has become indispensable in this context in order to minimize the casualties and damages.

2.3 Contributing Factors of Landslides in Sri Lanka

Landslides do not occur usually due to a single reason, but it is the net effect of several processes and factors, persisting for long periods on the hilly terrain. No single cause can be attributed to the occurrence of a landslide, or rockfall, but it is due to the interaction of a multitude of factors [19], either natural or man-made. Some of the natural causes include,The steepness of hillslope, type of rock material, deep weathering of rock material and the depth of the weathered rock, density of the joint pattern and the structure of the rock, thickness of colluvium deposits collected downslope due to gravity, poor drainage conditions leading to excessive water seepage in sub-strata, high intensity of the precipitation, earthquake as a triggering factor, flood and reservoirs in hilly areas etc.

It is very seldom for a landslide to occur off on a flat area or a plain, the reason being that there is no space or opportunity for any soil mass in the area to fall or slide anymore. However, landslides do occur in such terrain too, but that can only happen due to excessive weight or pressure placed on top of the hill slope. Slope surfaces with thick soil layers and slope angle between 15° and 45° [19] have been found to have a greater preponderance for landslides with a maximum tendency of hill slope of angle 26° to 35° [19] to the horizontal.

Soil and rock types having different characteristics or less cohesive properties are the ones most subjected to fall or slide. Various rock characteristics including the structure contribute directly to this looseness of a rock or soil and consequently to such sliding. Rock is composed of various minerals in different proportions. The rock is subjected to various natural elements such as sunlight, rain, wind, hot and cold temperatures and also to the penetration of plant root systems for long periods thus causing disintegration and weathering. Also, by the impact of various pressures rock material can be subjected to splitting, which spreads and in various quantities. The resulting material consisting of soil, clay etc., have their own characteristics. The above processes finally cause the deposition and settling in various layers or strata on the slopes thus leaving a conductive background for landslides. Similarly, the action of different temperature and pressures on the rock also create fractures in the rock and these fracture systems can be easily lead to separation of the boulders and rock fragments from the parent rock. The separated rock fragment and weathered components can be later transported to lower regions of the slope by erosion and rolling. The increase in the thickness of the overburden deposits also can lead to landslides.

The background for a landslide to move down a slope is created by the action of various environmental factors and undue human activities persisting for long periods. It is finally

triggered by intensive rainfall which leads to a landslide. The heavy rainfall not only causes water penetration into the subsoil layers, thereby losing the inter-layer cohesion but also increases the weight of soil mass. The penetrated water also acts as an easy lubricant flowing down-slope. The net effect of these processes is the sliding of the soil mass down the slope as a landslide. According to preliminary investigations carried out by researchers, it has been found that if a hill slope receives continuous rainfall of about 200 mm within a period of 3 days [21], the susceptibility to land sliding in such an area increases. It has been observed that as a result of the current poor, ill-planned land-use practices, even a rainfall of 75-100 mm for a 2 day period [19] is sufficient to trigger a landslide.

Earthquakes can influence the occurrence of a landslide directly or indirectly. The vibrations of an earthquake can break up the exciting bond between soil particles and similarly the bond between rock layers can also be weakened considerably. Sri Lanka has felt some tremors in the recent past and it is necessary to be alert about landslides that could be triggered by earthquakes. Scientists believe that an earthquake of above 4.5 [19] on the Richter scale can be considered as one that can cause a landslide. But the occurrence of landslides due to earthquakes are very rare in Sri Lanka.

Flood and reservoirs in the hill areas can influence the incidence of a landslide in several ways. One such instance is the failure of riverbanks scoured by the following of swollen rivers after heavy rains. The bottom areas of the banks are eroded by the flowing river water leaving the top areas of the riverbanks without the toe support, which can easily cause the top mass to come down. Flood water can also contribute significantly to the collapse of the hill slope. The groundwater level of the area usually rises after floods and consequently, the natural drainage pattern of the area is charged. This causes an increase in the internal water pressure within the slope subsurface while also blocking water from upper soil layers along sub-soil passages resulting in accumulation of water in the slope. As a result of the additional weight due to the water mass and the slope, susceptibility of landslides can increase in a region.

Undue human intervention strongly influences the incidence of landslides. Examples of some wanton land-use practices are denudation of forest areas, use of land without proper planning, construction on hilly terrain without due investigation or design, quarrying for metal without due investigation and adherence to norms, obstruction of natural water paths and storage of water on high ground. These activities can expose the top soil, and affect the stability of slope thereby causing soil cracks and when eroded with the rains they can ultimately result in landslides.

Other than the above-mentioned factors, several studies [6], [16], [30] have identified that factors such as aspect, Stream Power Index(SPI), Topographic Wetness Index(TWI), lithology, Normalized Difference Vegetation Index(NDVI), Sediment Transport Index(STI), Distance to faults, distance to roads, distance to rivers, plan curvature, profile curvature, etc. also contributes to landslides. Even though a wide variety of conditioning factors of landslides exists the applicability of these factors differ from region to region [31]. Therefore it is essential to identify the most suitable set of conditioning factors for a study area in the prediction of landslide susceptibility.

2.4 Landslide Evaluation Tools and Technologies

According to U.S. Geological Survey there are three types [1] of tools and technologies involved in landslide evaluation. They are Mapping, Remote Sensing and Real-time Monitoring.

2.4.1 Mapping

There are three types [1] of landslide maps that are useful for planners as well as general public when considering about criteria for landslide maps. They are,

1. Landslide Inventory Maps
2. Landslide Susceptibility Maps
3. Landslide Hazard Maps

The inventories denote areas that are identified as having failed by landslide processes. They range from simple inventories, which overview of the aerial extent of landslide occurrence, to the complex inventories, where more detailed layers of information including landslide classification,zone of depletion, active and inactive landslides, geological age, the rate of movement and depth and kind of material involved in sliding can be found.

A landslide susceptibility map goes beyond inventory maps by predicting the areas that have the potential for landsliding. As mentioned earlier, these areas are determined by considering the conditioning factors for landsliding (such as slope,geology, soil, elevation, etc.). The susceptibility maps are derived by overlaying data layers with an inventory map to identify geological units which have landslide-prone features compared to the previous landslides.

Landslide hazard maps demonstrate the likelihood of landsliding in various areas in the future. These maps are useful to predict the relative degree of hazard in a landslide area as they provide detailed information of type of the landslide, extent of slope subject to failure and probable maximum amount of ground movement etc.

Maps are useful and convenient for presenting information as they can present different kinds and combinations of information at different levels of detail. There are main 3 stages [1] involved in the preparation of maps.

1. Regional mapping: Synthesizes available data and identifies general problem area in a small scale, which is normally performed by a provincial, state or federal geological survey.

2. Community level mapping : A more detailed surface and subsurface mapping program in complex problem areas.

3. Site-specific large scale mapping: Concerned with the identification, analysis, and solution of actual site-specific problems, often presented in the size of a residential lot.

2.4.2 Remote Sensing

When the accessibility for the conditioning factors is difficult physically and those methods for in-field data gatherings are expensive to continue, the remote sensing methods came in to the topic. Some of the remote sensing methods utilized in landslide prediction are described below.

1. Aerial photography remote sensing : This technology can be used to identify the vegetation cover, topography,drainage pattern, soil drainage character, bedrock geology, surficial geology, landslide type and as well as relationships among the factors. This should be rely on a careful study of recent aerial photographs of the given area, as older slides may not be evident on changing terrain factors than recent photographs.

2. LiDAR imaging : LiDAR stands for Light Detection and Ranging, which uses a narrow laser beam to probe through dense ground cover and to produce accurate terrain maps. It includes the ability to eliminate the interference of forest cover which is present in traditional photography methods. The essential elements of LiDAR mapping system include a laser rangefinder mounted in an aircraft, a Global Positioning System (GPS), and an Internal Measurement Unit.

2.4.3 Real-time Monitoring

The immediate detection of landslide activity that is provided by real-time systems can be crucial in saving human lives and protecting property. The traditional on-field detection techniques are failed to observe the changes at the moment they occur. Hence, the continuous data provided by real-time landslides monitoring allows a better understanding of dynamic landslide behaviour and enables engineers and specialists to create more effective designs to prevent or halt landslides.

The monitoring process can be expensive and requires the experts knowledge in installation of monitoring systems and maintenance of them. These monitoring stations can be coordinated with a warning system, which would generate warning alerts to the public in a hazard event.

2.5 Landslide Susceptibility Zonation Maps

Landslide Research and Risk Management Division of NBRO has implemented a landslide hazard zonation mapping programme within the 10 landslide prone districts of Kalutara, Galle, Hambantota, Nuwara Eliya, Matale, Kandy, Kegalle, Ratnapura, Matara and Badulla. The maps which display the distribution of the severity of landslide hazard potential in a given area, were intended to be used with associated guidelines as a decision making tool for development of central highlands of the country. It is also used for identification of elements at landslide risk and can be utilized in relocation, rehabilitation, allocation of relief funds and insurance purposes also. Mapping is carried out at 1:50000 scale and at 1:10000 scale [20]. The process associated with the implementation of landslide hazard zonation maps is discussed below.

NBRO has identified following 7 conditioning factors which cause occurrence of landslides in Sri Lanka [21].

1. Slope
2. Geology
3. Hydrology
4. Land form
5. Soil Type & Thickness
6. Land Use
7. Rainfall

Slope, Geology, Hydrology, Landform data for a particular district is extracted from the respective district's contour map. Land form, land use and Soil Type & Thickness are extracted from field study in the region. Rainfall data is collected from the reading of the rain gauge stations installed in different areas covering all the landslide prone districts in Sri Lanka. After data relevant to these terrain factors are collected, following steps [32] will be used to build the landslide susceptibility zonation map.

1. Generating the map of State of Nature(SON) for each identified factor separately
2. Converting the each SON map to a digitized version which is called Digital Elevation Model (DEM)
3. The digitized coverage of these conditioning factors are integrated to create an inferred map of landslide potential
4. The polygons of this inferred map is dissolved into different hazard zones using the criteria given in table 2.2

Overall Hazard Rating (R)	Hazard Zone	Description
$R <= 40$	1	Safe slopes
$40 < R <= 55$	2	Landslides not likely to occur
$R <= 70$	3	Modest level of landslide
$70 < R$	4	Landslides are expected

Source: NBRO

Table 2.2: Criteria for Hazard Zonation [20]

NBRO uses factor weighing scale given in table 2.3 when generating the integrated hazard zonation map relevant to a region.

Factor	Weight
Slope	25%
Geology	20%
Hydrology	20%
Land form	10%
Soil Type & Thickness	10%
Landuse	10%

Source: NBRO

Table 2.3: Factor Weighing Scale [20]

Real-time rainfall data from 105 automated rain-gauges implemented in 12 districts are collected and used to issue early warning to vulnerable communities when rainfall intensities reach certain threshold values. The threshold values [20] are described in the table 2.4 given below.

Rainfall Threshold	Description
75 mm > continues within the next 24 hours	Possibility of landslides, rock falls, subsidence and cut slope failure
100 mm > continues within the next 24 hours	Danger of landslides and cut slope failures exist
150mm > continues within the next 24 hours or exceed 75 mm within 1 hour	Evacuate to a safe place

Source: NBRO

Table 2.4: Rainfall Thresholds [20]

2.6 Review of Similar Research

Landslide susceptibility mapping approaches are classified into two as qualitative or quantitative approaches [10]. Qualitative methods usually depend on expert opinions [33] whereas quantitative methods depend on the relationships between landslide controlling factors and landslides [34]. Figure 2.3 demonstrates a taxonomy of landslide susceptibility mapping methods by Hamid Reza Pourghasemi et al. [6].

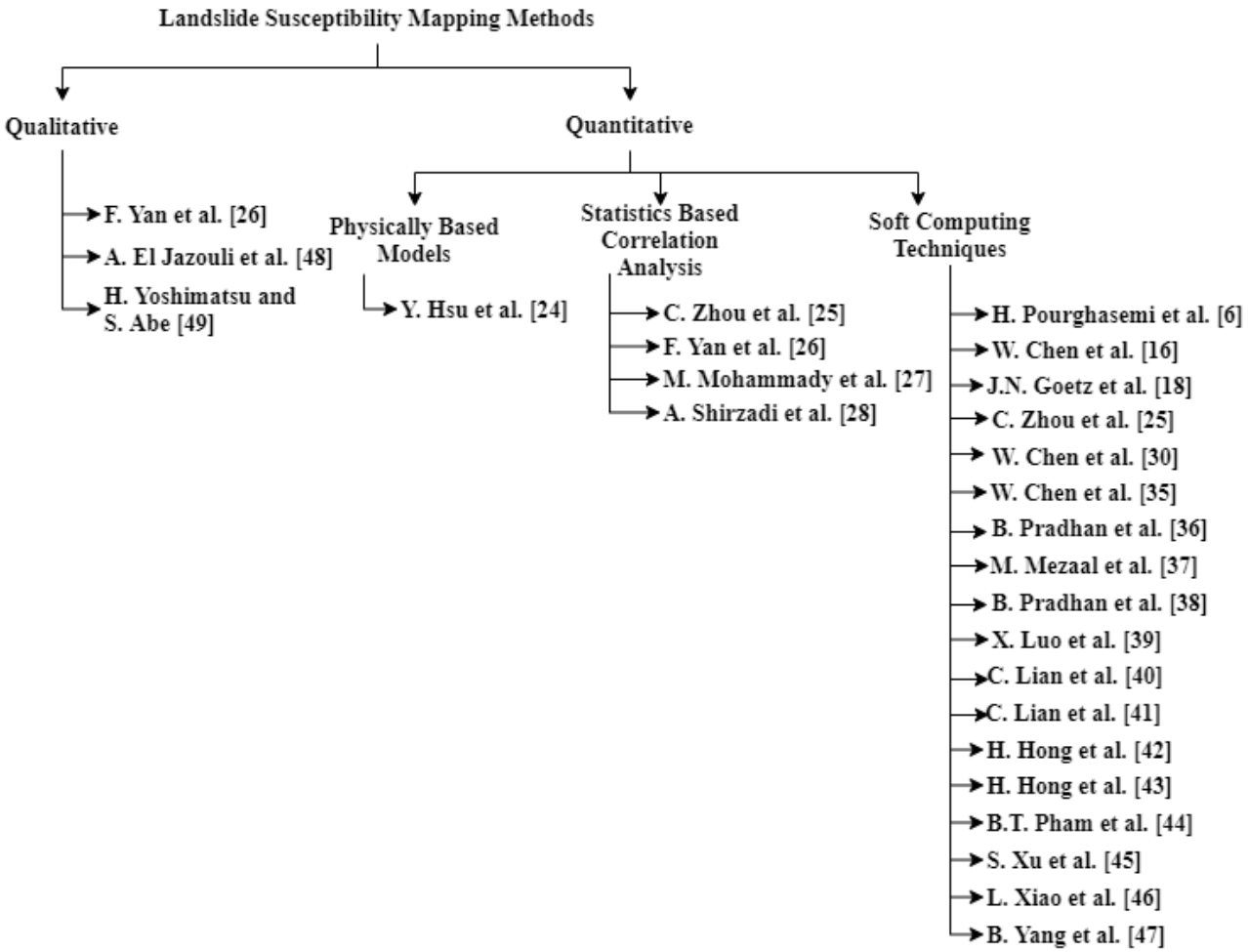


Figure 2.3: Taxonomy of Landslide Susceptibility Mapping Methods

Physically based models require detailed site specific geological data [6]. These models are expensive and not practical for large areas. Traditional statistical models, which assume an appropriate structural model and then focus on parameterizing it, are widely used for analysis of natural hazards such as landslides. Classification of each landslide conditioning factors in traditional statistical models is a key point that affects the quality of landslide susceptibility map [6]. In contrast, machine learning techniques, a powerful group of data driven tools, use algorithms to learn the relationship between a landslide occurrence and landslide related predictors, and avoids starting with an assumed structural model. ML-based models can effectively overcome the limitation of data dependent bivariate and multivariate statistical methods [6]. Machine learning techniques allow handle data from various measurement scales, any type of independent variable (i.e. ratio, interval, nominal, or ordinal), and without needing to define normally distributed transformed variables [10].

A comparison between four advanced machine learning techniques namely, Bayes' Net(BN), Radial Basis Function(RBF) Classifier, Logistic Model Tree(LMT) and Random Forest(RF)

for landslide susceptibility modeling in Chongren County, China has been studied by Wei Chen et al. [16]. The spatial database was constructed with 222 landslide locations. In the initial phase, the researchers have considered 15 conditioning factors where 3 of them; land use, plan curvature, and profile curvature were eliminated due to their non-predictive ability identified through the calculation of the Information Gain. The predictive capability of the BN. RBF Classifier, LMT and RF models were calculated and compared using the Receiver Operating Characteristic (ROC) and statistical measures, including sensitivity, specificity, and accuracy. The results showed that the RF model had the highest sensitivity, specificity, and accuracy values of 0.787, 0.716, and 0.752, respectively, for the training dataset. RF model yielded a high degree of fitting for both the training and validation datasets.

Coupling machine learning algorithms with spatial analytical techniques for landslide susceptibility modeling is a critical issue prevalent in the world. To address this issue, Hamid Reza and Omid Rahmati have presented with the comprehensive comparison [6] among the performances of ten advanced machine learning techniques(MLTs) including Artificial Neural Networks(ANN), Boosted Regression Tree(BRT), Classification and Regression Trees(CART), Generalized Linear Model(GLM), Generalized Additive Model(GAM), Multivariate Adaptive Regression Splines(MARS), Naive Bayes(NB), Quadratic Discriminant Analysis(QDA), Random Forest(RF) and Support Vector Machine(SVM) for modeling landslide susceptibility. Area under the ROC curve(AUC-ROC) approach has been utilized in evaluating the performance of the above-mentioned machine learning techniques and it has been found that the RF and BRT have the best performances compared to other MLTs with AUC values 83.7% and 80.7% respectively.

Another study conducted by Wei Chen et al. [30] has compared the landslide susceptibility predictive ability of Logistic Model Tree(LMT), Random Forest(RF), and Classification and Regression Tree (CART) models considering 171 landslide locations in Long County, China. The 12 landslide-related parameters used in the study are slope angle, slope aspect, plan curvature, profile curvature, altitude, NDVI, land use, distance to faults, distance to roads, distance to rivers, lithology, and rainfall. To obtain more accurate results, landslide conditioning factors with low or null predictive capability have been removed using the Linear Support Vector Machine(LSVM) method. Since all twelve landslide conditioning factors revealed positive predictive capability values, the twelve factors were used in the analysis for building the three models. The results obtained in model performance evaluation and comparison using ACC values, ROC curves, AUC values, Std. error, CI at 95%, and significance level P showed that the RF model has the highest predictive capability compared with the

LMT and CART models with a success rate of 0.837 and a prediction rate of 0.781.

To assist in the evaluation of landslide susceptibility modeling techniques to enhance a user's decision on which method is most suitable for a particular application, J.N. Goetz et al. [18] have conducted a study presenting a comparison of traditional statistical and novel machine learning models applied for regional-scale landslide susceptibility modeling. The modeling techniques applied were Generalized Linear Model(GLM), Generalized Additive Models(GAM), Weights Of Evidence(WOE), Support Vector Machine(SVM), Random Forest classification (RF), and Bootstrap aggregated classification trees(Bundling) with Penalized Discriminant Analysis (BPLDA). Slope angle, elevation, profile curvature, plan curvature, catchment area, catchment height, convergence index, topographic wetness index(TWI), slope aspect and surface roughness(SDS) were used as predictors in this study. The study demonstrated that there was generally little differentiation in prediction performance between statistical and machine learning landslide susceptibility modeling techniques. The researchers suggest that SVM, RF, and BPLDA may be particularly useful for high-dimensional prediction problems where a large number of highly correlated predictor variables are present.

C. Zhou et al. [25] have used SVM, ANN and LR in landslide susceptibility modeling in the Three Gorges Reservoir area in China. 12 landslide conditioning factors were considered and unimportant factors were selected and eliminated using information gain ratio. The performance of the models was evaluated using the ROC curve and the Friedman test. The results showed that SVM and ANN outperformed LR while SVM was found ideal for the case study area.

A study has been conducted combining the Adaptive neuro-fuzzy Inference System with Frequency Ratio (ANFIS-FR), Generalized Additive Model (GAM), and Support Vector Machine (SVM), for landslide susceptibility mapping in Hanyuan County, China [35]. A collinearity test and correlation analysis were applied between the 12 conditioning factors and landslides. The results of these analyses showed that there is no collinearity among different factors. The accuracy of the models was validated using the ROC curve. The results showed that the SVM model has the highest prediction rate of 0.875, followed by the ANFIS-FR and GAM models with prediction rates of 0.851 and 0.846, respectively.

Biswajeet Pradhan [36] compared the landslide susceptibility prediction performances of three different approaches, namely as Decision Tree (DT), Support Vector Machine (SVM) and Adaptive Neuro-Fuzzy Inference System (ANFIS) using 113 landslide locations in Penang Hill area, Malaysia. The results showed success-rate results for using ANFIS (AUC-94.21),

followed by SVM (AUC-91.67) and DT (AUC-88.36.). When considering prediction-rate results, eight landslide conditioning factors performed better in DT (83.07) SVM (81.46) and ANFIS (82.80) models. Since the adaptive neuro-fuzzy inference system (ANFIS) has not been applied commonly in the landslide susceptibility research, this research presented some promising results over its viability as a satisfactory model.

An optimized neural architecture for automatic landslide detection from high-resolution airborne laser scanning data, was proposed by Mustafa Ridha Mezaal, Biswajeet Pradhan et al [37]. In their study they addressed the drawbacks of traditional, time consuming and costly methods of analysis on landslide susceptibility; like field surveys, optical remote sensing and synthetic aperture techniques. In a solution, they proposed using Recurrent Neural Networks(RNN) and Multi-Layer Perceptron Neural Networks(MLP-NN) in landscape detection based on high-resolution LiDAR data. They have stated the main advantages of this approach are that it requires little or no prior knowledge compared to other traditional classification methods and its' ability to perform the nonlinear mathematical fitting for function approximation. The landslide inventory for the training process referred to the previously produced inventory by Pradhan and Lee [38] and a total number of 21 landslides had analyzed. The feature selection process had optimized by using a supervised approach and to rank features from most relevant to less, the correlation-based selection algorithm had utilized. The obtained models were evaluated using a 10-fold cross-validation method and in advance, the model had tested in another part of the area. The results obtained for the two models: RNN and MLP-NN are, 83.33% and 78.38% respectively. And the test accuracies for RNN and MLP-NN were 83.33% and 74.56% which indicated that the proposed models generated sufficiently accurate classification results.

Luo X et al. [39] evaluated a landslide inventory database of 493 landslides that occurred in the Shangli County, China under three models, Artificial Neural Network(ANN), Support Vector Machine (SVM) and one statistical model, Information Value Model (IVM). Initially, 16 conditioning factors were derived from Landsat 8 imagery and Global Digital Elevation Model(ASTER GDEM), and statistical measures like VIF and tolerances were used under multicollinearity analysis to filter out the best subset of conditioning factors. After running through the three models, the results showed that the ANN model achieved higher prediction capability which proves its capability of solving nonlinear and complex problems surpassing the performance of SVM and the statistical model, IVM on landslide data set. Also when the estimated landslide susceptibility map is compared with the ground-truth landslide map, the high-prone area was observed to be located in the middle area with multiple

fault distributions and the steeply sloped hill.

Cheng Lian et al. [40] proposed an ANNs switched prediction scheme to construct Prediction Interval(PI)s with a three-stage formulation to overcome the drawbacks of Artificial Neural Networks (ANN) in predicting mutational displacement points with time lags in landslide susceptibility mapping. In the first stage, K-means clustering was applied to divide the whole training dataset into two sub-training sets as stationary points and mutational points. In the second stage, a weighted ELM classifier was applied to construct the switched rules and in the third stage, bootstrap- and kernel-based ELMs were applied to construct candidate PIs for each sub-training set. The final PIs are constructed by switching between these two candidate PIs. In this study, three real-world landslide datasets from the Three Gorges region of China were used. In the final results, it was observed that the prediction accuracy of the mutational points has been significantly improved indicating the ability of the proposed method to construct the reliable quality PIs for landslide displacement.

Cheng Lian, Zhigang Zeng, Wei Yao, and Huiming Tang [41] conducted a study on landslide susceptibility prediction using a novel artificial neural network technique, extreme learning machine (ELM) with a kernel function. In this research, a convex combination of Gaussian kernel function and polynomial kernel function in ELM was used as the generalization performance of ELM with kernel function depends closely on the kernel parameters and kernel types. And a novel hybrid optimization algorithm based on the combination of Particle Swarm Optimization (PSO) and Gravitational Search Algorithm (GSA) was used to avoid blindness and inaccuracy in parameter selection.

