pennsylvania
DEPARTMENT OF TRANSPORTATION

Artificial Intelligence (AI) for Building a Landslide Inventory & Advanced Landslide Warning System in PA

FINAL REPORT

August 3, 2023

By Tong Qiu and Jun Xiong
Pennsylvania State University

PENNSTATE 1855

COMMONWEALTH OF PENNSYLVANIA
DEPARTMENT OF TRANSPORTATION

CONTRACT # PSUCIAMTIS2019
WORK ORDER # 02

DEPARTMENT OF TRANSPORTATION
UNITED STATES OF AMERICA

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA-PA-2024-003-CIAMTIS WO 02		

4. Title and Subtitle	5. Report Date
Artificial Intelligence (AI) for Building a Landslide Inventory & Advanced Landslide Warning System in PA	August 3, 2023
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Tong Qiu and Jun Xiong	LTI 2024-01

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
The Thomas D. Larson Pennsylvania Transportation Institute The Pennsylvania State University 201 Transportation Research Building University Park, PA 16802	**11. Contract or Grant No.** 3900039095/PSUCIAMTIS2019 WO 02

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
The Pennsylvania Department of Transportation Bureau of Planning and Research Commonwealth Keystone Building 400 North Street, 6th Floor Harrisburg, PA 17120-0064	Final Report 01/03/2022 – 08/03/2023
	14. Sponsoring Agency Code

15. Supplementary Notes
Heather Sorce served as the Project/Contract Administrator; and Beverly Miller served as the Technical Advisor.

16. Abstract
This report presents the results of a study aiming at developing artificial intelligence (AI) models for advanced warning of rainfall-induced landslides for unstable slopes above or below state-maintained roadways in Pennsylvania. Two landslide databases for spatial and spatiotemporal analyses of landslides in Pennsylvania and adjacent areas are compiled. Landslide susceptibility maps (LSMs) are generated for PennDOT Districts 11 and 12 and adjacent areas, including northern West Virginia and eastern Ohio. The results indicate that the spatiotemporal machine learning (ML) model can predict landslides, accounting for both spatial terrain factors and temporal rainfall factors, and the model outperforms pure spatial ML models with the same database size. The LSMs generated from this study highlight areas having very low to very high risk of landslide susceptibility with precipitation, which may be used to establish a hierarchy and mitigate risk for slopes at "very high risk" for landslide susceptibility. The maps may also be used for forecasting purposes. For example, they may be used as an aid for planning and programming purposes to address slopes with "very high" landslide susceptibility first. The maps may be used in the event of incoming storms to target slopes with a very high risk of landslide susceptibility so that mitigation or preventative measures (such as temporary road closure) can be employed to ensure safe travel and minimize damage. In addition, the maps may also help to target post-storm roadway/slope inspections to the most critical and high-risk locations first.

17. Key Words	18. Distribution Statement
Artificial intelligence, landslide, landslide database, landslide warning system, machine learning, precipitation, prediction	No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	102	$86,695.00

DISCLAIMER

This work was sponsored by the Pennsylvania Department of Transportation. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Commonwealth of Pennsylvania at the time of publication. This report does not constitute a standard, specification, or regulation.

Table of Contents

Glossary of Artificial Intelligence Terms

http://robotics.stanford.edu/~ronnyk/glossary.html
https://ml-cheatsheet.readthedocs.io/en/latest/glossary.html
https://developers.google.com/machine-learning/glossary

A

Accuracy

Percentage of correct predictions made by the model.

Algorithm

A method, function, or series of instructions used to generate a machine learning model. Examples include linear regression, decision trees, support vector machines, and neural networks.

Artificial Intelligence (AI)

A non-human program or model that can solve sophisticated tasks. For example, a program or model that translates text or a program or model that identifies diseases from radiologic images both exhibit artificial intelligence.

Formally, machine learning is a sub-field of artificial intelligence. However, in recent years, some organizations have begun using the terms artificial intelligence and machine learning interchangeably.

Attribute (field, variable, feature)

A quality describing an observation (e.g., color, size, weight). In Excel terms, these are column headers.

