

MACHINE LEARNING-BASED APPROACH FOR LANDSLIDE SUSCEPTIBILITY MAPPING USING MULTIMODAL DATA

Xianping Ma[†], Man-On Pun[†], Ming Liu[‡] and Yang Wang[#]

[†]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China

[‡]MizarVision, Shanghai, China

[#]Xi'an Leitong Technology Co. Ltd, Xi'an, China

ABSTRACT

In this work, a practical Machine Learning (ML) approach is proposed to produce Landslide Susceptibility Mapping (LSM) by exploiting multiple data sources. In contrast to conventional ML-based methods that consider a large number of factors, the proposed method fuses only three carefully selected factors, namely inventory Landslide and Debris Flow (LDF) data, rainfall data and forest change information to model and predict landslides and debris flow. More specifically, a Global Deforestation Detection Algorithm (GDDA) based on Synthetic Aperture Radar (SAR) and Convolutional Neural Network (CNN) are first developed to generate a hazard map by identifying areas of substantial forest changes. Capitalizing on rainfall forecast data derived from the Doppler Weather Radar (DWR) over the areas of severe deforestation detected by GDDA, an SVM-based classifier is established to predict the likelihood of the occurrence of landslide and debris flow. Using two real-world landslide and debris flows in 2021, we demonstrate that the proposed LSM system was able to provide early warning in advance.

1. INTRODUCTION

Landslide is one of the most hazardous geological disasters due to its devastating destructive power. Loose materials and runoff formed by landslides can inflict large debris flows. To mitigate the damages incurred by landslide, it is critical to accurately characterize the likelihood of landslide occurrence for a given area and subsequently, forecast the level of landslide hazard. To achieve this goal, LSM was recently proposed [1]. In LSM, the spatial distribution and occurrence probability of existing or potential landslides are analyzed by considering the Landslide Influencing Factors (LIF), including altitude, aspect, profile curvature, slope, vegetation cover, topographic wetness index, distance to faults/rivers/roads, soil, rainfall and land-use. More specifically, LSM can be generated quantitatively by exploiting the physical or statistical models. In general, the physical-model approach is con-

sidered more accurate as it models and simulates the physical formation process of landslides [1]. However, the physical-model approach requires detailed geological field surveys, which makes it difficult to develop large-scale physical models for a wide area. To cope this problem, the statistical-model approach was proposed to model the landslide occurrence in terms of LIF in a data-driven manner. Generally speaking, existing statistical-model methods can be categorized into classical and machine learning-based statistical methods [2, 3, 4]. Classical statistical methods, such as Frequency Ratio, Weight of Evidence and Information Value develop analytically tractable linear models of the landslide occurrence in terms of LIF. However, such linear models usually suffer from poor performance due to the complex mechanism of landslide occurrence. To circumvent this obstacle, machine learning-based nonlinear models were recently proposed using landslide inventory information and LIF. For instance, random forests [2], decision trees [3] and support vector machines [4] have been proposed to generate LSM. It has been demonstrated that the machine learning-based methods can provide superior performance as compared to classical statistical methods. Despite their many advantages, the existing machine-learning models cannot generate LSM on a global scale due to the computational complexity required. To cope with this problem, this work proposes to focus on selected LIFs to produce a large-scale LSM with practical computational complexity. Among the many LIFs, heavy rainfall is commonly considered as one of the most influential external factors. This is because that rainfall is directly associated with the cluster characteristics related to forest changing.