Haoyuan Honga et al. [42] done work on to investigate and compare the use of current state-of-the-art ensemble techniques namely as AdaBoost, Bagging, and Rotation Forest, for landslide susceptibility assessment with the base classifier of J48 Decision Tree (JDT). In this study, a landslide inventory map with 237 landslide locations in Guangchang district, China were evaluated under 15 conditioning factors. The results showed that all landslide models had high performance (AUC $>$ 0.8) while JDT with the Rotation Forest model showed the highest prediction capability (AUC = 0.855), followed by the JDT with the AdaBoost (0.850), Bagging (0.839), and the JDT (0.814). As a state-of-the-art technique, JDT with the Rotation Forest delivered promising results.

Potential applications of two new models such as two-class Kernel Logistic Regression (KLR) and Alternating Decision Tree (ADT) for landslide susceptibility mapping at the Yihuang area, China was explored by Haoyuan Hong, Biswajeet Pradhan, Chong Xua and Dieu Tien Bui in their study [43]. SVM has been used for comparison. The resulting models

were validated and compared using the ROC, Kappa index, and five statistical evaluation measures. The prediction capabilities were 81.1%, 84.2%, and 93.3% for the KLR, the SVM, and the ADT models, respectively. The result showed that the ADT model yielded better overall performance and accurate results than the KLR and SVM models.

The performance of five state-of-the-art hybrid machine learning techniques namely Reduced Bagging based Reduced Error Pruning Trees(BREPT), MultiBoost based Reduced Error Pruning Trees (MBREPT), Rotation Forest-based Reduced Error Pruning Trees(RFREPT), Random Subspace-based Reduced Error Pruning Trees(RSREPT), and Reduced Error Pruning Trees(REPT) was evaluated and the results were compared [44] for the selection of best landslide susceptibility model. ROC curve, Statistical Indexes, and Root Mean Square Error methods were used to validate the performance of these models. Analysis of model results indicated that the BREPT is the best model for landslide susceptibility assessment in comparison to other models.

Shiluo Xu, RuiqingNiu proposed an LSTM based methodology [45] to solve the problem of the hysteresis effects of triggering factors and landslide displacements in the rainfall-induced landslide displacements prediction task. In this study, data collected from 73 data points in the Baijiabao landslide site, China was used to research the total cumulative displacement of the site by dividing it into a trend and periodic components using empirical mode decomposition. The trend component was predicted using an S-curve estimation while the total periodic component was predicted using a Long Short-Term Memory Neural Network (LSTM). While the static deep learning approaches, such as BP and SVR, can only learn information at the current time step they tend to lean rules from a timing point and cannot make full use of historical data. As a result, the lag effect between trigger factors and landslide displacements cannot be addressed very well. LSTM connects hidden layers and obtains previous influences and information using a "state vector" and corresponds to the hysteresis effects of landslides. The results obtained from LSTM were more accurate than BP, SVR, and even the Elman method when the dataset is small. Most of the time unavailability of landslide historical time data limits the use of time as a factor in landslide susceptibility research.

A study was conducted by Liming Xiao, Yonghong Zhang and Gongzhuang Peng [46], in using integrated deep learning algorithms to predict the landslide susceptibility in China - Nepal highway. As for the concerned area, the study addresses the importance of hazard assessments more accurately in real-time. The instability factors concerned were: elevation, slope angle, slope aspect, plan curvature and vegetation index. The four machine learning

algorithms used were: Decision Tree (DT), Support Vector Machine (SVM), Back Propagation neural network (BPNN) and Long-Short term Memory algorithm. The results obtained from the experiments have compared and the prediction accuracy of BPNN, SVM, DT, and LSTM are 62.0%, 72.9%, 60.4%, and 81.2% respectively. So they found that the LSTM has outperformed with sufficient accuracy, as its capability to learn time series based patterns along with temporal dependencies.

The study of Beibei Yang, Kunlong Yin, Suzanne Lacasse, Zhongqiang Liu [47], presented time series analysis and long short-term memory neural network to predict landslide displacement in TGRA, a well-known as a landslide-prone area in China. They have identified the feature that landslide deformation patterns lead to accumulated displacement versus time showing a stepwise curve. The study proposed a dynamic model to predict landslide displacement based on time series analysis and Long-Short Term Memory network(LSTM). The final accumulated model was composed of two components: trend tern analysis and periodic term analysis in time series. The trend displacement was predicted using a cubic polynomial function. The periodic term displacement was predicted using the LSTM model based on landslide displacement factors. For the trend component, long-term deformation controlling factors such as; lithology, geological structure, progressive weathering, etc. have been considered causing landslide displacement as these factors were increased, approximately, monotonically with time and usually on a long-time scale. The final model was compared with Baishuihe and Bazimen landslide inventory to evaluate the success or failure of the model. According to that LSTM network given a model with 7.11 RMSE (root mean squared error) value while SVM based model given 21.83 of RMSE value. As a result, it was concluded that the LSTM based prediction model has outperformed in the domain of landslide susceptibility mapping rather than SVM based classifications.

Several statistical approaches [26], [27], [28] in landslide susceptibility mapping have been followed by some of the researchers. Among them are Frequency Ratio(FR), Dempster–Shafer, and Weights-of-Evidence. Y. Hsu et al. [24] explored the use of two physically based approaches, rainfall threshold-based method and real-time simulation in landslide forecasting. Among the quantitative approaches, physically-based approaches [28] are expensive and will not be practical for large landslides sites.

Qualitative studies [26], [48], [49] consider a set of conditioning factors and based on the importance, factors are weighted using methods like Analytic Hierarchy Process(AHP).

Table 2.5 shows a comparison between different approaches followed in the studies described above.

Reference	Conditioning Factors Used	Machine Learning Techniques Used	Best Performance
[16]	Slope, Aspect, Elevation, Plan Curvature, Profile Curvature, Lithology Stream Power Index, Sediment Transport Index, Topographic Wetness Index, Distance to Rivers, NDVI, Distance to Roads, Land Use, Distance to Faults, Rainfall	Bayes' Net Radial Basis Function Classifier Logistic Model Tree Random Forest	Random Forest
[6]	Altitude, Slope Angle, Slope Aspect, Slope Length, Plan Curvature, Profile Curvature, Driainage Density, Distance from Rivers, Distance from faults, Land Use, Lithology, Distance from Roads	Artificial Neural Networks Boosted Regression Tree Classification and Regression Trees General Linear Model Generalized Additive Model Multivariate Adaptive Regression Splines Naive Bayes Quadratic Discriminant Analysis Random Forest Support Vector Machine	Random Forest Boosted Regression Tree
[30]	Slope Angle, Slope Aspect Plan Curvature, Altitude Profile Curvature, NDVI Land Use, Distance to Faults, Distance to Roads, Distance to Rivers, Lithology, Rainfall	Logistic Model Tree Random Forest Classification and Regression Trees	Random Forest

Reference	Conditioning Factors Used	Machine Learning Techniques Used	Best Performance
[18]	Slope Angle, Elevation, Profile Curvature, Plan Curvature, Catchment Area, Convergence Index, Slope Aspect, TWI Catchment Height Surface Roughness	Generalized Linear Model Generalized Additive Models Weights Of Evidence Support Vector Machine Random Forest Bootstrap aggregated classification trees(Bundling) with Penalized Discriminant Analysis	Support Vector Machine Random Forest Bootstrap aggregated classification trees (Bundling) with Penalized Discriminant Analysis
[44]	Slope, Aspect, Geomorphology, Curvature, SFM, Land Cover, Distance to Roads, Overburden Depth, Distance to Rivers, Valley Depth,	Reduced Bagging based Reduced Error Pruning Trees MultiBoost based Reduced Error Pruning Trees Rotation Forest-based Reduced Error Pruning Trees Random Subspace-based Reduced Error Pruning Trees Reduced Error Pruning Trees	Reduced Bagging based Reduced Error Pruning Trees
[35]	Slope Aspect, Lithology Altitude, Slope Angle, TWI, Plan Curvature, Profile Curvature, Distance to Rivers, Distance to Faults, Distance to Roads, Land Use, NDVI	Adaptive Neuro-Fuzzy Inference System with Frequency Ratio Generalized Additive Model Support Vector Machine	Support Vector Machine

Reference	Conditioning Factors Used	Machine Learning Techniques Used	Best Performance
[39]	Slope Angle, Slope Aspect, Elevation, Road Density, Plan Curvature, Profile Curvature, Annual Rainfall, River Density, Distance to Rivers, Lithology, Distance to Faults, Watery Degree, Distance to Roads, NDVI, NDWI, Urban Land-use Index	Artificial Neural Networks Support Vector Machine Information Value Model	Artificial Neural Network
[43]	Slope, Aspect, Altitude, TWI, SPI, STI, Plan Curvature, Land Use, NDVI, Distance to Faults, Distance to Rivers, Distance to Roads, Lithology, Rainfall	Two-class Kernel Logistic Regression Alternating Decision Tree	Alternating Decision Tree
[46]	Elevation, Slope Angle, Slope Aspect, Plan Curvature, Vegetation Index, Built-up Index, Stream Power, Lithology, Precipitation Intensity, Cumulative Precipitation Index	Decision Tree Support Vector Machine Back Propagation Neural Network Long-Short Term Memory Algorithm	Long-Short Term Memory Algorithm

Table 2.5: Comparison of Results Obtained in Previous Research

According to the literature review, it was identified that Support Vector Machine(SVM), Artificial Neural Networks(ANN) and Random Forest(RF) have performed well in the prediction of landslide susceptibility in various regions of the world. Initially, they were selected as the three candidate machine learning algorithms for this study. The comparative studies [6], [18] done considering SVM, ANN and RF have indicated that RF outperforms other models in the prediction of landslide susceptibility. Therefore Random Forest machine learning algorithm was selected to implement the landslide susceptibility prediction model for Kalutara district employing the 12 conditioning factors identified.

Even though LSTM also has shown more than 80% accuracy in the prediction of landslides, since the dataset considered in this study did not contain the time aspect of the landslide events LSTM was not chosen as a candidate algorithm for the implementation of the landslide susceptibility model.

2.7 Summary

Landslides occur as a result of a combination of terrain factors and triggering events. The predictive capability of the factors about occurrence of landslide will depend on the particular landslide site and will differ from one terrain to another terrain. There are several machine learning techniques used in the past literature to predict landslide susceptibility. Selecting a suitable machine learning technique to build the model, that will yield high predictive capability is crucial.

Chapter 3

Methodology and Design

The methodology and design adopted in the implementation of the landslide susceptibility prediction model is described in this chapter. The study employed a hybrid approach that combined the design science research approach [31] and the quantitative research approach [50] to implement a landslide susceptibility prediction model for Kalutara District, Sri Lanka.

3.1 Overview of the Methodology

The implementation of the landslide susceptibility prediction model was carried out in six main steps. A detailed description of each step in the methodological approach is provided in the following sub sections. The process flow of this study is illustrated in figure 3.1.

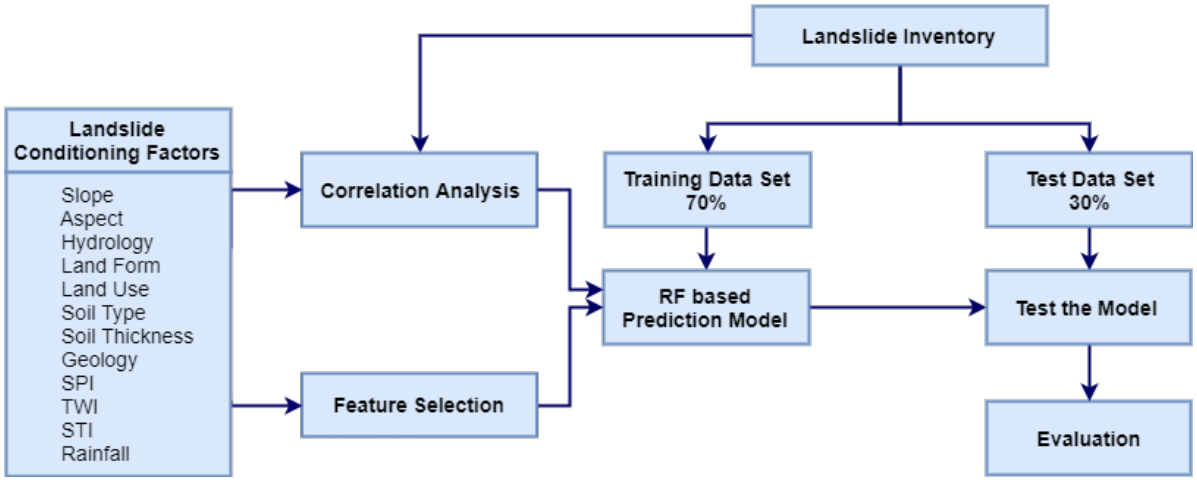


Figure 3.1: Process Flow of the Research

3.1.1 Landslide Inventory Mapping and Preparation of Training and Test Data Sets

The first important step in the prediction of landslide susceptibility would be to identify landslide locations that occurred in the past and create the landslide inventory map. A landslide inventory map depicts the spatial distribution of a single landslide event or multiple landslide events in a specific region over time. In this study, a detailed reliable landslide inventory map was created from the landslide database acquired from the NBRO. A total of 84 landslides were analyzed in this study. These points could not denote the entire regions covered by the respective landslides. The landslide locations were presented by discrete points placed at the centers of their head scarps. For the susceptibility analysis, the landslide locations were randomly split into two subsets with a ratio of 70:30 for training and testing respectively. In order to maintain the balance between all possible classifications for defining the problem (landslide locations and non-landslide locations), the dataset was included with the same number of (84) non-landslide locations randomly sampled from the landslide free area in Kalutara district. During the implementation, the entire data set was randomly split into 70:30 ratio where training dataset consisted of 117 instances while the test dataset consisted of 51 instances.

3.1.2 Preparation of Landslide Conditioning Factor Maps

The quality and the scale of conditioning factors affect the efficiency of the landslide susceptibility prediction [2]. Based on the knowledge obtained from the previous literature and expert consultation from the NBRO, initially, 12 conditioning factors were selected for the study, namely slope, aspect, hydrology, landform, land use, soil type, soil thickness, geology, Stream Power Index(SPI), Topographic Wetness Index(TWI), Sediment Transport Index(STI), and rainfall. A spatial database containing landslide related conditioning factors was constructed using data extracted from the contour map of Kalutara district, geospatial statistical data, and precipitation data acquired. The gathered raw data were processed in QGIS environment to generate thematic maps and to extract the data for each of the conditioning factors.

A Digital Elevation Model(DEM) generated using the contour map of the Kalutara District was utilized to extract the slope, aspect, SPI, STI, and TWI values for the area. The slope is the rise or fall of a land surface while the aspect is the compass direction that a slope faces. The slope map was reclassified into five equal interval classes ranging from 0° to

78.5384°. Aspect values of the study area ranged from 7° to 359° and were divided into five equal interval classes.

SPI is the amount of erosive power of water flow based on the presumption that discharge is proportional to a specific catchment area. TWI is the degree of accumulation of water at a site. STI describes the processes of slope failure and deposition. SPI and TWI were calculated using equation 3.1,3.2 and 3.3 respectively.

$$SPI = Atan\left(\frac{\beta}{b}\right) \tag{3.1}$$

$$TWI = \log_e\left(\frac{A}{btan\beta}\right) \tag{3.2}$$

$$STI = \left(\frac{A_s}{22.13}\right)^{0.6}\left(\frac{sin\beta}{0.0896}\right)^{1.3} \tag{3.3}$$

β (radian) is the slope at a given cell, b (m) is the cell width through which water flows, A (m^2) is the flow accumulation, and A_s is the upstream area.

STI values generated from DEM ranged from 0 to 20. Similarly, the values generated for SPI ranged from -2477 to 1325 while TWI values ranged from 9 to 17. SPI, TWI and STI values were also reclassified into five equal interval classes each.

Hydrology indicates the distance to waterways in the study area and it was divided into six classes, ranging from 2.3399m to 694.9425m. The natural features and shapes existent on the earth's surface are identified as the land-forms. The study area is comprised of seven types of land-forms including clay flats, clay swamps, scattered small hills, flat plains, marches, wetlands, etc. The land use varies from crops, forests to other non-agricultural uses distributed over ten classes. Alluvial, boulders, colluvium, residual, and rock exposure are the most prominent soil types in Kalutara district. The thickness of the soil in these types varies throughout the region. Geology varies within 7 classes, namely; Quartzite, Granite Biotite Gneiss, Charnochite, Charnockitic Gneiss, Khondalite, Quartzo Feldspathic, and Granulatic Gneiss.

Rainfall is considered as a triggering factor for the occurrence of landslides in this study. Annual rainfall data collected through 26 rain gauge stations in Kalutara District from 1984 to 2018 were obtained from the Department of Meteorology, Sri Lanka. It was divided into five classes for the analysis.

3.1.3 Correlation Analysis between Landslides and Conditioning Factors

To analyze the spatial relationship between the landslides and conditioning factors Frequency Ratio3.4 was used in the study. The ratio of the probabilities of landslide occurrence to non-landslide occurrence [51] for a given conditioning factor is indicated by the results obtained for frequency ratio. FR>1 suggests a strong relationship between the landslide event and conditioning factor while FR<1 suggests vice versa [52]. FR=1 represents an average value.

$$FR_i = \frac{N_{pix(S_i)}/N_{pix(N_i)}}{\sum N_{pix(S_i)}/\sum N_{pix(N_i)}} \tag{3.4}$$

$N_{pix(S_i)}$ is the number of landslide pixels in each conditioning factor class i. $N_{pix(N_i)}$ is the total number of pixels that have class i in the study area. $\sum N_{pix(S_i)}$ is the total number of landslide pixels in the study area. $\sum N_{pix(N_i)}$ is the total number of pixels in the study area.

3.1.4 Selection of Conditioning Factors

The predictive capability of the employed model may be reduced by features with a certain noise level. Therefore, the selection of conditioning factors with high predictive ability is an important step in landslide susceptibility prediction [9]. To achieve these, predictive abilities of the conditioning factors should be quantified and factors with low or null predictiveness should be removed [9]. This will result in a more accurate prediction of the resulting model.

Quantification of the predictive capability of the 12 landslide conditioning factors was carried out using the Information Gain Ratio. Information Gain Ratio is the ratio of information gain to the intrinsic information. Higher Information Gain Ratio indicates a higher predictive ability for the models [9]. The Information Gain Ratio for a certain landslide conditioning factor A is computed as follows using 3.5, 3.6, 3.7 and 3.8.

$$Info(S) = -\sum_{i=1}^{2} \frac{n(L_i, S)}{|S|} log_2 \frac{n(L_i, S)}{|S|} \tag{3.5}$$

$$Info(S, X) = \sum_{j=1}^{m} \frac{S_j}{|S|} Info(S) \tag{3.6}$$

$$SplitInfo(S, X) = -\sum_{j=1}^{m} \frac{S_j}{|S|} log_2 \frac{|S_j|}{|S|} \tag{3.7}$$

$$Information\ Gain\ Ratio(S, X) = \frac{Info(S) - Info(S, X)}{SpiltInfo(S, X)} \tag{3.8}$$

The training data set is denoted by S. n(L_i,S) is the number of class L_i samples in S. Info(S) is the Information Gain. Info(S,X) is the amount of information required to split S into m subsets related to X. Information generated by splitting S into m subsets is denoted by SplitInfo.

After the selection of the best set of conditioning factors using the information gain ratio, the feature values for landslide and non landslide points were extracted by employing a variety of vector and raster analysis methods and tools available in QGIS. The incomplete and null values resulted for some features, were eliminated from the dataset as a data pre-processing step.

3.1.5 Construction of Landslide Susceptibility Prediction Model

Initially, three candidate machine learning algorithms were identified through the previous literature to implement the model. They are Support Vector Machine(SVM) [39], [46], [53] Artificial Neural Networks(ANN) [39], [53] and Random Forest(RF) [6], [16], [18], [30], [16]. SVM is a supervised classifier that separates the feature space obtained by the input data set into classes, using a hyper-plane which creates the maximum margin [54]. ANNs consist of a chain of nodes called "Artificial Neurons", that are interconnected and able to identify the relationship patterns between input-output layers [55]. RF is an ensemble classifier that uses multiple decision trees to make the final prediction. Out of the identified candidate algorithms, Random Forest was selected to construct the model considering the performance [6], [16], [18], [30], [16] of the RF models in previous studies. Random Forest has also not been used in the prediction of landslides in Sri Lanka before the time of this study.

RF builds multiple bootstrap samples, known as training sets and constructs a classification rule (a tree) for each. In a random forest, each node is split using the best split among a subset of predictors that are randomly chosen by the node. The random feature selection at each node decreases the correlation between any pair of trees in the forest, decreasing the forest error rate. RF includes two powerful ideas in machine learning algorithms: random feature selection and bagging [30].

3.1.6 Evaluation of the Performance of the Model

For the evaluation of the model performance, initially, three evaluation approaches were inferred from the literature review [6], [2], [30] that most of the landslide prediction studies

have utilized these approaches to determine the model performance successfully.

1. Kappa Index
2. Receiver Operating Characteristic(ROC)
3. Statistical Evaluation Metrics: Specificity, Sensitivity, and Accuracy

The kappa Index is popularly used to quantify the magnitude of agreement between observers [56] in a study. The kappa value of 1 represents perfect agreement between true values and the classification whereas, the value of 0 represents no agreement [57]. Kappa value could be further quantified as follows: (0.81-1.0) Almost Perfect, (0.61-0.80) Substantial, (0.41-0.60) Fair, (0.21-0.40) Fair, and (0.0-0.20) Slight [58]. Kappa Index (equation 3.9) is used to measure the reliability of the classification approach.

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{3.9}$$

Pr(a) indicate probability of success of classification. Pr(e) indicate probability of success due to chance.

In an ROC curve,the sensitivity is plotted in the function of the 1-specificity for different cutoff points. The area under the ROC curve(AUC) provides the overall measure of test performance that reveals the capability of a model to predict landslide and non-landslide pixels. An AUC value of 1 indicates a perfect model, while an AUC value of 0 indicates a non-informative model [9] and a higher AUC value indicates a better predictive capability of a model. According to Tien Bui et al. [9], correlation of predictive capability and AUC could be quantified as follows: (0.9-1) Excellent, (0.8-0.9) Very Good, (0.7-0.8) Good, (0.6-0.7) Average, and (0.5-0.6) Poor [59].

Furthermore, three statistical evaluation measures; accuracy, sensitivity, and specificity were used to evaluate the performance of the trained landslide model. Accuracy(equation 3.10) is the proportion of landslide and non-landslide pixels that models correctly classified. Sensitivity (equation 3.11) is the proportion of landslide pixels that are correctly classified as landslide occurrences. Specificity(equation 3.12) is the proportion of the non-landslide pixels that are correctly classified as non-landslide.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.10}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.11}$$

$$Specificity = \frac{TN}{FP + TN} \tag{3.12}$$

TP(True Positive) indicate landslide instances classified as landslides. TN(True Negative) indicate non-landslide instances classified as non-landslides. FP(False Positive) and FN(False Negative) show non-landslide instances classified as landslides and landslide instances classified as non-landslides respectively.

3.2 Design Assumptions

The main scarp of the landslide represents the entire area covered by the respective landslide. The terrain features of the study area remain constant from 1984 to 2018 during which the occurrences of landslides were considered in this study. The rainfall measurement of a particular location is considered equal to its average rainfall measurement throughout the years from 1984 to 2018.

3.3 Random Forest Algorithm

RF is an ensemble classifier that uses multiple decision trees to make the final prediction. Out of the identified candidate algorithms, Random Forest was selected to construct the model considering the performance of the RF models in previous studies [6], [16], [18], [30], [16]. To the best of our knowledge, Random Forest has also not been used in the prediction of landslides in Sri Lanka before the time of this study.

RF implements a forest of random decision trees using the training set and assigns each tree as a classification rule. RF employs a subset of predictors that are randomly selected to perform the best possible split at each node. This random feature selection reduces the correlation among two decision trees and reduces the out of bag error of any tree in the forest. RF utilizes bagging to produce ensemble predictions from a forest of trees.

3.3.1 Mathematical Foundation

Random forest is an ensemble classifier which provides its final prediction based on the most voted class from the class predictions carried out by its multiple decision trees. RF employs two main decision tree algorithms [54] to define the rules and conditions to make these predictions, evaluate the node impurity and iteratively divide the dataset.

