CHAPTER 3:

METHODOLOGY

3.1 Introduction

This study employs Geographic Information Systems (GIS) and machine learning methods on an R-studio platform to formulate a strong theoretical model for predicting landslide susceptibility in the Selangor region of Malaysia. Landslide(mapping) is an essential geohazard that has been influenced by steepness in the terrain, rain intensity, lithology and land use changes in the region. Landslide risk assessment studies of the conventional method apply qualitative evaluation or simple statistical methods which might not be very accurate or feasible for large scale studies. This research overcomes these limitations by using systematic and data-driven research methodology that incorporate GIS as a tool for spatial data management and incorporate machine learning for predictive modelling of high-risk areas.

The process includes data collection, data processing and analysis and recommendation of the next course of action. The first phase in the data acquisition is data collection in which the geographic and environmental data like slope, elevation, geology, land use, rainfall, hydrology and historical landslide inventory data gathered from different sources. During the data preparation stage of a GIS project, the data is pre-processed and normalized for spatial referencing and compatibility to combine multiple data layers into composite analytical datasets. Data exploration and feature engineering applied GIS tools to assess the spatial patterns of the landslide susceptibility factors such as the drainage, steepness of the slopes, geology and land use.

In development of modelling, machine learning such as Random Forest (RF), Linear Regression (LR) and Extreme Gradient Boosting (XGBoost) are used. These machine learning applied to identify trends in the existing data and the risk of occurrence of a landslide in different areas. The predictive performance of these models using metrics such as ROC curves, AUC scores, accuracy, precision, F1-score, mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) are used for model evaluation to ensure the effectiveness and reliability of the data.

In visualization phase, the trained models are combined with GIS to produce landslide susceptibility maps that categorized the landslides areas into different risk levels. These maps then presented to stakeholders for decision-making and monitoring the landslides activities in future. This actionable awareness will providing valuable support for disaster management initiatives and risk mitigation strategies in landslide areas of Selangor, Malaysia.

3.2 Proposed Solution

This paper aims to identify the approach to mapping of the landslide susceptibility in Malaysia through GIS approach that incorporate modern machine learning techniques. They assess the ability of Random Forest and Extreme Gradient Boosting (XGBoost) in the performance for the prediction of landslides areas. Based on Acharya et al. (2023), the Random Forest model is the best model for computing the probability of landslides depending on the topographic, the amount of soil moisture, and rainfall using the least MSE than other algorithms when it is complemented by feature selection measures such as mutual information.

Regarding the problem of imbalanced datasets, Song et al. (2023) used oversampling and undersampling techniques that had a positive impact on all the model's recall and AUC values. Directly using the Logistic Regression, Random Forest and LightGBM models on equally balanced data helped in improving the degree of certainty for landslide susceptibility map predictions.

Wang et al. (2020) also indicate that Random Forest method provides high accuracy in GIS-based landslide mapping through various topographic and geologic indices extricated from satellites data. These conclusions validated the applicability of another machine learning algorithm, which is Random Forest, in providing accurate and usable susceptibility maps. Subsequently, Sharma and Sandhu (2023) again confirmed the high accuracy of Random Forest in terms of predicting landslide susceptibility and improving the performance on large and complicated datasets that focus on slope, lithology, land cover and rainfall as the crucial aspects for model training and validation. So, based on previous research data, Random Forest was chosen in this study along with Linear Regression and XGBoost to provide accurate prediction and tools for disaster prevention and risk reduction of the landslide areas in Malaysia.

