T22114

IMPROVEMENT OF LANDSLIDE PREDICTION SYSTEM BASED ON HYBRID NEURAL NETWORKS (PENANG ISLAND, MALAYSIA)

By

MUTASEM SHABIB HUSSIEN ALKHASAWNEH

Thesis submitted in fulfillment of the
requirements for the degree of
Doctor of Philosophy

January 2015

بسم الله الرَّحْمَنِ الرَّحِيمِ

﴿ يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنْكُمْ وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ وَاللَّهُ بِمَا تَعْمَلُونَ خَبِيرٌ ﴾

سورة المجادلة (11)

In the Name Of Allah, the Most Beneficent, the Most Merciful

﴿ **Allah will raise those who have believed among you and those who were given knowledge, by degrees. And Allah is acquainted with what you do** ﴾

Surah Al-Mujaadila (11)

DEDICATION

To the spirits of my mother; Muna Alkhasawneh who used to call me "Dr. Mutasem" when I was a kid ... but unfortunately, faith did not grant her the opportunity to live and celebrate these moments together with her ... may Allah SWT grant her Jannatul Firdaus by His mercy...

To my father Shabib Alkhasawneh I say, "Please wait for me to celebrate this achievement together ".

To my second mother Arwa Alkhasawneh for her prayers and tremendous sacrifices for my dad...

To my favorite father and mother in law Mustafa and Shama for their faithful, endless, and continuous prayer to Allah SWT asking success of my wife and I...

To all my brothers, sisters and beloved family ... to my childhood friend to the half of my soul Mahmoud Smadi (Abu Hashim).

To my other half ... to my lovely wife, Doaa, who offered me unconditional love and support throughout the course of this thesis, I wish you will pass your thesis too. To all my children; Arwa and my new baby Bushra who had to endure being confined in the house most of the time.

Mutasem Shabib Hussien Alkhasawneh

ACKNOWLEDGEMENTS

I wish to express my profound gratitude to god Almighty for His protection from the beginning of my study till now.

It is with immense gratitude that I acknowledge the help of my supervisor Assoc. Prof. Dr. Umi Kalthum Ngah and my first co-supervisor Dr. Tay Lea Tien who provided me inspiration, enthusiasm, and motivation, as well as for their support, patience, and encouragement throughout my graduate studies. It is not often that one finds an advisor that always finds the time for listening to the little problems and roadblocks that unavoidably crop up in the course of performing research. Their technical and editorial advice was essential to the completion of this thesis and has taught me innumerable lessons and insights on the workings of academic research in general. It was an honour to work with them. I also want to thank my second co-supervisor Associate Professor Dr. Nor Ashidi Mat Isa for his fruitful suggestions in my research.

It gives me great pleasure in acknowledging the support and help of my family father, mother in law, sisters and brothers, also my special gratitude to my sibling Kurdeah, Aber, UmAmmar, Fuad and Omar for looking after me in this long journey.

It gives me great pleasure in acknowledging the support and help of faithful friends; Dr. Mouhamad Subhi Albatah, Dr. Mahmoud Alya and Dr.Yazed Qaseamah. I would like to express the deepest appreciation to my first and most intimate friend Mahmoud Smadi, and all friends for providing me inspiration and motivation.

I would like to thank all staff at the School of Electrical and Electronics Engineering, Universiti Sains Malaysia for providing the necessary facilities, as well as the GA support.

Last, but not least, I would like to thank my wife Dooa Al-Shara for her personal support and great patience at all times. Her support and encouragement what made this thesis possible.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AUC	Area Under Curve
CF	Certainty Factor
CFNN	Cascade Forward Neural Network
CGB	Conjugate Gradient with Powell-Beale restarts
CGF	Conjugate gradient with Fletcher Reeves updates
CHAID	Chi-squared Automatic Interaction Detector
CI	Computational Intelligence
CGP	Conjugate Gradient with Polak–Ribiere updates
DA	Discriminant Analysis
DEM	Digital Elevation Model
DT	Decision Tree
ENN	Elman Neural Network
EXCHAID	Exhaustive Chi-squared Automatic Interaction Detector
FFNN	Feed Forward Neural Network
FR	Frequency Ratio
GA	Genetic Algorithm
GD	Gradient Descent
GDA	Gradient Descent with momentum and Adaptive learning rate
GIS	Geographic Information Systems
GDM	Gradient Descent with Momentum
GDX	Gradient Descent with momentum and adaptive learning rate
HMLP	Hybrid Multilayered Perceptron
HECFNN	Hybrid Elman Cascade Forward Neural Network
LM	Levenberg Marquardt
LR	Logistic Regression
MAE	Mean Absolute Error
MENN	Modified Elman Neural Network

MLP	Multilayered Perceptron
MSE	Mean Square Error
NB	Naïve Bayes
NDVI	Normalized Difference Vegetation Index
OSS	One Step Secant
QDA	Quadratic Discriminant Analysis
QUEST	Quick-Unbiased-Efficient Statistical Tree
R	Regression
RE	Relative Error
ROC	Receiver Operating Characteristics
Rp	Resilient back Propagation
SCG	Scaled Conjugate Gradient
SVM	Support Vector Machine
VLE	Vapor Liquid Equilibria

PENAMBAHBAIKAN SISTEM RAMALAN TANAH RUNTUH BERDASARKAN RANGKAIAN NEURAL GABUNGAN (PULAU PINANG, MALAYSIA)

