Proc. of the 9th International Symposium on Geotechnical Safety and Risk (ISGSR)
Edited by Zhongqiang Liu, Jian Dai and Kate Robinson
Published by Research Publishing, Singapore.
doi: 10.3850/GRF-25280825_isgsr-066-P036-cd

UNPRECEDENTED BREAKTHROUGH OF LANDSLIP WARNING SYSTEM IN HONG KONG: REAL-TIME, DATA-DRIVEN AND PERFORMANCE-BASED

Raymond W.M. Cheung[1], Florence W.Y. Ko[2], Edward K.H. Chu[3], D.S. Chang[4]

Geotechnical Engineering Office, Civil Engineering and Development Department, Government of HKSAR, Hong Kong SAR, China.
[1]E-mail: wmcheung@cedd.gov.hk;[2]E-mail: florenceko@cedd.gov.hk;
[3]E-mail: edwardkhchu@cedd.gov.hk;[4]E-mail: dchang@cedd.gov.hk

In managing the landslide risk in Hong Kong, the Geotechnical Engineering Office (GEO) of the Civil Engineering and Development Department has progressively established and maintained high quality inventories of territory-wide landslide-related datasets since the 1980s. Over the past forty-five years, the GEO has used these high-resolution spatio-temporal data to support the technical development of a landslide prediction model as part of the Landslip Warning System in Hong Kong. Amongst others, data-driven analyses using a conventional statistical approach have been pursued to establish rainfall-landslide correlations for man-made slopes. Recently, the GEO has explored the potential application of machine learning (ML) and big data analytics, using datasets from 1996 to 2023, for landslide prediction in Hong Kong.Several common ML algorithms such as XGBoost, Logistic Regression, and Neural Network, are being tested to establish the multivariate and non-linear correlation among a wide range ofpertinent features and the occurrence of landslides on man-made slopes. Domain knowledge of geotechnical and geological engineering was incorporated in the course of developing the ML model. This paper presents the modelling approach and workflow using the XGBoost algorithm through data pre-processing, algorithm selection, feature selection, model training and evaluation. The results indicate apromising predictive performance of the XGBoost model against various evaluation metrics compared with the conventional statistical model, and draw insight into contributing factors of landslide occurrence in Hong Kong.

Keywords: Machine learning, landslide prediction, rainfall-landslide correlation, landslip warning.

1. Introduction

Hong Kong is characterized by its mountainous landscape blanketing underneath a thick weathering profile and subtropical climate, which are conducive to frequent landslides threatening urban developed areas, disrupting infrastructure and endangering lives. The Geotechnical Engineering Office (GEO) of the Civil Engineering and Development Department has established and implemented a territory-wide Landslip Warning System (LWS), as a key component of the landslide risk management tools in the Slope Safety System in Hong Kong. It aims to provide timely alerts to the public and facilitate emergency responses with government departments during severe rainstorms. Since the 1980s, the GEO has maintainedextensive, high qualitydatabases of rainfall, man-made slopes, lithologyand reported landslides. These high-resolution spatio-temporal data have supported the development of landslide prediction models as part of Hong Kong's LWS, which traditionally uses statistical methods to establish correlations between rainfall and landslides on man-made slopes. Recently, the GEO has explored the application of machine learning (ML) and big data analytics, leveraging data from 1996 to 2023, to enhance capabilities of landslide prediction for man-made slopes. The study also considers alternative methods for characterizing rainfall severity and identifying additional factors contributing to landslide occurrence.

2. Relevant Databases in Hong Kong

The success of ML applications heavily depends on high-resolution spatio-temporal data of excellent quality. The GEO's databasesrelated to rainfall, man-made slopes, lithology and reported landslides, upon data preprocessing and cleansing, provide a solid foundation for evaluating the relationship between rainfall and landslides and developing a ML model for landslide prediction.

2.1*Rainfall data*

Based on theavailable rainfall data of over 120 raingauges at a 5-minute interval, 384 rainstorm events are identified from 1996 to 2023, mostly occurred between April and October, and most of the rainstorm periods range from one to three days. This data period provides the most reliable dataset in terms of data abundancy and quality. For each rainstorm event, spatial distributions of maximum rolling rainfall and antecedent rainfallare generatedand the rainfall-relatedfeatures for each registered man-made slopeare determined based on its location.

2.2 *Catalogue of Slopes*

Sizeable man-made slopes in Hong Kong are registered in a comprehensive Catalogue of Slopes. It contains a wide range of useful information of the man-made slopes, such as location, slope type, geometry, forming material,level of geotechnical input, maintenance responsibility and photographs. Currently, the Catalogue of Slopes contains about 60,000 registered man-made slopes, of which about one-third are owned by private parties and the remaining are government slopes.

