MethodsX 12 (2024) 102797

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/methodsx

Method for monitoring and forecasting landslide phenomenon based on machine learning

Van-Tinh Nguyen*, Quang-Anh Nguyen, Ngoc-Kien Nguyen

School of Mechanical Engineering, Hanoi University of Science and Technology, Vietnam

A R T I C L E I N F O

Method name:
Landslide monitoring method using IoT and machine learning

Keywords:
Landslide
Machine learning
Forecasting
Monitoring

A B S T R A C T

A landslide involves the downward movement of a mass of rock, debris, earth, or soil. Landslides happen when gravitational forces and other types of shear stresses on a slope surpass the shear strength of the materials. Additionally, landslides can be triggered by processes that weaken the shear strength of the slope's material. Shear strength primarily depends on two factors such as frictional strength, which is the resistance to movement between the interacting particles of the slope material, and cohesive strength, which is the bonding between those particles. A landslide is a terrible natural disaster that causes much damage to both human life and the economy. It often occurs in steep mountainous areas or hilly regions, ranging in scale from medium to large. It progresses slowly (20–50 mm/year), but when it occurs, it can move at a speed of 3 m/s. Therefore, early detection or prevention of this disaster is an essential and significant task. This paper developed a method to collect and analyze data, with the purpose of determining the possibility of landslide occurrences to reduce its potential losses.

- The proposed method is convenient for users to grasp information of landslide phenomenon.
- A machine learning model is applied to forecast landslide phenomenon.
- Internet of things (IoT) system is utilized to manage and send a warning text to individual email address and mobile devices.

Specifications table

Subject area:	Engineering
More specific subject area:	Landslide warning system
Name of your method:	Landslide monitoring method using IoT and machine learning
Name and reference of original method:	N/A
Resource availability:	N/A

Background

A landslide is the phenomenon of the movement of soil, dust, and rocks sliding down from slopes in hills and mountains. It is a natural disaster that causes massive damage to human life, the economy, and the environment. Various types of landslides include rockfalls, topples, and debris flows… According to USGS (United States Geological Survey), an average of 25–50 people is killed by landslides each year in the United States [1], and worldwide, the number can go up to thousands. In Vietnam, landslides often occur in hilly areas during the rainy season. Vietnam is one of the countries that have suffered the most significant damage from this

* Corresponding author.
E-mail address: tinh.nguyenvan@hust.edu.vn (V.-T. Nguyen).

https://doi.org/10.1016/j.mex.2024.102797
Received 26 February 2024; Accepted 8 June 2024
Available online 9 June 2024

disaster. Just in 2023, landslides in Vietnam caused the death of 49 people and about 1000 billion VND (about 41 million dollars) to the economy [2]. From 2000 to present, the average number is 15 large landslides per year. The results of geological investigations and surveys have been verified at 15,000 locations across 21 provinces and cities where geological incidents have occurred. Among them, there are 13,233 points identified as rock landslides [3]. The resulting damage is indeed serious and impactful. In August 2023, Vietnam had to suffer many severe landslides, leading to the consequence that 13 people died and eight injuries. In addition, approximately 82 houses were destroyed and uninhabitable, while over a thousand others suffered varying degrees of damage. Regarding infrastructure, the number of structures that had been affected and damaged is about 261 irrigation systems, 26 schools, and four health facilities. In agriculture, 1277 hectares of crops have been ruined, 30.43 hectares of aquaculture area have suffered damage, and about 2052 poultry and 255 cattle have died or been swept away. The traffic was also under impact, with about 709 landslide points identified on highways and local streets [4]. Therefore, it is necessary to monitor and detect the occurrence of landslides early to minimize the potential damage.