ABSTRAK

Tanah runtuh adalah salah satu bencana alam paling agresif yang menyebabkan kehilangan nyawa dan kerosakan bernilai berbilion-bilion dolar di seluruh dunia setiap tahun. Kejadian tanah runtuh ini mengancam keselamatan dan nyawa manusia, alam sekitar, sumber asli dan hartanah. Ia adalah salah satu bencana alam yang berlaku agak kerap semasa musim hujan lebat di Malaysia amnya dan khususnya di Pulau Pinang. Pelbagai kajian tentang tanah runtuh telah dijalankan di Pulau Pinang. Walau bagaimanapun, banyak isu-isu serius yang berkaitan dengan tanah runtuh belum diselesaikan lagi. Isu-isu ini termasuk pengekstrakan faktor baru yang menyebabkan tanah runtuh, penyiasatan faktor optimum yang menyebabkan tanah runtuh dan menghasilkan peta bahaya tanah runtuh yang lebih tepat untuk Pulau Pinang. Di samping itu sehingga kini masih tiada sistem pintar yang jelas ramalan bahaya tanah runtuh sama ada di Pulau Pinang mahupun di seluruh dunia. Oleh sebab itu, satu sistem pintar pemetaan bahaya tanah runtuh telah dicadangkan. Ia terdiri daripada tiga peringkat; pengekstrakan faktor, pemilihan faktor optimum dan Rangkaian Neural Buatan (ANN) sebagai alat analisis. Dua puluh satu faktor telah digunakan dalam kajian ini yang mana sembilan faktor dikumpulkan daripada pelbagai agensi-agensi kerajaan. Faktor selebihnya (dua belas) diekstrak daripada Model Digital Elevation (DEM), yang mana tujuh daripada faktor-faktor ini digunakan buat pertama kali di Pulau Pinang. Dalam fasa pemilihan faktor, enam teknik pemilihan faktor digunakan untuk memilih faktor yang paling penting dalam ramalan kejadian tanah runtuh itu. Teknik-teknik ini adalah kaedah Zhou, Klasifikasi

Ketepatan Multilayer Perceptron (MLP), Analisis mendiskriminasi (DA) dan tiga jenis Pohon Keputusan: Pengesan Interaktif Automatik Kuasa Dua (CHAID), CHAID Menyeluruh (EXCHAID) dan Quick-Unbiased-Efficient Statistical Tree (QUEST). Cascade Forward Neural Network (CFNN), Rangkaian Neural Elman (ENN) dan rangkaian baharu yang dicadangkan iaitu Hibrid ENN CFNN (HECFNN) adalah alat analisis. HECFNN adalah struktur rangkaian neural baharu yang dicadangkan dalam kajian ini. Ia adalah gabungan di antara CFNN dan ENN. Struktur yang dicadangkan dapat menyelesaikan prestasi rendah rangkaian neural Elman dengan data statik dan HECFNN juga sesuai untuk masalah linear dan tak linear. Selain itu, ia mempunyai prestasi ketepatan dalam pengelasan dan ramalan yang lebih tinggi berbanding dengan CFNN dan ENN. Prestasi HECFNN ditunjukkan dengan menggunakan enam set data tanda aras. Keputusan yang diperolehi dengan menggunakan pelbagai teknik dalam sistem bahaya tanah runtuh menunjukkan bahawa EXCHAID adalah teknik pemilihan faktor yang lebih tepat manakala 14 faktor yang dipilih dianggap sebagai faktor yang paling penting. HECFNN adalah alat analisa yang paling tepat dibandingkan dengan ENN dan CFNN. Sistem pintar tanah runtuh ini diuji sebagai pengelas dan peramal. Pada peringkat klasifikasi, 13786 sampel telah digunakan dan sampel-sampel itu mengklasifikasikan hasil kepada kategori "tanah runtuh" dan "tiada tanah runtuh" untuk kawasan yang terdedah dengan purata ketepatan sebanyak 95.66%. Sistem ini juga meramalkan bahaya tanah runtuh Pulau Pinang dengan ketepatan sebanyak 94.16%.

IMPROVEMENT OF LANDSLIDE PREDICTION SYSTEM BASED ON HYBRID NEURAL NETWORKS (PENANG ISLAND, MALAYSIA)

ABSTRACT

Landslides are one of the most aggressive natural disasters that cause loss of lives and of billions dollars in damages annually worldwide. They pose a threat to the safety of human lives, the environment, resources and property. It is one of the natural disasters that occur quite often in Malaysia and particularly in Penang Island during heavy rainy seasons. Numerous researches on landslides studies have been done based on Penang Island. However, many issues seriously related to landslides have not been solved yet. These issues include the extraction of new factors which cause landslides, investigation on the optimum factors which cause landslides and the generation of an accurate landslide hazard map for Penang Island. In addition to that, the landslide hazard prediction intelligent system, either for Penang Island or for the entire world is still being investigated up to this date. For that reason, an intelligent landslide hazard mapping system is proposed. It consists of three stages: factor extraction, factor selection and Artificial Neural Network (ANNs) as an analysis tool. Twenty one factors are used in this study where nine factors were collected from different governmental agents. The rest of the factors (twelve) were extracted from the Digital Elevation Models (DEM), seven of these factors were extracted and used for the first time on Penang Island. In the factor selection phase, six factor selection techniques are employed to select the most important factors in the landslide prediction. These techniques are Zhou's method, Multilayer Perceptron (MLP) Classification Accuracy, Discriminant Analysis (DA) and three types of the Decision Tree: Chi-square Automatic Interaction Detector (CHAID), Exhaustive Chi-square Automatic Interaction Detector (EXCHAID) and Quick-Unbiased-

