Landslide identification in Hong Kong using generative adversarial networks based on inventory data

Xiaoyi Wu[a]*, and Anran Zheng[b]

[a] Stuart Weitzman School of Design, University of Pennsylvania, Philadelphia, United States
[b] Herbert Wertheim College of Engineering, University of Florida, Gainesville, United States
* xiaoyiwu@upenn.alumni.edu

Keywords: landslide identification, generative adversarial networks, machine learning, deep learning, GIS, Hong Kong

Introduction

Landslides are one of the most destructive geological hazards, which happen frequently at mountainous places with intensive rainfall events and active geological activities, disturbing human activities. Consequently, accurate land slide prediction and susceptibility analysis become necessary and critical to mitigate loss of resources and human life. In recent years, machine learning algorithms have been used in landslide susceptibility mapping but these models still need to be assessed and compared to enhance their performance for landslide prediction. The scarcity of geological hazard data remains one of the challenges for these data-driven approaches. The imbalanced distribution of datasets usually leads to poor performance in landslide prediction. To correct imbalanced landslide datasets used in mapping landslide susceptibility, this study applies a new proposed approach, based on the generative adversarial network (GAN), to synthesize landslide inventory data and increase prediction capability in the minority class.

Method

In this research, landslide inventory data of the most recent 10 years was collected for Hong Kong, China, one of the most frequent landslide prone regions. Based on previous studies, multiple geo-environmental factors, including elevation, slope, aspect, curvature, vegetation density, land use and land cover (LULC), were integrated. The whole dataset was split into a training dataset and testing dataset with a ratio of 30/70. To verify the capability of GAN, the proposed model was compared with traditional machine learning models that have been commonly used in landslide susceptibility mapping, including K-nearest neighbours (KNN), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR). KNN, RF, DT and LR with original training data were used as the benchmark. Our proposed model was trained using KNN, RF, DT and LR with new data generated from the original data by GAN. Model performance was evaluated using the overall accuracy (OA), under the receiver operating characteristic curve (AUROC), confusion matrix and F1-score. After the models were compared and measured, the maps of landslide susceptibility in different categories with proposed methods were drawn to analyse the spatial distribution.

Results

The model performance is summarised in Table 1. Class 1 is the channelized debris flow, Class 2 is open hillslope landslide, and Class 3 is the coastal landslide. The OA remained similar as the benchmark model, as OA of original KNN, RF, DT and LR model could achieve the OA values of (0.72, 0.77, 0.76, 0.73), while (0.71, 0.75, 0.70, 0.68) after adding synthetic data generated from GAN. However, the F1-score in the minority class (Class 1) were improved significantly in models with GAN. F1-scores of models with GAN in Class 1, which covers 30% of the whole dataset, are (0.39, 0.58. 0.60, 0.38), while (0, 0.57, 0.53, 0.55) without GAN. The prediction performance in Class 2 is low due to extreme data scarcity. Class 3 only covers 2% of the whole dataset.

Model		Accuracy	F1 score on Y label		
			Class 1	Class 2	Class 3
Propose model with GAN	KNN	0.71	0.39	0.82	0.00
	RF	0.74	0.58	0.82	0.00
	DT	0.70	0.61	0.77	0.00
	LR	0.68	0.39	0.79	0.00
Benchmark model	KNN	0.72	0.00	0.46	0.81
	RF	0.77	0.58	0.85	0.53
	DT	0.76	0.53	0.82	0.50
	LR	0.73	0.53	0.82	0.50

Table 1: Model Performance

Landslide susceptibility mapping of four machine learning models with GAN was generated from the study area. According to the Figure 1, the proposed GAN model captures the spatial distribution pattern of landslide accidents in different categories. Open hillslope landslide, which accounts for most types of landslides in Hong Kong in recent years, tend to happen in high-elevation mountainous areas in the northwestern New Territories, western Lantau Island, and western Hong Kong Island. While channelized debris flows, which are destructive landslides with high velocities and large volume, tend to happen at the northwestern New Territories. Coastal landslides are distributed along the south shore.