2.3 *Geological maps*

In this ML study, the lithology of the Hong Kong territory is categorized into intrusive, volcanic and sedimentary groups based on the Geological Map of Hong Kong in 1:100 000 scale(GEO, 2000). The lithology classification map is applied to determine the lithology of each registered man-made slope, as a supplementary slope information.

2.4*Reported landslides*

The GEO has maintaineda computerized database of about 10,000 reported landslide records to date since 1984, comprising the information of location, failure date, reporting date, failure volume and affected facilities. There are 2,696 reported landslides on registered man-made slopes associated with the 384 rainstorm events from 1996 to 2023. As the information in the Catalogue of Slopes has been continuously updated, the slope-related features for each registered man-made slopeat the time of each rainstorm event can be determined.

3. Machine Learning-based Landslide Prediction for Man-made Slopes

3.1*Study objectives*

In this study, the GEO explores the ML application and big data analytics to re-examine the relationship between rainfall and landslides on man-made slopesbased on the available data from 1996 to 2023, with an aim to developing a ML model for prediction of the number of landslides on man-made slopes of a rainstorm event in Hong Kong. The rainstorm events from 1996 to 2023 are selected for this study as the landslide data in that period is more representative of the current slope safety system implemented since 1995.

The current LWS 4.0 model (Yu, 2004; Chung et al. 2023) adopts maximum rolling 24-hour rainfall as a key parameter, however, experiences suggest that other parameters, such as maximum rolling 4-hour rainfall and rainstorm period could also be prominent factors contributing to the occurrence of landslides. ML tools shed light on the relevance of various maximum rolling rainfalls and antecedent rainfalls to identify better predictors. Additionally, integrating more detailed slope information, such as slope angle and level of geotechnical input, into the ML model could enhance prediction accuracy. The high-quality data inventories maintained since the 1980s are instrumental in this endeavor, while domain knowledge-basedfeature selection and model evaluation are necessary to ensure the physical relevance, reliability and accuracy of the selected features.

3.2*Modelling approach*

The ML modelis a slope-based binary classification analysis, involving data pre-processing, feature selection, algorithm selection, model training and evaluation, to give a predicted probability of a landslide on a man-made slope under a rainstorm event. As landslidesarerare events as compared to non-landslide events, the approach should be able to handle highly imbalanced data of reported landslides during rainstorm events, in which the data imbalance ratio of this modelis about 1:8,200.No data resampling is conducted, so the predicted probability of a landslide on a man-made slope can be directly taken as the probability of the positive class of an individual slope.

Figure 1 shows the overall workflow, data resampling, performance evaluation and model applicationof the study. Thepre-processed dataset of the 384 rainstorm events from 1996 to 2023 is split into a training dataset and 2 testing datasets. The rainstorm events in 1997 and 2016 are extracted to form Testing Dataset 1, which comprises about 10% of rainstorm events in the whole dataset, with a similar data imbalanced ratio.Another testing dataset, namely Testing Dataset 2, is formed by randomly extracting 10% data points from the remaining rainstorm events in a stratified manner. Its ratio of positive to negative cases remains the same as that of the whole dataset. After the extraction of the two testing datasets, the remaining data points form the training dataset. The ML model is trained with the training dataset, and hyper-parameters of the machine learning algorithm are tuned using five-fold cross validation.Given the real-time rainfall data collected by over 120 automatic raingeuges of the GEO Raingauge System, the trained ML model can perform a real-time landslide prediction for man-made slopes to give aprediction of the spatio-temporal distribution of landslidesduring a rainstorm event.

3.3*Algorithm selection*

The study references recentliterature and similar machine learning studies (Tehrani et al. 2022; Xiao et al. 2022), acknowledging that ML algorithm performance varies across different applications, and algorithm selection should consider factors including algorithm interpretability, balance between bias and variance, suitability for handling correlated features, and computational efficiency. XGBoost algorithm (Chen and Guestrin, 2016), a tree-based method known for strong predictive capabilities and high interpretability in similar ML applications of landslide prediction, is chosen as an initial adoption for model training. AnXGBoost model is developed, utilisingthe pre-processed datasetof the 384 rainstorm events, which is resampled for training, validation and testing, to explore the multivariate and non-linear correlations between features and landslides on man-made slopes.