Efficient Statistical Tree (QUEST). Cascade Forward Neural Network (CFNN), Elman Neural Network (ENN) and new proposed Hybrid ENN and CFNN (HECFNN) were the analyses tools where HECFNN is a new neural network structure proposed in this research. It is a hybrid between the CFNN and ENN. The proposed structure solves the low performance of Elman neural network with static data and HECFNN is also suitable for linear and nonlinear problem. In addition to that, it has higher performance accuracy in the classification and prediction as compared to the CFNN and ENN. The performance of the HECFNN was demonstrated using six benchmarked data sets. The result obtained by applying various techniques in the landslide hazard prediction system shows that EXCHAID is the more precise factor selection technique while 14 factors chosen are considered as the most important factors. The HECFNN is the most precise analysis tool compared with the Elman and CFNN Network. The landslide intelligent system was tested as a classifier and predictor. In the classification stage, 13786 samples from the whole study area were used and the sample classifies "landslide" and "no landslide" prone areas with an average accuracy of 95.66%. The system also predicts the landslide hazard map of Penang Island with an accuracy of 94.16%.

CHAPTER 1
INTRODUCTION

1.1 Introduction

In the last few decades, huge advancements have been observed in the application of the Computational Intelligent System (CIS), particularly in the areas of classifications and predictions. CIS have been used in different applications such as in agriculture, material, environmental science and water resources. Classification tends to group samples or objects into groups or categories according to their characteristics. For example, CIS is used in medical sciences such as for diagnoses of patients for disease detection while in finance such as classifying the credit history of possible clients into different levels of bad, good or excellent. On the other hand, prediction tends to speculate the expected outcome in the future, based on available knowledge or experience. Prediction of the weather for the few coming days (sunny, windy, rainy, etc.) and prediction of the economic growth (slow, moderate, good, etc.) are some examples of using CIS in predictions. In most cases, the differences between the classification and prediction are very slight. It may be based on the user requirements and applications.

CIS refers to the ability to understand, to comprehend and to learn from experience (Byrd and Hauser, 1991). Classifications or prediction procedure using CIS follows three major stages; pre-processing, processing and analysis (Al-batah, 2009). At each stage, one or more operations are performed to enhance the process and improve the performance of the CIS output.

The pre-processing stage includes detecting the object of interest in the scene. In addition, pre-processing also removes unnecessary and noisy data, which affects the object of interest. Extracting the factors (feature or attributes) of the object is a crucial step in CIS (Sabeh, 2012). Factor extraction includes the traditional and new proposed factors. Traditional factors are the ones used in the previous studies such as slope angle and slope aspect, while the new factors are the factors used for the first time in the landslide hazard mapping such as tangential and cross section curvature. Unwanted factors can complicate and slow down the process in the system, thus processing is the second stage of the CIS is very necessary.

In the processing stage, factor selection is the major step where the relevant factors will be chosen for the CIS construction. Using the selected factors in the prediction and classification process can result in many advantages such as reducing the data dimension since the irrelevant factors will be ignored. The irrelevant factors usually reduce the accuracy and increase the classification or prediction cost (Peng, H.et al 2005). By eliminating the irrelevant factors, the CIS accuracy will be improved and the processing cost will be decreased. Another advantage of factor selection is that it can be used to determine the varying and ranking of the importance of the various factors (Al-batah, 2009). Analysis stage is the last stage in the CIS where the selected factors are fed as input into the analytical tool. ANN is one of the most common analytical tools used in classification and predictions. ANN has gained this reputation because of its ability to learn from the examples of real situation (Ruizheng, 1998, Martin and Peter 2009).

1.2 Background of Landslide Hazards

Landslides continue to be one of the worst natural disasters around the world. The term 'landslide' has many definitions. Based on Cruden (1991), any movement of debris or a mass of rock or earth down a slope is considered as a landslide. Varnes (1984) considered any probability of movement of the earth downward or outwards under the effect of the gravity, rain and slope as a landslide. Landslides cause losses in billions of dollars in damages and claim as many as thousands of lives each year worldwide.

Based on estimates from the Red Crescent Societies and Red Cross, landslides kill 1550 people in average every year (Natural Disaster, 2006). In the summer of 1998, multiple major landslides followed by heavy rain hit Bangladesh and China. More than 1100 people were killed in the former and around 4000 died in the latter. On October 30th, 1998, a major landslip around the volcano of Casitas buried around 2,200 people and caused millions of dollars in property losses. More than 1500 were killed in March of 1998 in Pakistan after a landslip and when floods hit the southwestern part of the country. In November 2001, a major landslide left Bab El-Oued, Algeria with more than 1000 people either dead or missing and a quarter of the country sinking in the mud and debris. In late February 2005, landslide occurred in Bandung, Indonesia killed more than 140 people. On August 10th, 2010, China suffered the worst landslide in decades which killed 702 people and left thousands of people missing (The BBC News, 2010).

