Faculty of Computer Science and Information Technology

Extreme Gradient Boosting (XGBoost) and Random Forest (RF) Hybrid Ensemble with Bayesian Optimization in Landslide Susceptibility Mapping

Dorothy Anak Martin Atok

Master of Science
2025

Extreme Gradient Boosting (XGBoost) and Random Forest (RF) Hybrid Ensemble with Bayesian Optimization (BO) in Landslide Susceptibility Mapping

Dorothy Anak Martin Atok

A thesis submitted

In fulfilment of the requirements for the degree of Master of Science

(Computer Science)

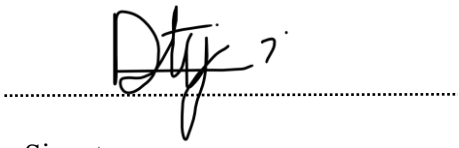
Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2025

DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Malaysia Sarawak. Except where due acknowledgements have been made, the work is that of the author alone. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.



Signature

Name : Dorothy Anak Martin Atok

Matric No. : 21020026

Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak

Date : 11/04/2025

ACKNOWLEDGEMENT

First, I acknowledge the blessings bestowed upon me by the Almighty God that enabled me to be motivated and persistent in this endeavour. Thanks to my amazing parents and family for supporting me financially and morally. I would also like to thank my best friends for constantly boosting my motivation and making me believe I can complete this roller-coaster journey. The research and thesis writing process was not easy, but thanks to my fantastic supervisor, Dr Chai Soo See, who gave me the freedom to explore the subject of my study without any pressure and was always there to guide me. Lastly, thank you to all the Faculty of Computer Science and Information Technology and the Centre of Graduate Studies members who have helped me whenever I have doubts.

Thank You.

ABSTRACT

Many models have been developed in the past for predicting landslide vulnerability. The generated models have certain shortcomings, especially when it comes to the problems of overfitting and overestimation. Therefore, the objective of this study is to assess and enhance the performance of Extreme Gradient Boosting (XGBoost) and Random Forest (RF) in predicting the susceptibility of landslides through the combination of the two models and the use of the Bayesian Optimization (BO) algorithm. The selected study area for this case study is Penang Island, which is a landslide-prone region in Malaysia. The topographical, hydrological, anthropogenic, and external factors that influence landslides were considered in this study. Geographic Information Systems (GIS) have been used to create the spatial databases of all the landslide conditioning elements. With an area under the curve (AUC) of success rate (SR) as high as 100.0% and prediction rate (PR) of 97.1%, the results indicate that optimized XGBoost performed best, followed by random forest (SR:99.3%, PR:95.9%). and stacked models (SR:96.8%, PR:95.6%) The final AUC findings showed that stacking and optimizing the hyperparameters can increase the algorithms' performance accuracy by overcoming the overfitting and overestimation issues. Although there are many applications for landslide susceptibility maps, they are crucial for site selection, engineering structure and disaster avoidance.

Keywords: Bayesian optimization, extreme gradient boosting, geographic information system, landslide susceptibility mapping, random forest

Penggabungan Extreme Gradient Boosting (XGBoost) dan Random Forest (RF) dengan implementasi Bayesian Optimization (BO) dalam penghasilan peta kerentanan tanah runtuh

ABSTRAK

Banyak model telah dibangunkan pada masa lalu untuk meramalkan kawasan yang berisiko tinggi untuk mengalami tanah runtuh. Model yang dihasilkan mempunyai kelemahan tertentu, terutamanya apabila ia melibatkan masalah ramalan model terlampau sama dengan set data pelatih dan masalah aggaran berlebihan oleh model. Oleh itu, objektif kajian ini adalah untuk menilai dan meningkatkan prestasi Teknik "Extreme Gradient Boosting" (XGBoost) dan "Random Forest" (RF) dalam meramalkan kejadian tanah runtuh melalui gabungan kedua-dua model dan penggunaan "Bayesian optimization" (BO). Lokasi kajian kes adalah di Pulau Pinang, salah satu kawasan yang kerap mengalami tanah runtuh di Malaysia. Faktor topografi, hidrologi, antropogenik dan luaran yang mempengaruhi tanah runtuh telah dipertimbangkan dalam kajian ini. Sistem Maklumat Geografi (GIS) telah digunakan untuk mencipta data spatial bagi semua faktor tanah runtuh. Dengan kawasan di bawah lengkung (AUC) kadar kejayaan (SR) setinggi 100.0% dan kadar ramalan (PR) 97.1%, keputusan menunjukkan bahawa XGBoost yang dioptimumkan menunjukkan prestasi terbaik, diikuti oleh RF (SR:99.3%, PR :95.9%) dan model hibrid (SR:96.8%, PR:95.6%) Penemuan akhir AUC menunjukkan bahawa menyusun dan mengoptimumkan hiperparameter boleh meningkatkan ketepatan prestasi algoritma dengan mengatasi isu ramalan model terlampau sama dengan set data pelatih dan masalah aggaran berlebihan oleh model.