1. Classification and Regression Tree(CART)

   CART uses Gini Index[x6] to measure the quality of the split when splitting nodes

of the decision tree. It measures the probability of a data sample being incorrectly classified when it is chosen randomly. The following equation 3.13 is used to calculate the gini index.

$$G = \sum_{k=1}^{k} P_k(1 - P_k) \tag{3.13}$$

k is the number of classes, which in this study was 2(landslide, non landslide) and Pk is the proportion of the number of elements in class k. The purity of the node is considered high when the value of the Gini index is small. This suggests that the node primarily contains samples belonging to a single class [55].

2. Iterative Dichotomiser (ID3)

   ID3 consists of an iterative structure which uses a decision tree formed using a subset of the training dataset chosen at random to classify the remaining objects in the training set [60]. ID3 uses Entropy and Information Gain to evaluate the node impurity or the quality of the split at each node in the decision tree. The Entropy is defined by equation 3.14

$$H(S) = -\sum_{k=1}^{k} P_k log P_k \tag{3.14}$$

   P_k is proportion of the number of elements in class k to the number of elements in current dataset S. Entropy also takes a minimum value when the node purity is high, similar to Gini index. Information Gain is the decrease in entropy calculated using equation 3.15

$$Gain(S, a) = H(S)_{t1} - H(S|a)_{t2} \tag{3.15}$$

   a is the attribute evaluated to split, H is the entropy, t1 is the prior state and t2 is the state after the split. The splits are carried out considering the decrease in entropy between the parent node and the weighted average entropy of its children.

3.3.2 Bagging

In the training phase, each tree in the forest is grown by randomly selecting samples (with replacement) from the training data set. This is known as "Bootstrapping". This results in training each model(tree) using different sets of samples from the training data set. Therefore the ensemble prediction from the forest of trees tends to be more accurate than an individual prediction from a single decision tree. In the testing phase, the final prediction will be made by averaging the prediction of each decision tree. This process is known as Bootstrap Aggregation in Random Forests. It is demonstrated in Figure 3.2.

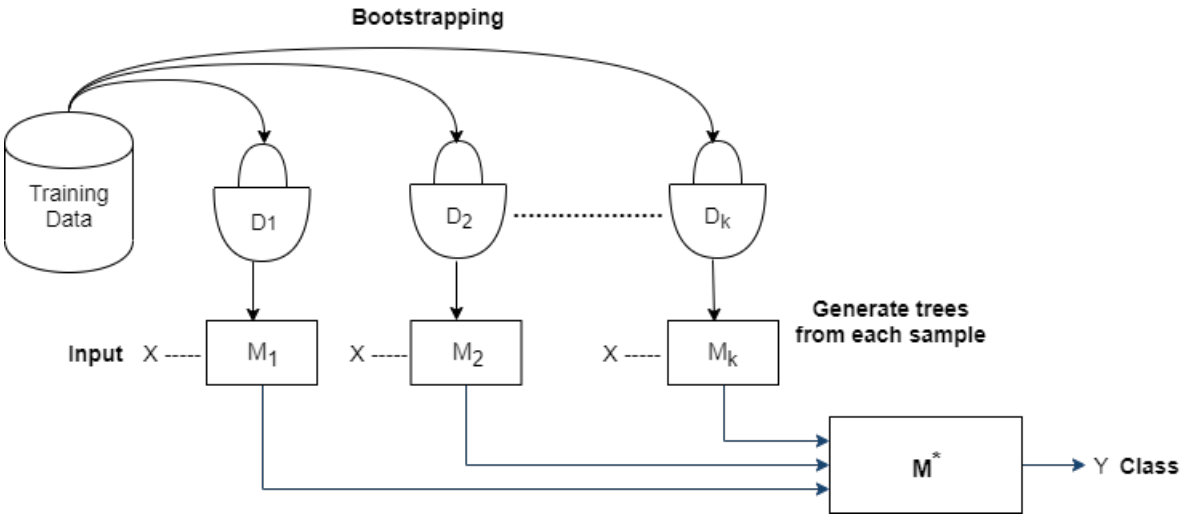


Figure 3.2: Boostrap Aggregation

Bagging improves accuracy when random features are used [61]. It also gives ongoing estimates of the generalization error of the combined ensemble of trees, as well as estimates for the strength and correlation [61].

3.3.3 Random Feature Selection

In Random Forest a node is split considering a random subset of all features available. From the feature subset, the feature that produces the most separation between the data in the left node and the right node is selected. Feature randomness minimizes the correlation while maintaining strength [61]. From the 12 conditioning factors, a random subset of factors is selected to split the nodes in each tree. Figure 3.3 demonstrates the random feature selection carried out by the random forest algorithm with respect to landslide susceptibility.

Random forest algorithm has the ability to preserve its performance even when a large proportion of data is missing and when the original data set presence with many outliers. So in this study, these characteristics of RF were benefited as the nature of the original data set used.

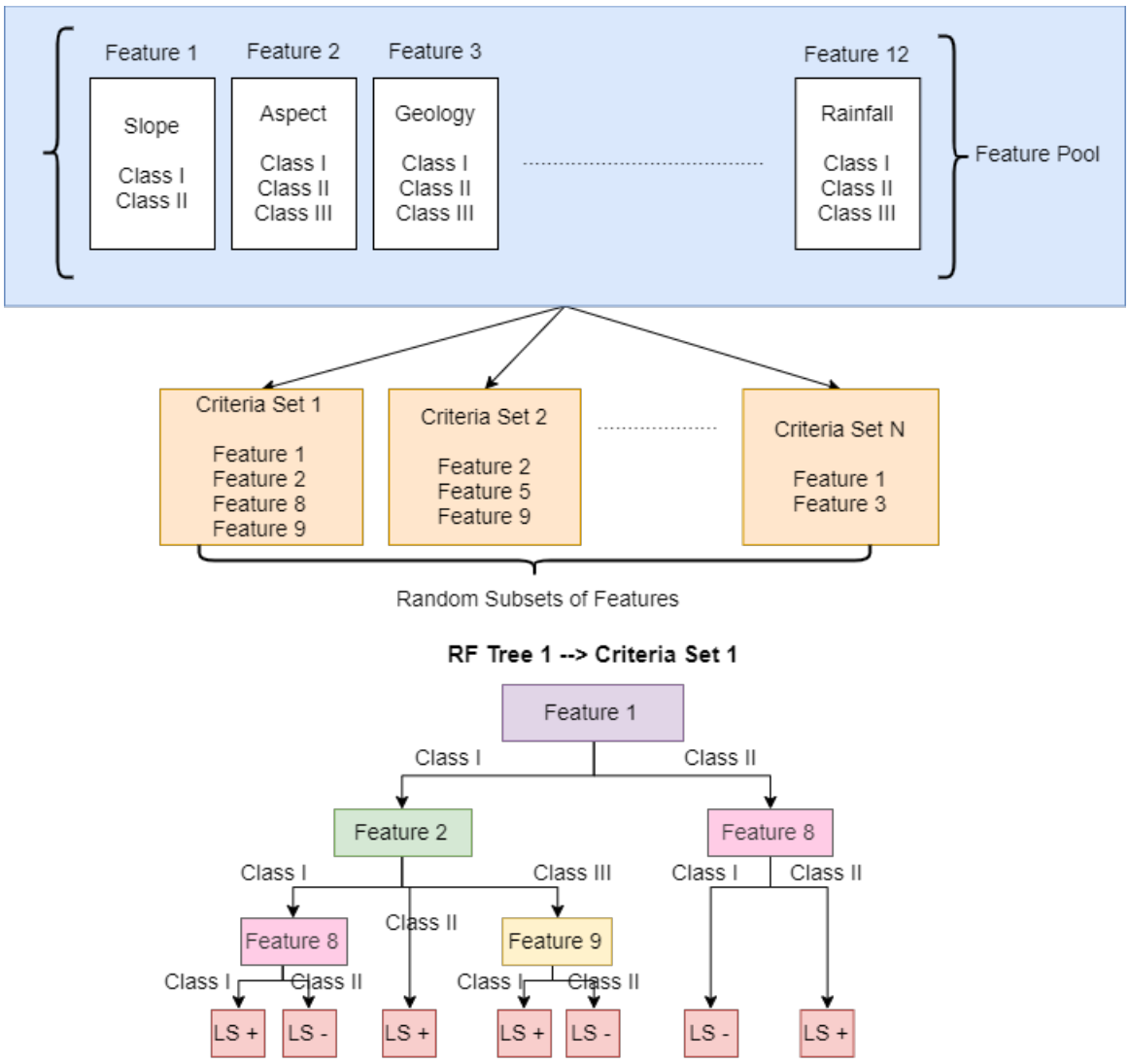


Figure 3.3: Random Feature Selection in Random Forest

The effectiveness of Random Forest algorithm in landslide studies identified through the investigation of past research was the principle reason behind the selection of Random Forest algorithm to implement the prediction model. Random Forest algorithm also employs two powerful concepts in machine learning which are bagging and random feature selection. These concepts also play a huge role in improving the effectiveness of the application of Random Forest in this study.

A summary of the research design used in the study is demonstrated in table 3.1.

Research Design Aspect	Type
Research Type	Quantitative
Research Purpose	Exploratory
Approach	Design Science and Quantitative
Data Sources	NBRO Department of Meteorology
Tools	QGIS 2.8.1 VS Code Spyder
Result Evaluation	ROC Accuracy, Sensitivity, Specificity
Result Visualisation	Django based web application

Table 3.1: Summary of Research Design

3.4 Summary

Under this chapter, the six main steps followed to design and thus investigate the research questions are discussed, namely; constructing landslide inventory map and preparation of training and test data sets, preparation of landslide conditioning factor maps, correlation analysis between landslides and conditioning factors, selection of conditioning factors, construction of landslide susceptibility prediction model and evaluation of the performance of the model.

Chapter 4

Implementation

This chapter discusses the steps followed in the implementation of the prediction model. Spatial analysis with respect to the landslide inventory was carried out using QGIS 2.8.1, an open source Geographic Information System. Preparation of the dataset consisting of landslide locations and their respective features took major portion of the time in the implementation phase due to heavy processing that had to be carried out to extract the necessary features for each landslide location. The comprehensive set of tools and plugins provided by QGIS assisted most part of the implementation. Python language was utilized in application of the Random Forest algorithm on the dataset to predict landslide susceptibility. Visual Studio Code, a free source-code editor developed by Microsoft for Windows and Spyder which is an open source cross-platform integrated development environment for scientific programming in the Python language were used for editing, refining and debugging the python code.

4.1 QGIS

Quantum GIS or QGIS is a popular Geographic Information System available on all major platforms with a steadily growing user base which easily exceeds 100,000 users even by conservative estimates. The QGIS project provides one of today's most popular applications for working with spatial data. The multitude of user requirements has led to a diverse ecosystem Quantum GIS is used all around the world for tasks as diverse as forestry and city planning, bushfire mapping and paleontological surveys [62] etc. One feature that makes it popular is its flexibility to scale with user requirements: from a simple data viewer, to data collection, editing and analysis, to serving data on the web – on as many machines as needed and without any licensing issues.

QGIS Desktop application is considered the core of the QGIS ecosystem. It is a classic desktop GIS application with powerful tools to view, edit and analyze spatial data. Additionally, there is an optional QGIS Browser application which acts as a data catalog and viewer. It facilitates browsing through big data archives and lists of web services and offers drag-and-drop of layers to QGIS Desktop. QGIS 2.8.1 desktop application was used to manipulate maps in the study.

QGIS supports a wide variety of file formats, database systems (such as PostGIS, Oracle Spatial, or MS SQL Server) and OGC standards compliant services such as WMS, WFS and WPS. This ensures that Quantum GIS and other, even proprietary GIS can be used side-by-side and complement one another.

QGIS is designed to be very modular. Users can both reduce and increase complexity and functionality of the application by either removing unneeded user interface elements or activating additional functionality via the plugin system.

4.2 Landslide Inventory

Preparation of landslide inventory is the basis for landslide prediction. A vector map demonstrating 84 landslide locations in Kalutara district was provided by the NBRO and QGIS was used to generate a landslide inventory map where landslide locations were depicted using points. To identify landslide free area, symmetrical difference function under geoprocessing tools in QGIS vector analysis was used. Then the same number of non-landslide locations were randomly sampled from the landslide free area. The final landslide inventory map consisted of a total of 168 landslide locations and non-landslide locations. Figure 4.1 demonstrates the landslide inventory map prepared for Kalutara district with the landslide and non landslide locations.

To extract the latitude and longitude of each of the landslide and non landslide points several steps were followed. First, OpenStreetMap raster layer was added to QGIS map view with EPSG:3857 coordinate system using OpenLayers Plugin. Then the landslide inventory map was saved with EPSG:4326 as the CRS and added the map layer to the canvas. As the final step the field calculator was used to generate latitude and longitude values from the maps corresponding to each point.

For the landslide susceptibility prediction, inventory map was randomly split into two subsets containing 118(70%) landslides and non landslide instances for training data set and 50(30%) landslide and non landslide instances for testing data set.

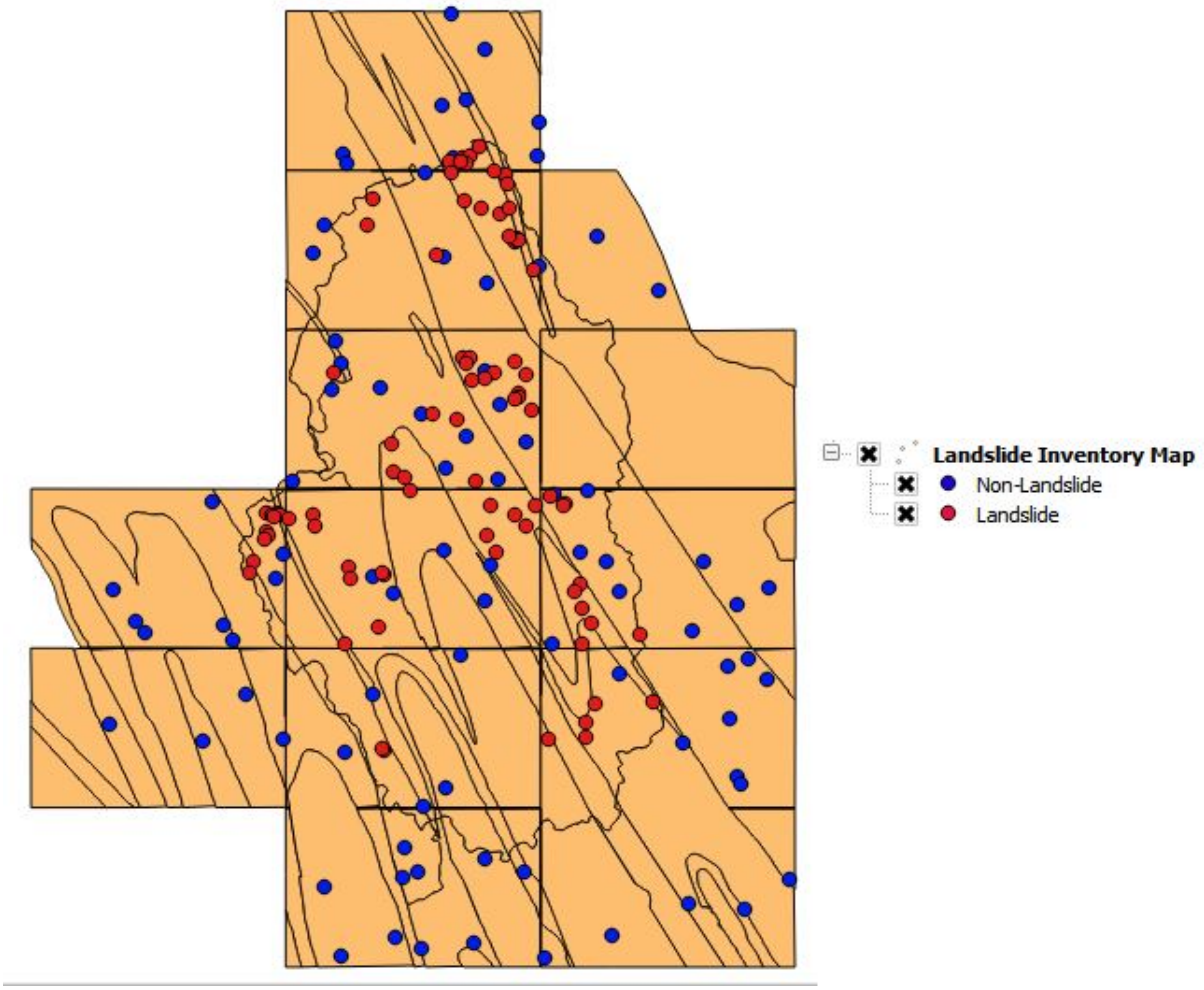


Figure 4.1: Landslide Inventory

4.3 Thematic Maps of Conditioning Factors

In this study 12 conditioning factors were considered which are slope, aspect, geology, hydrology, landform, land use, soil type, soil thickness, SPI, STI, TWI and rainfall. Thematic maps were generated for all the conditioning factors using QGIS. Factors including slope, aspect, SPI, STI, TWI, hydrology and rainfall having continuous values were divided into five equal interval classes each. Geology, Landform, Land use and Soil Type & Thickness consisted of several categorical classes each. The spatial distribution of each of these classes belonging to the conditioning factors in the entire study area was visualized using the thematic maps.

4.3.1 Raster Maps

Using the contour lines extracted from the topographic map of 1:10000 scale, the Digital Elevation Model(DEM) was created with a grid size of 10X10 m. Since contour map was

available in the vector format it was converted to raster format using Rasterize function in QGIS and DEM was generated using the DEM function available under Raster Analysis in QGIS. DEM data was used to construct maps of the geomorphometric factors analyzed in the study; slope, aspect, SPI, STI, and TWI. Figure 4.2 shows the DEM generated for Kalutara district.

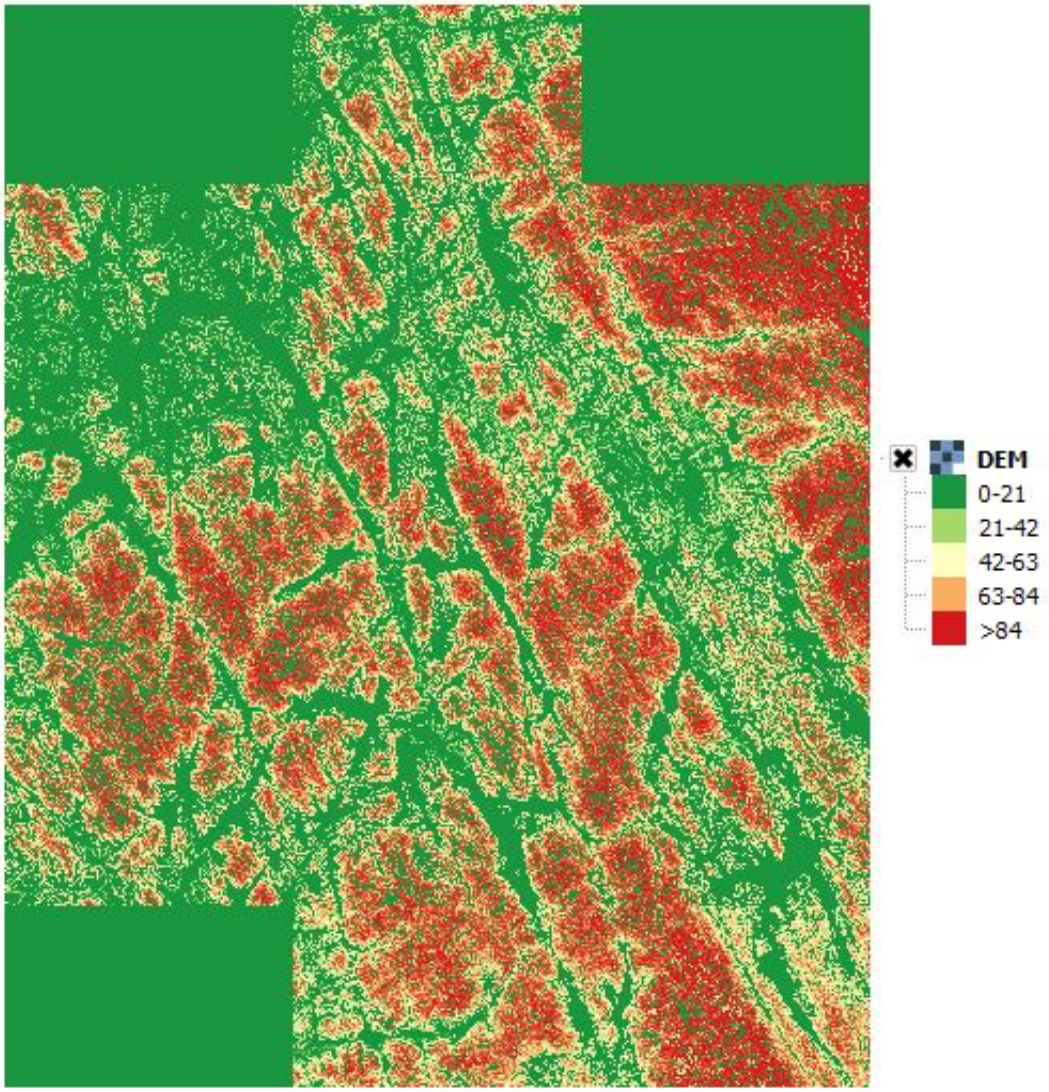


Figure 4.2: Digital Elevation Model(DEM)

Figure 4.3 demonstrates the thematic maps generated for slope, aspect, SPI, TWI and STI. Slope, SPI, TWI and STI maps were reclassified into five classes while Aspect map was reclassified into 9 classes. The colors represented by each class is given in the legend for each map.

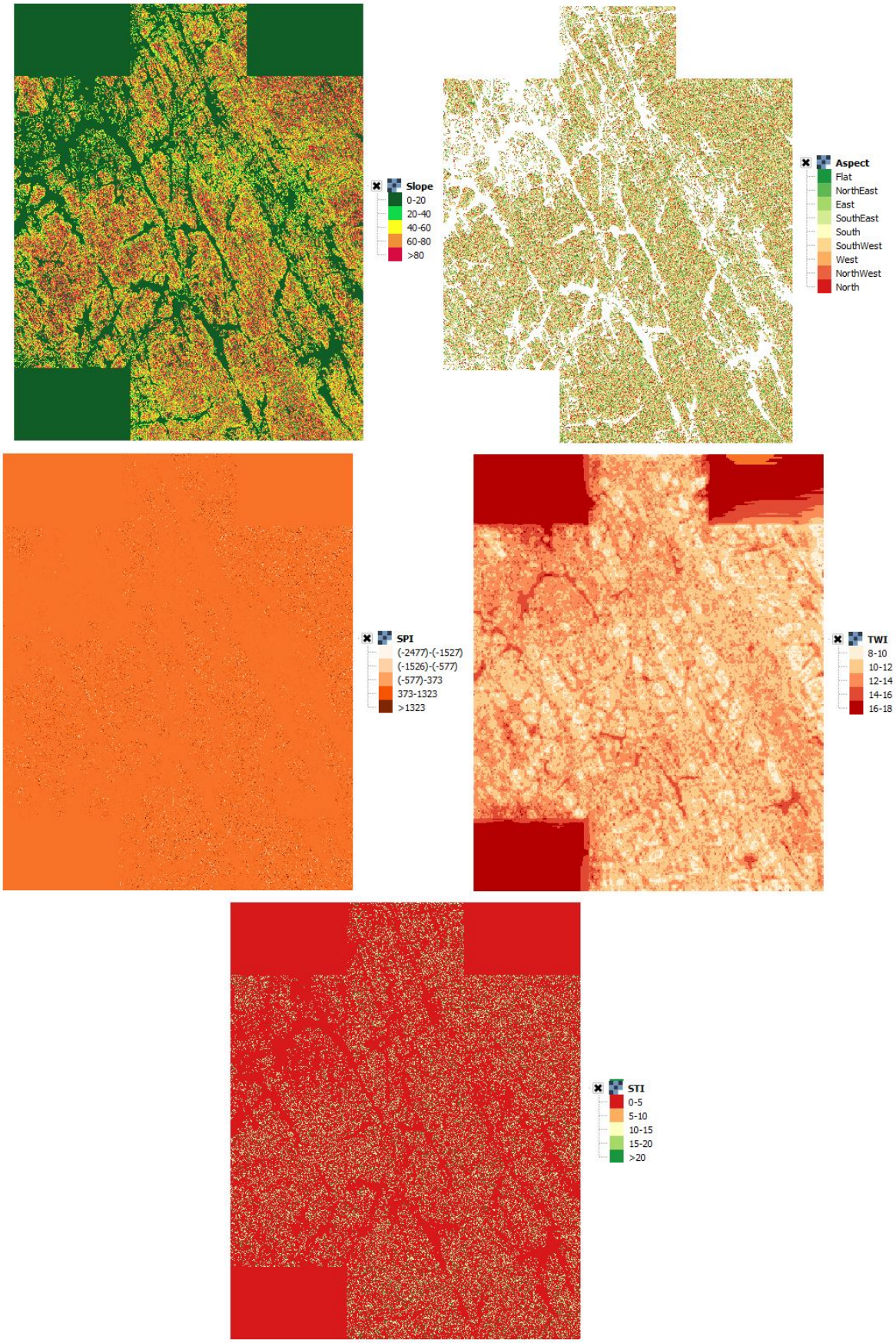


Figure 4.3: Slope, Aspect, SPI, TWI, and STI Thematic Maps

The rainfall dataset provided by the Meteorological Department of Sri Lanka contained

location(latitude and longitude) information and annual rainfall data collected from the 26 rain gauge stations in Kalutara district from 1984 to 2018. To generate a map for rainfall, the average rainfall values were calculated for each station. Using the location information and the calculated average rainfall values a map was generated in vector format where the 26 rain gauge stations were depicted as points on the map.