Landslide is a major threat in Malaysia too, Particularly Penang Island, especially during the heavy rainy seasons. The island has a long history of landslides. In 1995, 60 massive landslides hit Penang Hill area following a heavy storm. Soil erosion had occurred which damaged the pathway near the Penang Botanical Gardens, causing trees to be uprooted and damaging some properties (The Malaya Mail Online, 2013). In September 2008, 14 landslides occurred in Teluk Bahang, while 3 along Jalan Tun Sardon. Landslides also cut off power supply and closing the roads and threated drivers along Jalan Tun Sardon (Asia One, 2008). Again in 2013, 13 landslides badly hit the area of Penang Hill damaging the roads and incurring costs around 2 million Malaysian Ringgit for repair (The Malay Mail Online, 2013).

Figure 1.1: Damage caused by landslide at Penang Island (Google source, 2013)

1.3 Motivations

As mentioned, landslides are one of the devastating phenomena, causing huge damage and loss of lives. The occurrence is caused by different factors such as geological, topographic, physical, and human causes (human disregard for sustainable developments) (Hutchinson, 1995). Digital Elevation Model (DEM) is one of the data source for extracting most of the factors, which are expected to have a linkage with landslide occurrence (Thompson, et. al, 2001). DEM is the representation of continuous elevation values over a topographic surface by a regular array of z-values, referenced to a common datum (Support Esri, 2014). Most of the studies on landslide hazard in Penang Island does not consider factor extraction as a part of the work methodology, hence some of the factors were obtained from various government agencies (Pradhan et al., 2010). In addition to that, the factor analysis of the study area in the previous works only focused on a few number of factors with little efforts to study the effect of extra factors such as the topography factors with 10 factors Pradhan et al., (2008), 9 factors Pradhan and Lee, (2010), Lim et al., (2011) with 12 factors, Pang et al., (2012) with 12 factors, 10 factors with Biswajeet Pradhan and Lee,(2009) and 8 factors with Lee and Pradhan, (2006).

Elman Neural Network, Elman (1990) is commonly used for real time problems such as monitoring in nuclear power plant (Şeker, 2003), motors speed estimations (Toqeer, 2003), electroencephalogram signals (Guler, 2005) and many other applications. In such applications, ENN shows good performance. However, ENN has not been used to predict landslides before. In this study, the performance of ENN in prediction the landslide will be investigated.

Cascade Forward Neural Network has shown good performance in many application such as food industry (Sumit and Kumar, 2011), in constructions (Sharma, 2010) and many other application. The incredible results achieved by the CFNN make it very popular. Even though, CFNN has never been used in the landslide hazard predictions. CFNN will be used to predict the landslide hazard map of Penang Island.

1.4 Problem Statements

A counter argument suggests that the topographic factors on the Penang Island need further investigations (Oh and Pradhan, 2011, Pradhan et al., 2010, Pradhan et al., 2008). The in-depth analyses on these factors are therefore required for two reasons:

- To determine the most important factors (the highest risk factors) that cause landslides.
- To produce an improved accuracy version of the landslide hazard map for Penang Island using these important factors.

Since the nature of earth is not the same and the factors triggering the landslide are not consistent, the ideal methodology for assessment and prediction of landslide occurrence is still under investigation. In all of the previous landslide hazard analysis, the extracted factors are used in the analytical tools without going through any factor selection process. In fact, not all the extracted factors are necessary for further analysis, whereas some factors are irrelevant or uncorrelated. Moreover, the unnecessary factors could increase the analysis complexity and decrease the performance of the analytical system by using as input, a large number of unnecessary data for analysis. The advantage of using the factor selection is not limited to a decrease in complexity and an increase in the performance of the system, but it is considered as a useful tool to determine the optimum factors. In other words, through factor selection, the most contributing factors can be determined and precautionary steps to reduce these high risk factors can be taken.

The strategy of factor extraction, factor selection and the analysis tool can be used to help mitigate hazards to people and facilities. This strategy can be referred in the developing plans to prevent landslide hazards, such as in locating, monitoring, and facility sites.

1.5 Research Objectives

The Penang Island landslide hazard analysis is proposed based on the factor extraction, factor selection and artificial intelligent analysis tool. Through these processes, a comprehensive idea on the high risky areas in Penang Island can be made available to the planer and decision maker, which in turn can assist in reducing the unnecessary cost of lives and properties. This research aims to achieve the following objectives:

1. To investigate various factor extraction techniques and to further determine new landslide causative factors.
2. To determine the best factor selection method, to identify and determine the ranking of the highest risk factors which affect landslides predictions.
3. To introduce a modified neural network structure, that would not only be able to predict the occurrences of the landside but also to classify the pattern of classification for different data sets.
4. To identify the best tool that can predict the occurrence of landslides and to produce an improved accuracy version of the landslide hazard map for Penang Island.

1.6 Research Scope

With the motivation and the problems described in section 1.2, intensive research is carried out which focuses on the landslide hazard mapping analysis. The main purpose of this work is to define a suitable landslide hazard intelligent prediction strategy for Penang Island. In addition to that, this strategy could be used to predict future landside occurrences with respect to the location in the study area (landslide hazard map). The proposed predictions strategy would be able to classify the study area into four levels, i.e. no hazard, moderate hazard prone, hazard prone and high hazard prone. The suggested strategy would then be more consistent, accurate, fast and automatic.