In fact, the primary factors that lead to landslides are rain and soil conditions in the area. In hilly and mountainous regions, heavy rainfall has the potential to destabilize soil and rocks, leading to landslides. Additionally, deforestation can also contribute to this instability, as the absence of tree roots to anchor the soil may result in various types of landslides. Numerous methods and equipment are available to analyze or detect landslides worldwide, such as remote monitoring (analyzing images from satellites) [5–7], topo-geodetic survey (geological analysis) [8]. For example, Matteo Del Soldato and his colleagues researched on satellite interferometry for landslide detection in Italy [9]. Likewise, Hiroshi Kimura and Yasushi Yamaguchi integrated satellite radar interferometry with the Digital Elevation Model Elimination (DEME) method [10]. In addition, Amrita Mohan applied a convolutional neural network (CNN) to detect the landslide phenomenon [11]. Mohammad Azarafza addressed the application of a deep convolutional neural network (CNN–DNN) on forecasting landslide with geo-hazard assessment and analysis of the landslide phenomenon happened in the past in Isfahan, Iran [12]. Likewise, research on applying CNN to categorize the shape of rock blocks and evaluate the level of impact on slopes was presented by Honghu Zhu [13]. The other researchers focus on the geological analysis method such as Paul Sestras proposed multi-instrumental approach to slope failure monitoring in a landslide susceptible newly built-up area [8]. Another paper from Arzu Arslan and Kelam Mehmet Abdullah indicated a method based on optical fiber to monitor the movement of slope [14–17], which has very quick, sensitive, and continuous responses to the movement of soil to predict the chance of landslides. These methods are expensive and complicated to apply worldwide. The developing countries require more suitable and affordable solutions. Besides, there are many algorithms that can predict landslides based on analyzing natural conditions, such as Guzzetti's research on rainfall thresholds for the possible occurrence of landslides [18]. Thiebes developed the stability model for sloping roofs based on physics, integrating stability and hydrological models [19]. This paper applied the random forest algorithm combined with weather data collected from sensor systems and the weather forecasting center to train a prediction model of landslide phenomenon in Vietnam. The proposed method is convenient to deploy and apply because it uses affordable electronic equipment but still has an effective response and predictions with a trained model using the decision tree algorithm. However, this method relies heavily on the authenticity and completeness of the data that provide for training and the environmental conditions such as soil quality, air temperature, and moisture.

Method details

This method is constructed with numerous sensors and embedded components as follows:

- NodeMCU ESP8266;
- Solar and lithium battery;
- DC-DC converter 5 V module LM2596;
- Real-time clock module DS3231;
- Relay 5 V module;
- Mini pump 365 12 V module;
- Digital weighing scale;
- Soil moisture sensor module;
- Temperature sensor module DHT11.

Determining threshold to detect landslide

Rain is recognized as the main reason that leads to landslides, and scientists all over the world have tried to determine exactly the amount of rainfall that triggers the landslides. In 1980, Nel Caine listed 73 conditions of rainfall intensity and duration that led to the occurrence of landslides and debris flows worldwide [20]. This data is a global threshold of rainfall intensity and duration (ID) for the occurrence of landslides. The threshold curve proposed by Caine was described by Eqs. (1) and (2).

$$I_1 = 0,1 + 8,5 \times D - 0,65(\mathrm{mm}) \tag{1}$$

$$I_2 = 14,82 \times D - 0,39(\mathrm{mm}) \tag{2}$$

Where I_1 is warning of intense rainfall threshold, I_2 is the dangerous intense rainfall threshold, and D is the time of rain (hour).

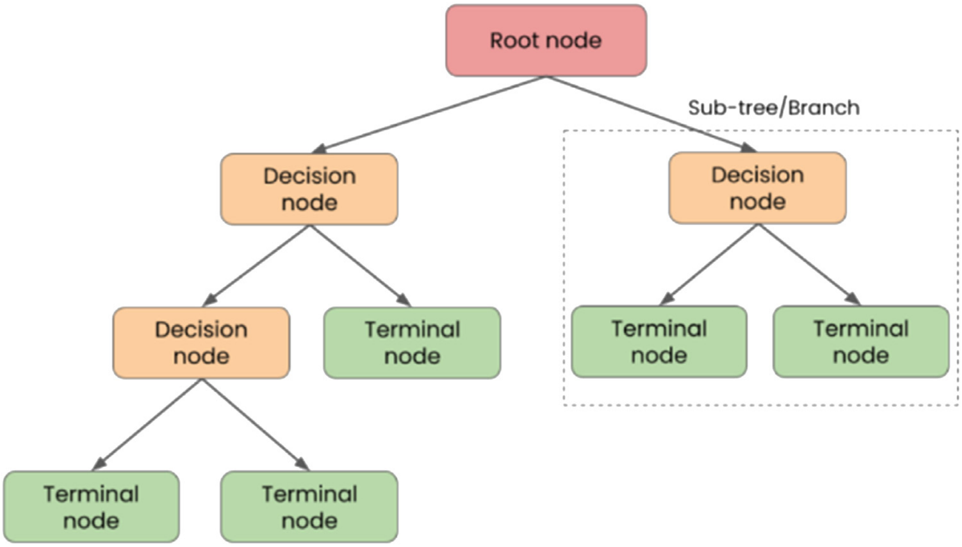


Fig. 1. Decision tree model [21].

1	Day	Time	luongmua	thoigianmua	nguycosatlo
2	1/1/2023	0:00	2.6	1	0
3	1/1/2023	1:00	5.23	2	0
4	1/1/2023	2:00	4.55	3	0
5	1/1/2023	3:00	4.44	4	0

Fig. 2. The collected data.

Forecasting landslide phenomenon using decision tree and random forest

Decision Tree as shown in Fig. 1 is a supervised learning algorithm that can address both regression and classification tasks.