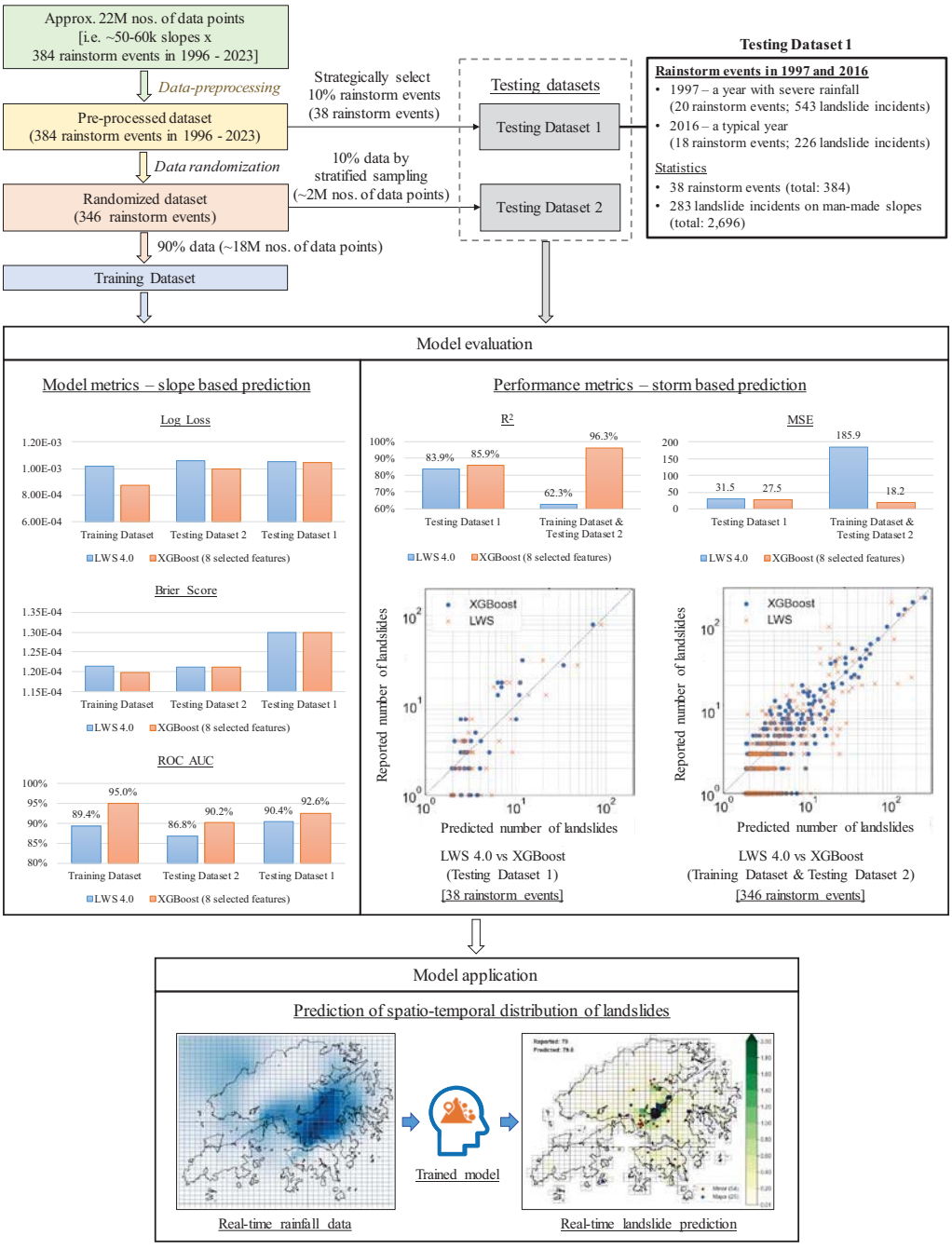

Fig. 1. Overall workflow, data resampling, performance evaluation and model application of the study

3.4*Feature selection*

ML is a powerful tool for developing prediction models with optimalaccuracy. However, our domain knowledge and experiences on landslide process, as well as engineering judgement, should be integratedinto the analysis in

order to build a physically meaningful model. Selection of features pertinent to prediction modelling for landslides on man-made slopes is therefore crucial. The study evaluates 21features based on their statistical significance and physical relevance, considering factors such as maximum rolling rainfall, antecedent rainfall, rainstorm period, slope type, slope and wall geometry, lithology and level of geotechnical input. Aligning with domain knowledge of geotechnical and geological engineering on landslide occurrence supported by ananalysis ofstatistical significance, eight features that are highly influential in landslide occurrence are therefore selected to develop the XGBoost model. These features include 4-hour and 24-hour maximum rolling rainfalls, slope type, rainstorm period, slope angle, level of geotechnical input, slope forming material and 7-day antecedent rainfall.