Inspired by this observation, we will first propose a Global Deforestation Detection Algorithm (GDDA) established by exploiting rainfall data derived from SAR Doppler Weather Radars. In sharp contrast to visible-light images that may be obscured by clouds, SAR systems are less sensitive to clouds by penetrating through clouds. As a result, such SAR data can provide high-resolution ground information in all weather conditions. This work investigated more than 1,000 doppler weather radar distributed in more than 90 countries/regions to predict future rainfall trends based on the current rainfall

Corresponding authors: Man-On Pun (SimonPun@cuhk.edu.cn); Ming Liu (liuming@mizarvision.com)

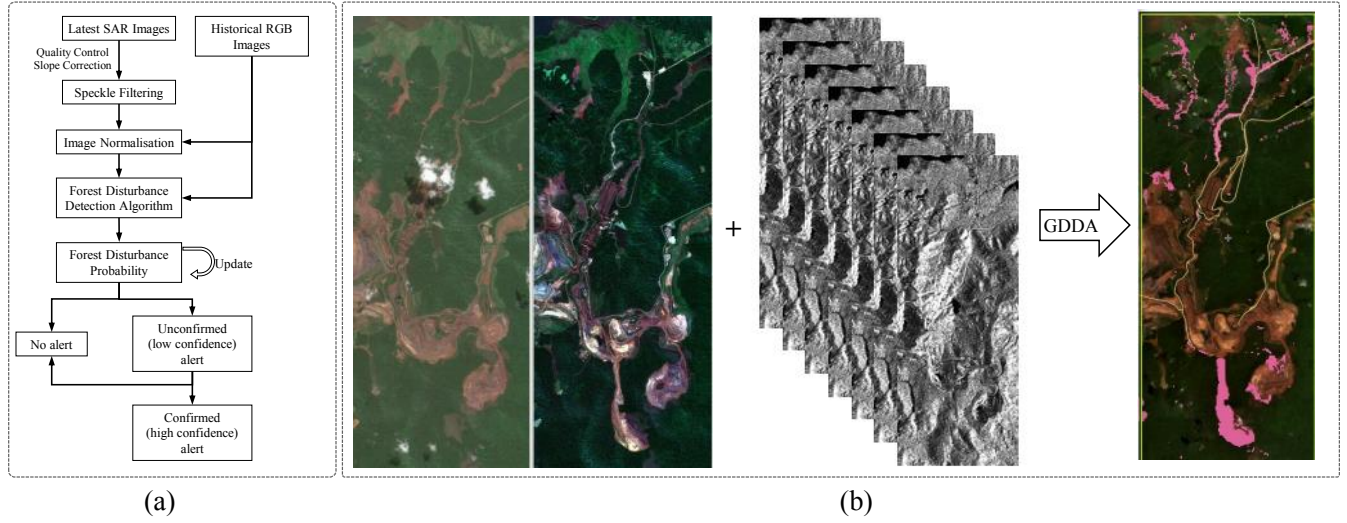


Fig. 1. Schematic for the proposed GDDA: (a) presents the detailed deforestation detection algorithm. (b) shows the inputs of RGB images and time-series SAR data and gives a simple result of GDDA. The results are superimposed with RGB images. The pink points represent the areas with a high probability of forest changes, namely the areas with high sensitivity to landslides.

situations in real-time. Furthermore, inventory information used in this investigation is the distribution of 11,033 reported landslide and debris flows caused by rainfall from 2007 to 2019. By exploiting these three data sources, namely inventory data, rainfall data and forest change information, an applied SVM-based method is proposed to generate the LSM. The main contributions of this work are summarized as follows:

- A SAR and CNN-based Global Deforestation Detection Algorithm (GDDA) is proposed to obtain global-scale LSM showing landslide and debris flows sensitive areas with spatial distribution characteristics;
- An applied LSM system is proposed to effectively fuse inventory landslide and debris flows data, rainfall data and forest change information. Two successful real cases show that it can accurately predict LDF and make great guidance for reducing geological disasters.

2. PROPOSED LSM

In this section, we will first provide an overview of the proposed LSM framework before elaborating the details of each key component.