The root node is the starting point of the decision tree, created by dividing the data based on a specific condition. The decision node is where the decision tree tests the data on a specific feature and divides the data into two or more subsets based on the test result. This node represents the decision rules used for classifying or regressing the data.

The terminal node (Leaf node) is the result of the decision tree and contains a class or regression value. In the decision node, computers require data to evaluate and compare different splitting conditions. Indexes for evaluation have been introduced, including entropy and information gain.

Random forest model is a popular machine-learning algorithm used for prediction and classification tasks. It is an ensemble model constructed from multiple decision trees [22]. The main idea behind Random Forest is to combine predictions from many individual decision trees to produce a more accurate and stable prediction. To build every individual decision tree, this research used a bootstrap technique to create a dataset. Each dataset has the same size as the original dataset but may contain duplicate or missing data points to ensure diversity and randomness in the data for each decision tree. The result of random forest is combined from multiple decision trees. When predicting the outcome for a new data point, the decision trees built from the previous step will "vote" to determine the result. Each decision tree contributes its own result, and the result is the one chosen by the majority. Data described in Fig. 2 is soil moisture and daily rainfall collected from the sensor system in the area.

Data collected in the data table include day, time, rainfall (luongmua) (mm), the time of rain (thoigianmua) (hour), and the potential of landslide happening (nguycosatlo) (0 means there is no risk of landslide, 1 means there is a risk of landslide). This research applied Eqs. (1) and (2) to calculate the probability of a landslide as shown in Figs. 3 and 4.

The Panda library in Python was applied to read data from data files in a machine-learning model and then assign labels to input data as presented in Fig. 5.

The training data for the machine learning model was collected in 9 months (May 2022 to February 2023) from the Vrain system [23] at the rainfall measurement station in Quang Ninh. The total data points were 7296 as shown in Fig. 6.

Next, this research used the scikit-learn library in Python to create the Random Forest model, then divided the dataset into two parts: The training part and the test part (80–20). n_estimators and random_state were set to 100 and 42, respectively as described in Fig. 7.

```python
        if rainfall > (Icanhbao[rainfalltime-1]):
            nguycosatlo = 1
        if rainfall < Icanhbao[rainfalltime-1]:
            nguycosatlo = 0
        data.append([hour.strftime('%Y-%m-%d'), hour.strftime('%H:%M'), rainfall,rainfalltime, nguycosatlo])
```

Fig. 3. Determine the risk of landslide.

```
AI_TRAIN1.py   AI_TEST.py   File Data.py   data_train.csv
Plugins supporting *.csv files found.
Day,Time,luongmua,thoigianmua,nguycosatlo
2022-05-01,00:00,5.48,1,0
2022-05-01,01:00,3.01,2,0
2022-05-01,02:00,2.0,3,0
```

Fig. 4. File collected data.

```python
# Bước 2: Đọc dữ liệu và chia thành tập huấn luyện và tập kiểm tra
# (dữ liệu lượng mưa và độ ẩm đất trong 9 tháng trong file "data_train.csv")
data = pd.read_csv("data_train.csv")

X = data[["luongmua", "thoigianmua"]]
y = data["nguycosatlo"]
```

Fig. 5. Loading data to model.

After training the model, the test file was used to evaluate the performance of the model based on using indexes such as accuracy, sensitivity, specificity, or F1-score to assess the model's classification ability as presented in Fig. 8.

Accuracy: Accuracy is the percentage of the number of correct predictions to the total number of the data in the test set. This index can be used to assess the overall performance of the model simply.

Marco avg: Marco average calculates the average value of Precision, Recall, and F1-score across all classes by computing the average of the values for each class without considering the number of data points in each class.

Weighted avg: Weighted average computes the average value of Precision, Recall, and F1-score based on the total number of data points in each class, weighing the value of each class based on its presence in the dataset.

F1-score (F): This index is used to assess the performance of the model when precision and recall have the same value. This index is calculated as described in Eq. (3).