3.3 Research Operational Framework

3.3.1 Research Design Diagram

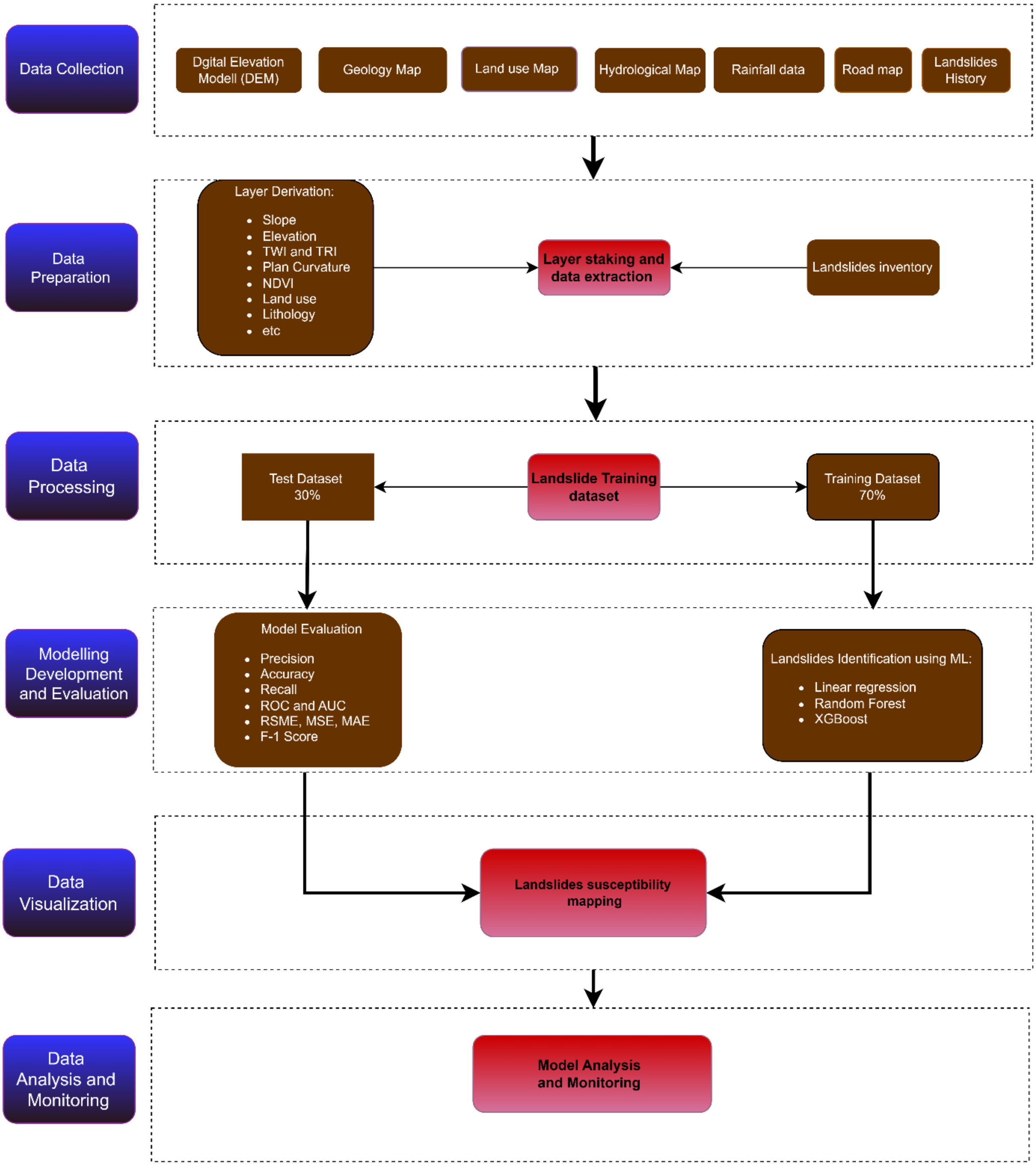

Figure 3.1 : Landslide susceptibility analysis flowchart

3.3.2 Description of Work Breakdown (Step)

1. Data Collection

Firstly, the landslide susceptibility maps produced from the combination of multisources, multilayer spatial and environmental data obtained from the existing databases and the Internet. Elevation data is obtained from Google Earth Pro and other supplementary geographic information is obtained from Landsat 8. The road network data is derived from Openstreetmap; the historical landslide point data is downloaded from NASA Global Landslides Catalogue. Such datasets are important for solving the task of factor which are contributing to the landslide risks. In this study, the topographic parameters are determined using the slope and elevation of the terrain, which were obtained from a Digital Elevation Model (DEM) from USGS open-source data. Geological maps show rocks and faults, while land use indicate changes that humans have made like deforestation, or development of land into urban centers. Drainage patterns and groundwater flow data together with rainfall data is used to study the effects of water in slope stability. Anthropogenic pressures are integrated into road network data while historical data on landslides provide general information on distribution and trend. This strong method of data collection sources ensures that all important variables are included in the analysis type.