B

Baseline

A model used as a reference point for comparing how well another model (typically, a more complex one) is performing. For example, a logistic regression model might serve as a good baseline for a deep model.

For a particular problem, the baseline helps model developers quantify the minimal expected performance that a new model must achieve for the new model to be useful.

Batch

The set of examples used in one training iteration. The batch size determines the number of examples in a batch.

Bias metric

What is the average difference between your predictions and the correct value for that observation?

- **Low bias** could mean every prediction is correct. It could also mean half of your predictions are above their actual values and half are below, in equal proportion, resulting in a low average difference.
- **High bias** (with low variance) suggests your model may be underfitting and you're using the wrong architecture for the job.

Bias term

Allow models to represent patterns that do not pass through the origin. For example, if all my features were 0, would my output also be zero? Is it possible there is some base value upon which my features have an effect? Bias terms typically accompany weights and are attached to neurons or filters.

C

Categorical Variables

Variables with a discrete set of possible values. Can be ordinal (order matters) or nominal (order doesn't matter).

Classification

Predicting a categorical output.

- **Binary classification** predicts one of two possible outcomes (e.g., is email spam or not spam? will landslide occur at a location or not?)
- **Multi-class classification** predicts one of multiple possible outcomes (e.g., is this a photo of a cat, dog, horse, or human?)

Classification Threshold

The lowest probability value at which we're comfortable asserting a positive classification. For example, if the predicted probability of having a landslide is > 50%, return True, otherwise return False; in this case, 50% is the classification threshold for landslide occurrence.

Classifier

A mapping from unlabeled instances to (discrete) classes. Classifiers have a form (e.g., decision tree) plus an interpretation procedure (including how to handle unknowns, etc.). Some classifiers also provide probability estimates (scores), which can be the threshold to yield a discrete class decision thereby taking into account a utility function.

Confusion Matrix

A table that describes the performance of a classification model by grouping predictions into 4 categories.

- **True Positives**: we correctly predicted that landslide occurred at a location.
- **True Negatives**: we correctly predicted that landslide did not occur at a location.
- **False Positives**: we incorrectly predicted that landslide occurred at a location
- **False Negatives**: we incorrectly predicted that landslide did not occur at a location

Continuous Variables

Variables with a range of possible values defined by a number scale (e.g., sales, lifespan).

Convergence

A state reached during the training of a model when the loss changes very little between each iteration.

Cross-validation

A method for estimating the accuracy (or error) of an inducer by dividing the data into k mutually exclusive subsets (the "folds") of approximately equal size. The inducer is trained and tested k times. Each time it is trained on the data set minus a fold and tested on that fold. The accuracy estimate is the average accuracy for the k folds.

D

Data set

A schema and a set of instances matching the schema. Generally, no ordering on instances is assumed. Most machine learning work uses a single fixed-format table.

Commonly (but not exclusively) organized in one of the following formats:

- a spreadsheet
- a file in CSV (comma-separated values) format

Deep Learning

Deep Learning is derived from a machine learning algorithm called perceptron or multi-layer perceptron that is gaining more and more attention nowadays because of its success in different fields such as computer vision to signal processing and medical diagnosis to self-driving cars. Like other AI algorithms, deep learning is based on decades of research. Nowadays, we have more and more data and cheap computing power that makes this algorithm really powerful in achieving state-of-the-art accuracy. In the modern world, this algorithm is known as artificial neural network. Deep learning is much more accurate and robust compared to traditional artificial neural networks. But it is highly influenced by machine learning's neural network and perceptron networks.

Dimension

Dimension for machine learning and data scientist is different from physics. Here, dimension of data means how many features you have in your data ocean (dataset). e.g., in case of object detection application, flatten image size and color channel (e.g., 28*28*3) is

a feature of the input set. In the case of house price prediction (maybe) house size is the data-set so we call it one-dimensional data.

E

Epoch

An epoch describes the number of times the algorithm sees the entire data set.