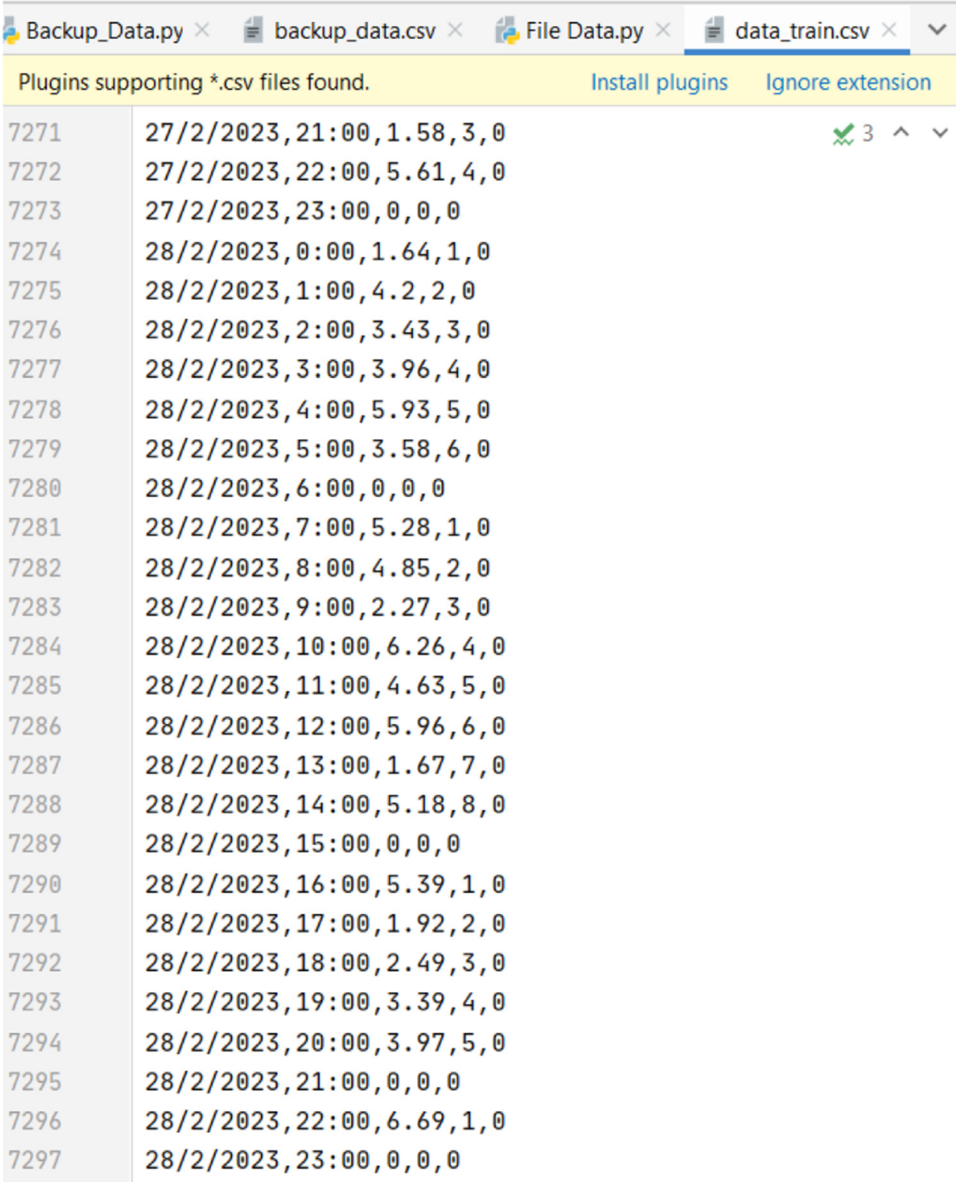
$$F = \frac{2 \times (P + R)}{(P \times R)} \tag{3}$$

Where P is the percentage of the number of true positive predictions to the total number of positive predictions. This index indicates the model's ability to accurately predict positive cases; R is the percentage of the number of true positive predictions to the total number of positive data in the data set.

Data processing

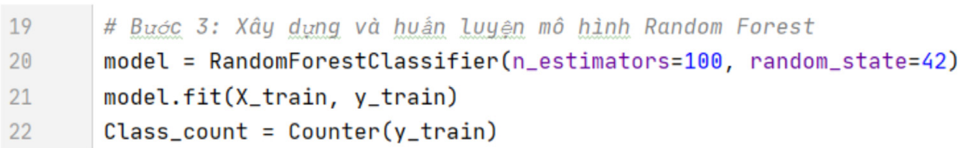
A database is a collection of organized data related to each other, typically stored and accessed electronically from a computer system. When databases become more complex, they are often developed using formal design techniques and modeling. This research used Firebase, which is a real-time database service provided by Google, operates on the cloud platform. It helps developers quickly develop mobile applications by simplifying interactions with the database. To connect the sensor system to the database, this research utilized API (Application Programming Interface). API could provide access to one of its functions or utilities so it can exchange data between specific applications.

The system's API operates with the API Websocket protocol. This protocol supports communication between the client and the server. The server can send callback messages to connected clients, making this type of API more efficient than API REST. A Web API is a method used to allow different applications to communicate with each other by exchanging data back and forth. This data is returned by the Web API in the form of JSON or XML through HTTP or HTTPS protocols. FireBase database has a token to connect from the microcontroller as described in Figs. 9 and 10.