In order to visualize and get the rainfall distribution in the entire district, interpolation of the point map was carried out using Inverse Distance Weighting method where the final output was a raster map as shown in figure 4.4.

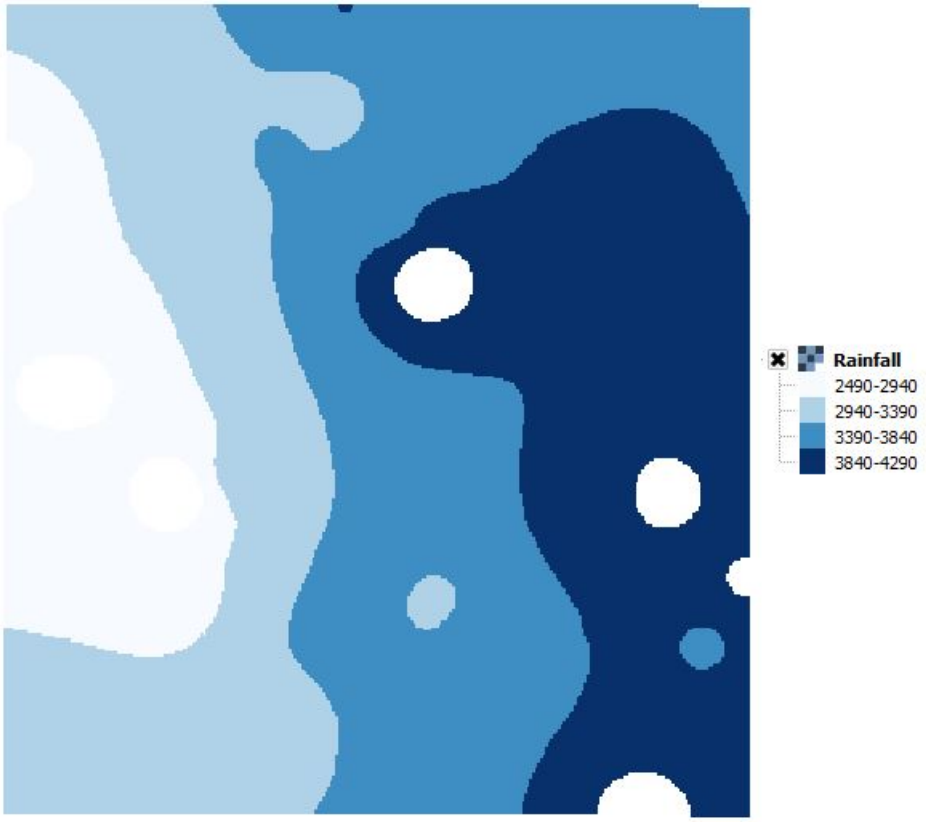


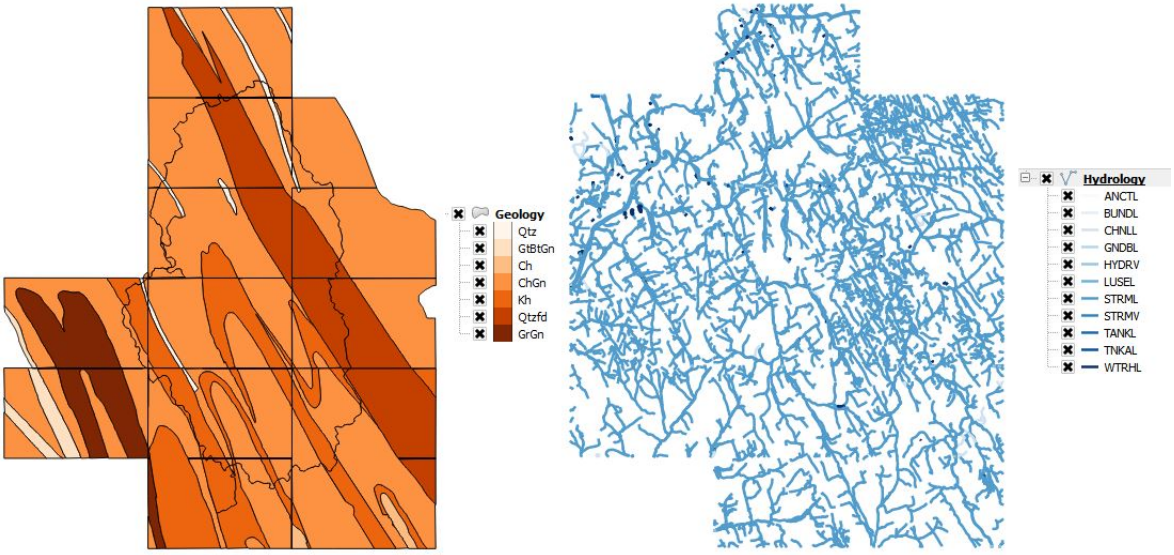
Figure 4.4: Rainfall Thematic Map

Since the generated rainfall map was following WGS 84 coordinate reference system (CRS) which was different from the previous maps having WGS 84/UTM zone 44N CRS, the map was reprojected to follow the same CRS as the others. This reprojection was essential to get rainfall values corresponding to each landslide and non landslide point in the construction of the feature pool.

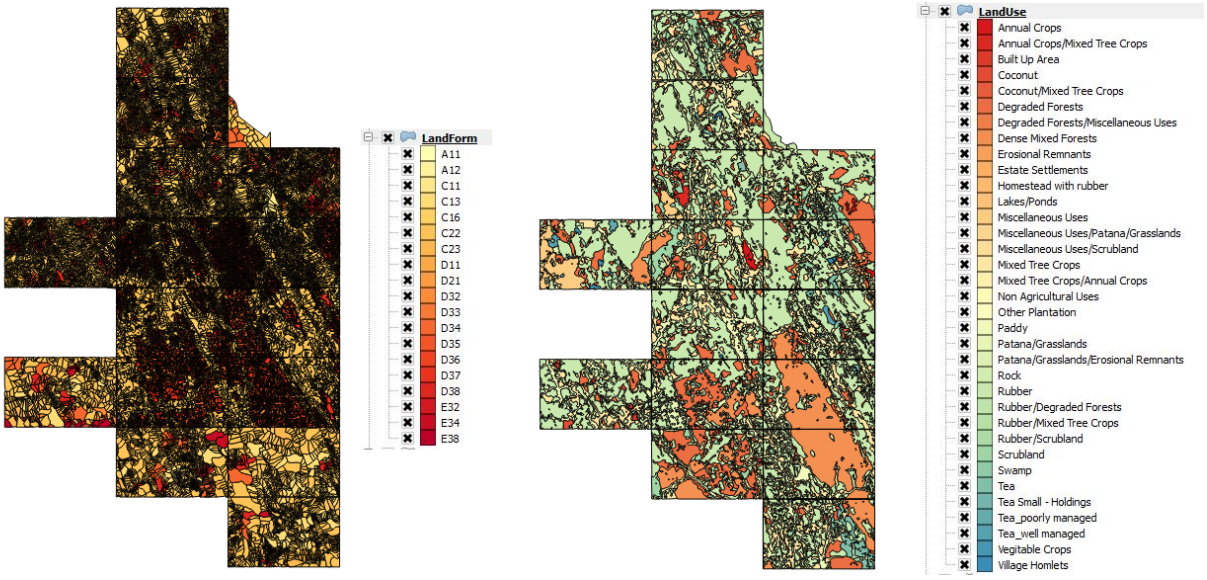
4.3.2 Vector Maps

The thematic maps generated in vector format for geology, hydrology, land form, land use, and soil type and soil thickness are given in figure 4.5.

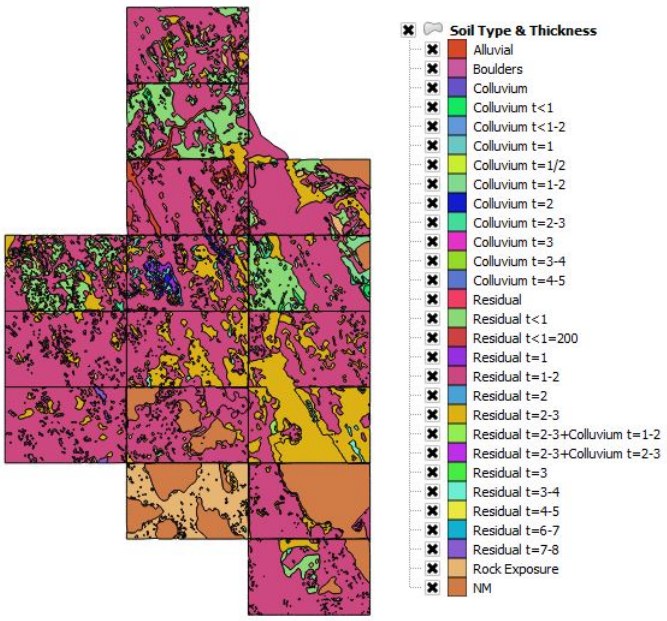
7 Geology categories, 11 Hydrology categories, 19 Land Form categories, 35 Land Use categories, and 28 Soil Type and Thickness categories were identified to be distributed through out Kalutara district and was visualized using the following maps.



(a) (b)



(c) (d)



(e)

Figure 4.5: (a)Geology, (b)Hydrology, (c)Land Form, (d)Land Use, and (e)Soil Type and Thickness Thematic Maps

The conditioning factor maps which were in vector format were converted into a pixel(raster) format with a spatial resolution of 10X10 m in order to get the pixel counts corresponding to the classes of each landslide conditioning factor to calculate Frequency Ratio in the preliminary study.

4.4 Frequency Ratio and Information Gain Ratio Calculation

Using the raster maps generated for the 12 conditioning factors as discussed in the previous section, pixel counts were calculated using raster functionalities available in QGIS. The number of landslide pixels in each class of the conditioning factors, total number of pixels belonging to each class, number of landslide pixels in study area, and total pixel count in the study area were obtained.

Using Clip by Mask Layer raster extraction available in GDAL, each raster layer was clipped by providing the landslide polygon map as the mask. In GIS, clip is to overlay a polygon on one or more target features (layers) and extract from the target feature (or features) only the target feature data that lies within the area outlined by the clip polygon. The clipped raster layer was then reclassified using r.reclassify, a Grass GIS geo-algorithm

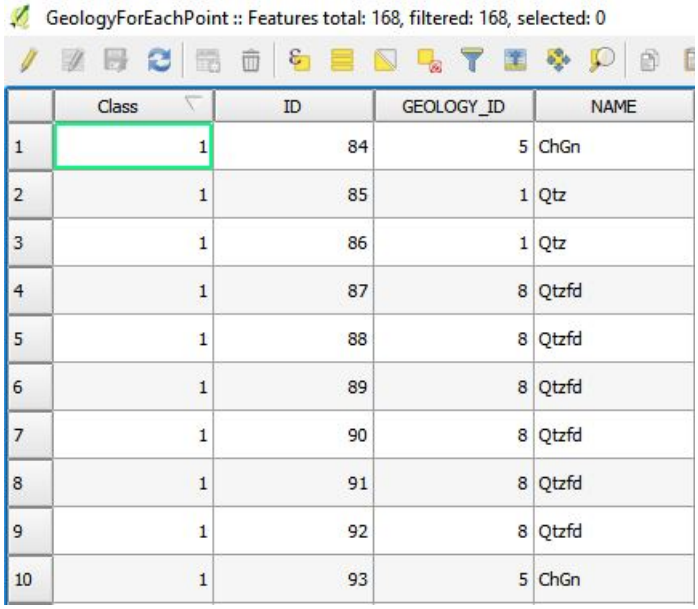
available in QGIS to obtain the classes for each conditioning factor. The reclassified raster was used to report landslide pixel counts for each class of the respective conditioning factor using r.report function. For example the clipped slope map was reclassified into five classes as 0-16, 16-32, 32-48, 48-64 and 64-80 and the number of landslide pixels in each of these classes were generated. The original raster map of the slope was reclassified in the same manner to obtain the total number of pixels in each class.

These values were assigned to the frequency ratio equation and the frequency ratio values were obtained for each class of the conditioning factors. The results obtained for frequency ratio is given in the Results Chapter. From the results obtained from the correlation analysis it was evident that only certain classes of conditioning factors have a high correlation to landslide occurrence.

Similarly, pixel counts were generated from QGIS to get values for Information Gain and Intrinsic Information for each conditioning factor. Information Gain and Intrinsic Information values were used to calculate the Information Gain Ratio. The results obtained for information gain ratio is discussed in the Results Chapter. Information Gain Ratio values indicated that all the 12 conditioning factors initially considered in the study have a positive landslide predictive capability there by demonstrating the relevance of their usage in the prediction. Therefore 12 conditioning factors and the identified classes of the factors were considered in the preparation of the feature pool.

4.5 Feature Pool

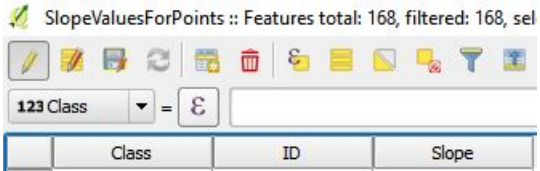
Landslide conditioning factors are the features considered in the landslide susceptibility prediction. To extract features for each landslide and non landslide point, raster and vector functions in QGIS was used. The vector maps of Geology, Landform, Land use, Soil Type & Thickness were intersected with the landslide inventory map separately to generate a new points map containing feature values for each point. Figure 4.6 demonstrates a sample of the attribute table generated for vector map containing geology values for each landslide and non landslide point. This attribute table values were saved to a CSV file in order to prepare the feature pool. Similarly, using the maps generated for land form, land use, soil type and thickness features corresponding to each point were extracted to a CSV file.



	Class	ID	GEOLOGY_ID	NAME
1	1	84	5	ChGn
2	1	85	1	Qtz
3	1	86	1	Qtz
4	1	87	8	Qtzfd
5	1	88	8	Qtzfd
6	1	89	8	Qtzfd
7	1	90	8	Qtzfd
8	1	91	8	Qtzfd
9	1	92	8	Qtzfd
10	1	93	5	ChGn

Figure 4.6: Attribute Table - Geology

Since Slope, Aspect, SPI, TWI, STI, and Rainfall maps were available in raster format and landslide inventory map was available in vector format, a raster-vector processing was carried out to extract the features. A geo-algorithm available in QGIS to add grid values to points was utilized. A sample of the attribute table of the output vector map generated by the algorithm for slope is shown in figure 4.7. The same process was carried out to get values of aspect, SPI, TWI, STI and rainfall for all landslide and non landslide points.

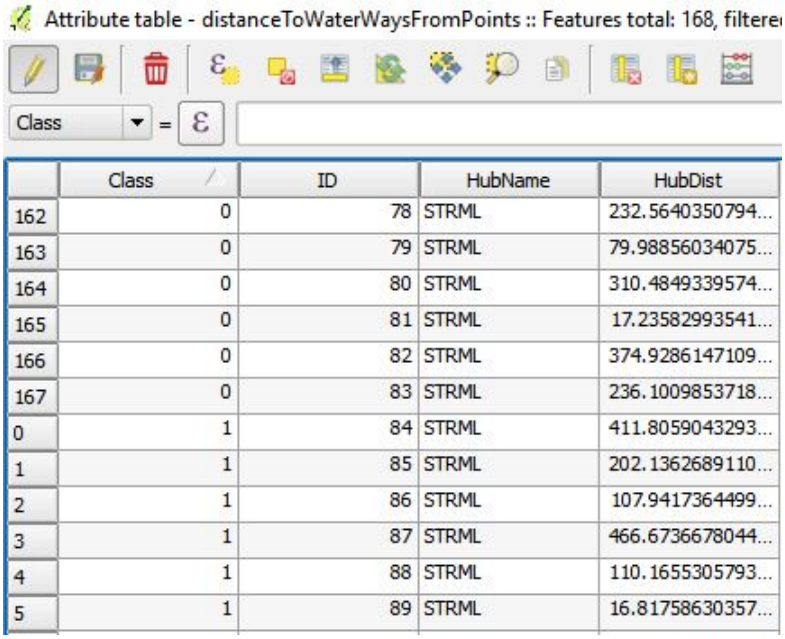


	Class	ID	Slope
1	1	84	7.9263467789
2	1	85	69.7618484500
3	1	86	53.9376983640
4	1	87	20.4804210660
5	1	88	45.3223381040
6	1	89	18.4628047940
7	1	90	37.4714393620
8	1	91	61.4172401430
9	1	92	63.8777389530
10	1	93	51.8706359860

Figure 4.7: Attribute Table - Slope

In order to get the distance from waterways to each landslide and non landslide point

using the Hydrology vector map, "Distance to nearest hub" function under vector analysis in QGIS was used. It created a new vector map with the distances utilizing a geo algorithm to calculate the distance in meters . Figure 4.8 demonstrates the attribute table of the output vector map.



	Class	ID	HubName	HubDist
162	0	78	STRML	232.5640350794...
163	0	79	STRML	79.98856034075...
164	0	80	STRML	310.4849339574...
165	0	81	STRML	17.23582993541...
166	0	82	STRML	374.9286147109...
167	0	83	STRML	236.1009853718...
0	1	84	STRML	411.8059043293...
1	1	85	STRML	202.1362689110...
2	1	86	STRML	107.9417364499...
3	1	87	STRML	466.6736678044...
4	1	88	STRML	110.1655305793...
5	1	89	STRML	16.81758630357...

Figure 4.8: Attribute Table - Hydrology

Considering the continuous values obtained for Slope, Aspect, SPI, TWI, STI, Hydrology and Rainfall at each landslide and non-landslide point, they were divided into equal interval classes and labeled into categories. Then the labeled categories were included in the feature pool as the Slope, SPI, TWI, STI, Hydrology, and Rainfall values for each landslide and non landslide point. Table 4.1 shows the classes and the labels given to each class of Slope, SPI, TWI, STI, Hydrology and Rainfall respectively.

Conditioning Factor	Class	Label
Slope	$0 <= \text{Slope} < 15.7077$	Slope Category1
	$15.7077 <= \text{Slope} < 31.4153$	Slope Category2
	$31.4153 <= \text{Slope} < 47.1230$	Slope Category3
	$47.1230 <= \text{Slope} < 62.8307$	Slope Category4
	$62.8307 <= \text{Slope} < 78.5384$	Slope Category5
	$78.5384 <= \text{Slope}$	Slope Category6

Conditioning Factor	Class	Label
SPI	-1382.9666 <= SPI < 204.0868	SPI Category1
	204.0868 <= SPI < 1791.1402	SPI Category2
	1791.1402 <= SPI < 3378.1936	SPI Category3
	3378.1936 <= SPI < 4965.2469	SPI Category4
	4965.2469 <= SPI < 6552.3003	SPI Category5
	6552.3003 <= SPI	SPI Category6
TWI	7.7498 <= TWI < 9.1141	TWI Category1
	9.1141<= TWI < 10.4784	TWI Category2
	10.4784 <= TWI < 11.8427	TWI Category3
	11.8427 <= TWI < 13.2069	TWI Category4
	13.2069 <= TWI < 14.5712	TWI Category5
	14.5712 <= TWI	TWI Category6
STI	0 <= STI < 6.1262	STI Category1
	6.1262 <= STI < 12.2524	STI Category2
	12.2524 <= STI < 18.3786	STI Category3
	18.3786 <= STI < 24.5049	STI Category4
	24.5049 <= STI < 30.6311	STI Category5
	30.6311 <= STI	STI Category6
Hydrology	2.3399 <= Hydrology < 140.8605	Hydrology Category1
	140.8605 <= Hydrology < 279.3810	Hydrology Category2
	279.3810 <= Hydrology < 417.9015	Hydrology Category3
	417.9015 <= Hydrology < 556.4220	Hydrology Category4
	556.4220 <= Hydrology < 694.9425	Hydrology Category5
	694.9425 <= Hydrology	Hydrology Category6
Rainfall	3215.7876 <= Rainfall < 3465.4297	Rainfall Category1
	3465.4297 <= Rainfall < 3715.0718	Rainfall Category2
	3715.0718 <= Rainfall < 3965.7139	Rainfall Category3
	3965.7139 <= Rainfall < 4214.3560	Rainfall Category4
	4214.3560 <= Rainfall < 4463.9980	Rainfall Category5
	4463.9980 <= Rainfall	Rainfall Category6

Table 4.1: Reclassification of Slope, SPI, TWI, STI, Hydrology and Rainfall

To include these labels in the attribute tables of the previously generated vector maps, field calculator in QGIS was used where an expression was defined to add respective labels of the categories to each point in a new column. Figure 4.9 demonstrates the field calculator parameters given to add the categories for slope. Similar expressions were defined to generate the categories of aspect, SPI, TWI, SPI, Hydrology and Rainfall in their respective attribute tables.

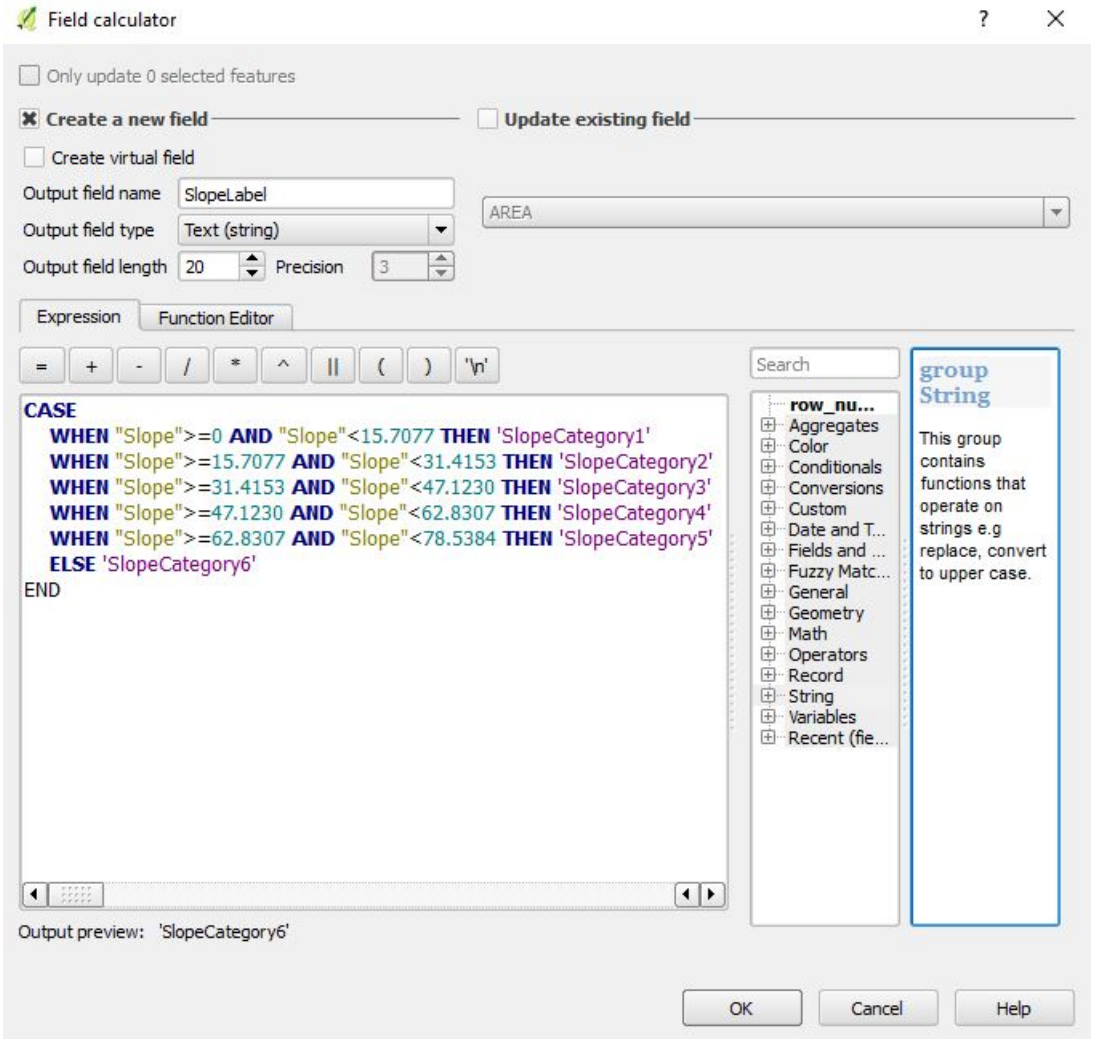


Figure 4.9: Field Calculator for Slope

This categorization was carried out to assist in the visualization of the decision trees in the random forest. Rather than taking the distinct continuous values of each of the above-mentioned attributes(features) as labels, considering the categories as labels stand to reason when the nodes are split in the decision tree. Once the above steps were completed and all the corresponding features were obtained at each landslide and non landslide point, a separate CSV file was created to include these values so that the entire feature pool was maintained in a single file. Figure 4.10 contains a sample of the final csv file used in the

random forest classification.