LSM Framework

Despite the fact that many factors may influence LSM, we propose to focus on a few key factors in order to construct a global-scale LSM. Clearly, rainfall and surface forest change are two critical factors strongly related to LSM. In addition, rainfall data and remote sensing image data are more accessible than other LIFs such as soil moisture and geological rock formation. Equipped with remote sensing image data such as time-series SAR data and true-color images, we propose to

train a CNN model called GDDA to detect forest change to identify spatially sensitive areas at the global scale. The detailed information about GDDA is discussed in the next paragraph. After that, multimodal data consisting of the GDDA output, rainfall data and the global inventory data is input to train an SVM-based classifier to perform large-scale LSM early warning prediction.

GDDA

As shown in Fig. 1, GDDA receives two inputs which are the latest SAR images and historical true-color images, also known as Red-Green-Blue (RGB) images. Thanks to sophisticated SAR technology, state-of-the-art SAR systems can provide high-quality datasets with quality control and slope correction. According to the direction in the electromagnetic field vibrates along propagation, there are two polarization modes for SAR systems, namely horizontal and vertical polarization modes. We collect the vertically polarized remote sensing images as one of the inputs to GDDA. Since the amplitude of SAR images is a part of the polar form of a complex pixel, we denote it by I_t in SAR data at time t for $n = 1, 2, \dots$. Thus the GDDA algorithm works as follows:

$$I_{\Delta}(t, n) = I_{t+n} - I_t, \quad (1)$$

$$I_{DD} = f(I_{\Delta}(t, n), I_{RGB}) \quad (2)$$

where Eq. (2) computes the energy change for the same pixel from time t to time $t + n$ while $f(\cdot)$ stands for a CNN-based model trained on the true-color images and the energy change computed from Eq. (2). Utilizing the proposed GDDA, we can calculate the forest disturbance probability in a global scale using the latest data.

Rainfall data and Inventory data

Next, we illustrate the utilization of time-series rainfall data in

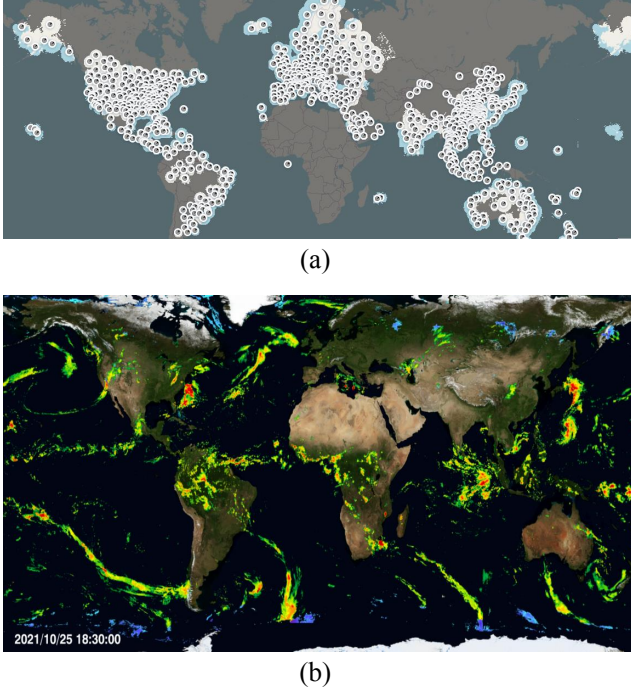


Fig. 2. Rainfall data. (a) shows historical rainfall data on a global scale, and (b) presents rainfall prediction made by DWR.

our proposed LSM system. Rainfall information is compiled from Doppler Weather Radar across the world. DWR is an active remote sensing detection tool with important applications in measuring clouds, rainfall and the occurrence of various severe convective weather by exploiting the Doppler effect. We collected data from more than 1,000 DWRs in over 90 countries/regions. The data enables us to monitor current rainfall conditions in real-time and predict future rainfall trends. Let us denote $R_{x,y}$ the rainfall probability at geographical location (x, y) . Fig. 2 shows an example of DWR data at a global scale, where the size of the marker is proportional to the rainfall capacity.