Backup_Data.py backup_data.csv File Data.py data_train.csv
Plugins supporting *.csv files found. Install plugins Ignore extension

```
27/2/2023,21:00,1.58,3,0
27/2/2023,22:00,5.61,4,0
27/2/2023,23:00,0,0,0
28/2/2023,0:00,1.64,1,0
28/2/2023,1:00,4.2,2,0
28/2/2023,2:00,3.43,3,0
28/2/2023,3:00,3.96,4,0
28/2/2023,4:00,5.93,5,0
28/2/2023,5:00,3.58,6,0
28/2/2023,6:00,0,0,0
28/2/2023,7:00,5.28,1,0
28/2/2023,8:00,4.85,2,0
28/2/2023,9:00,2.27,3,0
28/2/2023,10:00,6.26,4,0
28/2/2023,11:00,4.63,5,0
28/2/2023,12:00,5.96,6,0
28/2/2023,13:00,1.67,7,0
28/2/2023,14:00,5.18,8,0
28/2/2023,15:00,0,0,0
28/2/2023,16:00,5.39,1,0
28/2/2023,17:00,1.92,2,0
28/2/2023,18:00,2.49,3,0
28/2/2023,19:00,3.39,4,0
28/2/2023,20:00,3.97,5,0
28/2/2023,21:00,0,0,0
28/2/2023,22:00,6.69,1,0
28/2/2023,23:00,0,0,0
```

Fig. 6. Data for training.



```
# Bước 3: Xây dựng và huấn luyện mô hình Random Forest
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
Class_count = Counter(y_train)
```

Fig. 7. Create the model.

Validation of the method

When the proposed system is connected to the internet, it can send the data to the cloud, including environment temperature, humidity, soil moisture, rainfall, and the real-time location of the system (longitude and latitude). For example, this paper tested the system at the location of (21.0047537, 105.8438662) as shown in Fig. 11.

```
              precision    recall  f1-score   support

           0       0.78      0.81      0.79      1443
           1       0.94      0.79      0.86        17

    accuracy                           0.79      1460
   macro avg       0.87      0.80      0.83      1460
weighted avg       0.86      0.74      0.79      1460

Báo cáo so luot du lieu huan luyen: 5836
```

Fig. 8. Performance of the model.

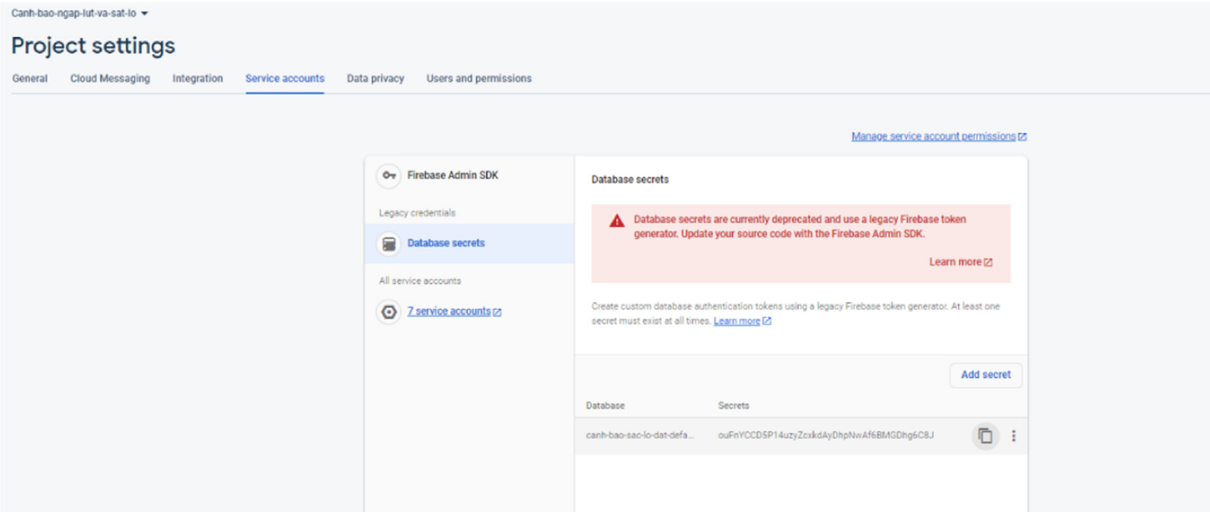


Fig. 9. Token of database.

```
#define FIREBASE_HOST "canh-bao-sac-lo-dat-default-rtdb.firebaseio.com"
#define FIREBASE_AUTH "ouFnYCCD5P14uzyZcxkdAyDhpNwAf6BMGDhg6C8J"
```

Fig. 10. Connect from microcontroller.