LandUse	LandForm	SoilTypeAndThickness	Geology	Slope	Aspect	SPI	TWI	STI	Rainfall	DistanceToWaterways	Class
Scrubland	C22	Residual t=2-3	ChGn	SlopeCategory3	AspectCategory5	SPICategory1	TWICategory3	STICategory2	RainfallCategory2	DistanceTWCategory1	0
Rubber	C22	Residual t=1-2	ChGn	SlopeCategory4	AspectCategory4	SPICategory1	TWICategory3	STICategory1	RainfallCategory4	DistanceTWCategory1	0
Rubber	C22	Residual t<1	ChGn	SlopeCategory2	AspectCategory1	SPICategory1	TWICategory3	STICategory1	RainfallCategory3	DistanceTWCategory1	1
Tea	C23	Rock Exposure	Qtzfd	SlopeCategory5	AspectCategory3	SPICategory1	TWICategory3	STICategory1	RainfallCategory3	DistanceTWCategory2	1
Rubber	C22	Residual t=1-2	Qtzfd	SlopeCategory4	AspectCategory2	SPICategory1	TWICategory4	STICategory1	RainfallCategory3	DistanceTWCategory1	1
Scrubland	C22	Residual t=1-2	ChGn	SlopeCategory2	AspectCategory1	SPICategory1	TWICategory4	STICategory1	RainfallCategory4	DistanceTWCategory2	0
Village Homlets	C22	Residual t=1-2	Qtzfd	SlopeCategory2	AspectCategory2	SPICategory1	TWICategory2	STICategory1	RainfallCategory3	DistanceTWCategory4	1
Mixed Tree Crops	C22	Residual t=1-2	ChGn	SlopeCategory1	AspectCategory3	SPICategory1	TWICategory2	STICategory1	RainfallCategory3	DistanceTWCategory1	1
Degraded Forests	D37	Colluvium t=2	ChGn	SlopeCategory5	AspectCategory2	SPICategory1	TWICategory2	STICategory1	RainfallCategory2	DistanceTWCategory4	1

Figure 4.10: Final Dataset

4.6 Prediction Model

Random Forest Classifier available in Scikit Learn which is a free software machine learning library for the Python programming language was used to execute random forest algorithm on the final data set prepared using the steps discussed above.

4.6.1 Preprocessing

As the initial step in fitting the data for random forest classification, categorical data was normalized such that they contain only values between 0 and n_classes-1. LabelEncoder utility class available in sklearn was used to carry our this task. The python function written to achieve this given in listing 4.1.

```
from sklearn import preprocessing
def encodeDataset(data):
    encode_data = preprocessing.LabelEncoder()
    dataset['LandUse'] = encode_data.fit_transform(data.LandUse.astype(str))
    dataset['LandForm'] = encode_data.fit_transform(data.LandForm.astype(str))
    dataset['SoilTypeAndThickness'] = encode_data.fit_transform(data.SoilTypeAndThickness.astype(str))
    dataset['Geology'] = encode_data.fit_transform(data.Geology.astype(str))
    dataset['Slope'] = encode_data.fit_transform(data.Slope.astype(str))
    dataset['Aspect'] = encode_data.fit_transform(data.Aspect.astype(str))
    dataset['SPI'] = encode_data.fit_transform(data.SPI.astype(str))
    dataset['TWI'] = encode_data.fit_transform(data.TWI.astype(str))
    dataset['STI'] = encode_data.fit_transform(data.STI.astype(str))
```

```
13     dataset['Rainfall'] = encode_data.fit_transform(data.Rainfall.astype(
       str))
14     dataset['Hydrology'] = encode_data.fit_transform(data.Hydrology.astype
       (str))
15     data=data.fillna(-999)
16     return data
```

Code Listing 4.1: Encode Data

4.6.2 Training and Testing

The 12 features were assigned to X while the class(landslide or non landslide) was assigned to Y. The data set was split in the ratio of 70:30 for the training and testing respectively using the code snippet given below(Listing 4.2).

```
1 from sklearn.model_selection import train_test_split
2
3 X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,
4 random_state=1,shuffle='true')
```

Code Listing 4.2: Split Data

X and y train sets were utilized in the training phase to fit to random forest classifier in sklearn. There are several hyper parameters taken by the classifier as inputs. Some of them are,

1. n_estimators: Number of decision trees.
2. criterion: The function to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain.
3. max_depth: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
4. min_samples_split: The minimum number of samples required to split an internal node
5. min_samples_leaf: The minimum number of samples required to be at a leaf node.
6. max_features: The number of features to consider when looking for the best split.
7. bootstrap: Whether bootstrap samples are used when building trees.

8. oob_score: Whether to use out-of-bag samples to estimate the generalization accuracy.

9. random_state: Controls both the randomness of the bootstrapping of the samples used when building trees and the sampling of the features to consider when looking for the best split at each node

10. class_weight: Weights associated with classes in the form class_label: weight.

In order to get the model with the best predictive performance a set of these hyper parameters were tuned. The optimal values were experimentally obtained by testing different possible combinations. An example of a combination tried during the training phase is demonstrated in code listing 4.3

```
#accuracy=74.51%
classifier=RandomForestClassifier(n_estimators=50,criterion='entropy',
    max_depth=10,min_samples_split=4,min_samples_leaf=1,max_features='sqrt'
    ,class_weight='balanced',bootstrap='true',random_state=0,oob_score='
    true')
```

Code Listing 4.3: RF Classifier Hyper-Parameters

There were differences in accuracy of the model when the values of the hyper-parameters were changed. Considering these differences in the model performance the optimal set of hyper-parameters were selected for the model. Further analysis on the hyper-parameters and the optimal set of hyper-parameters selected for the classifier is discussed in Chapter 5. After providing the parameters to the classifier, it was fit to x and y training sets.

Once the random forest classifier was fit to x and y training sets, x test set was used to carry out the prediction using the classifier in the testing phase. Listing 4.4 provides the code snippet used for model fitting and prediction of landslide susceptibility.

```
classifier.fit(X_train,y_train)
y_pred=classifier.predict(X_test)
```

Code Listing 4.4: Model Fitting and Prediction

4.6.3 Model Assessment

In order to assess the performance of the landslide susceptibility prediction model confusion matrix and the ROC curve was generated. Using the confusion matrix values for accuracy, sensitivity, and specificity were calculated. ROC curve was used to calculate the AUC value. Listing 4.5 was used to calculate the above mentioned performance measures.

```
from sklearn import metrics
from sklearn.metrics import roc_curve

cm=confusion_matrix(y_test,y_pred)

Accuracy=float(cm[0,0]+cm[1,1])/float(cm[1,0]+cm[1,1]+cm[0,0]+cm[0,1])
Sensitivity = float(cm[0,0])/float(cm[0,0]+cm[0,1])
Specificity=float(cm[1,1])/float(cm[1,0]+cm[1,1])

probs = classifier.predict_proba(X_test)
probs = probs[:, 1]
fper, tper, thresholds = roc_curve(y_test, probs)
AUC_value=metrics.auc(fper,tper)

print('kappa index: ', metrics.cohen_kappa_score(y_test,y_pred,weights='
    quadratic'))
```

Code Listing 4.5: Performance Measures

4.7 Summary

This chapter elaborates the implementation of the research design in terms of technicality associated. The topics discussed in the chapter includes the use of QGIS as a geographic information system tool, to generate thematic maps for conditioning factors, manipulation of the maps to extract the required features and implementation of the landslide susceptibility prediction model using python.

Chapter 5

Evaluation and Results

This research was aimed at identifying the applicability of 12 conditioning factors to predict landslides in Kalutara district using Random Forest machine learning technique. Spatial analysis of the data acquired from the National Building Research Organization and Meteorological Department of Sri Lanka was carried out using QGIS from which the feature pool for the random forest classifier was prepared. After the data preparation, implementation and training of the model was carried out using scikitlearn.

This chapter unravels the results of the entire work process of the study. It provides detailed description of the thematic maps generated from the initial data set, analysis on the results of the pilot study and assessment on the performance of Random Forest based prediction model.

The pilot study was carried out to identify the spatial relationships between each class of conditioning factors and landslide events and to quantify the predictive capability of individual conditioning factors. Analysis on the prediction model include discussion on different approaches adopted in implementing the model and performance evaluation using statistical measures and Area Under the ROC curve.

5.1 Evaluation of the Thematic Maps

As discussed in the previous section thematic maps were generated in vector format for geology, land form, land use, hydrology, soil type and soil thickness. An analysis on each of those maps are given below.

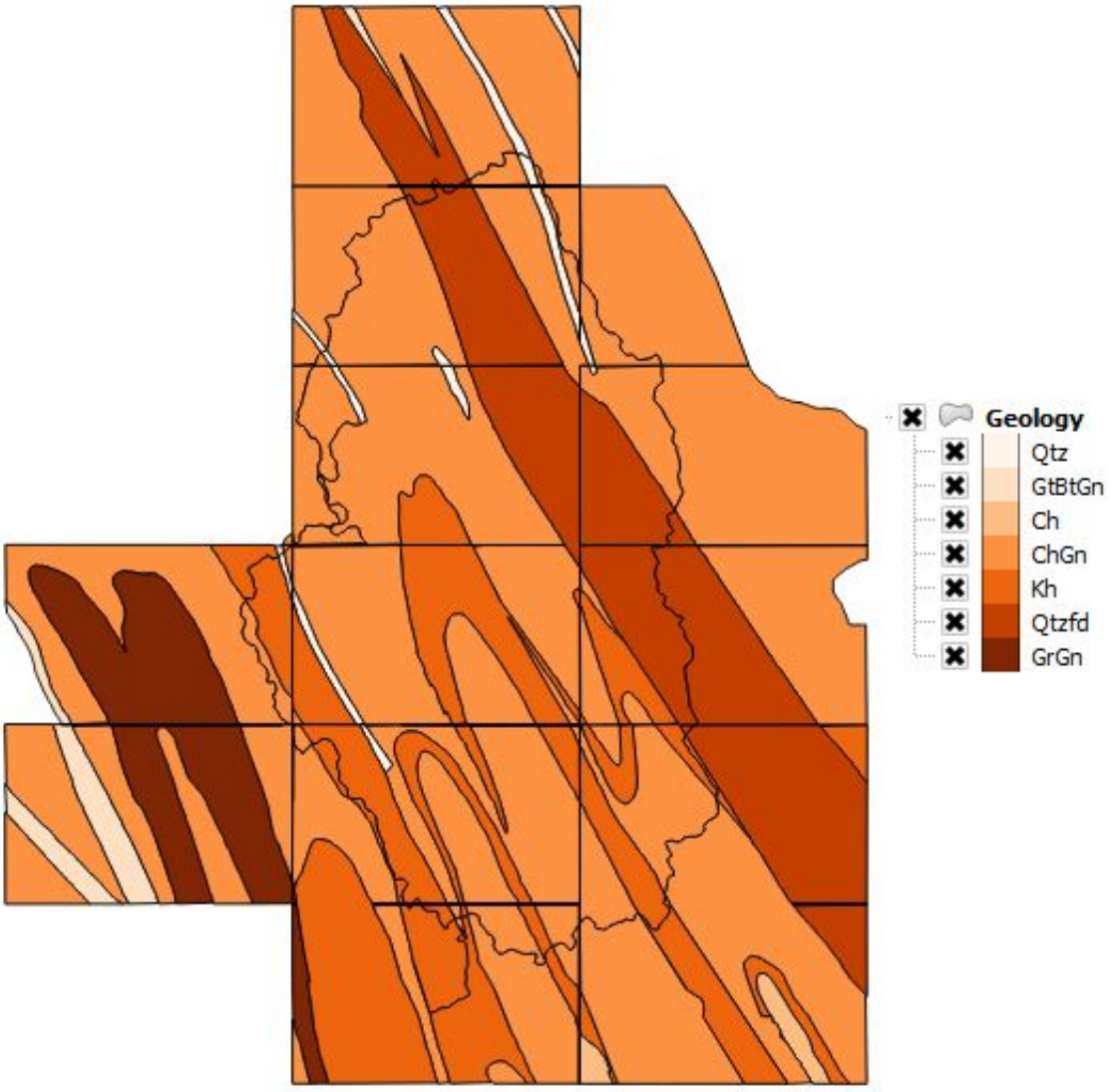


Figure 5.1: Geology Thematic Map

Kalutara district consists of 7 types of geologies as depicted in figure 5.1. They are Quartzite(Qtz), Granite Biotite Gneiss(GtBtGn), Charnochite(Ch), Charnockitic Gneiss(ChGn), Khondalite(Kh), Quartzo Feldspathic(Qtzfd), and Granulatic Gneiss (GrGn). The spatial distribution of the geology demonstrates that the study area mostly consists of lithologies containing Charnockitic Gneiss followed by Khondalite.

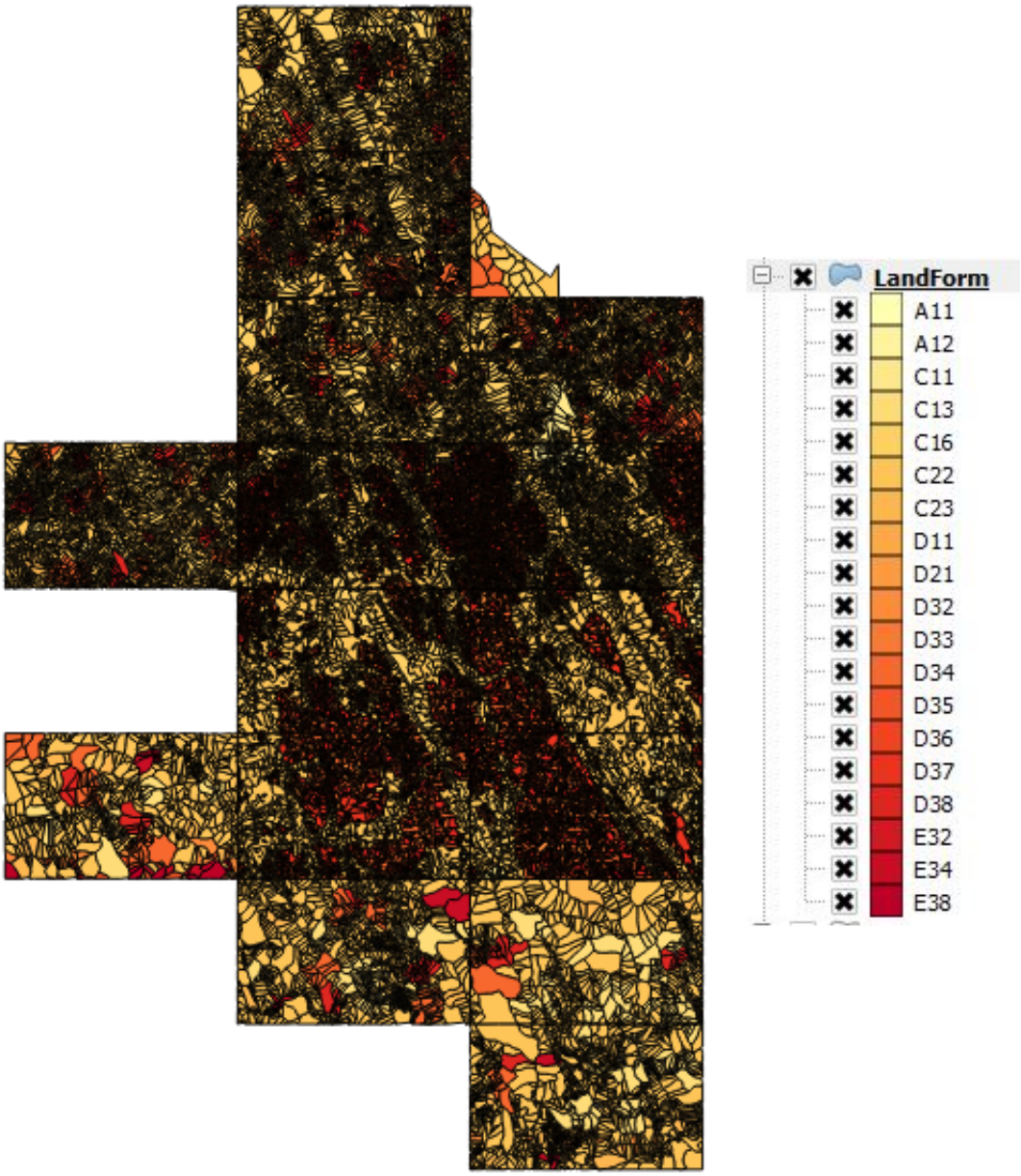


Figure 5.2: Land Form Thematic Map

The thematic map(Figure 5.2) generated for land form indicate five types of Erosion Landforms and eight types of Slope Landforms in Kalutara. Colluvial benches(A11), river/stream capture zones(A12), dissected gullied surfaces(C11), fault fracture zones(C13), and landslide scars(C16) are among the erosion landforms. Slope landforms present in the study area include dissected plateaus(C22), straight slopes(C23), undulating land(D11), isolated hillocks(D21), complex slope (D32), corrugated slope(D33), drainage basins(D34), corrugated and complex slope (D35), straight mountain slope(D36), mountain slopes(D37), and talus/screen slopes (D38), complex hill(E32), dissected hill slope(E34), and complex corrugated and dissected slope(E38). Dissected plateaus, drainage basins, and talus/screen slopes are the most common land forms found in the region. Complex hills have the least

significance in the land form of the area.

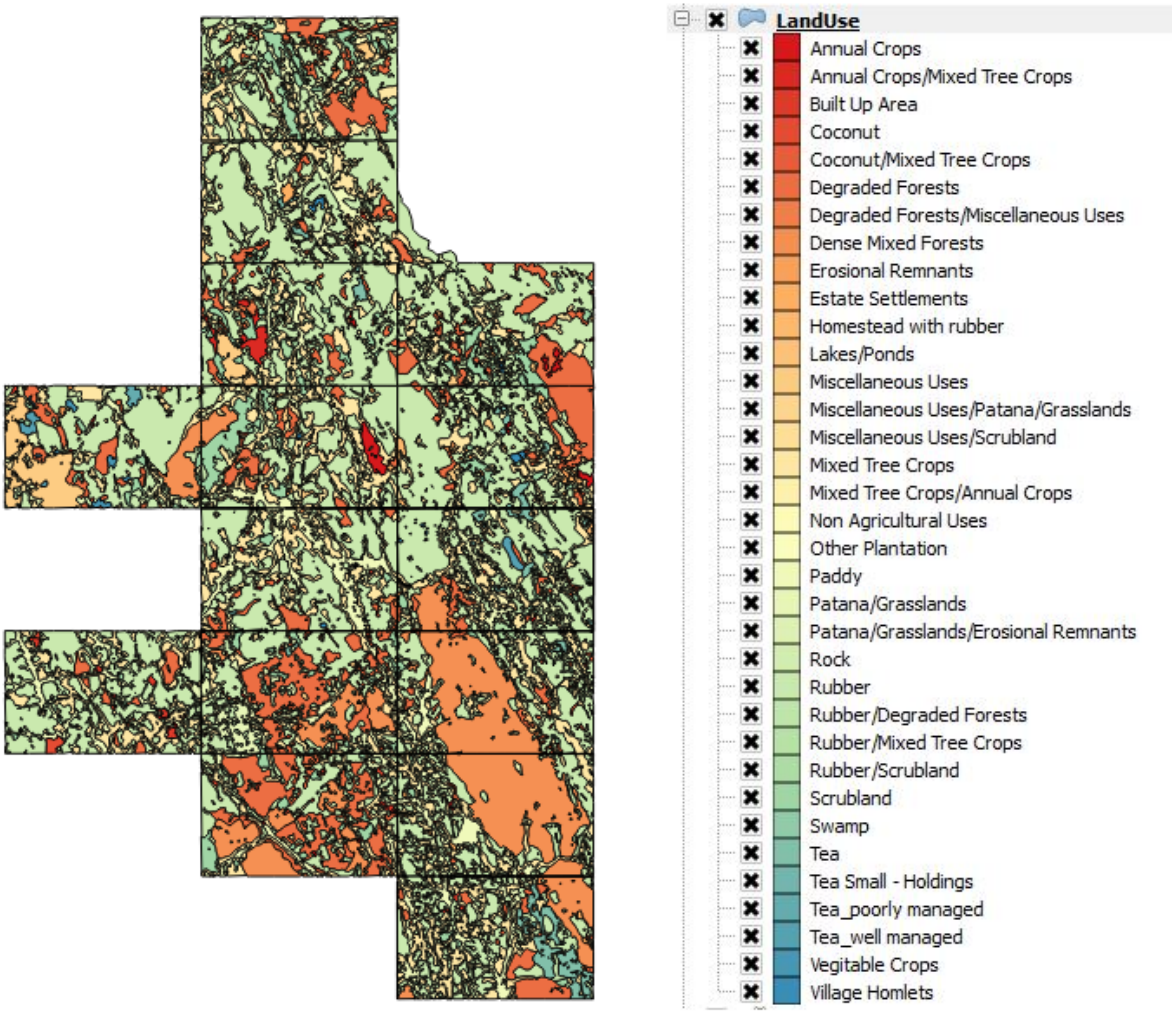


Figure 5.3: Land Use Thematic Map

The land use in Kalutara, is rapidly changing from rubber plantation to other commercial crops. According to the land use thematic map(Figure 5.3) it can be seen that mixed tree crops, paddy and rubber cover a major part of land use in Kalutara district. Other types of crops such as coconut, tea, annual crops and vegitable crops are also present. Forest cover in the area mainly consists degraded forests and scrublands. Densed mixed forests, patana\grasslands, erosional remnants, swamps, and lakes\ponds can also be found. Land is also being used for other miscellaneous purposes, non-agricultural uses, estate settlements and village homelets.