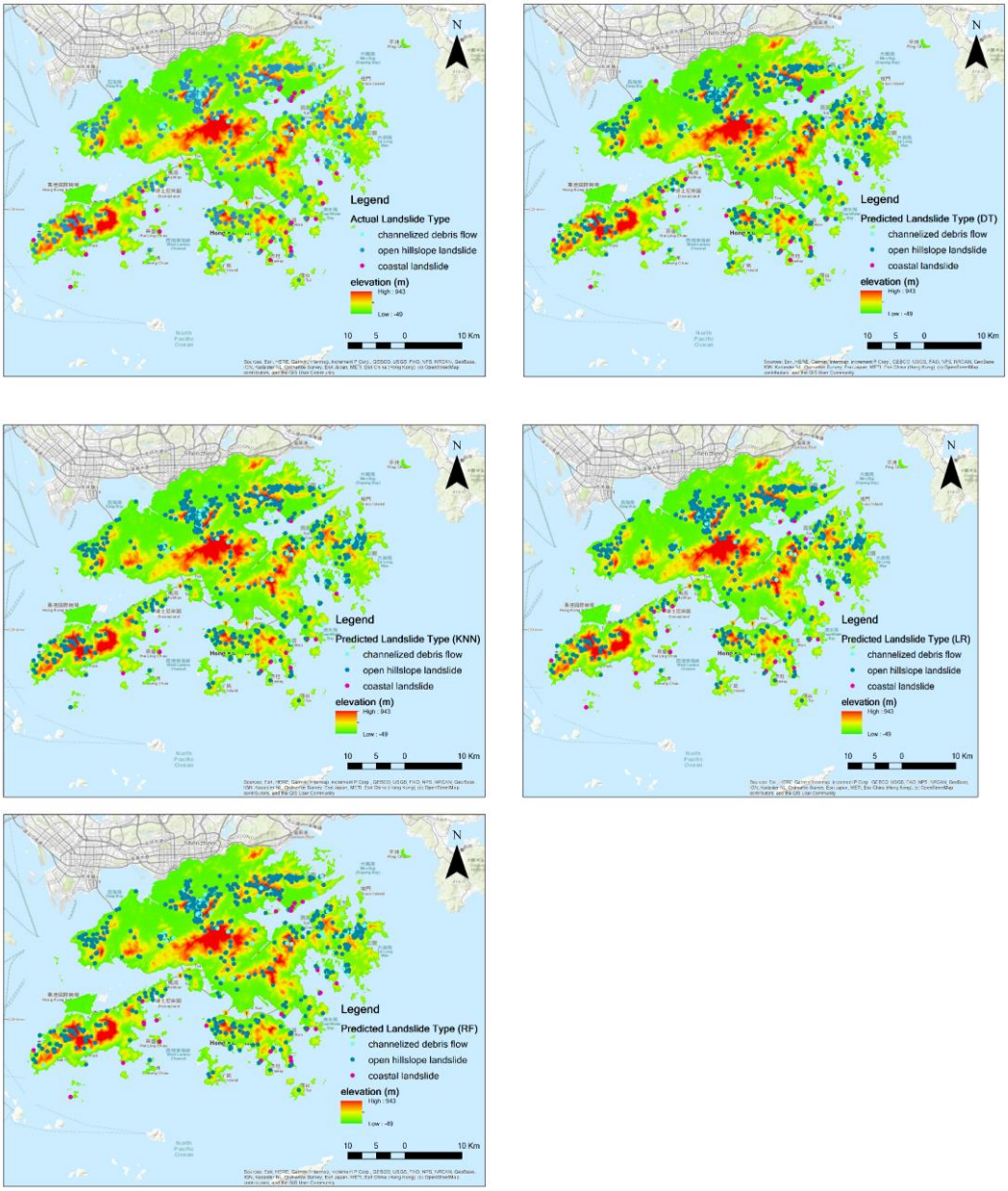

Figure 1: comparison of actual and predicted four machine learning models (DT, KNN, LR, RF) with GAN synthetic data

Discussion and Conclusion

Landslide susceptibility analysis is important for hazard risk prevention in dense urban development cities such as Hong Kong. Machine learning has been widely used to improve the accuracy of landslide predictions. However, data imbalance remains one main challenge in these data-driven methods. This study applied a new proposed approach, which was based on the generative adversarial network (GAN), to generate synthetic inventory data that makes up the limitations of the original dataset. In this research, landslide inventory data was identified in Hong Kong, China, a region with one of the most frequent landslides as well as dense city structure. The proposed model with synthetic data by GAN was compared with four traditional machine learning models. The results reveal that GAN did not increase the accuracies in landslide

susceptibility mapping in Hong Kong Island with available datasets, but it improved the predictive capability in extremely hazardous and scarce landslide category. Future studies could consider utilizing GAN as landslide prediction models by adding more significant conditioning factors.

Acknowledgements:

We appreciate suggestions and encouragements by Prof Tao Wang and Dr Julia Koschinsky to submit this work to AutoCarto 2022.