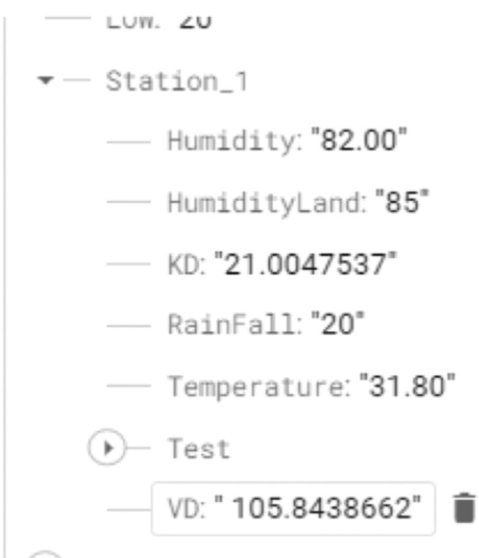


Fig. 11. Data sent to the database.

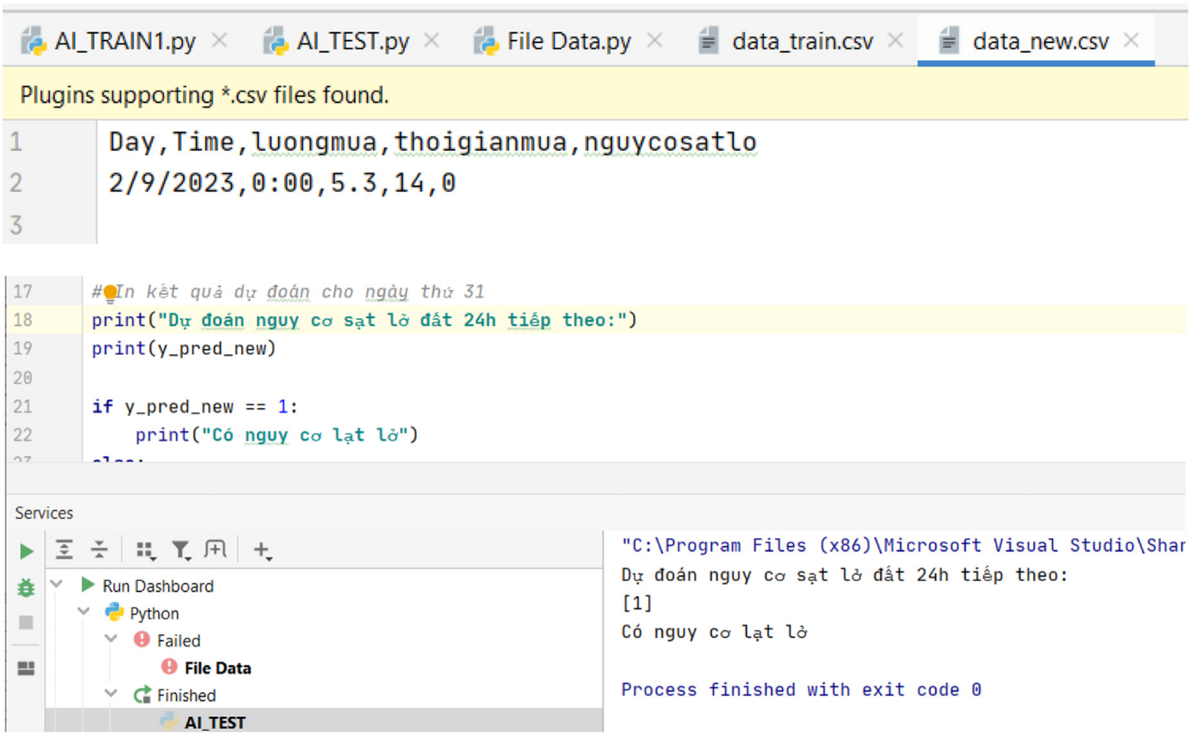


Fig. 12. Test system with high-risk data.