F

False Positive Rate

Defined as

$$FPR = 1 - Specificity = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

The False Positive Rate forms the x-axis of the **ROC curve**.

Feature

With respect to a dataset, a feature represents an attribute and value combination. Color is an attribute. “Color is blue” is a feature. In Excel terms, features are similar to cells. The term feature has other definitions in different contexts.

Feature Selection

Feature selection is the process of selecting relevant features from a dataset for creating a Machine Learning model.

Feature Vector

A list of features describing an observation with multiple attributes. In Excel, we call this a row.

G

Gradient Accumulation

A mechanism to split the batch of samples used for training a neural network into several mini-batches of samples that will be run sequentially. This is used to enable using large batch sizes that require more GPU memory than available.

H

Hyperparameters

Hyperparameters are higher-level properties of a model such as how fast it can learn (learning rate) or the complexity of a model. The depth of trees in a Decision Tree or the number of hidden layers in a neural network are examples of hyperparameters.

I

Instance (example, case, record)

A data point, row, or sample in a dataset. Another term for observation.

L

Label

The "answer" portion of observation in supervised learning. For example, in a dataset used to classify flowers into different species, the features might include the petal length and petal width, while the label would be the flower's species.

Learning Rate

The size of the update steps to take during optimization loops like Gradient Descent. With a high learning rate, we can cover more ground with each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing. With a very low learning rate, we can confidently move in the direction of the negative gradient since we are recalculating it so frequently. A low learning rate is more precise, but calculating the gradient is time-consuming, so it will take a very long time to get to the bottom.

Loss

Loss = actual value (from the dataset)-predicted value (from ML model) The lower the loss, the better the model (unless the model has over-fitted to the training data). The loss is calculated on training and validation and its interpretation is how well the model is doing for these two sets. Unlike accuracy, loss is not a percentage. It is a summation of the errors made for each example in training or validation sets.

M

Machine Learning (ML)

Mitchell (1997) provides a succinct definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." In simple language, machine learning is a field in which human-made algorithms can learn by themselves or predict the future for unseen data.

Model

A data structure that stores a representation of a dataset (weights and biases). Models are created/learned when you train an algorithm on a dataset.

N

Neural Networks

Neural Networks are mathematical algorithms modeled after the brain's architecture, designed to recognize patterns and relationships in data.

Normalization

Restriction of the values of weights in regression to avoid overfitting and improve the computation speed.

Noise

Any irrelevant information or randomness in a dataset obscures the underlying pattern.

O

Observation

A data point, row, or sample in a dataset. Another term for instance.

Outlier

An observation that deviates significantly from other observations in the dataset.

Overfitting

Overfitting occurs when your model learns the training data too well and incorporates details and noise specific to your dataset. You can tell a model is overfitting when it performs great on your training/validation set, but poorly on your test set (or new real-world data).

P

Parameters

Parameters are properties of training data learned by training a machine learning model or classifier. They are adjusted using optimization algorithms and are unique to each experiment.

Examples of parameters include:

- weights in an artificial neural network
- support vectors in a support vector machine
- coefficients in a linear or logistic regression

Precision

In the context of binary classification (Yes/No), precision measures the model's performance at classifying positive observations (i.e., "Yes"). In other words, when a

positive value is predicted, how often is the prediction correct? We could game this metric by only returning positive for the single observation we are most confident in.

$$P = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

R

Recall

Also called sensitivity. In the context of binary classification (Yes/No), recall measures how "sensitive" the classifier is at detecting positive instances. In other words, for all the true observations in our sample, how many did we "catch." We could game this metric by always classifying observations as positive.

$$R = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Recall vs. Precision

Say we are analyzing Brain scans and trying to predict whether a person has a tumor (True) or not (False). We feed it into our model and our model starts guessing.

- **Precision** is the % of True guesses that were actually correct. If we guess 1 image is True out of 100 images and that image is actually True, then our precision is 100%. Our results aren't helpful, however, because we missed 10 brain tumors. We were super precise when we tried, but we didn't try hard enough.
- **Recall**, or Sensitivity, provides another lens through which to view how good our model is. Again, let's say there are 100 images, 10 with brain tumors, and we correctly guessed 1 had a brain tumor. Precision is 100%, but recall is 10%. Perfect recall requires that we catch all 10 tumors.