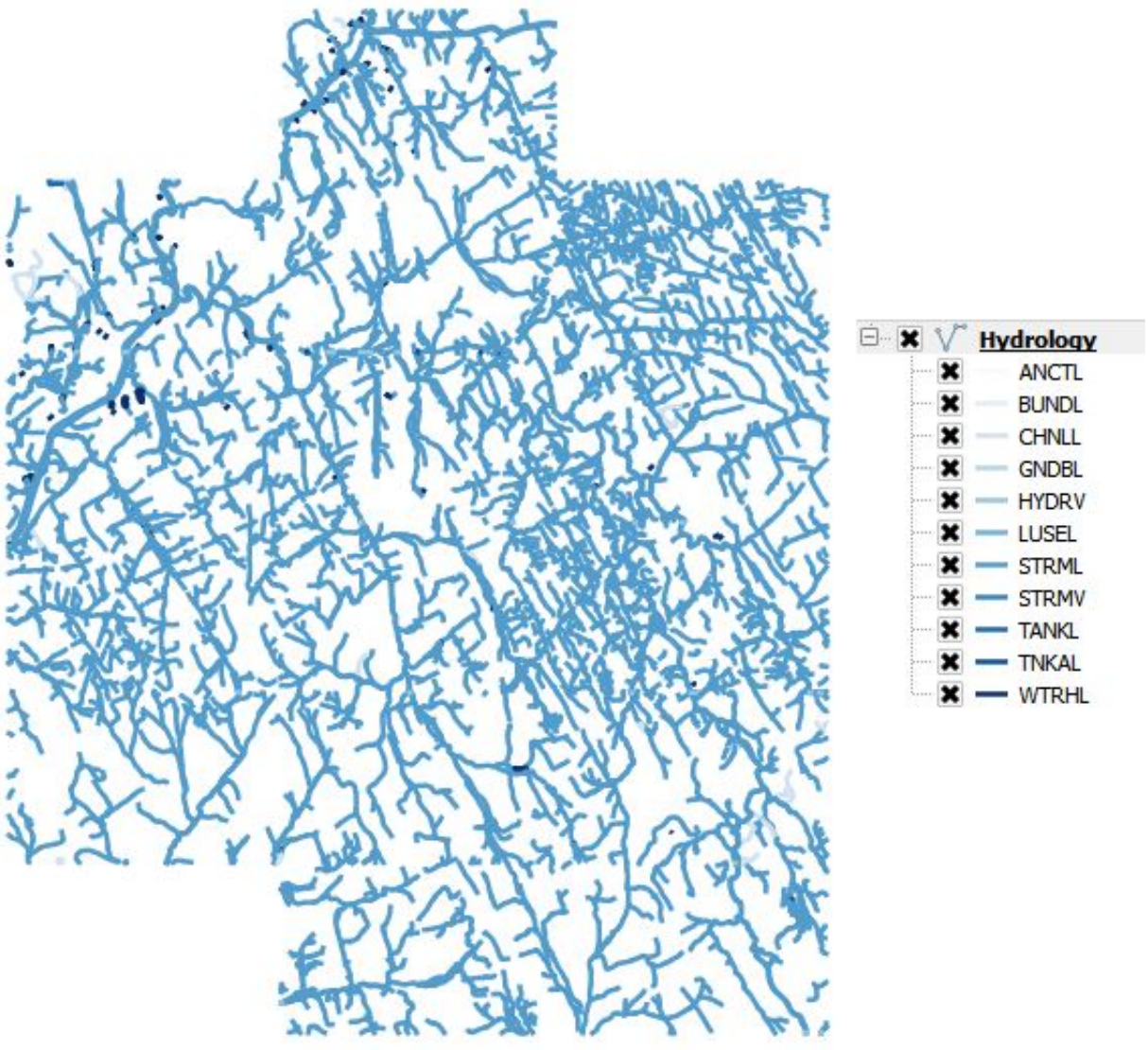


Figure 5.4: Hydrology Thematic Map

Figure 5.4 demonstrated the hydrology map with 11 categories of water sources dispersed in the study area. They are lake streams(STRML), rivers(STRMV), channels(CHNLL), ground water sources(GNDBL), tanks(TANKL), watersheds(WTRHL), lake bunds (BUNDL), lake anicuts(ANCTL), river basins(HYDRV), lagoons(LUSEL), and artificial tanks (TNKAL). Significant portion of the hydrology of the district constitute of lake streams.

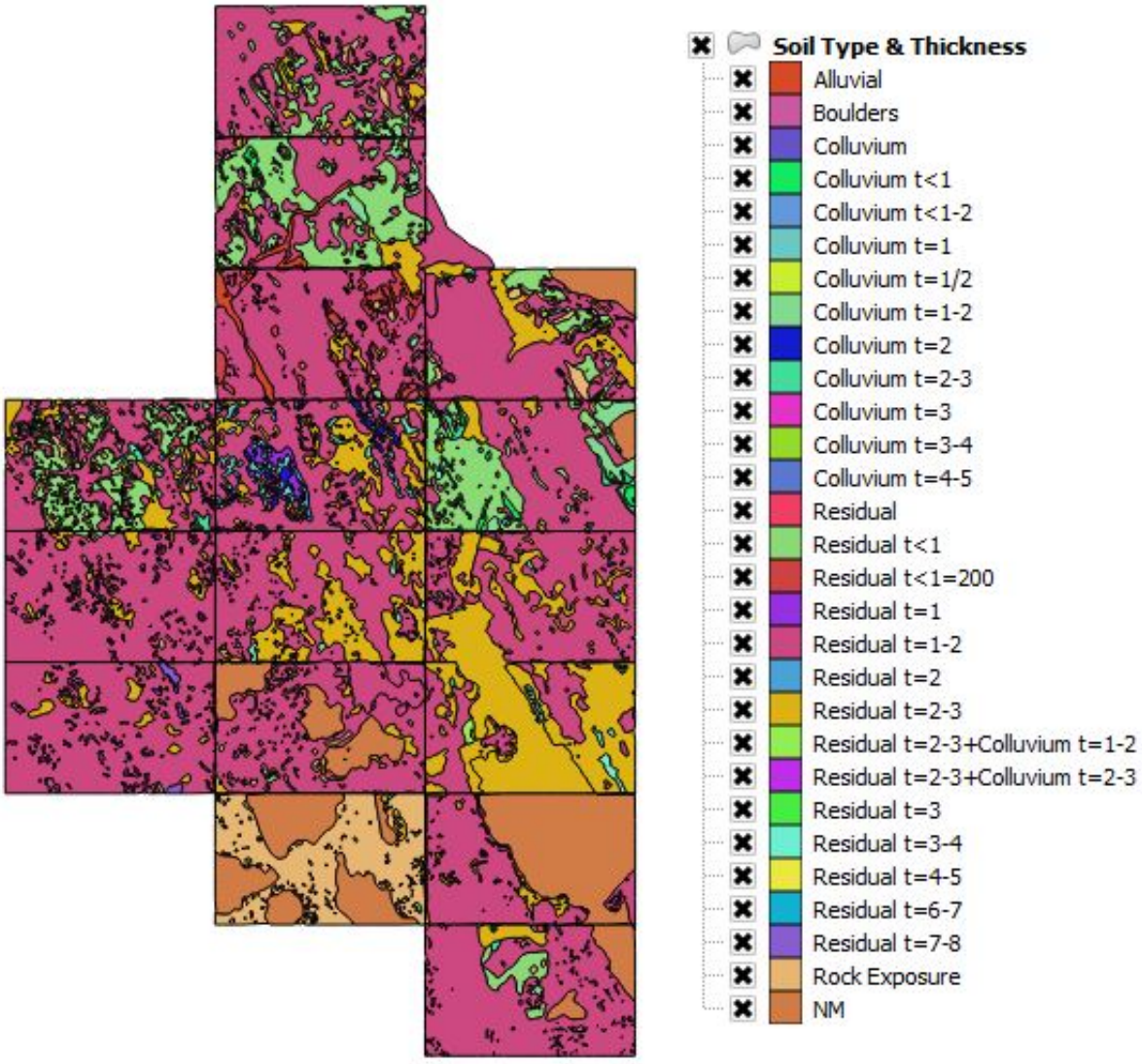


Figure 5.5: Soil Type and Thickness Thematic Map

Since Kalutara belong to the wet zone of Sri Lanka, the climatic influences have affected the dominance of red-yellow podzolic soil types [63] in the study area. Thematic map(Figure 5.5) generated for soil type comprises of five types of soil which are alluvial, boulders, colluvium, residual and rock exposure. Residual soils cover major part of the study area. Alluvial soils have 1m thickness, colluvium soils have 1-5m of thickness and residual soils have 1-8m of thickness while boulders and rock exposure having no thickness in the district.

Thematic maps generated in raster format for slope, aspect, SPI, TWI, STI and rainfall demonstrated continuous values. Therefore they were reclassified into equal interval classes. Analysis on the above mentioned raster maps are given below.

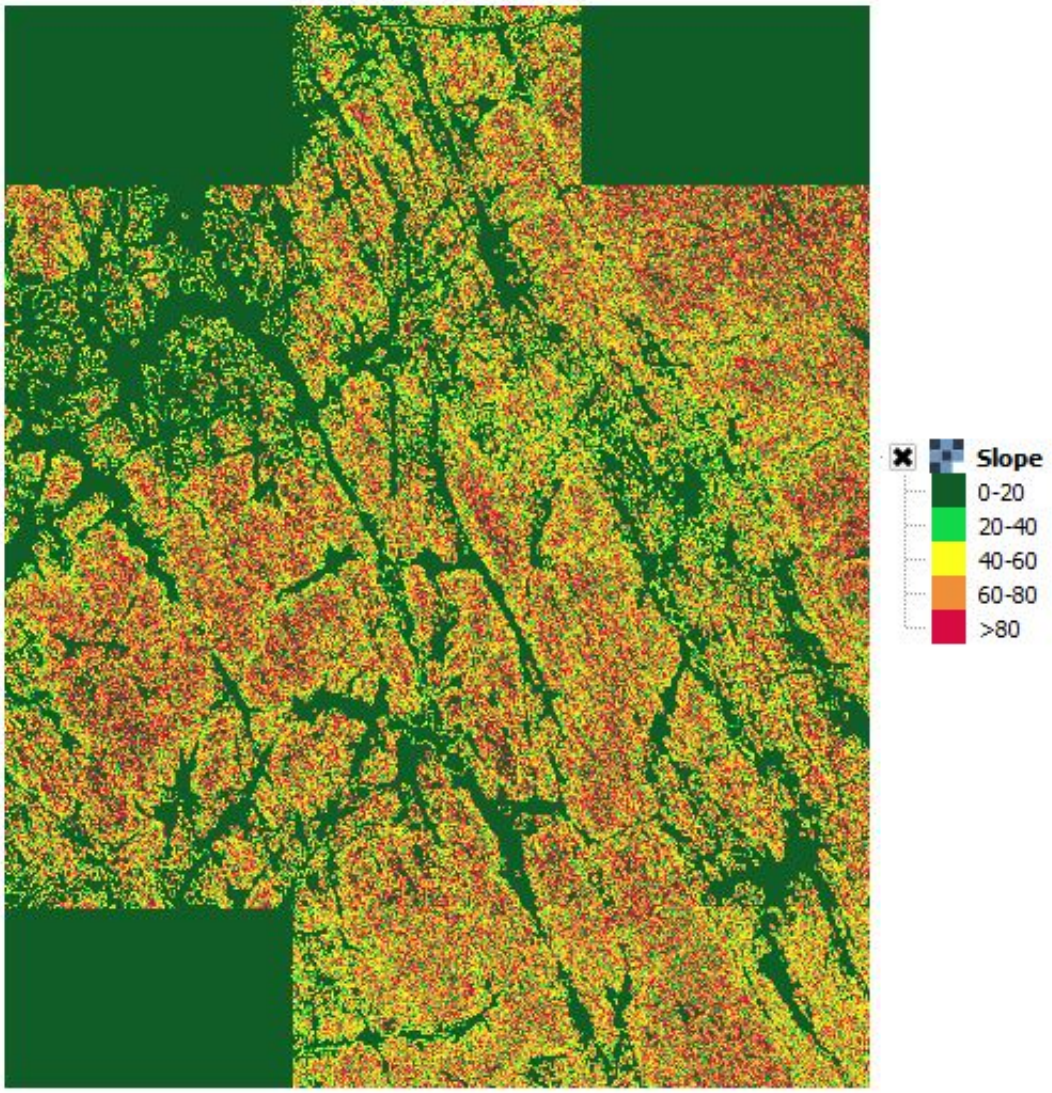


Figure 5.6: Slope Thematic Map

Slope ranged from 0°-80°and it was classified into five classes as 0°-20°, 20°-40°, 40°-60°, 60°-80°, >80°. It was evident that larger part of the study area consisted of slope ranging from 0° to 20° followed by slopes ranging from 60° to 80° and 40° to 60°.

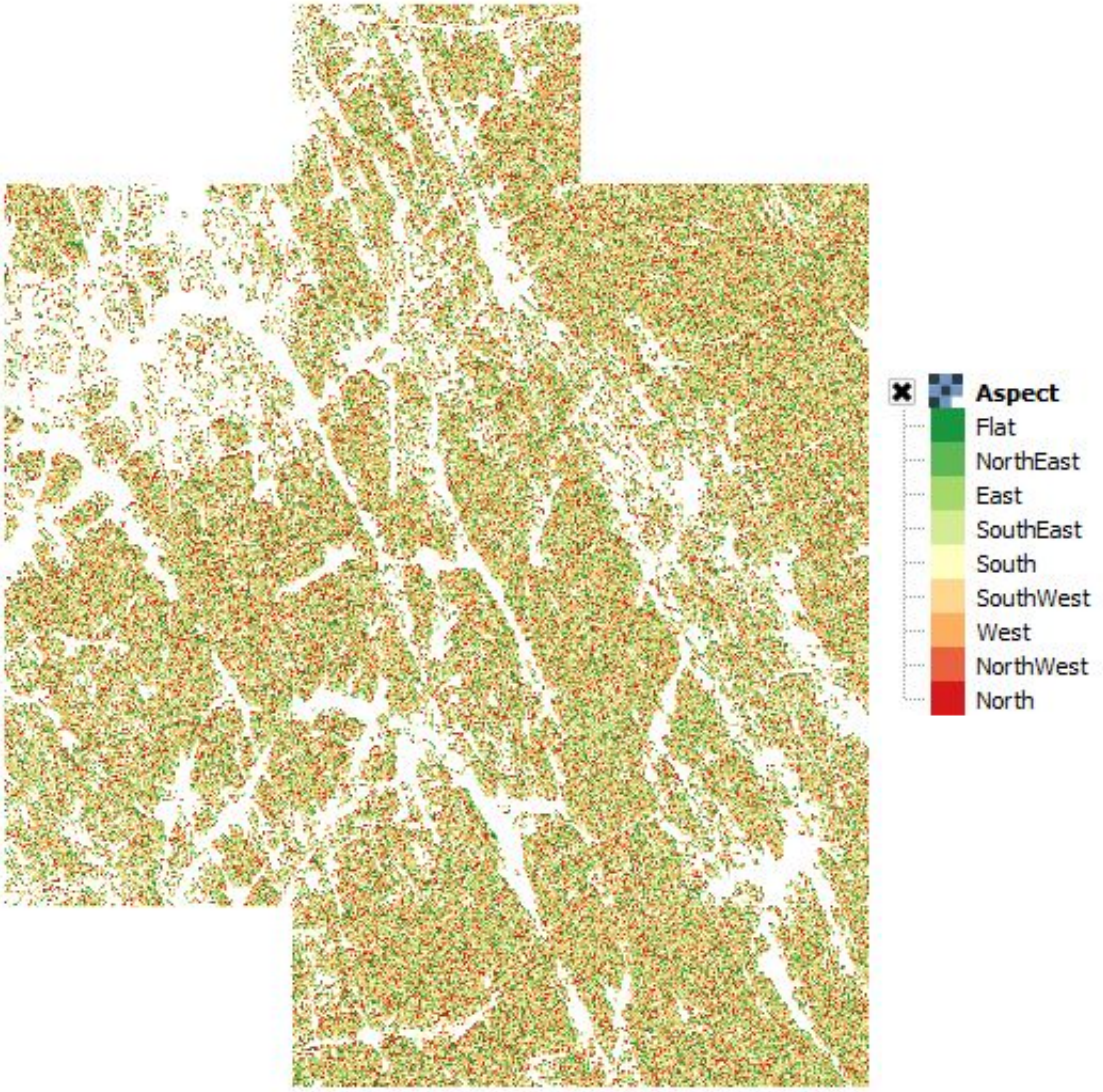


Figure 5.7: Aspect Thematic Map

Aspect map was grouped into nine classses as Flat(0°-22.5°), NorthEast(22.5°-67.5°), East(67.5°-112.5°), SouthEast(112.5°-157.5°), South(157.5°-202.5°), SouthWest (202.5°-247.5°), West(247.5°-292.5°), NorthWest(292.5°-337.5°), and North (337.5°-360°). The aspect distribution indicated that Kalutara district mostly has West, East, South West and North East aspects.

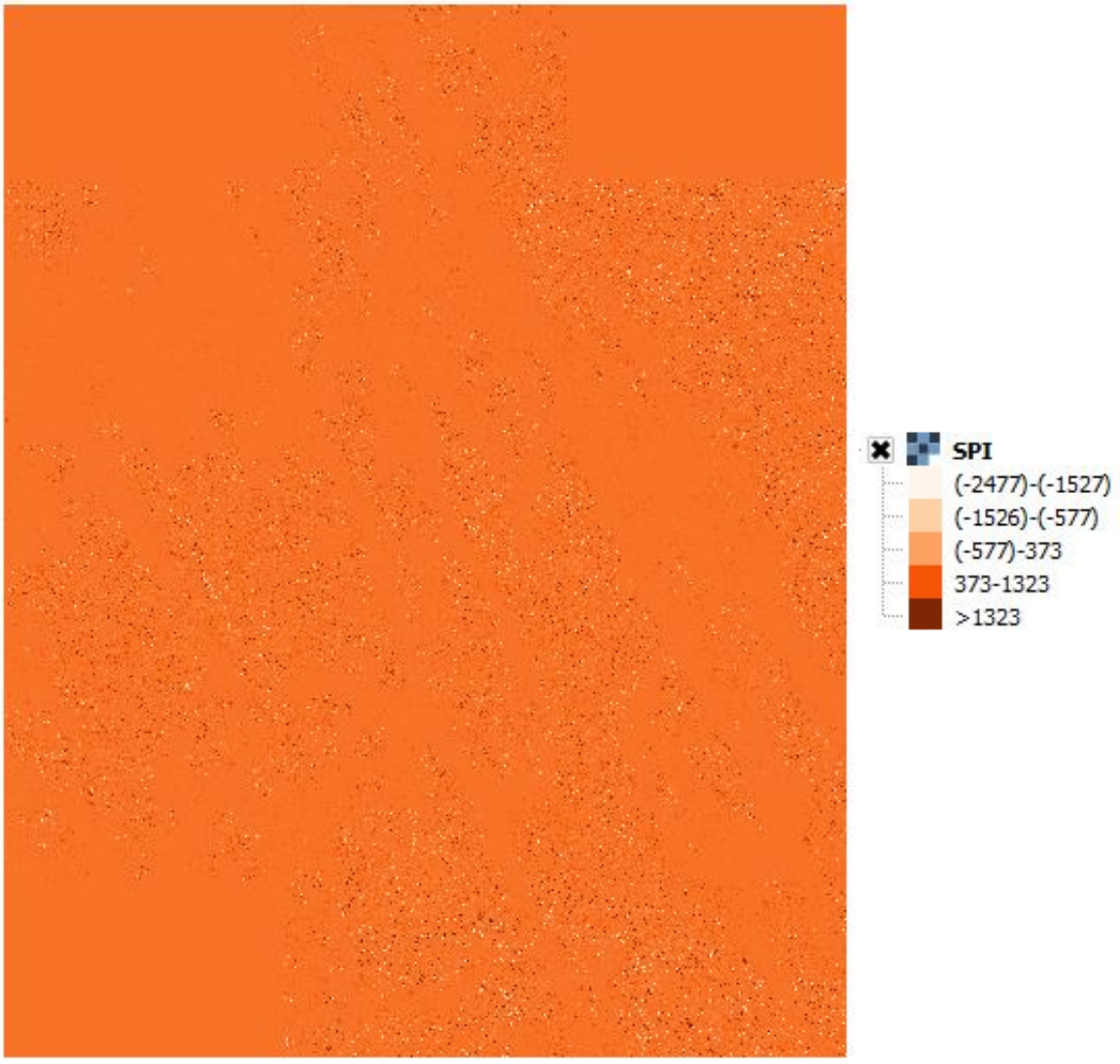


Figure 5.8: SPI Thematic Map

SPI map consisted of five equal interval classes including (-2477)-(-1527), (-1527)-(-577), (-577)-373, 373-1323, and >1323. Stream power index values generated for the study area mostly belonged to 373-1323 class.

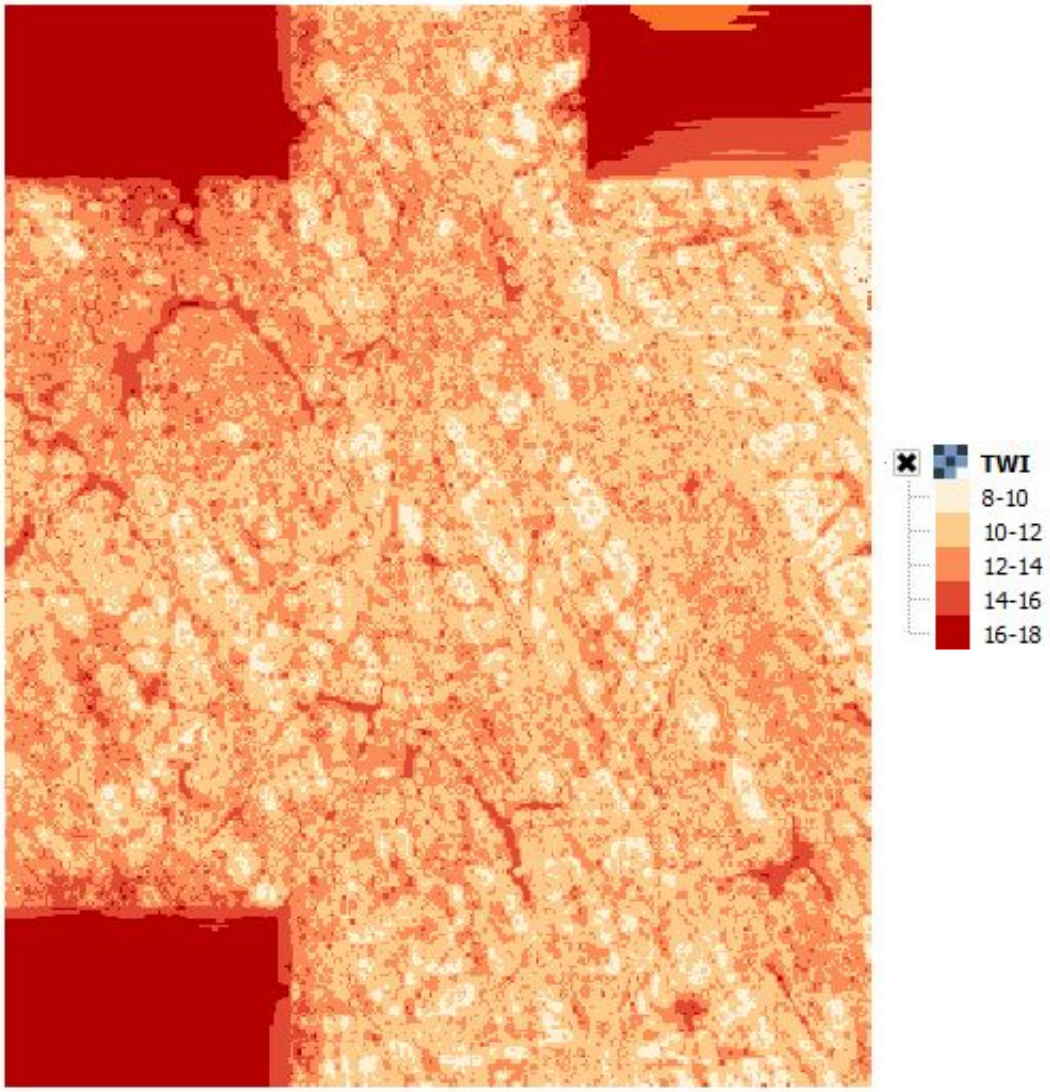


Figure 5.9: TWI Thematic Map

TWI values were classified into five classes as 8-10, 10-12, 12-14, 14-16, and 16-18 in the map. From the distribution of TWI vaues in the map, it was evident that the degree of water accumulation in the district was between 10 to 12 and 12 to 14.