The study is based on variance data including the DEM of the Penang Island with a 5-meter resolution. The DEM is used to extract the topographic factors such as slope angle and curvatures and many more factors. In the landslide data analysis, new analysis tools are proposed to improve the landslide hazard analysis system prediction accuracy.

In this study, it is worth to mention that the image processing techniques are not involved in the factor extraction. The proposed system is developed and tested in the Matlab R2010a with Intel® Core ™ i7-2600 CPU @ 3.40 GHz and 16 GB RAM environment.

1.7 Thesis Outline

This thesis is organized in seven chapters as follows. Chapter one which is this chapter, explains the general background of the study, established the problem statement, objectives and scope of the study.

Chapter 2 is on the literature review, which generally discusses landside hazard in worldwide context and the study area of Penang Island in particular. A comprehensive review is also given on the techniques applied to predict the occurrences of the landslide hazards. This include, the techniques previously used to extract factors related to landslides.

Chapter 3 includes two parts i.e. a brief description for the methodology used in this work and the methodology applied to extract the factors causing landslides in the study area.

Chapter 4 introduces six different methods used in the factor selections. Two of these methods are based on ANNs while another three are based on the Decision Tree (DT) techniques, while the last method is based on discriminant analysis (DA). A comparison between the factors selected is shown, whereby the optimum factors are identified for the next stage in the analysis.

In Chapter 5, a modified neural network structure called the Hybrid Cascade Forward and Elman neural network (HECFNN) is proposed. In this chapter, the structure of the proposed network is introduced. In addition, the performance of the developed model (HECFNN) is evaluated through a number of experiments, which were conducted using well-known benchmark data sets and the results are compared with those from other methods published in the literature.

Chapter 6 describes the landslide hazard system by using the results from chapters 4 and 5. The best factors determined from chapter 4 are used as input and fed into three of ANN architectural models, which have already been explained in

Chapter 5. They are then trained, tested and the result are analysed with supportive discussions.

Chapter 7 presents the conclusion and highlights the contributions of this work, and recommends some possible extensions to this work.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

In the last chapter, the background on landslide has been provided. Therefore, in this chapter previous studies related to the subject matter will be revised.

This chapter is divided into different sections in order to provide comprehensive basic knowledge on landslide hazard mapping analysis and the factors which contribute toward the occurrence of landslides. The chapter starts with an overview of the impact of landslides over the entire world and some previous studies related to the prediction of landslides in section 2.2. Section 2.3 introduces the study area, reviews the landslide disaster history in Penang Island and gives a review on the previous works only on the study area. In section 2.4, the methods of factors extraction based on the factors which cause the landslides in the study area are presented. Details on the factor selection methods include classification accuracy, Zhou's method; Decision Tree (DT) and Discriminant Analyses (DA) are presented in section 2.5. An overview of related ANN theory which includes architectural design and some neural network applications are reviewed in section 2.6. Finally, the summary of this chapter is drawn in section 2.7.

2.2 Previous Studies on Intelligent Methods for Landslides

To predict the landslide occurrence, different methods have been applied and developed. These methods are divided into qualitative and quantitative methods. These two methods vary because of various reasons. For qualitative methods, parameters such as the methodologies used, the direct field mapping, geomorphologic analysis and methods based on human judgments are some examples. Meanwhile, methods such as deterministic analyses, artificial intelligence, probabilistic approaches and statistical methods represent the quantitative methods and are based on mathematical models. Nevertheless, the general agreement about the ideal method for producing landslide susceptibility map has not been reached yet (Murat and Candan, 2004). Initiation of landslide mapping began in the 1970's (Fell et al., 2008). Then, particularly in 1980's, in line with the achievements in computer technology and Geographic Information Systems (GIS), there was a boom related to landslide mappings in the scientific literature. The early 1990's showed the beginning of GIS applications for few cases on landslides. In some cases, the GIS package demonstrated the ability to achieve major analysis on landslide mappings whereas the usage of GIS was partial in other cases. Figure 2.1 shows the damages caused by landslides in different parts of the world.

Figure 2.1: Damages caused by landslides (Google source, 2013)

Previous works on landslide analysis using the quantitative methods include Lee et al. (2001) who proposed a study to introduce the landslide hazard mapping using Multilayers perceptron (MLP) neural network in Yong, Korea as a case study area. Seven landslide causative factors were collected and extracted from a special database. These factors include curvature, slope, soil effective thickness and texture, drainage and timber age and diameters. The back propagation algorithm was used twice in this study, the first to create the landslide map, while the second to determine the weights of each factor in the landslide map. The verification results between the susceptibility index and existing landslide location data shows good agreement and satisfactory output results. The landslide hazard prediction map was divided into five classes of hazard prone i.e. very low, low, medium, high and very high hazard. In order to determine each of the factor's weight and to obtain susceptibility based on the ratio values, a neural network is used to implement a three-layer feed-forward. Topographic slope had the highest value while the lowest is topographic curvature. The study suffered a setback in using only a few number of the landslide occurrence factors (only seven).