Finally, inventory data has been collected for the distribution of 11,033 reported rainfall-generated landslides and debris flows from 2007 to 2019. These datasets form the basis for LDF warnings on a global scale. We argue that landslides are more likely to occur repeatedly in the same region. We denote by $L_{x,y}$ the landslide history of the geographical location with $L_{x,y} = 1$ indicating landslides occurred in recent years and $L_{x,y} = 0$ otherwise. Fig. 3 depicts the distribution of LDF mentioned above.

SVM Classifier

SVM is a supervised learning algorithm based on statistical learning theory. Assuming that a training dataset contains a set of landslide-influencing factors as input values and locations as output values, the goal of the SVM training algorithm is to find an optimal hyperplane. This hyperplane divides



Fig. 3. Inventory landslide and debris flow caused by rainfall.

the dataset into two categories: landslide and non-landslide. The process of maximizing the separation of the two classified classes leads to two parallel hyperplanes called boundary planes. The minimum perpendicular distance between the hyperplane and the training observations is called the margin. SVM classifies the data using the separating hyperplane by maximizing the margin. After the algorithm converges, an optimal hyperplane is obtained.

For the sake of training and computational efficiency, we use a simple yet efficient SVM to utilize the multimodal data for the final LSM generation. In the three-dimensional space consisting of deforestation probability, rainfall probability, and inventory information, we train an SVM classifier to find a segmentation hyperplane with strong robustness. The process can be formulated as:

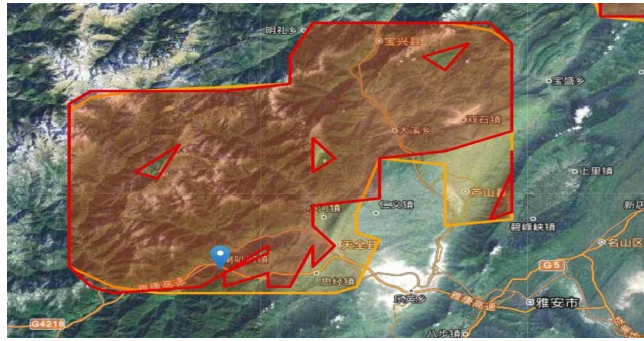
$$P_{x,y} = SVM(I_{DD}, R_{x,y}, L_{x,y}) \quad (3)$$

where $P_{x,y}$ represents the probability of LDF occurrence predicted by SVM classifier at a certain location based on multimodal data I_{DD} , $R_{x,y}$ and $L_{x,y}$.

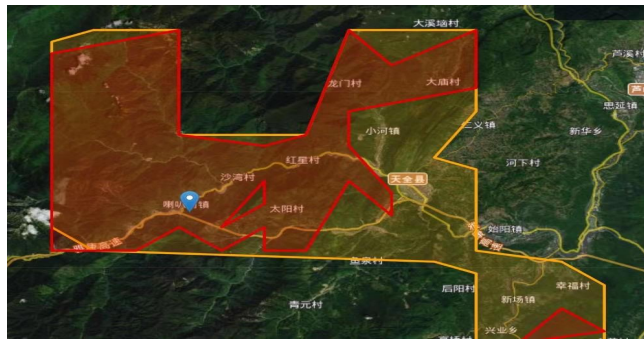
3. RESULTS AND DISCUSSIONS

In this section, we will use two real-world cases to demonstrate that the proposed LSM system can make accurate early warnings for LDF.