Regression

Predicting a continuous output (e.g., price, sales).

Regularization

Regularization is a technique utilized to combat the overfitting problem. This is achieved by adding a complexity term to the loss function that gives a bigger loss for more complex models.

ROC (Receiver Operating Characteristic) Curve

A plot of the true positive rate against the false positive rate at all classification thresholds. This is used to evaluate the performance of a classification model at different classification thresholds. The area under the ROC curve can be interpreted as the probability that the model correctly distinguishes between a randomly chosen positive observation (e.g., "spam") and a randomly chosen negative observation (e.g., "not spam").

S

Segmentation

It is the process of partitioning a data set into multiple distinct sets. This separation is done such that the members of the same set are similar to each other and different from the members of other sets.

Specificity

In the context of binary classification (Yes/No), specificity measures the model's performance at classifying negative observations (i.e., "No"). In other words, when the correct label is negative, how often is the prediction correct? We could game this metric if we predict everything as negative.

$$S = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

Supervised Learning

Training a model using a labeled dataset.

T

Test Set

A set of observations used at the end of model training and validation to assess the predictive power of your model. How generalizable is your model to unseen data?

Training Set

A set of observations used to generate machine learning models.

Transfer Learning

A machine learning method where a model developed for a task is reused as the starting point for a model on a second task. In transfer learning, we take the pre-trained weights of an already trained model (one that has been trained on millions of images belonging to 1000's of classes, on several high-power GPUs for several days) and use these already learned features to predict new classes.

True Positive Rate

Another term for recall, i.e.

$$TPR = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

The True Positive Rate forms the y-axis of the ROC curve.

U

Underfitting

Underfitting occurs when your model over-generalizes and fails to incorporate relevant variations in your data that would give your model more predictive power. You can tell a model is underfitting when it performs poorly on both training and test sets.

Universal Approximation Theorem

A neural network with one hidden layer can approximate any continuous function but only for inputs in a specific range. If you train a network on inputs between -2 and 2, then it will work well for inputs in the same range, but you can't expect it to generalize to other inputs without retraining the model or adding more hidden neurons.

V

Validation Set

A set of observations used during model training to provide feedback on how well the current parameters generalize beyond the training set. If the training error decreases but the validation error increases, your model is likely overfitting, and you should pause training.

Variance

How tightly packed are your predictions for a particular observation relative to each other?

- **Low variance** suggests your model is internally consistent, with predictions varying little from each other after every iteration.
- **High variance** (with low bias) suggests your model may be overfitting and reading too deeply into the noise found in every training set.

1 Introduction

1.1 Background

Artificial intelligence (AI) has become an important technology to solve many problems in a variety of areas, like healthcare, entertainment, banking, education, science, and almost everything else in the world today. Based on the success of AI in various domains, it is also becoming more significant and effective in the fields of engineering and natural science, including hazard prediction. Among various natural hazards, landslides are a significant geologic hazard throughout most of southwestern Pennsylvania and in certain other parts of the state (Delano and Wilshusen 2001). A landslide is the gravity-driven movement of an unstable mass of rock, unconsolidated soil, or debris down a slope. According to the mode of material movement, landslides can be categorized into three main types: falling, sliding, and flowing (Cruden and Varnes 1996). Combinations of these types are also common.

Landslides could destroy utilities, structures, and transportation routes, causing travel delays and other negative effects. In Pennsylvania, landslides cause much damage each year. In a 1986 study, more than 700 recent and active landslides in Allegheny County were identified. U.S. Geological Survey (USGS) landslide-inventory maps indicated thousands of landslides in Allegheny and Washington Counties. A study by the Pennsylvania Geological Survey also included the identification of 480 recent and active landslides and nearly 1000 old or unknown landslides in the Williamsport area in north-central Pennsylvania. In Pennsylvania, especially in the Pittsburgh region, many landslides are repaired incompletely or not at all. For most landslide events in Pennsylvania, the cost is not expensive but still significant. Cost estimates of several hundred thousand dollars for stabilization and repair of a landslide affecting two or three properties are typical. When maintenance costs surpass the value of the property, abandonment is often the solution. The state transportation department in large municipalities has incurred significant expenses as a result of landslide damage and additional construction costs for new roads in landslide-prone areas (Delano and Wilshusen 2001).