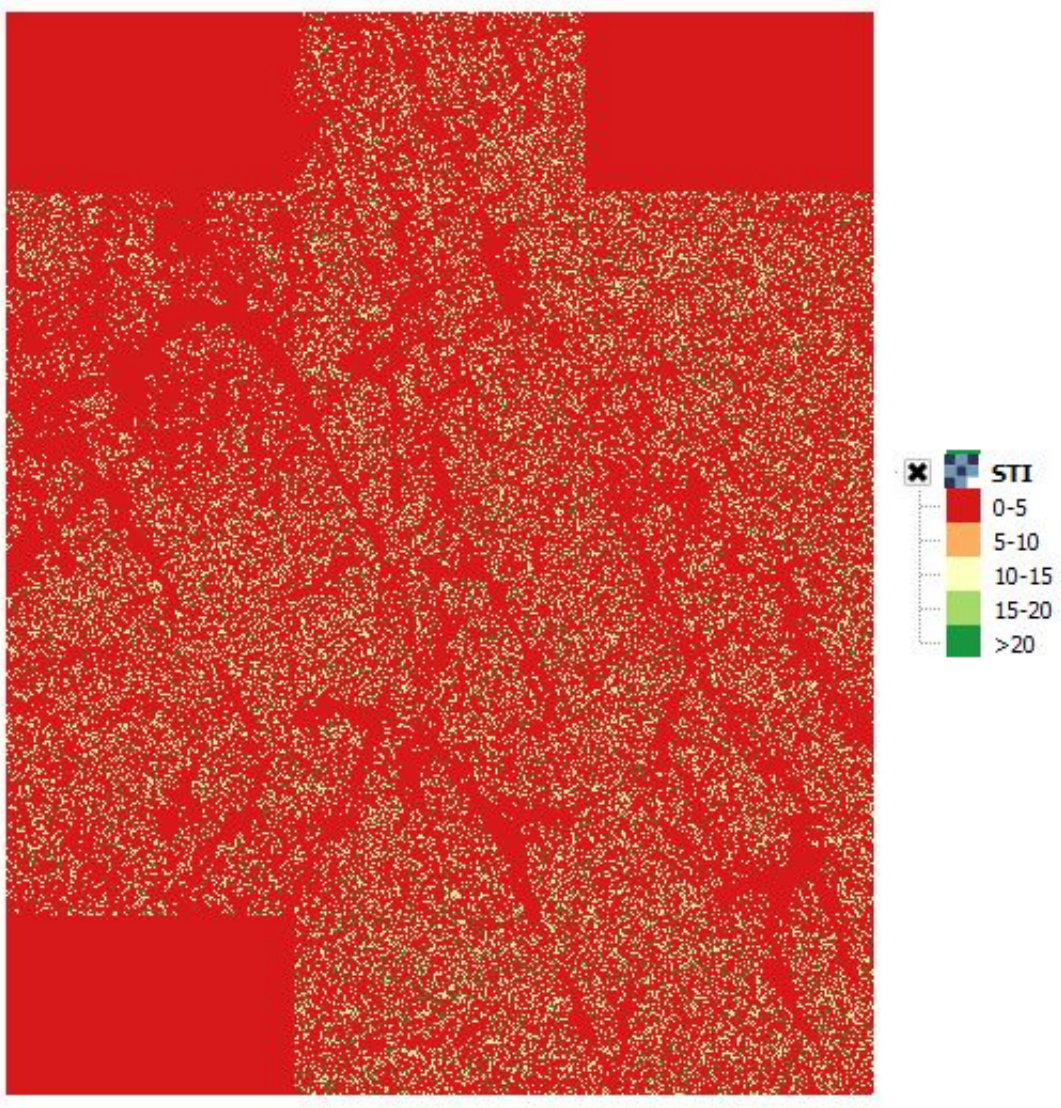


Figure 5.10: STI Thematic Map

As shown in figure 5.10 0-5, 5-10, 10-15, 15-20, >20 were the classes used for the reclassification of the STI thematic map. Predominant extent of the study area consisted of STI values ranging from 0 to 5.

Figure 5.11 demonstrates the thematic map generated for rainfall in Kalutara district. It was classified into 4 classes as 2490mm-2940mm, 2940mm-3390mm, 3390mm-3840mm, and 3840mm-4290mm. Rainfall distribution indicated that average rainfall of the study area mostly ranges from 3390mm to 4290mm. South West monsoon season contribute to 73% [64] of the annual rainfall in Kalutara district. The minimum rainfall values experienced in the study area was between 2490mm and 2940mm.

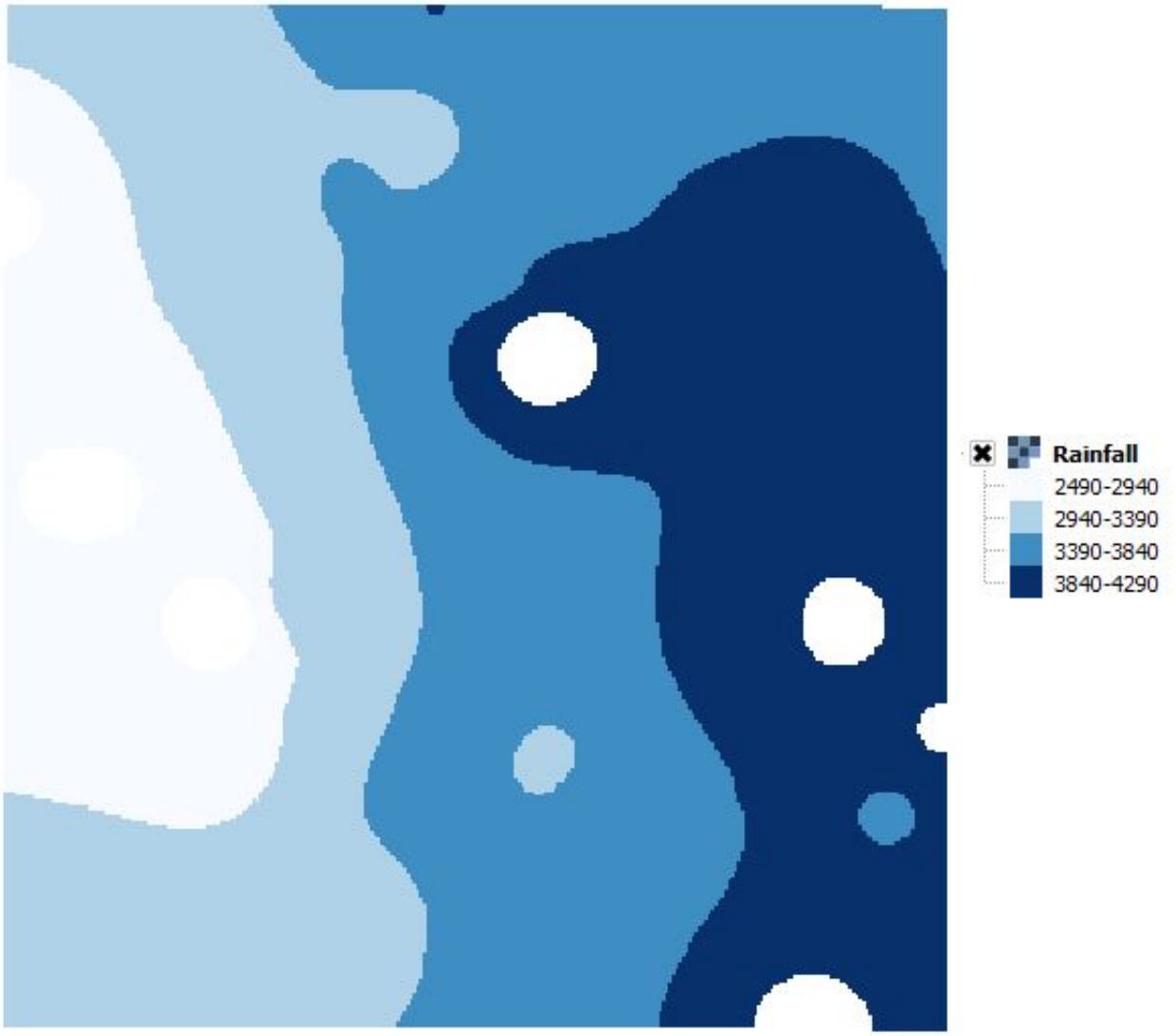


Figure 5.11: Rainfall Thematic Map

Through evaluation of the thematic maps as discussed above, a comprehensive understanding of the topographical, geological, morphological and hydrological features of the study area was obtained.

5.2 Pilot Study

An initial study was carried out to analyse the spatial relationship between each class of landslide conditioning factors and landslide occurrence and to quantify the predictive capability of conditioning factors. The results obtained in the pilot study assisted in identifying the highly correlated classes of conditioning factors with landslide occurrence and elimination of conditioning factors having null predictive capability in the study area. It paved way in improving the quality of the constructed feature pool thereby enhancing the reliability

of the prediction model. Following subsections provide the results and description of the evaluation of results of the pilot study.

5.2.1 Correlation Analysis

Correlation between landslide locations and conditioning factors were revealed using Frequency Ratio(Equation 3.4). Table 5.1 demonstrates the frequency ratio values obtained for each class of conditioning factors. When FR>1 it indicates a higher correlation, while FR<1 indicates a lower correlation. An average value is informed by FR=1.

Conditioning Factor	Class	Frequency Ratio
Slope	0-16	0.242
	16-32	1.908
	32-48	2.179
	48-64	1.739
	64-80	1.265
Aspect	Flat	0.971
	NorthEast	0.948
	East	1.012
	SouthEast	0.974
	South	1.123
	SouthWest	1.058
	West	1.046
	NorthWest	0.929
	North	0.828
SPI	(-2480)-(-1719)	0.378
	(-1719)-(-958)	2.144
	(-958)-(-197)	1.933
	(-197)-564	0.970
	564-1325	0.918
TWI	8-10	2.039
	10-12	1.682
	12-14	0.502
	14-16	0
	16-18	0

Conditioning Factor	Class	Frequency Ratio
STI	0-5	0.887
	5-10	1.287
	10-15	1.690
	15-20	1.739
	20-25	2.579
Soil Type	Colluvium	2.99
	Alluvial	2.76
	Residual	1.02
	Boulders	0.68
	Rock Exposure	0.35
Soil Thickness	Rs t=4-5	0.597
	Rs t=3-4	0.140
	Rs t=2-3	0.132
	coll t=1	0.108
	coll t=2	0.024
Landform	D38	0.107
	D32	0.082
	D37	0.069
	C22	0.043
	C23	0.036
	C16	0.028
	D34	0.022
Land Use	Mixed Tree Crops	0.005
	Paddy	0.002
	Annual Crops	1.524
	Scrub Land	0.053
	Rubber	0.006
	Tea	0.052
	Degraded Forest	0.051
	Dense Mixed Forests	0.058
	Village Home-lets	10.583
	Rubber/Scrub Land	1.253

Conditioning Factor	Class	Frequency Ratio
Geology	QtzFd	2.29
	Qtz	1.55
	ChGn	1.07
	GrGn	0.07
	GtBtGn	0
	Ch	0
	Kh	0
Hydrology	33-270	4.235
	270-573	1.447
	573-843	0.548
	843-1113	0.115
	1113-1383	0
Rainfall	2490-2940	0
	2940-3390	0
	3390-3840	0.489
	3840-4290	3.094

Table 5.1: Frequency Ratio values for Conditioning Factors

For slope angles between 0 to 16, the frequency ratio was 0.242, which indicates low probability of landslide occurrence. Since all other slope classes depict FR values greater than 1, it can be concluded that they have high correlation to landslide events. The maximum ratio value is obtained for slope angle between 32 and 48.

The probability of landslide occurrence in aspects facing south, south east, west and east is the highest with frequency ratio values of 1.123, 1.058, 1.046 and 1.012 respectively. North facing aspects(FR=0.828) demonstrates the least susceptibility to landslides. Other types of aspects including flat, north east, south east and north west also have FR values close to 1, which may suggest a considerable probability of occurring landslides in such aspects.

It can also be seen that SPI values(FR=0.378) ranging from (-2480) to (-1719) are least susceptible to landslides. The highest correlation with landslide occurrence is indicated by SPI values between (-1719) to (-958) with a FR value if 2.144. When SPI is between (-958) to (-197) also the probability of landslide occurrence can be considered high due to its FR value(1.933).

The results indicate that TWI between 8 and 12 has highest susceptibility to landslides in the study area with FR values 2.039 and 1.682. TWI greater than 14 show the least probability of landslide occurrence. When STI values are between 20 and 25, it demonstrates an FR value of 2.579 suggesting a high probability of landslides. Classes of STI including 5-10(FR=1.287), 10-15(FR=1.690) and 15-20(1.739) also have a considerably high correlation to landslide events since they have FR values greater than 1.

In terms of soil type, colluvium has the highest FR(2.99). Therefore, the probability of landslides occurring with colluvium soil type is greater than that of the other soil types.

The relationship between landslides and landform shows that the landform classes of D38 and D32 have FR values of 0.107 and 0.082 and the greatest potential for landslide occurrence among the landform classes. The results also showed that land use for village home-lets has the highest(FR=10.583) probability for landslide occurrence. Residual soil thickness in the range of 4-5m and colluvium soil type with 2m thickness have the highest(FR=0.597) and lowest(FR=0.024) susceptibility to landslide incidence, respectively.

Resulting FR values indicate that Quartzo Feldspathic geology class demonstrates an increased susceptibility (FR = 2.29) to landslide occurrence compared with the other classes of geology. Moreover, geology with GraniteBiotite Gneiss, Charnochite, and Khondalite, has the lowest FR (0.000), indicating a low probability of landslide occurrence.

Frequency Ratio values calculated for hydrology indicate that here is high probability of landslide occurrence when the distance to waterways is between 33m to 270mm(FR=4.235). The least correlation(FR=0) between landslide occurrence is for distance to waterways from 1113m to 1383m. Rainfall between 3840mm to 4290mm demonstrate the highest probability of landslides with frequency ratio of 3.094. Rainfall values between 2490mm and 3390mm have the least susceptibility to landslides as indicated by the FR value of 0.

5.2.2 Attribute Relevance Analysis

In landslide modeling, as all the conditioning factors in the initial dataset may not have the equal predictive ability and in some scenarios the presence of noisy parameters may cause reduction of model performances. Therefore it is important to quantify the predictive capabilities of conditioning factors and remove factors having null predictiveness. In order to achieve this the Information Gain Ratio(IGR) was utilized. It is calculated using equation 3.5, 3.6, 3.7 and 3.8 as discussed in Chapter 3. Table 5.2 shows the calculated entropy values, Information Gain, Split Information Gain values and Information Gain Ratios for

the selected 12 parameters.

Conditioning Factor	Info(S)	Info(S,A)	Split.Info(S,A)	Info.Gain(S,A)
Geology	0.1562	0.0272	0.8858	0.1456
Soil Thickness	0.1562	0.0418	0.7294	0.1568
Soil Type	0.1562	0.0964	1.7069	0.0035
Landform	0.1562	1.5762	14.5067	0.1010
Land use	0.1562	0.0941	2.4816	0.0250
Hydrology	0.1562	0.0241	1.8320	0.0721
Rainfall	0.1562	0.0035	2.3333	0.0654
Slope	0.1562	0.0121	2.341	0.0615
Aspect	0.1562	0.0881	1.8765	0.0362
SPI	0.1562	0.0144	1.0916	0.1299
TWI	0.1562	0.0552	1.7336	0.0582
STI	0.1562	0.1103	2.4816	0.0184

Table 5.2: Information Gain Ratio for Conditioning Factors

It can be observed that highest information gain ratio was given by soil thickness(0.1568) followed by geology (0.1456), SPI(0.1299), landform(0.1010), hydrology (0.0721), rainfall(0.0654) and slope(0.0615). The minimum information gain ratios were demonstrated by TWI(0.0582), land use(0.0250), aspect(0.0362), STI(0.0184) and soil type(0.0035).

The results obtained for the information gain ratio indicated that none of the features have zero information gain. It informs that none of the conditioning factors show null predictiveness to the landslide occurrence. Therefore the set of 12 landslide conditioning factors initially considered in the study was utilized in the implementation of the machine learning model using Random Forest.

5.3 Results of Landslide Susceptibility Prediction Model

Based on the results obtained for the correlation analysis and analysis of the predictive capability of conditioning factors in the pilot study, a total of 12 factors were selected and used to implement the Random forest based prediction model.

When fitting the random forest classifier to the data, one of the most important factor that needs to be considered is tuning the hyper-paramters so that the best performance of the model is achieved. In order to find the optimal set of hyper-paramters for the classifier, model performance was tested for different combinations of the paramters. Following code listing 5.1 demonstrates a sample of such combinations tried during the training phase and the model accuracy achieved in each of those instances.

```
from sklearn.ensemble import RandomForestClassifier
#accuracy=64.71%
classifier=RandomForestClassifier(n_estimators=35,criterion='entropy',
    max_depth=10,min_samples_split=2,min_samples_leaf=1,max_features='sqrt'
    ,class_weight='balanced',bootstrap='true',random_state=0,oob_score='
    true')
```

```
#accuracy=66.67%
classifier=RandomForestClassifier(n_estimators=35,criterion='entropy',
    max_depth=20,min_samples_split=4,min_samples_leaf=1,max_features='sqrt'
    ,class_weight='balanced',bootstrap='true',random_state=0,oob_score='
    true')
```

```
#accuracy=68.63%
classifier=RandomForestClassifier(n_estimators=40,criterion='entropy',
    max_depth=10,min_samples_split=4,min_samples_leaf=1,max_features='sqrt'
    ,class_weight='balanced',bootstrap='true',random_state=0,oob_score='
    true')
```

```
#accuracy=70.59%
classifier=RandomForestClassifier(n_estimators=20,criterion='entropy',
    max_depth=20,min_samples_split=2,min_samples_leaf=1,max_features='sqrt'
    ,class_weight='balanced',bootstrap='true',random_state=0,oob_score='
    true')
```

```
#accuracy=72.55%
classifier=RandomForestClassifier(n_estimators=50,criterion='entropy',
    max_depth=20,min_samples_split=4,min_samples_leaf=1,max_features='sqrt'
    ,class_weight='balanced',bootstrap='true',random_state=0,oob_score='
    true')
```

```
#accuracy=74.51%
classifier=RandomForestClassifier(n_estimators=50,criterion='entropy',
    max_depth=10,min_samples_split=4,min_samples_leaf=1,max_features='sqrt'
    ,class_weight='balanced',bootstrap='true',random_state=0,oob_score='
    true')
```

```
#accuracy=76.47%
classifier=RandomForestClassifier(n_estimators=30,criterion='entropy',
    max_depth=20,min_samples_split=2,min_samples_leaf=1,max_features='sqrt'
    ,class_weight='balanced',bootstrap='true',random_state=0,oob_score='
    true')
```

```
#accuracy=80.39%
classifier=RandomForestClassifier(n_estimators=35,criterion='entropy',
    max_depth=20,min_samples_split=2,min_samples_leaf=1,max_features='sqrt'
    ,class_weight='balanced',bootstrap='true',random_state=0,oob_score='
    true')
```

Code Listing 5.1: RF hyper-parameters

Initially the the parameters were set to n_estimators:10, criterion:'gini', max_depth:2 and the observed sensitivity, specificity and accuracy values were 64.0%, 57.69%, 61.53% respectively. Then the depth of the decision trees were increased by setting max_depth:10 and a sensitivity of 64.0%, a specificity of 84.0% and an accuracy of 74.0% were recorded. With criterion:'entropy', depth with max_depth:20 and n_estimators:25, the overall performance of the model could improved further with a sensitivity of 74.5%, a specificity of 80.0% and an accuracy of 69.23%. The performance of the model decreased when the value of n_estimators were decreased or increased to a value less than 35 without changing values of other parameters. When the value of max_depth and min_samples_split was increased it also affected the model performance to drop. It was identified that the best model performance was achieved when n_estimators =35, criterion='entropy', max_depth=20, min_samples_split=2, min_samples_leaf =1, max_features='sqrt', class_weight ='balanced', bootstrap='true', random_state =0, and oob_score='true'. Therefore the following set of optimal hyper parameters given in table 5.3 was used for the random forest classifier.

Hyper-parameter	Final Value
n_estimators	35
criterion	entropy
max_depth	20
min_samples_split	2
min_samples_leaf	1
max_features	sqrt
class_weight	balanced
bootstrap	true
random_state	0
oob_score	true

Table 5.3: Optimal hyper-parameters

The classifier with optimal hyper-parameters was used fit the training data, it was used to predict the landslide susceptibility using the test data. Figure 5.12 demonstrates the confusion matrix generated from predictions obtained through the application of random forest classification on the test data set.

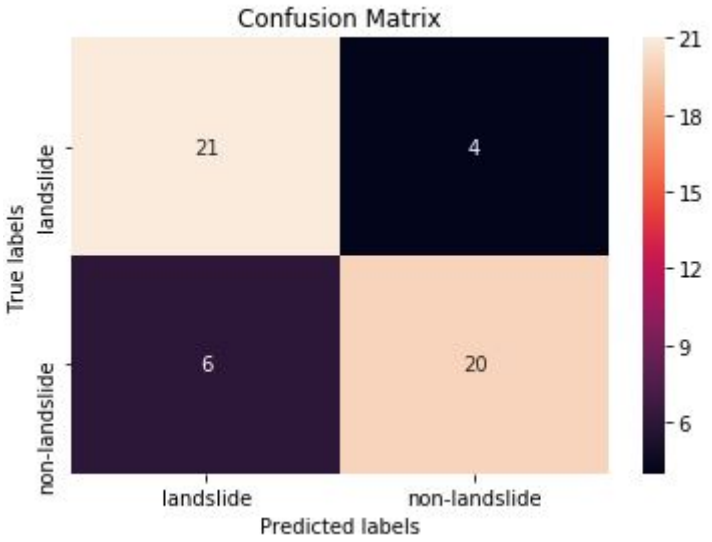


Figure 5.12: Confusion Matrix

Sensitivity, specificity and accuracy values calculated using the confusion matrix for the final model are given in table 5.4.

Performance Measure	Value Obtained
Accuracy	80.39%
Sensitivity	84.00%
Specificity	76.92%
Kappa Index	0.61

Table 5.4: Confusion matrix for prediction model

the confusion matrix included the values generated for the true positives(TP),true negatives(TN),false positives(FP) and false negatives(FN).The values obtained from the confusion matrix used to calculate various statistical indices,such as accuracy, sensitivity, specificity etc. The model showed 80.39% accuracy on testing data, while total of true positive classifications were 21 which indicated the correctly classified pixels as landslide class. Out of 51 test instances 20 of true negative classifications were given, indicating correctly classified non-landslide points to the non-landslide class.This indicates that the model was able to correctly identify the potential landslide pixels out of non-landslide pixels in the study area with a confidence of over 80% accuracy. Meanwhile, a total of 10 instances were missclassified including, of 4 false positives and 6 false negatives out of 51. An important fact about these indicators are that, depending on the application domain and the risk carried out in each prediction,sometimes the cost of false positives can be higher than the cost of false negatives. this is because,the cost of emergency preparation activities that need to be executed is relatively high in such disaster mitigation strategies.So maintaining a lower false negative(FN) rate is also as important as maintaining a high TP and TN values in the model.

As indicated by the Kappa Index value of 0.61, the the proposed approach shows a nearly substantial reliability. However, landslide points classified as false negatives and false positives are relatively low. There is a possibility of field data collection errors in the study area as the reliability of the model not being excellent (Kappa $>$ 0.81).

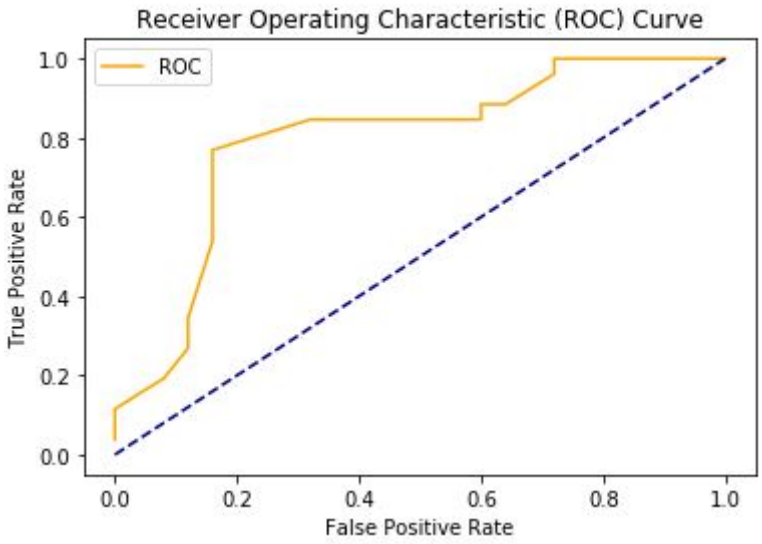


Figure 5.13: Receiver Operating Characteristic

The prediction capability of the model was evaluated using the area under the ROC curve, plotting the true positive rate(Sensitivity) against the false positive rate(1-Specificity) values referring to the confusion matrix as shown in the figure 5.13. The area under the curve obtained a value of 0.7946 (79.46%) which indicated a substantial agreement between the observed landslide points and the predicted landslide points. The higher AUC value was given along with the highest probability given by the model for the correctly classified landslide points to the classified none-landslide points. So the AUC value was closer to 1 (or 100%) in the final model, which proves the high predictive capability of the model.