2. Data Preparation

For main data collected, it undergoes pre-processing to generate landslide inventory map that shows the previous occurrences of the landslides. Elevation points are processed into DEM and analyzed with GIS software to obtain the thematic layers critical to modeling of landslides. The DEM is utilized to determine the slope, elevation, model-derived topographic wetness index (TWI), and plan curvature and referring to the hydrological and morphological features of the land surface. Water body area is extracted from remote sensing imagery as a measure of water surface area that is another stabilizing factor where vegetation indices which Normalised Difference Vegetation Index (NDVI) is are used. Further aggregation layers involve lithology to incorporate human effects on the slope steadiness. These datasets are in spatial matching format, normalized, and integrated together into the standardized format for further model development.

3. Data Processing

The obtained data is pre-processed and partitioned further into training (70%) and test (30%) sets for the further creating of a machine learning model. This way, the data will only consist of records with the same fields and format, the features' values will be in the right scale and there will be no errors in any of the records. Only well-processed data can guarantee relatively accurate and dependable assessment of landslide susceptibility. A training subset is employed to train the machine learning models and a testing subset is used to assess the generality of the models. According to the principles of data processing, the subsequent modelling phase is made more reliable and accurate by strictly following defined procedures.

4. Model Development and Evaluation

Landslide susceptibility is conducted using machine learning algorithms in the predictive models include Linear Regression, Random Forest, and XGBoost. These models literally calculate the associations between the input parameters like slope, rain and geology and previously reported landslides. The measures of performance use precision to establish accuracy of positive predictions, recall, F1-score, RSME, MSE and MAE which show the model's potential to detect the real landslides. Receiver Operating Characteristic (ROC) graphs and the associated area under the curve (AUC) also supports model efficiency when differentiating between high susceptibility of landslides and low susceptibility zones. It enables more accurate results and confidence of the identified models in the assessment of landslide risks.

5. Data Visualization

The results of the modeling are displayed in the form of the landslide susceptibility maps where regions with high, moderate and low susceptibility to landslides are easily identified. These maps categorise the regions into high, moderate and low risk areas and are decisive for the stakeholders. Instead, envisioning the idea of a spatial narrative, these maps provide a clear picture of the areas under landslide risk so that effective changes like infrastructural developments and protection procedures can be implemented in the risky zones. Through visualization, large datasets are easy to understand and utilize by professionals in different fields technical or otherwise.

6. Data Analysis and Monitoring

The final stage aimed on the assessment of the effectiveness of and the potential for employing the produced landslide susceptibility maps. This consists of plotting the maps against reports of the recent occurrence of landslides to establish the extent of predictive accuracies of the maps. Constant supervision of areas that were identified as high-risk enable the authorities to address possible hazard in future. Further, the chance of model updating with new data provides needed flexibility in changing environmental conditions. This cyclical process improves the sustainability of the methodology for evaluating risk of landslides and managing it in long term.

2.3 Data and Tools

i. Geological Information System (GIS)

GIS focuses on making full use of geographic data (data input and output, analysis, storage, and updating) beyond the limitations of traditional paper maps. Geographic and attribute data are sometimes referred to as spatial data. Geographical data refers to topographic feature distribution, positions, and configurations (elevations, rivers, etc) as well as characteristics of human activities and social environments (buildings, land use, vegetation, and population). A GIS provides a few functionalities that assist users in making decisions and analysing geographical data. Visualization employing selective overlay of spatial data or legends, statistical processing using spatial analysis and buffering of chosen information are examples of such functions. GIS plays an essential role in disaster information management since it reduces equipment investment and improves information sharing.

The primary dataset for this study comes straight from secondary data comes from published materials such as maps, photos, and aerial photographs. The scanned base map served as a guide for digitising the Arc Info Geological Information System on-screen (ArcGIS). All of this information was recorded as a shape file in the database and then transformed to raster files using the ArcGIS software. GIS data sets may be evaluated and updated frequently. TABLE 3.1 lists the GIS data sets developed for this study. The digital thematic maps were produced including geology map, slope gradient map, slope aspect map, topographic wetness index map, landslide susceptibility map, etc.