Real Case 1 On September 26, 2021, the reconstructed road in the reservoir area of the Guolangqiao Hydropower Station in Labahe Town, Tianquan County, Ya'an, Sichuan, China, was flooded by heavy rains, which induced a blockage-type mountain torrent that left 62 people trapped. On-site traffic roads, electricity, satellite communications, and so on were interrupted. In the early morning of the 26th, the disaster-stricken point was included in the red warning area in the LSM system and rescuers responded immediately. On the 28th, with the post-disaster rescue, this area was still a red area in the system indicating a very high incidence of secondary disasters probability. The proposed system successfully provided a decision-making basis for rescuers to prevent secondary disasters. Fig. 4 shows the predictions at the two moments mentioned above for the occurrence of LDF in the



(a)



(b)

Fig. 4. Real Case 1: (a) Early warning of ground disasters in Ya'an, Sichuan, and (b) early warning of subsequent secondary disasters. Red areas represent high risk while yellow areas represent moderate risk.

area.

Real Case 2 On September 30, 2021, the proposed LSM system showed that many roads in Heishui County, Aba Prefecture, Sichuan, China were located in red areas with a high risk of LDF. Later, a section of the surrounding road actually collapsed. Fig. 5 shows the predictions of the road disaster risk and the site map of a road collapse in Heishui County.

In summary, the above two cases have shown that the machine learning framework based on multimodal data, i.e. rainfall data, forest regression data and historical data can effectively predict LDF disasters. Furthermore, due to its real-time monitoring and continuous update characteristics, it can also provide early warning for possible secondary disasters. Lightweight CNN and SVM can effectively achieve a balance between accuracy and efficiency.

4. ACKNOWLEDGEMENTS

This work was supported by the Shenzhen Science and Technology Innovation Committee under Grant No. JCYJ 20190813170803617.s

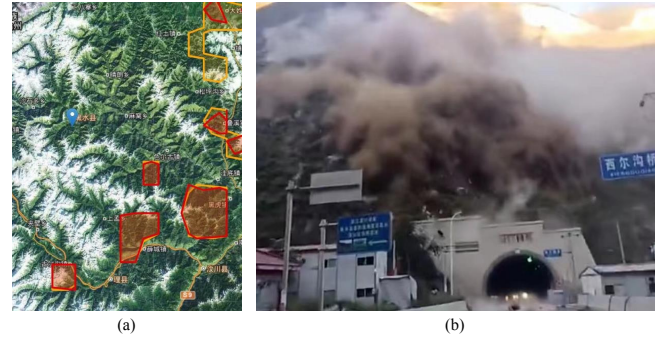


Fig. 5. Real Case 2: (a) Road disaster risk map and (b) site map of road collapse in Heishui County.

5. CONCLUSION

This work has presented a unique and practical perspective on constructing a disaster early warning system by focusing on three multimodal data related to the development of geological structure and environmental external factors. By exploiting the spatial and temporal distribution characteristics of the LDF, a GDDA algorithm is proposed to utilize SAR and RGB images to detect spatially sensitive regions from the perspective of forest change. After that, a highly efficient SVM-based classifier is established to classify the multimodal data space for robust prediction. Empowered with our LSM system, the geological disasters that happened at Ya'an and Heishui in 2021 were forewarned in advance, which provided valuable guidance for disaster prevention and relief.

References

- [1] G. G. Anagnostopoulos, S. Fatichi, and P. Burlando, "An advanced process-based distributed model for the investigation of rainfall-induced landslides: The effect of process representation and boundary conditions," *Water Resources Research*, vol. 51, no. 9, pp. 7501–7523, 2015.
- [2] F. Catani, D. Lagomarsino, S. Segoni, and V. Tofani, "Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues," *Natural Hazards and Earth System Sciences*, vol. 13, no. 11, pp. 2815–2831, 2013.
- [3] B. T. Pham, D. T. Bui, M. Dholakia, I. Prakash, and H. V. Pham, "A comparative study of least square support vector machines and multiclass alternating decision trees for spatial prediction of rainfall-induced landslides in a tropical cyclones area," *Geotechnical and Geological Engineering*, vol. 34, no. 6, pp. 1807–1824, 2016.
- [4] X. Yao, L. Tham, and F. Dai, "Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of hong kong, china," *Geomorphology*, vol. 101, no. 4, pp. 572–582, 2008.