3.5*Model evaluationand performance*

To ensure a robust evaluation of model accuracy and reliability, an assessment of model metrics, including log loss, Brier Score and Area under Receiver Operating Characteristic Curve (ROC AUC),is employed to compare the XGBoostmodel's performance against the LWS 4.0 model's using the two testing datasets.In addition, as Testing Dataset 1possesses data points of the rainstorm events in 1997 and 2016, a predicted number of landslides for each rainstorm event can be calculated by summing the predicted landslide probability of all slopes of the rainstorm event. Performance metrics, including R^2 and Mean Squared Error, can be calculated to compare the predicted numbers of landslideswith the actual ones for the rainstorm events.

The predictive performance of the XGBoost model against the various evaluation metrics is promising compared to the LWS 4.0 model, as presented in **Fig.1**. Using the two testing datasets, the XGBoost model consistently gives better model metrics, suggesting its accuracy and reliability in overall performance. In particular, the XGBoost model demonstrates superior accuracy in both performance metrics for predicting the number of landslides of each rainstorm event using Testing Dataset 1. Furthermore, the predicted numbers of landslides are plotted against the reported number of landslides for the rainstorm events in **Fig.1**. A point closer to the diagonal line in the plot means a better prediction. The predicted numbers using XGBoost'smodelare generally closer to the actual, compared with the LWS 4.0 model.

The above findings demonstrate the XGBoost model's ability to handle non-linear relationships and incorporate multiple variables for a more nuanced prediction of landslides. It is worth noting that, although the same databases are used for developing the LWS 4.0 and XGBoost models, the XGBoost's performance metrics obtained using Training Dataset and Testing Dataset 2are significantly better than those of LWS 4.0 model. It may imply that the three features considered in the LWS 4.0 model(i.e. maximum rolling 24-hour rainfall, slope type and forming material) may not be sufficient to build a performing model.

3.6*Insight into Contributing Factors*

The study provides insight into the contributing factors of landslide occurrence, showing the relative feature importance of the trained XGBoost model. The analysis revealed that while maximum rolling rainfall and slope type remain primary triggers aligning with our domain knowledge and consistent with the LWS 4.0 model, other factors such as rainstorm period, slope angle and level of geotechnical input also play crucial roles. It highlights the importance of considering multiple rainfall and slope characteristics that offer a deeper understanding of the complex interactions leading to landslides on man-made slopes. The insight underscores the value of a multivariate approach in enhancing prediction accuracy and informing risk management strategies. By identifying the most influential factors, the GEO will employ a ML model to better anticipate landslide events for the LWS operation.

4. Conclusion

The GEO has stepped into the application of ML and big data analytics to enhance the capability to predict the number of landslides on man-made slopes during a rainstorm in Hong Kong. Integrating the ML techniques into the landslide prediction model coupled with the relevant domain knowledge shows improved prediction accuracy and reliability of the XGBoostmodel, which enables a real-time prediction ofthe spatio-temporal distribution of landslides during a rainstorm event. It also offers valuable insight into the factors contributing to landslide occurrence, suggesting an excellent potential of the ML application in uncovering hidden knowledge in landslide triggers and enhancing landslide prediction ability for the LWS operation, so as to elevate public safety and resilience against future challenges on extreme rainfall events.

Acknowledgements

This paper is published with the permission of the Head of the Geotechnical Engineering Office and the Director of Civil Engineering and Development, the Government of the Hong Kong Special Administrative Region, China.

References

1. Chen, T. andGuestrin, C. (2016). "XGBoost: A scalable tree boosting system."*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA: 785-794.
2. Chung, P.W.K., So, S.T.C. and Chu, E.K.H. (2023). Landslip warning system in Hong Kong - over 40 years of evolution.*Current Chinese Science*, 2023, 3, 123-140.
3. GEO (2000). *The Quaternary Geology of Hong Kong*.Geotechnical Engineering Office, Hong Kong, 210 p.
4. Tehrani, F.S.,Calvello, M., Liu, Z.Q., Zhang, L.M.andLacasse, S. (2022). "Machine learning and landslide studies: recent advances and applications." *Natural Hazards*, 114: 1197-1245.
5. Xiao, T., Zhang, L.M., Cheung, R.W.M. and Lacanne, L. (2022) "Predicting spatio-temporal man-made slope failures induced by rainfall in Hong Kong using machine learning techniques." *Géotechnique*, 73(9): 749-765.
6. Yu, Y.F.(2004)."Correlations between rainfall, landslide frequency and slope information for registered man-made slopes."*GEO Report No. 144*. Geotechnical Engineering Office, Hong Kong, 109 p.