Probabilistic Neural Network (PNN) and Multi Layered Perception (MLP) were used by Ermini et al. (2005) to produce landslide hazard maps. Five factors were considered in this study. These are namely lithology, profile curvature, slope angle, land cover and upslope. The size of the case study area is 17 km^2, located in Riomaggiore, Italy, which is considered as an ideal space for performing tests on landslide hazard analysis. The five factors used in the analysis are considered the classic controlling variables which control the landslide hazard. All the input factors were converted into binary variable strings which consist of 19 positions. These factors were used as input to the ANN. Satisfactory results with preference for MLP were shown by comparison with the recent landslide inventory of the study area, and the ANN has the ability to predict the hazard mapping with satisfactory results. The major weakness of the study is that only a few numbers of factors i.e. five were used.

Yesilnacar and Topal (2005) proposed a comparison of neural networks and logistic regression methods. In this study, neural networks have again proved that they are more realistic than any other techniques for landslide susceptibility hazard mapping. The goal of the study was to produce a landslide mapping for natural gas pipeline in the study area which covers 290 km^2 representing the area around the gas pipeline, located in Marmara and Black Sea regions of Turkey. Collecting and preparing the data are one of the major steps in landslide susceptibility mapping. In the study, the landslide inventory map was prepared based on the previous inventory map and extensive fieldwork. Logistic regression (LR) and MLP were used to analyze some probable landslide causing factors such as slope angle and slope length, topographic wetness watershed basins index, surface area ratio, curvature plane and profile, distance from road drainage and fault line, elevation, density of drainage and fault, land cover and use and the stream power index. In the study, two landslide susceptibility maps were produced by LR and NNs. Validation data set and the field check were used to evaluate the two maps. On a 1:25000 scale map, 112 landslides were found. The pre-processing in the study was done by putting all the independent variables on hold. In some subsequent steps when the variables determined are felt to be significant, they will be added to the system while the others will be withheld. The study did not include factors such as general curvature and vegetation index into the model. The data set nature played an important rule in the accuracy of this comparison. Three of the accuracy indicators used in this study includes percentage of relevant susceptibility level in whole area (PTA), percentage of relevant susceptibility level in landslide bodies (PLB), percentage of relevant susceptibility level in seed cells (PSC). The ratio of PTA/PSC and receive operating characteristics (ROC) curves were calculated. The values of PTA, PSC and the ratio of PTA/PSC should be below the value of PLB and PSC. The results of these indicators showed once again that the neural feed forward network with back propagation algorithm performs better than the logistic regression model.

The back propagation learning algorithm was used in the landslide susceptibility mapping by Caniani et al. (2008) in Potenza, Italy. In the study, 920 landslides were recognized, which represent the earth flow, rotational slide and rotational slides, spreading over 46 km^2. The space of study area represents only 26% of the entire area of Potenza. Three layers of neural network input layer, hidden layer and output layer were connected to each other respectively. The back propagation algorithm with the three layers was used as the learning algorithm. The landslide causes factors used are such as geomorphology, geological, metrological and hydrologic conditions, which include lithology, slope aspect and angle, elevation, topographic index and topographical shape and land use. All the morphometric parameters were derived from the digital elevation model (DEM) of the area of Potenza with a resolution of 20 meters. The work was divided into two phases, i.e. the training phase and the validation phase. In the study, 32% of the landslide site was selected for training phase and the rest of the landslide site was used for the validation phase. The weights of each factor on the seven factors were calculated. Slope aspect, slope gradient, elevation and lithology had the highest weight. The verification step found those factors were the most effective factors, which could lead to landslide susceptibility. ANN showed good performance by classifying 80% of the landslide pixels correctly. However, in this study, not all of the possible factors, which cause landslides, were considered. The training data sets collected were also from a small part of the study area i.e. not all the study area was represented in the training stage.

Pradhan et al. (2008) partially applied 10 of the factors linked to the landslides to calculate the weight of each factor using the MLP neural network and predict the landslide hazard map. Field survey and aerial photographs were used to identify the landslide location of part of Cameron Highlands in Malaysia. 324 landslides were found in the study area. In addition to that, the database of the study area was divided into three parts to assemble the access to the map database. Again, the MLP with one input one output and one hidden layer was used; the weight of each factor between the layers was calculated by applying the back propagation algorithm. The aim of the study was to calculate the weight of each factor. Once the weight of each factor is calculated, it can be used in the classification of the new data which have never been used in the neural network before. The rate curve was created by finding the error value between the actual output value and the neural network output value. The area under the curve (AUC) was used to detect the NNs prediction accuracy. The Matlab software was used to implement the feed forward neural network, the relative importance of every factor between the weights showed that the slope factor has the highest value among all the factors which is 2.05, this is then followed by the distance from drainage which is 1.4 and then geology, whose value is 1.1. The network accuracy prediction is 83.45%. In this study, the importance of 10 factors were calculated in the first stage of the work (Pradhan et al, 2008). In the second stage, all the factors were used to predict the Cameron Highland landslide hazard map, i.e. both the important and not important factors were used in the hazard prediction map.

Marjanovic et al. (2009) focused on using support vector machine (SVM), Neighbor (k-NN) algorithms and Analytical Hierarchy Process (AHP) for the weighting of the influences of different input parameters. Seven factors were used to predict the landslide map. These factors include elevation, slope angle, aspect and distance from road, vegetation cover, lithology and rainfall to represent the natural factors of the slope stability. The study area was the North West slopes of the Fruška Gora Mountain, in the vicinity of Novi Sad, North West Serbia which represents 40 km^2 of hilly landscape. The research was divided into two parts i.e. the expert's opinion in multi-criteria analysis and the machine learning feature of SVM and K-NN algorithm. Multi-criteria analysis is a widespread tool for various types of assessments, especially for spatial implications. It implements a procedure where several inputs fused a single outcome of the modeled phenomenon. However, these geo parameter inputs have different levels of importance for the phenomenon. These need to be scaled in some fashion. Therefore, in this case, the Analytical Hierarchy Process was useful. SVM method reached the highest accuracy of 88.00%.