Kata Kunci: *Bayesian optimization, extreme gradient boosting, peta ramalan tanah runtuh, random forest, sistem maklumat geografi*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AB	AdaBoost
ADTree	Alternating decision tree
AHP	Analytical Hierarchy Process
AUC	Area under curve
BGAM	Boosted Generalized Additive Model
BO	Bayesian Optimization
CF	Certainty Factor
CNN	Convulational Neural Network
DEM	Digital Elevation Model
FOP	Factor Optimization Process
FR	Frequency Ratio
GP	Gaussian process
IDW	Inverse Distance Weighted
KNN	K-Nearest Neighbor
LaGRiSU	Landslide Grid and Slope Unit
LR	Logistic Regression
LSM	Landslide Susceptibility Mapping
ML	Machine Learning
NDVI	Normalized Difference Vegetation Index
NSMP	National Slope Master Plan
OOB	Out of Bag
RF	Random Forest

RFE	Recursive Factor Elimination
RNN	Recurrent Neural Network
ROC	Receiver operating characteristics
RUSLE	Revised Universal Soil Loss Equation
SPI	Stream Power Index
SVM	Support Vector Machine
TOL	Tolerance
TWI	Topographic Wetness Index
UCB	Upper Confidence Bound
USGS	United States Geological Survey
USLE	Universal Soil Loss Equations
VIF	Variance Inflation Factor
XGB	Extreme Gradient Boosting

CHAPTER 1

INTRODUCTION

1.1 Background

Landslides are natural hazards that present a major threat to the property, human life, and the environment where they occur because of their potential for destruction (Ayalew & Yamagishi, 2005; Chen et al., 2018). Penang Island is a region that consistently grapples with the occurrence of landslides. Penang Hill, situated in a hilly terrain, stands out as a construction project subject to ongoing, vigilant monitoring by regulatory authorities. The intensified scrutiny is a proactive measure undertaken in response to the increased susceptibility of Penang Hill to potential landslide events. Minor landslides have been observed to occur frequently on Penang Hill, which is particularly during periods of heavy rainfall.

To lower risks and losses from landslides, one of the first tasks completed by the Slope Engineering Malaysia upon its establishment in 2004 was the creation of a National Slope Master Plan (NSMP). A few of the NSMP action plans (2009-2023) that have been put into practice involve developing susceptibility, hazard, and risk maps, which are then followed by tasks like inventorying slopes, executing public awareness and education campaigns, and establishing agency committees to create efficient policies and collaboration for slope mitigation (Kementerian Kerja Raya, 2009).