Table 3.1 : list of the GIS data sets developed for this study

NAME	TYPES	SOURCE
DEM (toporaster)	Raster	Generated from USGS GIS data
Contour	Shapefile	Generated from DEM
Hydrology (Drainage / River)	Shapefile	Generated from DEM
Geology	Shapefile	Digitized from Geological maps of Jabatan Mineralogi dan Geoscience
Relict landslides	Shapefile	Digitized from aerial photos and satellite Imagery of NASA Global Landslides Catalogue
Road	Shapefile	Open Street Map
Elevation	Shapefile	SRTM.tif
Slope gradient	Raster	SRTM.tif
Slope curvature	Raster	SRTM.tif
Slope aspect	Raster	SRTM.tif
Distance to lineament	Shapefile	SRTM.tif
Lineament	Shapefile	Digitized from satellite imagery and shaded relief map
Terrain Roughness	Shapefile	SRTM.tif
Topographic Wetness Index (TWI)	Raster	SRTM.tif
Topographic Ruggedness Index (TRI)	Raster	SRTM.tif
Normalized Difference Vegetation Index (NDVI)	Raster	LANDSAT8
Stream Order	Shapefile	SRTM.tif
Distance to Drainage	Shapefile	Digitized from stream shapefile
Drainage Density	Shapefile	Digitized from stream shapefile
Distance to Road	Shapefile	Digitized from road shapefile
Road Density	Shapefile	Digitized from road shapefile

ii. R- studio

R Studio is a powerful statistical programming environment in which tools and libraries exist to construct machine learning models including the Random Forest Regression, Linear Regression and XGBOOST to analyze the susceptibility of landslides. Random Forest (RF) models are primarily delivered through the 'randomForest' package aimed at training RF models for large geospatial datasets. This package enables the researcher to combine many decision trees and therefore improve the model by decreasing the rate of overfitting. Bravo-López et al (2023) integrates RF in R Studio to produce landslide susceptibility maps based on diverse conditioning factors of rainfall, slope gradient and land use. Another advantage with the use of R Studio is that one can handle spatial data through libraries such as `sp" and `raster` hence making it easier to incorporate RF into geospatial applications.

Linear Regression, a conventional approach to statistical modeling is also encouraged significantly in R Studio using functions including the *'lm()'*. Although Linear Regression is one of the basic models it serves purpose during initial data exploration and when comparing simple model with more complex ones. In the objective of landslide susceptibility analysis, R Studio allows using R tools to add multivariate regression tests that estimate the effect of geology and environment factors in the susceptibility models. Linear regression, using R Studio software, was used as a benchmark model to run and compare its accuracy with more complex artificial models and machine learning precisely due to its inability to accommodate multiple nonlinear interaction that are simultaneously present in the process of landslide formation.

The extreme gradient boosting technique or XGBOOST is one of the sophisticated machine learning algorithms installed from the R Studio through the `*xgboost*' package. This package facilitates control of hyperparameters such as the learning rate and the maximum depth of a tree which is so important when improving the model. Hasanuzzaman, and Shit (2024) have used the XGBOOST in R Studio in combining numerous datasets and recognizing various significant variables causing landslides. Due to the capability of parallel processing the `*xgboost*` that is available in the R Studio enhances capability of handling large data and therefore is effective in high-resolution susceptibility mapping. Hussain et al. (2023) used this integration in a work involving landslide susceptibility in Pakistan and showed that combining their RF and XGBOOST models produced better results. The graphics libraries including the *`ggplot2'* by R studio also aids in interpretability by making it easier for the researcher to present the created maps depicting susceptibility and features of importance.

3.4 Performance Measure

3.4.1 Machine Learning

i. **Linear Regression:** Multiple linear regression analysis is used to determine the pattern of association between the landslide probability and potential predictors. In this study, ordinary least absolute deviation regression is employed rather than a single measure since probability of landslide has large variability. The linear function is expressed in equation 3.1.

$$Y=a+b_1X_1+b_2X_2+...+b_nX_n \qquad \textbf{(eq 3.1)}$$

where, Y is the dependent variable (landslide probability in our case) , X_1 represents independent variables (all predictors), a is the constant and b_1 is the regression coefficient of the variable X_1.

ii. **Random Forest Regression**: Random Forest Regression is an ensemble model based on decision trees, at the creation of which bootstrap samples of training data are used. Bagging (bootstrap aggregation) works to reduce variance and over training thus improving the models generalization capabilities. The function is expressed in **equation 3.2.**

(eq 3.2)

$$\hat{y} = \frac{1}{T}\sum_{k=1}^{K} f_t(x)$$

where; K is number of trees, $f_t(x)$ is prediction of the *t*-th tree, and $\hat{y}$ is final prediction. This method effectively balances bias and variance, delivering robust predictions for landslide susceptibility.