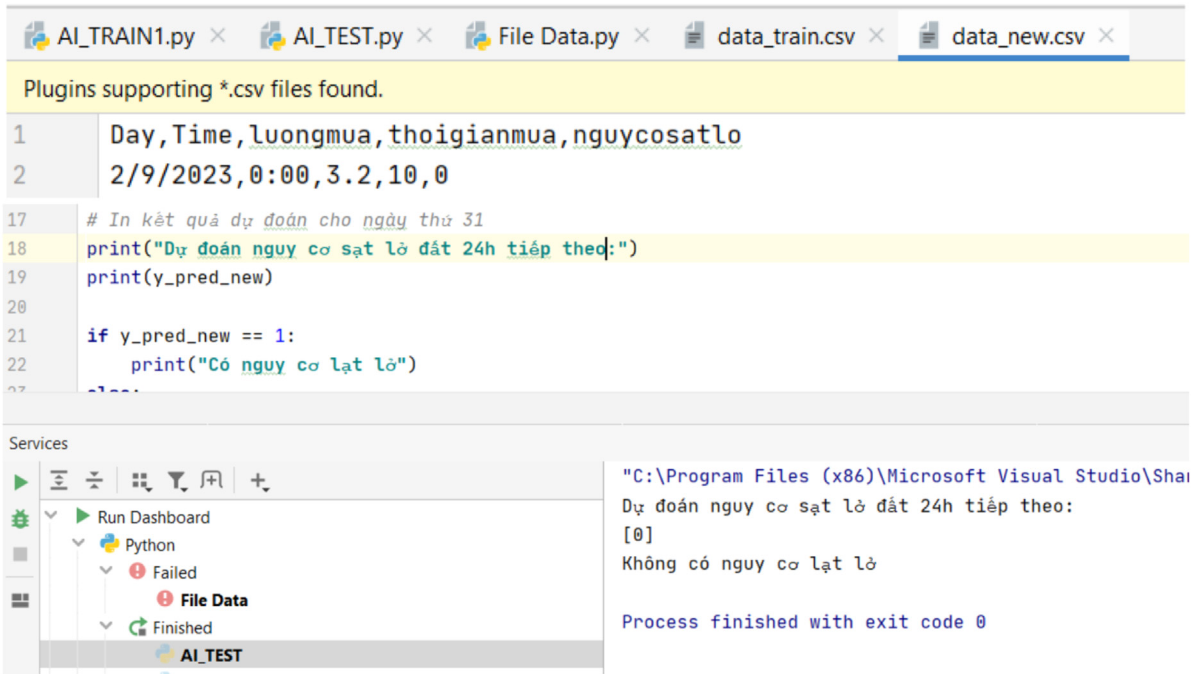


Fig. 13. Test system with low-risk data.

Fig. 14. Test system in real life.

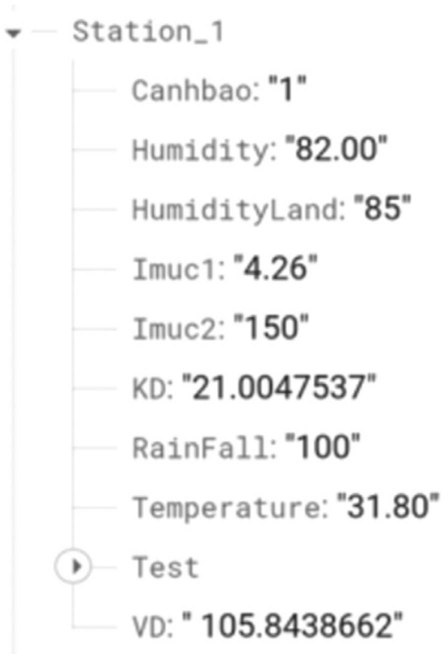


Fig. 15. Final data is sent to the user.

After receiving data from the system, the machine learning model will evaluate the risk of landslides. Firstly, this paper tested with high-risk data: rainfall is 5,3 mm and the time of rain is 14 h as presented in Fig. 12.

In addition, this paper tested with low-risk data: the rainfall is 3.2 mm and the time of rain is 10 h as shown in Fig. 13. Fig. 14 performed the working state of the device in real life and the analyzed data was sent to the user as illustrated in Fig. 15.

Conclusion

A landslide is a terrible disaster, early prediction of landslide phenomenon becomes an essential task to reduce consequences for the humans and the economy. Using a machine learning model to automatically determine the risk of landslides can save time and human resources but still get high performance. This paper proposed a method for monitoring and forecasting landslide phenomena based on machine learning. The proposed method was effectively validated at the location of (21.0047537, 105.8438662) in Quang Ninh, Vietnam.

Ethics statements

The platform's data redistribution policies were complied with: https://vrain.vn/.