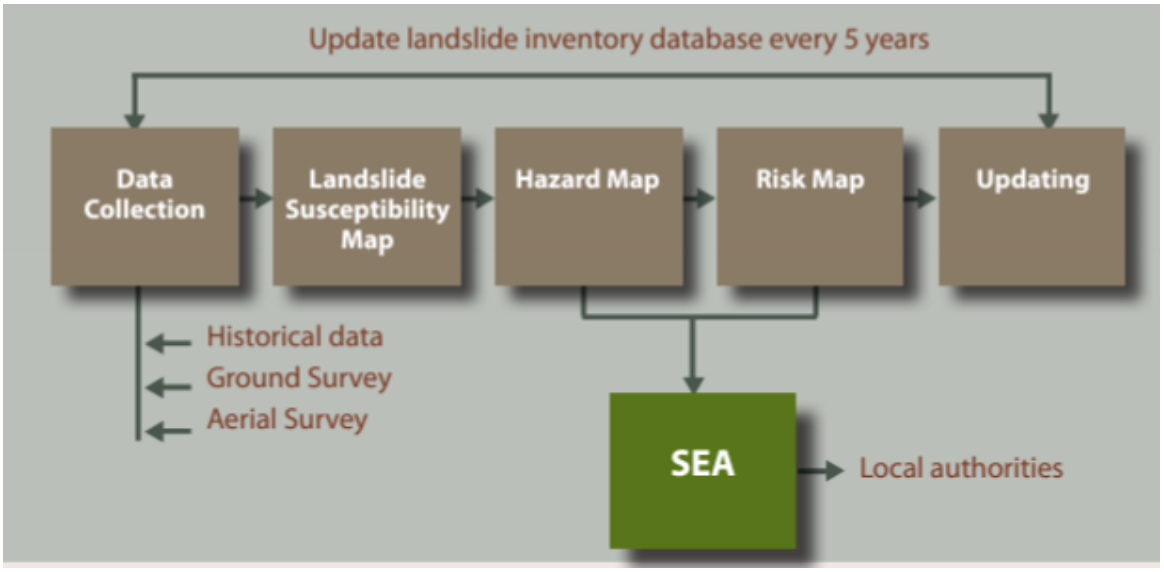


Figure 1.1: Implementation Process of hazard mapping assessment by Malaysia NSMP (2019-2023)

The Public Works Department (PWD) of Malaysia was tasked with cataloguing slopes and creating hazard and risk maps in high-risk regions, particularly cities, and towns where development has occasionally occurred without appropriate planning, design, and construction, according to NSMP. The budget for further hazard and risk mapping for landslide-prone areas, which spans an estimated 1,700 km^2, has been granted and encompasses approximately 95% of Malaysia's high-risk landslide-prone areas, which include major cities such as Penang Island and Kuala Lumpur (Kementerian Kerja Raya, 2009). In addition, the landslide susceptibility map (LSM) has also been included in the production of landslide risk assessments and mapping in guidelines prepared by the Association of Southeast Asian Nations (ASEAN) (ASEAN, 2021).

In addition, the Penang State Government has identified highland areas with an elevation over 76 m and a slope greater than 25º as a sensitive environment where development needs to be limited in their Structure Plan for 2030 (Jabatan Perancang Bandar dan Desa Negeri Pulau Pinang, 2019). The state's population is growing along with its

economy, which has increased demand for housing and infrastructure. As a result, development has spread into more vulnerable geohazard-prone locations, including hillsides and low-lying coastlines. Land development inside high-land and sloped areas has proceeded despite the inherent risks, particularly for residential construction, which raises the likelihood and potential effects of landslides (Lee & Pradhan, 2006).

The primary working hypothesis in prognostic techniques based on historical landslide events is that future landslides will have similar distribution patterns to previous ones (Youssef & Pourghasemi, 2021). An LSM can be created based on the information available to predict the occurrence of future landslides. Statistical, analytical, and soft computing tools for generating landslide susceptibility models exist.

It is essential to select suitable techniques for assessing landslide hazards and zoning. Geographic Information Systems (GIS) and machine learning (ML) models are powerful tools for LSM studies. ML model is a computational algorithm which includes Random Forest (RF) or Extreme Gradient Boosting (XGBoost) that is fed data to spot trends and extrapolate future trends. More specifically, within the context of landslide susceptibility study, the term refers to the predictive models which have been developed to analyze the landslide susceptibility by learning from data that is fed into them such as landslide inventories and landslide conditioning factors and producing results in the form of classification of susceptibility. As with the rest of machine learning, these models are part of it but only as tools to be applied to resolve certain issues through obtaining results from given data.