Young et al. (2003) have done comparisons between frequency ratio (FR), logistic regression (LR) and artificial neural network (ANN). A small study area was chosen in the Republic of Korea. With a size of 8.13 km^2, the landslide occurred in 300 locations. Five landslide factors were used, i.e. the topographical factor, hydrological factor, soil factor, forest factor and land cover factor. AUC analysis was built with each model to assess the performance of FR, ANN and LR. Analysis result showed that there was a high correlation between the maps using LR and ANN methods which exhibits the highest correlation coefficient of 0.829 while the lowest coefficient of 0.619 was found between LR and FR methods. Each model has some advantages and disadvantages. FR can be applied in a simple way, whereas an LR method needs data conversion to be read by the statistical software program. LR method has limitations on calculation in the program when the data is massive. ANN attains the highest accuracy.

Murat and Candan (2004) carried out fuzzy logic as a new methodology to create the landslide map to the West Black Sea Region of Turkey with 275.4 km2. The study area contained two hundred and sixty six landslides. Different factors have been involved in this study. These include slope shape, slope angles, slope aspect, elevation, distances to drainage network and geological factors. The study has also employed software program to utilize the fuzzy relations to produce the landslide susceptibility map automatically. The map of the case study area was classified into very high, high, moderate, low and very low or no susceptibility. From the results obtained in the study, the fuzzy logic showed good performance in producing the landslide susceptibility map. In addition, the approach was considered as a useful tool because its results were obtained from the available data of landslide. The study needs more factors to be investigated.

Yilmaz (2009) compared the landslide susceptibility mapping produced by three different methods; MLP network, FR and LR. Eight factors were considered, namely slope angle, drainage system, geology, faults, slope aspect, topographical elevation, Topographic Wetness Index (TWI) and Stream Power Index (SPI). Factors used in this study were obtained using the ArcGIS software. ArcGIS is a comprehensive system that allows people to collect, organize, manage, analyze, communicate, and distribute geographic information (Resources Arcgis, 2014). The case study is situated at Kat (Tokat—Turkey). The landslide susceptibility produced by the MLP showed the best accuracy compared to FR and LR with values of 0.826, 0.842 and 0.852 for frequency ratio, logistic regression and artificial neural network, respectively.

Furthermore, some researchers have studied the factors which affect the landslides individually. Bibalani et al. (2007) studied the link between the vegetation cover factor and soil stability, in the study area located in the northwest of Iran. Gasim et al. (2010) have produced a study to determine geomorphology and geological factors of the Bukit Bauk, Malaysia. The geological and geomorphology factors were considered among of the important factors causing landslide hazards. The term geomorphology refers to the study of the characteristics, origin, and development of landforms (Perillo, 1995).

Dieu et al. (2012) have applied three models; Naïve Bayes (NB), support vector machines (SVM) and the Decision Tree (DT) to produce the landslide hazard map of Hoa Binh, Vietnam (4,660 km^2). In the study, 10 factors included were slope angle, relief amplitude, distance to the rivers, distance to roads, rainfall, distance to faults line, land use, slope aspect, soil types and lithology. Only three landslide hazard maps were predicted from each model (NB, SVM and DT). The prediction accuracy of the three predicted maps were compared. The results showed that the prediction accuracy achieved using the SVM has the best, followed by the NB and the hazard map predicted using DT was the worst. The study did not implement any factors selection methods before producing the landslide hazard map.

Table 2.1 summarizes some of the methods and techniques introduced in this chapter for landslide analysis over the entire world. The advantages and the disadvantages of methods mentioned in Table 2.1 are summarized in Table 2.2.

Table 2.1: Some previous work on landslide hazard mapping in worldwide (excluding Penang Island)

References	Analysis Method	Study Area	Number of Factors	Analysis Concept
Lee et al. (2001)	MLP and GIS	Yongin, Korea	6	-Produce landslide hazard map. -Factor importance.
Young et al. (2003)	MLP, FR, LR	Republic of Korea	8	-Produce landslide hazard map.
Murat and Candan (2004)	Fuzzy logic and GIS	Turkey	13	- Produce landslide hazard map.
Ermini et al. (2005)	MLP, PNN and GIS	Riomaggior, Italy	17	-Produce landslide hazard map.
Yesilnacar and Topal (2005)	MLP, LR and GIS	Hendek, Turkey	19	- Produce landslide hazard map. -Factor importance.
Bibalani et al. (2007)	Finite Element Method	Iran	2	-Study link between vegetation cover and soil stability.
Caniani et al. (2008)	MLP and GIS	Potenza, Italy	7	- Produce landslide hazard map. - Factors weight.
Pradhan et al. (2008)	MLP and GIS	Cameron Highland, Malaysia	10	-Factors weight.
Yilmaz (2009)	MLP, FR, LR and GIS	Turkey	8	- Produce landslide hazard map.
Gasim et al. (2010)	Schmidt Net	Malaysia, Bukit Bauk,	2	-Factors weight.
Dieu et al., (2012)	NB, SVM, DT and GIS	Vietnam	10	- Produce Landslide hazard map.

