A Study on Performance Comparisons between KNN, Random Forest and XGBoost in Prediction of Landslide Susceptibility in Kota Kinabalu, Malaysia

Dorothy Martin
Faculty of Computer Science and Information Technology
University of Malaysia Sarawak
Kota Samarahan,Malaysia
dorothymartinatok@gmail.com

Soo See Chai
Faculty of Computer Science and Information Technology
University of Malaysia Sarawak
Kota Samarahan,Malaysia
sschai@unimas.my

2022 IEEE 13th Control and System Graduate Research Colloquium (ICSGRC) | 978-1-6654-6806-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSGRC55096.2022.9845146

***Abstract* – One of the most natural catastrophes in Malaysia, landslides, has resulted in several fatalities, infrastructure damage and economic losses. Over time, researchers have used various methods to forecast the vulnerability to landslides. Unfortunately, the most accurate algorithm which can be used to develop a landslide susceptibility model is still lacking. Therefore, the current study aims to evaluate how well Kota Kinabalu, Sabah's landslide susceptibility, can be predicted using three different machine learning techniques: K-Nearest Neighbor (KNN), Random Forest, and Extreme Gradient Boosting (XGBoost). The research areas had 242 landslide locations, and the inventory data was arbitrarily separated into training and testing datasets in a 7/3 ratio. As prediction parameters, ten spatial databases of landslides conditioning factors were employed. The area under the curve (AUC) was utilized as the models' performance metric. With an AUC score of 87.52 %, the final analysis showed that KNN had the highest prediction accuracy, followed by Random Forest (84.34 %) and XGBoost (78.07%). According to the AUC findings, KNN, Random Forest, and XGBoost performed consistently well in forecasting landslide susceptibility. The final forecast map can be a helpful tool for urban planning and development and for aiding the authorities in creating a strategic mitigation plan.**

Keywords— Landslides susceptibility map, K-Nearest Neighbors(KNN),Random Forest, Extreme Gradient Boosting(XGBoost)

I. INTRODUCTION

Landslide is one of major natural disasters that frequently occur in Malaysia. Major disasters have a big influence on the economy. They disrupt livelihoods, cause human deaths, and damage to property and infrastructure. Malaysia usually experience landslides due to massive rainfall events as well as rapid deforestation which exposed the soil to erosion [1].Since, landslides cannot be avoided from occurring, numerous researchers have conducted research on landslide susceptibility in attempt to reduce the impact of landslides. These models can be grouped into three categories namely statistical, soft computing, and analytic approaches.

In Malaysia, past researches had conducted study regarding the landslide susceptibility mapping by using various modelling approaches such as Factor Analysis Model (FAM) [2], Analytical Hierarchy Process(AHP) [3], Probability-Frequency Ratio Model [4], Adaptive Neuro-Fuzzy Inference System (ANFIS) [5], [6], logistic regression [7] , Decision Trees and Support Vector Machine (SVM) [6]. Analytic methods are difficult to apply when the subject area is large. As a result, the employment of statistical and soft computing methods has gradually expanded. Furthermore, these approaches are quite simple to implement in a geographic information system (GIS).

Numerous studies have moved their focus to machine learning in landslide susceptibility prediction during the past ten years as a result of significant advances in computing. The complex relationship between landslide proneness and factors can be tackled using machine learning, a quantitative approach. Each machine learning model has its limitations and strengths, and the features of different research areas influence its behavioral patterns. Therefore, access to machine learning for landslide susceptibility map comparison is highly desirable.

Three machine learning methods, K-Nearest Neighbors (KNN), Random Forest, and Extreme Gradient Boosting (XGBoost), were used in the current study. These models were implemented for various purposes, including that they had never been used to predict landslides in Malaysia. The three machine learnings can also use remote sensing data rather than intensive field surveys.

Random Forest and Extreme Gradient Boosting (XGBoost) have outperformed single learning algorithms. The current study project will use Random Forest and XGBoost due to their reliable performance. Similarly, KNN was chosen because of its ease of implementation, fast computation time, and the output is easy to be interpreted. In previous studies, the KNN had shown to be a reliable predictor of landslides [8], [9].

The selection and preparation of the database of landslide conditioning factors is a crucial stage in achieving the high accuracy of the landslide susceptibility model in predicting landslide-sensitive areas. In the current study, the landslide conditioning factors were chosen based on data gathered from works of literature connected to the subject location, such as [8]–[10]. The ten landslides conditioning factors emphasized on this study areas are Digital Elevation Model (DEM), slope length, slope angle, distance from road and stream, normalized vegetation index , profile and plan curvatures, topographic wetness index and stream power index.

For the local government and town planners, the expected landslide susceptibility prediction results will be beneficial for planning better mitigation measures to deal with landslide events and identifying favorable places for future development. A good and reliable landslide susceptibility prediction model and map can be accomplished by integrating machine learning and geographic information systems (GIS).