ML models, such as Random Forest and Extreme Gradient Boosting, are used in landslide susceptibility mapping (LSM) to predict areas at risk for future landslides. These

models learn patterns from data and are trained on historical occurrences and environmental variables. Studies like Youssef and Pourghasemi (2021) and Pham et al. (2021) have successfully created accurate susceptibility maps using these models. Despite a decade of comprehensive studies employing GIS and ML models, there remains ample opportunity for refinement within the current models due to certain drawbacks such as overfitting and overestimation which can affect the models' performance (Agrawal & Dixit, 2023; Park & Kim, 2019). This enhances the model's accuracy and performance in developing a highly accurate and reliable LSM. Thus, this research study will focus on addressing issues of the existing models used in LSM and proposing potential solutions to improve the performance of the models.

1.2 Problem Statement

Machine learning (ML) models have been used by researchers to generate landslide susceptibility maps in the past decade. Somehow, the variability in geo-environmental conditions across various study areas is also affecting the performance of the machine learning models (Dou et al., 2019). Each region has different characteristics of topography, hydrology, and environmental factors. Different models generate different evaluation results because each model has its own implementation method, benefits, and limitations (Merghadi et al., 2020). Besides the quality of landslide inventory, the landslide conditioning factors spatial databases play a great role in mapping landslide susceptibility (Tien Bui et al., 2012).

The existing data-driven susceptibility mapping models, despite a decade of study utilizing both GIS and ML models, face limitations in global applicability. This constraint arises from the substantial variability observed in landslide conditioning parameters. Challenges are further compounded by restrictions in accessing up-to-date and accurate data,

particularly in areas that are inherently difficult to reach. These constraints collectively impede the seamless development of a universally applicable landslide susceptibility map (Can et al., 2021). Thus, it is very important to choose a robust machine learning model with consistent performance across various study areas with the various landslide conditioning factors less affected by the available data quality. This will save more time, cost, and resources if there is a robust model that can predict and able to produce a reliable landslide susceptibility map with minimum input of landslide conditioning factors (Dou et al., 2019).

Furthermore, an overlooked challenge in landslide susceptibility mapping, evident in prior research such as Agrawal and Dixit (2023) and Rabby et al. (2020), is the issue of overestimation. In the context of LSM, overestimation in ML model is the model's tendency to predict more areas as susceptible to landslides than are actually at risk, potentially inflating the number of high-risk zones. This occurs when the model becomes too sensitive to certain patterns or features in the training data, leading to an exaggerated number of false positives (areas predicted as landslide-prone but are not). Overestimation tends to occur even when the ML model has high prediction performance (Agrawal & Dixit, 2023). This problem will produce a less reliable and accurate landslide susceptibility map (Rabby et al., 2020). Thus, it will also cause cost and time inefficiency for the town planner and local authorities.

Moreover, with contemporary machine learning prediction techniques, striking a balance between model simplicity and predictive accuracy proves challenging. This is because machine learning tends to perform better when its complexity increases, and this will lead to the occurrence of overfitting (Wu et al., 2019). The persistent issue of overfitting remains a critical concern, necessitating further refinement and the development of more reliable solutions. Developing an optimum model that is less complex but accurate is very

important. This can be achieved through a good hyperparameter optimization technique (Sun et al., 2020).

Numerous ML models have been developed over the years, but ensemble learning models such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) has consistently outperformed them. XGBoost and RF are two effective machine-learning techniques employed in LSM research. The XGBoost is the boosting method of ensemble learning that has recently been employed in constructing landslide susceptibility models. The RF bagging strategy has been widely used for the past seven years.

Although the XGBoost and RF have been proven to be reliable models in predicting landslide susceptibility, they still have drawbacks which requires improvement, especially in dealing with overfitting (Park & Kim, 2019) and overestimation issues (Agrawal & Dixit, 2023; Rabby et al., 2020). Hence, integrating two different ensemble learning models can further enhance the single homogenous models' ability to make predictions and deal with geo-environmental variability (Susan et al., 2021). Moreover, the overestimation problem can be overcome by stacking two different models (Rabby & Li, 2020).

However, the practical applications of the proposed hybrid model may vary across different geo-environmental contexts. While the hybrid model could provide better adaptability and robustness in diverse environments, its limitations must also be considered. These include potential challenges in handling highly complex or large-scale datasets, variations in model performance across different geographical regions, and the need for careful selection and preprocessing of landslide conditioning factors to ensure accurate predictions (Liu & Li, 2020).