Table 2.2: The advantages and disadvantages of analyzing methods

References	Advantages	Disadvantages
Lee et al. (2001)	- High data resolution.	- Few factors (6) and few samples for studying (200) samples - No factors selection - Expensive using GIS program to produce the final hazard map
Young et al. (2003)	- mixed models statics and ANNs - Low data dimension.	- Few factors (7). - no factors selection
Murat and Candan (2004)	- Satisfactory factors	- No factors selection - the hazard map divided to landslide or no landslide (0,1)
Ermini et al. (2005)	- Satisfactory input factors.	- No factors selection. - Complex. Changing the data to binary - Expensive using GIS program to produce the final hazard map.
Yesilnacar and Topal (2005)	- Satisfactory input factors.	-No comparison between the hazard map before and after factors selection. - Expensive using GIS program to produce the final hazard map.
Bibalani et al. (2007)	-Find the factors importance.	- few samples. - Low resolutions
Caniani et al. (2008)	-Low data dimension	- Few factors (7). - Few sample for studying. -No comparison between the hazard map before and after factors selection.
Pradhan et al. (2008)	- Satisfactory input factors.	-No factor selection -Expensive using GIS program to produce the final hazard map.
Yilmaz (2009)	- High data resolution.	- Few factors (8). -Expensive using ArcGIS program to produce the final hazard map.
Gasim et al. (2010)	-Produced the landslide	-Few factors (7). - Few sample for studying
Dieu et al., (2012)	- Satisfactory factors	- No factors selection - Expensive using ArcGIS to produce the final hazard map.

In this section, a review of the different methods used in the landslides hazard mapping in different study areas is presented.

2.3 Study Area

As this study is focused on Penang Island (Figure 2.2), this section dwells on the background and information related to it. Penang Island lies between 5°15' to 5°30' N latitude and 100° 10' to 100° 20' E longitude. It occupies an area of 285 km^2 and is one of the thirteen states of Malaysia. The island is bounded to the north and east by the state of Kedah, to the south by the state of Perak, and to the west by the Straits of Malacca and Sumatra, Indonesia.

Figure 2.2: Map of the study area (Google source)

Penang state consists of both the island of Penang and a coastal strip on the mainland, which is known as Province Wellesley. The study area in this thesis is state of Penang Island. It experiences frequent landslides that threaten lives and damage properties. These landslides occur quite frequently during the rainy seasons (Oh and Pradhan, 2011 and Lim et al.;2011). Figure 2.3 shows some damages caused by the previous landslide in Penang Island (New Straits Time, 2012). The study area has a tropical climate with high temperatures of 29 °C to 32 °C. The average amount of rainfall varies from 2254 mm to 2903 mm annually and the humidity ranging from 65% to 96%. The topographic elevations of the Penang Island areas vary between 1 m and 820 m above sea level, and the slope angle ranges from 0° to 87°. Flat lands make up to 43.28% of the island. Geological data from the Department of Mineral and Geosciences show that Ferringhi granite, Batu Maung granite, clay, and sand granite represent more than 72% of the study area's geology. Vegetation cover consists mainly of forests and fruit plantations.

Figure 2.3: Damage caused by landslide Penang Island. (New Straits Times, 2012)

Pradhan et al. (2010) combined Geo Information Technologies (GIT) and neural networks to produce landslide susceptibility mapping for three study areas in Malaysia, which are Penang Island, Selangor and Cameron Highlands. 15 different factors were used in this study. They are slope aspect, slope angle, plan curvature, altitude, stream power index, topographic wetness, power index, distance from road, distance from drainage, distance from fault line, geology, land use, Normalized Difference Vegetation Index (NDVI), soil texture, soil material and topography. In the study, the cross validation methods were applied, i.e. the weights of each landslide causative factor were used among the three study areas. MLP neural network with back-error propagation for training were employed to train the MLP. The amount of data collected to train the MLP was based on the availability of data in the three study areas i.e. for Penang Island, 579 cases of landslides, 409 cases of landslides for Selangor and 405 for Cameron Highlands. During the MLP training stage, the data was divided into two parts, 80% for training and 20 % for testing. The trained and tested network was then applied on a different study area, i.e. the trained network through Penang Island data was used to test the Selangor study area while the trained network from Selangor data was used to test Cameron Highlands and so on. As a result, the study shows that the best accuracy trends are achieved while using appropriate weights for the study area itself. The accuracies of landslide hazard map for the three study areas were Penang Island 84.43%, Selangor 86.15%, and Cameron Highlands 89.32%. In addition to landslide hazard map, Pradhan et al. (2010) also used the MLP layers weights to determine the importance of each factor. The importance of each factor varies from area to area. The factor weightage for the three study areas show that; topographic factors such as slope angle and slope aspect are among the most important factors, which cause the landslide to occur. Moreover, Pradhan et al. (2010) stated the importance of factors in descending order for Penang Island i.e. slope angle, soil texture, slope aspect, topography wetness index, distance from road, distance from drainage, stream power index, plan curvature, land use, NDVI, topography, geology, soil material and distance from fault line. This study is considered as an interesting study since it approves that the best accuracy is achieved