The results obtained from RF based landslide prediction model, were used to implement a "Landslide Susceptibility Map"(LSM) for the study area, Kalutara district.The landslide susceptibility indices(LSIs) were generated and reclassified the mapping area according to different susceptibility classes.The LSIs were calculated based on the values obtained from the RF model and the indices were classified into five classes using natural breaks method. The identified classes are: very high, high, moderate, low and very low as shown in the figure 5.14.

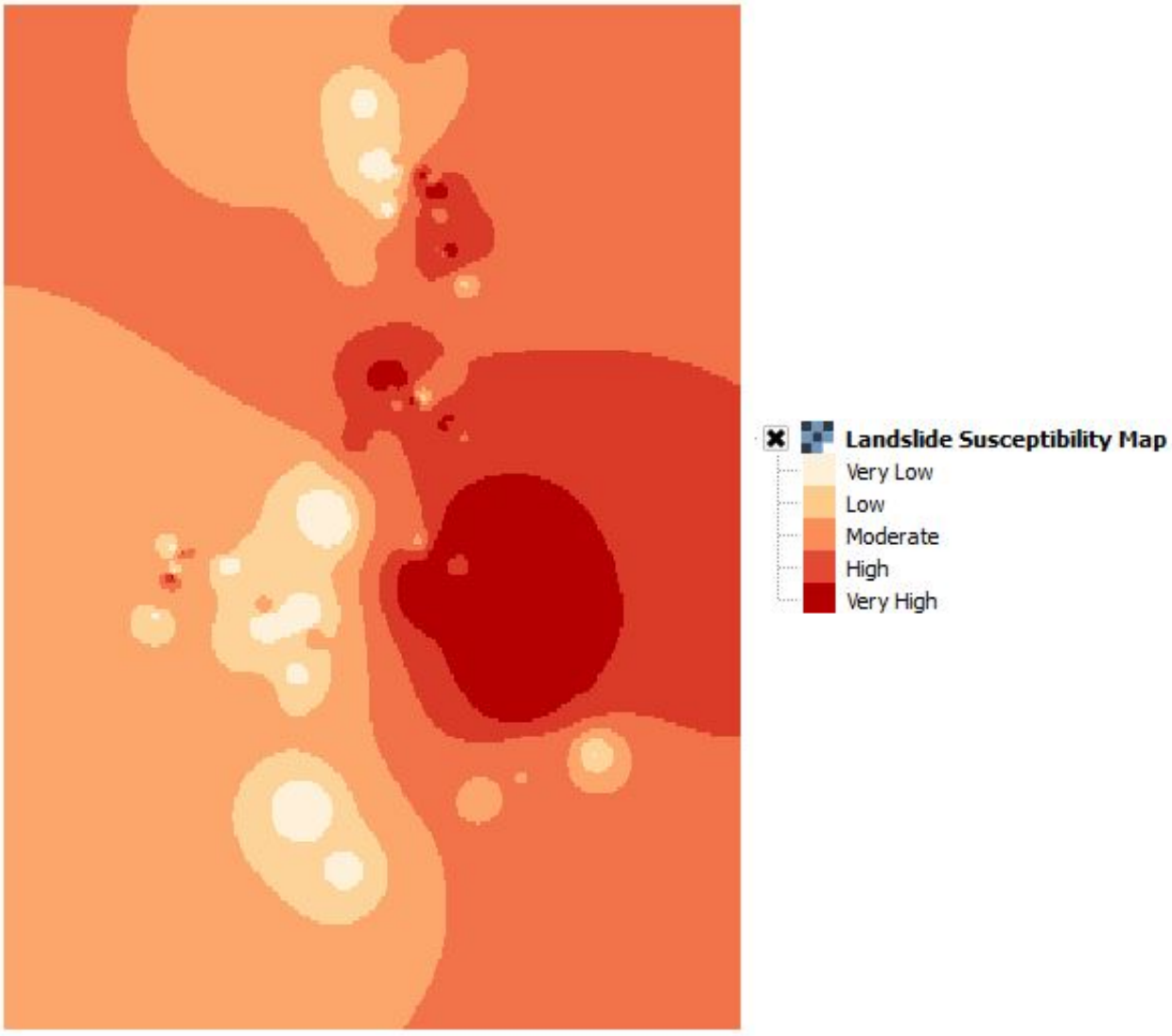


Figure 5.14: Landslide susceptibility map -Kalutara District

5.4 Summary

This chapter provides a comprehensive analysis of the results obtained in the study. It evaluates the maps generated in the implementation process, results obtained in the pilot study and finally assesses the performance of the landslide susceptibility prediction model implemented using Random Forest. The results indicate that the model provides an accuracy of 80.39% by employing the 12 conditioning factors in the prediction. Since the model has a considerable accuracy it can be concluded that the applicability of slope, aspect, SPI, TWI, STI, hydrology, geology, land form, land use, soil type, soil thickness and rainfall in the prediction of landslide susceptibility in Kalutara district is successful.

Chapter 6

Conclusion

In this study, the primary focus was to investigate a suitable approach to predict landslides occurrences in Sri Lanka making use of both terrain and triggering factors of landslide causation. In the process of tackling the research question, *'How to predict landslide susceptibility using a machine learning employing the data extracted from contour maps, geospatial statistical data, and precipitation data?'*, a sound understanding of how the existing landslide prediction system works at NBRO, limitations and possible improvements were identified first. Due to the increase in the reported landslide incidents in Kalutara district over the last three years, it was selected as the case study. The related contour maps data and geospatial statistical data were obtained from NBRO and average rainfall data was obtained from the Department of Meteorology, Sri Lanka. In order to capture, store, manipulate, interpret and analyze the geospatial data and the relationships between the data, capabilities of a geographical information system was employed. In this study an open source GIS tool, QGIS was used to achieve this.

According to past literature, it was identified that the role of conditioning factors over landslide occurrence is crucial and not the same set factors affect the occurrence of landslides throughout the world but site-specific. When tackling with the sub research questions *'How to determine the spatial relationship between landslide conditioning factors and landslide occurrence'* and *'How to eliminate landslide conditioning factors having low or null predictive capability in the given study area'*, with the use of Frequency Ratio and Information Gain Ratio, it was deduced that only certain classes of the initial set of landslide conditioning factors were having an impact over landslide occurrence while none of the factors demonstrated null predictiveness. These classes were considered during data pre-processing operations.

The need to explore on *'What are the machine learning algorithms that can be used*

to predict landslide susceptibility with high accuracy?' was fulfilled by reviewing similar research work done in the landslide prediction domain and Random Forest was identified to be demonstrating promising performance as a machine learning algorithm when several factor classes have an impact on the final prediction.

The model showed 80.39% accuracy on test data. Out of 25 instances of true landslide occurrences in the test data, 21 were correctly predicted as landslides and out of the 26 non-landslide instances, 20 were correctly predicted as non-landslide occurrences. Depending on the application domain and the risk carried out in each prediction, the cost of false positives may be higher than the cost of false negatives.

Furthermore, to answer *'How to evaluate the proposed approach and assess the accuracy of the proposed model'*, area under the ROC curve(AUC), specificity and sensitivity measures were calculated. AUC value was 0.7946 (79.46%) indicating a substantial agreement between the observed landslide instances and the predicted landslide instances. Since the value of AUC is closer to 1 (or 100%) in the final model, it proved a high predictive capability of the model.

Understanding and prediction of landslide susceptibility with high accuracy will help for better decision making and thus plan risk mitigation actions in the future. It will also assist in the reduction of the destruction of human lives and property.

6.1 Future Work

As future work, a natural extension of this study entails working to improve the accuracy on the current model using other parameter fine tuning methods. Continuous investigating on novel suitable machine learning algorithms to give out better prediction accuracy with high true positive rate is also a viable research scope left to uncover.

In this study, only 12 factors were considered for predicting landslide susceptibility. Investigating the possibility of using more than 12 landslide conditioning factors, to build the model to predict the landslide susceptibility on the given study area is important as well.

Another possible avenue would be application of the existing model to detect new landslides distributed in other landslide-prone districts in Sri Lanka and include them as part of the training samples. A higher number of landslides will increase the robustness and generalization of the model.

References

[1] L. Highland and P. Bobrowsky, The Landslide Handbook - A Guide to Understanding Landslides. Reston, Virginia: U.S. Geological Survey, Circular 1325, 2008, pp. 1-74.

[2] M. Jebur, B. Pradhan and M. Tehrany, "Manifestation of LiDAR-Derived Parameters in the Spatial Prediction of Landslides Using Novel Ensemble Evidential Belief Functions and Support Vector Machine Models in GIS", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 2, pp. 674-690, 2015.

[3] K. Christopher, E. Arusei and M. Kupti, "The causes and socio-economy impacts of landslidein Kerio Valley, Kenya", Agricultural Science and Soil Sciences 4, 58-66, 2016.

[4] N.M.T De Silva, “Landslide flow path modelling;A Case Study on Aranayaka Landslide”, MCS Dissertation, University of Colombo School of Computing, Colombo, 2018.

[5] E. Perera, D. Jayawardana, P. Jayasinghe, R. Bandara and N. Alahakoon, "Direct impacts of landslides on socio-economic systems: a case study from Aranayake, Sri Lanka", Geoenvironmental Disasters, vol. 5, no. 1, 2018.

[6] H. Pourghasemi and O. Rahmati, "Prediction of the landslide susceptibility: Which algorithm, which precision?", CATENA, vol. 162, pp. 177-192, 2018.

[7] G. Metternicht, L. Hurni and R. Gogu, "Remote sensing of landslides: An analysis of the potential contribution to geo-spatial systems for hazard assessment in mountainous environments", Remote Sensing of Environment, vol.98, no. 2-3, pp. 284-303, 2005.

[8] L. Donati and M. Turrini, "An objective method to rank the importance of the factors predisposing to landslides with the GIS methodology: application to an area of the Apennines (Valnerina; Perugia, Italy)", Engineering Geology, vol. 63, no. 3-4, pp. 277-289, 2002.

[9] D. Tien Bui, T. Tuan, H. Klempe, B. Pradhan and I. Revhaug, "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree", Landslides, vol. 13, no. 2, pp. 361-378, 2015.

[10] P. Vorpahl, H. Elsenbeer, M. Marker and B. Schroder, "How can statistical models help to determine driving factors of landslides?", Ecol. Model., vol. 239, pp. 27–39, 2012.

[11] I. Yilmaz, "A case study from koyulhisar (sivas-turkey) for landslide susceptibility mapping by artificial neural networks," Bull. Eng. Geol. Environ., vol. 68, no. 3, p. 297–306, 2009.

[12] M. Jebur, B. Pradhan and M. Tehrany, "Optimization of landslide conditioning factors using very high-resolution airborne laser scanning (LiDAR) data at catchment scale", Remote Sensing of Environment, vol. 152, pp. 150–165,2014.

[13] H. Hemasinghe, R. Rangali, N. Deshapriya and L. Samarakoon, "Landslide susceptibility mapping using logistic regression model (a case study in Badulla District, Sri Lanka)", Procedia Engineering, vol. 212, pp. 1046-1053, 2018.

[14] Karunanayake, K.B.A.A.M. and Wijayanayake, W.M.J.I., “Predicting landslides in hill country of Sri Lanka using data mining techniques”, in International Research Symposium on Pure and Applied Sciences, Faculty of Science, University of Kelaniya, Sri Lanka, 2016, p 76.

[15] C. Madawala, B. Kumara and L. Indrathilaka, "Novel machine learning ensemble approach for landslide prediction", in International Research Conference On Smart Computing And Systems Engineering, Faculty of Science, University of Kelaniya, Sri Lanka, 2019, p. 78.

[16] W. Chen, J. Peng, H. Hong, H. Shahabi, B. Pradhan, J. Liu, A. Zhu, X. Pei and Z. Duan., "Landslide susceptibility modeling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China", Science of The Total Environment, vol. 626, pp. 1121-1135, 2018.

[17] Q. He et al., "Landslide spatial modelling using novel bivariate statistical based Naïve Bayes, RBF Classifier, and RBF Network machine learning algorithms", Science of The Total Environment, vol. 663, pp. 1-15, 2019.

[18] J. Goetz, A. Brenning, H. Petschko and P. Leopold, "Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling", Computers & Geosciences, vol. 81, pp. 1-11, 2015.

[19] R. Bandara, "Landslides in Sri Lanka", Vidurava, no. 22, pp. 9-13, 2005.

[20] "National Building Research Organisation, Sri Lanka", Nbro.gov.lk. [Online]. Available: https://www.nbro.gov.lk/.

[21] G. Jayatissa, "Landslides, Landslide Disaster Risk Reduction and Slope Stability", 2019.

[22] B. Pradhan, S. Lee and M. Buchroithner, "A GIS-based back-propagation neural network model and its cross-application and validation for landslide susceptibility analyses", Computers, Environment and Urban Systems, vol. 34, no. 3, pp. 216-235, 2010.

[23] B. Pradhan, E. Sezer, C. Gokceoglu and M. Buchroithner, "Landslide Susceptibility Mapping by Neuro-Fuzzy Approach in a Landslide-Prone Area (Cameron Highlands, Malaysia)", IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 12, pp. 4164-4177, 2010.

[24] Y. Hsu, Y. Chang, C. Chang, J. Yang and Y. Tung, "Physical-based rainfall-triggered shallow landslide forecasting", Smart Water, vol. 3, no. 1, 2018.

[25] C. Zhou et al., "Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China", Computers & Geosciences, vol. 112, pp. 23-37, 2018.

[26] F. Yan, Q. Zhang, S. Ye and B. Ren, "A novel hybrid approach for landslide susceptibility mapping integrating analytical hierarchy process and normalized frequency ratio methods with the cloud model", 2019.

[27] M. Mohammady, H. Pourghasemi and B. Pradhan, "Landslide susceptibility mapping at Golestan Province, Iran: A comparison between Frequency Ratio, Dempster Shafer, and Weights of Evidence models", Journal of Asian Earth Sciences, vol. 61, pp. 221-236, 2012.

[28] A. Shirzadi, K. Chapi, H. Shahabi, K. Solaimani, A. Kavian and B. Ahmad, "Rockfall susceptibility assessment along a mountainous road: an evaluation of bivariate statistic, analytical hierarchy process and frequency ratio", Environmental Earth Sciences, vol. 76, no. 4, 2017

[29] "The physical impact of the disaster. In Natural Disaster management", 1999.

[30] W. Chen, X. Xie, J. Wang, B. Pradhan, H. Hongd, D.T. Bui, Z.Duan,J. Ma, "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility", CATENA, vol. 151, pp. 147-160, 2017.

[31] P. Offermann, O. Levina, M. Schönherr and U. Bub, "Outline of a design science research process", Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09, 2009.

[32] K. Weerasinghe, Landslide Hazard Zonation Mapping using GIS. NBRO, 1997, pp. 1-8.

[33] R. Soeters and C. van Westen, "Slope Instability Recognition Analysis and Zonation", Onlinepubs.trb.org, 2019. [Online].Available:http://onlinepubs.trb.org/Onlinepubs/sr/sr247/sr247-008.pdf.

[34] F. Guzzetti, A. Carrara, M. Cardinali and P. Reichenbach, "Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy", Geomorphology, vol. 31, no. 1-4, pp. 181-216, 1999.

[35] W. Chen,H.R. Pourghasemi, M. Panahi, A. Kornejady, J. Wang, X. Xie, S. Cao, "Spatial prediction of landslide susceptibility using an adaptive neuro-fuzzy inference system combined with frequency ratio, generalized additive model, and support vector machine techniques", Geomorphology, vol. 297, pp. 69-85, 2017.

[36] B. Pradhan, "A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS", Computers & Geosciences, vol. 51, pp. 350-365, 2013.

[37] M. Mezaal, B. Pradhan, M. Sameen, H. Mohd Shafri and Z. Yusoff, "Optimized Neural Architecture for Automatic Landslide Detection from High-Resolution Airborne Laser Scanning Data", Applied Sciences, vol. 7, no. 7, p. 730, 2017.

[38] B. Pradhan and S. Lee, "Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia", Landslides, vol. 7, no. 1, pp. 13-30, 2009.

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

[40] C. Lian, C. Chen, Z. Zeng, W. Yao and H. Tang, "Prediction Intervals for Landslide Displacement Based on Switched Neural Networks", IEEE Transactions on Reliability, vol. 65, no. 3, pp. 1483-1495, 2016.

[41] C. Lian, Z. Zeng, W. Yao and H. Tang, "Displacement prediction of landslide based on PSOGSA-ELM with mixed kernel", 2013 Sixth International Conference on Advanced Computational Intelligence (ICACI), 2013.

[42] H. Hong, J. Liu, D. Bui, B. Pradhan, T. Acharya, B. Pham, A. Zhu, W. Chen and B. Ahmad, "Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)", CATENA, vol. 163, pp. 399-413, 2018.

[43] H. Hong, B. Pradhan, C. Xu and D. Tien Bui, "Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines", CATENA, vol. 133, pp. 266-281, 2015. Available: 10.1016/j.catena.2015.05.019

[44] B.T. Pham, I. Prakash, S.K. Singh, A. Shirzadi, H. Shahabi, T. Tran, D.T. Bui, "Landslide susceptibility modeling using Reduced Error Pruning Trees and different ensemble techniques: Hybrid machine learning approaches", CATENA, vol. 175, pp. 203-218, 2019.

[45] S. Xu and R. Niu, "Displacement prediction of Baijiabao landslide based on empirical mode decomposition and long short-term memory neural network in Three Gorges area, China", Computers & Geosciences, vol. 111, pp. 87-96, 2018.

[46] L. Xiao, Y. Zhang and G. Peng, "Landslide Susceptibility Assessment Using Integrated Deep Learning Algorithm along the China-Nepal Highway", Sensors, vol. 18, no. 12, p. 4436, 2018.

[47] B. Yang, K. Yin, S. Lacasse and Z. Liu, "Time series analysis and long short-term memory neural network to predict landslide displacement", Landslides, vol. 16, no. 4, pp. 677-694, 2019.

[48] A. El Jazouli, A. Barakat and R. Khellouk, "GIS-multicriteria evaluation using AHP for landslide susceptibility mapping in Oum Er Rbia high basin (Morocco)", Geoenvironmental Disasters, vol. 6, no. 1, 2019.

[49] H. Yoshimatsu and S. Abe, "A review of landslide hazards in Japan and assessment of their susceptibility using an analytical hierarchic process (AHP) method", Landslides, vol. 3, no. 2, pp. 149-158, 2006

[50] G. Morgan, J. Gliner and R. Harmon, "Quantitative Research Approaches", Journal of the American Academy of Child & Adolescent Psychiatry, vol. 38, no. 12, pp. 1595-1597, 1999.

[51] M. Mohammady, H. Pourghasemi and B. Pradhan, "Landslide susceptibility mapping at Golestan Province, Iran: A comparison between Frequency Ratio, Dempster Shafer, and Weights of Evidence models", Journal of Asian Earth Sciences, vol. 61, pp. 221-236, 2012.

[52] A. Shirzadi, K. Chapi, H. Shahabi, K. Solaimani, A. Kavian and B. Ahmad, "Rockfall susceptibility assessment along a mountainous road: an evaluation of bivariate statistic, analytical hierarchy process and frequency ratio", Environmental Earth Sciences, vol. 76, no. 4, 2017.

[53] C. Zhou, K. Yin, Y. Cao, B. Ahmed, Y. Li, F. Catani, H. Reza and Pourghasemie, "Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China", Computers & Geosciences, vol. 112, pp. 23-37, 2018.

[54] M. Herrera, "Landslide Detection using Random Forest Classifier", M.Sc. in Geomatics for the Built Environment, Delft University of Technology, 2019.

[55] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and regression trees. Monterey, CA : Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[56] A.J. Viera and J.M. Garrett, "Understanding interobserver agreement: the kappa statistic", Fam med, 37(5), pp.360-363, 2005.

[57] K. Pawłuszek, S. Marczak, A. Borkowski and P. Tarolli, "Multi-Aspect Analysis of Object-Oriented Landslide Detection Based on an Extended Set of LiDAR-Derived Terrain Features", 2020.

[58] J. Landis and G. Koch, "The Measurement of Observer Agreement for Categorical Data", Biometrics, vol. 33, no. 1, p. 159, 1977. Available: 10.2307/2529310.

[59] H. Pourghasemi, H. Moradi and S. Fatemi Aghda, "Landslide susceptibility mapping by binary logistic regression, analytical hierarchy process, and statistical index models and assessment of their performances", Natural Hazards, vol. 69, no. 1, pp. 749-779, 2013. Available: 10.1007/s11069-013-0728-5.

[60] J. Quinlan, "Induction of decision trees", Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.

[61] Tin Kam Ho, "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, Quebec, Canada, 1995, pp. 278-282 vol.1.doi: 10.1109/ICDAR.1995.598994

[62] A. Graser, "QGIS: Introducing the Quantum GIS Ecosystem", GIS Lounge, 2012. [Online]. Available: https://www.gislounge.com/introducing-the-quantum-gis-ecosystem/.

[63] "Sri Lanka - The land", Encyclopedia Britannica, 2020. [Online]. Available: https://www.britannica.com/place/Sri-Lanka/The-land. [Accessed: 20- Jan- 2020].

[64] P. Nagamuthu, "TRENDS OF RAINFALL IN DRY ZONE AND WET ZONE OF SRI LANKA-BASED ON KILINOCHCHI AND KALUTARA DISTRICT", in International Symposium of South Eastern University of Sri Lanka, South Eastern University of Sri Lanka, 2017.

Appendices

Appendix A

Prediction Model

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import preprocessing
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_curve

features_list=['LandUse','LandForm','SoilTypeAndThickness','Geology','
    Slope','Aspect','SPI','TWI','STI','Rainfall','Hydrology']

dataset = pd.read_csv(r'C:\Users\Dell\Desktop\FeaturesForAllPoints.csv',
    usecols=fields)

def encodeDataset(data):
    encode_data = preprocessing.LabelEncoder()
    dataset['LandUse'] = encode_data.fit_transform(data.LandUse.astype(str
    ))
    dataset['LandForm'] = encode_data.fit_transform(data.LandForm.astype(
    str))
    dataset['SoilTypeAndThickness'] = encode_data.fit_transform(data.
    SoilTypeAndThickness.astype(str))
    dataset['Geology'] = encode_data.fit_transform(data.Geology.astype(str
    ))
    dataset['Slope'] = encode_data.fit_transform(data.Slope.astype(str))
```

```
    dataset['Aspect'] = encode_data.fit_transform(data.Aspect.astype(str))
    dataset['SPI'] = encode_data.fit_transform(data.SPI.astype(str))
    dataset['TWI'] = encode_data.fit_transform(data.TWI.astype(str))
    dataset['STI'] = encode_data.fit_transform(data.STI.astype(str))
    dataset['Rainfall'] = encode_data.fit_transform(data.Rainfall.astype(
   str))
    dataset['Hydrology'] = encode_data.fit_transform(data.Hydrology.astype
   (str))
    data=data.fillna(-999)
    return data

dataset=encodeDataset(dataset)

X = dataset.iloc[:, 0:11].values

y=dataset.iloc[:,11].values

X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,
   random_state=1,shuffle='true')

classifier=RandomForestClassifier(n_estimators=35,criterion='entropy',
   max_depth=30,min_samples_split=2,min_samples_leaf=1,max_features='sqrt'
   ,class_weight='balanced',bootstrap='true',random_state=0,oob_score='
   true')
classifier.fit(X_train,y_train)
y_pred=classifier.predict(X_test)

cm=confusion_matrix(y_test,y_pred)
print(cm)

Accuracy=float(cm[0,0]+cm[1,1])/float(cm[1,0]+cm[1,1]+cm[0,0]+cm[0,1])
Sensitivity = float(cm[0,0])/float(cm[0,0]+cm[0,1])
Specificity=float(cm[1,1])/float(cm[1,0]+cm[1,1])

def plot_roc_cur(fper, tper):
    plt.plot(fper, tper, color='orange', label='ROC')
    plt.plot([0, 1], [0, 1], color='darkblue', linestyle='--')
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title('Receiver Operating Characteristic (ROC) Curve')
```

```
    plt.legend()
    plt.show()

probs = classifier.predict_proba(X_test)
probs = probs[:, 1]
fper, tper, thresholds = roc_curve(y_test, probs)
plot_roc_cur(fper, tper)

print('Kappa Index: ', metrics.cohen_kappa_score(y_test,y_pred,weights='
    quadratic'))
```