CRediT authorship contribution statement

Van-Tinh Nguyen: Data curation, Writing – review & editing. **Quang-Anh Nguyen:** Writing – original draft. **Ngoc-Kien Nguyen:** Conceptualization, Methodology, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Death result from landslides each year in the United States. https://www.usgs.gov/faqs/how-many-deaths-result-landslides-each-year.
[2] Landslides in Vietnam. https://kinhtedothi.vn/lu-quet-sat-lo-dat-lam-49-nguoi-chet-thiet-hai-khoang-1-000-ty-dong-ke-tu-dau-nam.html.
[3] Results of geological investigations in Vietnam, (2020). https://quangngaitv.vn/tin-tuc-n10206/gia-tang-tinh-trang-sat-lo-dat.html.
[4] The effects of heavy rain and landslides, (2023). https://dangcongsan.vn/xa-hoi/mua-lon-sat-lo-dat-gay-nhieu-thiet-hai-ve-nguoi-va-tai-san-643953.html.
[5] F. Fiorucci, M. Cardinali, R. Carlà, M. Rossi, A.C. Mondini, L. Santurri, F. Ardizzone, F. Guzzetti, Seasonal landslide mapping and estimation of landslide mobilization rates using aerial and satellite images, Geomorphology 129 (1–2) (2011) 59–70 Iss.
[6] K.S. Cheng, C. Wei, S.C. Chang, Locating landslides using multi-temporal satellite images, Adv. Space Res. 33 (3) (2004) 296–301 Iss.
[7] T. Blaschke, B. Feizizadeh, D. Hölbling, Object-based image analysis and digital terrain analysis for locating landslides in the urmia lake basin, Iran, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7 (2014) 12 Iss.
[8] P. Sestras, Ș. Bilașco, S. Roșca, I. Veres, N. Ilies, A. Hysa, V. Spalević, S.M. Cîmpeanu, Multi-instrumental approach to slope failure monitoring in a landslide susceptible newly Built-Up area: topo-geodetic survey, UAV 3D modelling and ground-penetrating radar, Remote Sens, 14 (22) (2022) 5822.
[9] L. Solari, M. Del Soldato, F. Raspini, A. Barra, S. Bianchini, P. Confuorto, N. Casagli, M. Crosetto, Review of satellite interferometry for landslide detection in Italy, Remote. Sens. 12 (8) (2020) 1351.
[10] Hiroshi Kimura, Yasushi Yamaguchi, Detection of landslide areas using satellite radar interferometry, Photogrammetric Eng. Remote Sensing 66 (2020) 3.
[11] A. Mohan, A.K. Singh, B. Kumar, R. Dwivedi (2020). Review On Remote Sensing Methods For Landslide Detection Using Machine and Deep Learning. Transactions on Emerging Telecommunications Technologies.
[12] M. Azarafza, M. Azarafza, H. Akgün, P.M. Atkinson, R. Derakhshani, Deep learning-based landslide susceptibility mapping, Sci. Rep. 11 (2021) 1–16 Article No. 24112.
[13] H. Zhu, M. Azarafza, H. Akgün, A deep learning-based key-block classification framework for discontinuous rock slopes, J. Rock Mech. Geotechnical Eng. 14 (4) (2022) 1131–1139 Issue.
[14] A. Arslan, H. Akgün, M.K. Koçkar, A.M. Eker, and M.A. Kelam, Slope Movement Monitoring With Optical Fiber Technology, EUROCK 2015 & 64th Geomechanics Colloquium, Future Development of Rock Mechanics, W. Schubert, A. Kluckner (Eds.), October 7-10, Salzburg, Austria, (2015) pp. 419–426, Austrian Society for Geomechanics.
[15] M.A. Kelam, A. Arslan, A.M. Eker, H. Akgün, M.K Koçkar, et al., Optical fiber technology to monitor slope movement, in: G. Lollino, D. Giordan, G.B. Crosta, J. Corominas, R. Azzam, J. Wasowski, et al. (Eds.), Engineering Geology for Society and Territory-Volume 2, Landslide Processes, Springer International Publishing, Switzerland, 2015, pp. 289–293. ISBN 978-3-319-09057-3.
[16] A.A. Kelam, M.K. Koçkar, H. Akgün, Utilization of optical fiber system for mass movement monitoring, Disaster Sci. Eng. 2 (1) (2016) 19–24.
[17] A.A. Kelam, H. Akgün, M.K. Koçkar, Application of an optical fiber-based system for mass movement monitoring, Environ. Earth. Sci. 81 (5) (2022) 1–18 IssueArticle 170.
[18] M.T. Brunetti, S. Peruccacci, M. Rossi, S. Luciani, D. Valigi, F. Guzzetti, Rainfall thresholds for the possible occurrence of landslides in Italy, Natural Hazards and Earth System Sci. 10 (3) (2010) 447–458 issNHESS, 10(2010).
[19] B. Thiebes, R. Bell, T. Glade, S. Jäger, J. Mayer, M. Anderson, L. Holcombe, Integration of a limit-equilibrium model into a landslide early warning system, Landslides 11 (2014) 859–875.
[20] N. Caine, The rainfall intensity: duration control of shallow landslides and debris flows, Geografiska Annaler. Series A, Phy. Geography 62 (1980) 23–27 No. 1/2.
[21] Decision tree model. https://medium.com/@burak_karakaya/wandering-through-the-branches-of-the-knowledge-tree-decision-trees-8eabc0e94c6e.
[22] K. Ho Tin, Random Decision Forests, in: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 1995, pp. 14–16. pp. 278–282.
[23] The data of rainfall measurement station in Quang Ninh, (2023). https://www.vrain.vn/landing.