iii. **XGBoost Regression:** XGBoost is an advanced model that develops an instance of gradient boosting decision trees in an iterative form to take the best value of the

(eq 3.3)

loss function. Each new tree opens errors of the previous trees to ensure accurate predictions on any given data set. The objective function includes a training loss term and a regularization term to control model complexity:

$$Obj = - \sum_{i=1}^{n} \text{loss}\,(y_i - \hat{y}_i) + \sum_{k-1}^{K} \Omega\,(f_k)$$

The final prediction of this combination is calculated using the equation 3.4:

$$Y_{xgb}\,(x) = \sum_{k-1}^{K} tree_k\,(x), tree_k \in T \qquad \text{(eq 3.4)}$$

3.4.2 Performance Metrics:

i. RMSE, MSE and MAE

The performance of the models is determined and compared to mean measures including Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE). These metrics quantify the differences between predicted and actual values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \text{(eq 3.5)}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{(eq 3.6)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(|y_i - \hat{y}_i|)^2 \qquad \text{(eq 3.7)}$$

Here, y_i denotes the expected value and $\hat{y}_i$ is the predicted value. These metrics provide insights into model accuracy and error distribution.

ii. ROC and AUC

Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) are used to assess the models' classification performance. The AUC quantifies the ability of a model to distinguish between positive (landslide-prone) and negative (non-landslide) areas, with values ranging from 0.5 (random prediction) to 1 (perfect prediction):can be calculated by the integral trapezoidal rule. The equation is written as follows:

AUC (Area Under Curve): Measures classification performance on ROC curve.

$$AUC = \frac{(\sum TP + \sum TN)}{(P + N)} \qquad \text{(eq 3.8)}$$

where TP (true) and TN (true negative) denote the correctly classified raster cells, P expresses the total number of landslide raster cells, and N represents the total number of non-landslide raster cells.

iii. *Confusion Matrix*

In this study, the performance of the evaluation model for landslide susceptibility was calculated with the confusion matrix which is a popular method used in binary classification for model assessment. Three statistical measurements including precision, recall, accuracy and F1- Score were used to assess the efficacy of the specific model. It expressed as follows:

- **Precision**: Proportion of correctly identified positive cases. (eq 3.9.1)

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**: Sensitivity or True Positive Rate.

$$Recall = \frac{TP}{TP + FN} \qquad \text{(eq 3.9.2)}$$

- **Accuracy:**

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{(eq 3.9.3)}$$

- **F-1 Score :**

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad \text{(eq 3.9.4)}$$

where TP is for the correct predicted number of landslide raster, TN is the correct predicted number of non-landslide raster, FP is the number of incorrectly predicted landslide raster and FN is the number of wrong predicted non-landslide raster cells.

In order to calculate these indices, concepts of four kinds of predicted samples for classification learning need to be clarified: Detection accuracy is the ratio of TP+TN to TP+TN+FP+FN and is defined by the four ways the test can be classified: (1) true positive (TP): the patient has the disease and is predicted as positive; (2) false positive (FP): the patient has no disease and is predicted as positive; (3) true negative (TN): the patient has no disease while is predicted as negative; and the prediction disagrees with the actual class.

2.5 Gantt Chart

Task	Months											
	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept
1. Desk Study	X	X										
2. Proposal Development												
2.1 Research Question and Objective		X	X									
2.2 Literature Review			X									
2.3 Proposal Submission and Approval				X	X							
Project 1 :												
3. Data Collection												
3.1 GIS Data Acquisition					X							
3.2 Historical Landslide Data						X						
3.3 Field Surveys							X					
4. Data Processing												
4.1 Data Cleaning							X					
4.2 GIS Layer Preparation							X					
4.3 Feature Selection							X					
5. Model Development												
5.1 Algorithm Selection								X				
5.2 Training Machine Learning Model								X				
5.3 Model Validation and Testing								X	X			
Project 2:												
6. **Results Analysis**												
6.1 Performance Metrics Analysis									X	X		
6.2 Visualization of Results									X	X		
7. Thesis Writing	X	X	X	X	X	X	X	X	X	X		
8. Thesis Presentation and Defense											X	
9. Thesis Submission											X	
10. Thesis Revision and Final Submission												X