It is expected that landslides will occur more frequently in the future due to increased urbanization, deforestation, and precipitation intensity due to global climate change. The repair and mitigation work, roadway reconstruction costs, travel delays, and other side effects could be significantly reduced if an advanced warning system of landslides could be provided to, and

implemented by transportation officials to address landslides before they affect the safety, cause inconvenience, and increase cost to the public.

1.2 Research objectives and tasks

The study aims to use machine learning (ML) techniques for landslide susceptibility assessment in Pennsylvania, particularly in the southwest regions (i.e., PennDOT Districts 11 and 12). A warning system for rainfall-induced landslides is developed. The warning system is based on the analysis combining spatial and temporal prediction, where the probability of landslide occurrence is predicted both in time and space.

For the spatial analysis, a landslide susceptibility map (LSM) is developed to identify areas subject to landslide risks ranked from low to high. The LSM takes into account where landslides may potentially occur and what causes them. In the present study, different ML algorithms are applied to find the underlying relationships between landslide occurrence and spatial causative factors (e.g., slope angle, elevation, soil properties, etc.). The probability of landslide occurrence is obtained from ML algorithms, and the susceptibility map is constructed using the model with the best performance as evaluated by comprehensive evaluation matrices.

For the temporal analysis, the risk of landslides is highly related to precipitation, which is a variable of time. In different physiographic and climatic regions worldwide, rainfall is recognized as one of the most common triggers for landslides. In the present study, to predict landslides on a temporal scale, cumulative precipitation factors are included as temporal features. By considering both static spatial and time-varying precipitation factors, spatiotemporal LSM is conducted with ML techniques.

To develop the warning system and achieve the proposed goals, the study consists of four main tasks. Tasks 1 and 2 focus on compiling a database and collecting pertinent information for existing rainfall-induced landslides in Pennsylvania, which serve as the ground truth for training ML models. The results of Task 1 and Task 2 are presented in Section 2 and Section 3, respectively. Task 3 focuses on developing, training, and testing ML models to predict the occurrence of landslides both in space and time. The results of Task 3 are presented in Sections 4 through 7. Task 4 focuses on developing the final project report. Section 8 presents the major conclusions of this

study and limitations of the approaches used. Section 9 presents recommendations and instructions for generating LSMs using the ML models developed in this study.

2 Landslide Database Compilation

2.1 Data sources

In the present study, two databases of rainfall-induced landslides in Pennsylvania and adjacent areas, particularly those in the southwest regions of Pennsylvania (i.e., PennDOT Districts 11 and 12), are compiled based on Google Earth and other sources of satellite images. The two databases serve distinct purposes: one for spatial analysis and the other for spatiotemporal analysis. For the spatial analysis, a larger landslide database was compiled without event dates. Since spatial analysis focuses on the relationships between landslide spatial distribution and relevant static topographic and geotechnical factors (e.g., slope angle, soil strength parameters), event dates and other time variables are not necessary. For the spatiotemporal analysis, a small landslide database with event dates was compiled. Landslides recorded with event dates are rare to collect, resulting in a smaller size of database; however, every landslide event in this database contains more detailed information.

2.1.1 Data sources for spatial analysis

For spatial analysis, the following data sources are used for the compilation of the database. In the 1970s-1980s, John S. Pomeroy and other researchers of USGS published a series of Topo sheets to identify landslides and related features based on topography from aerial photographs by multiplex methods (Pomeroy and William 1979). The maps show the area that was active or has recent evidence of a slide in several counties of southwestern Pennsylvania. A typical sheet created by John S. Pomeroy is shown in Figure 1.

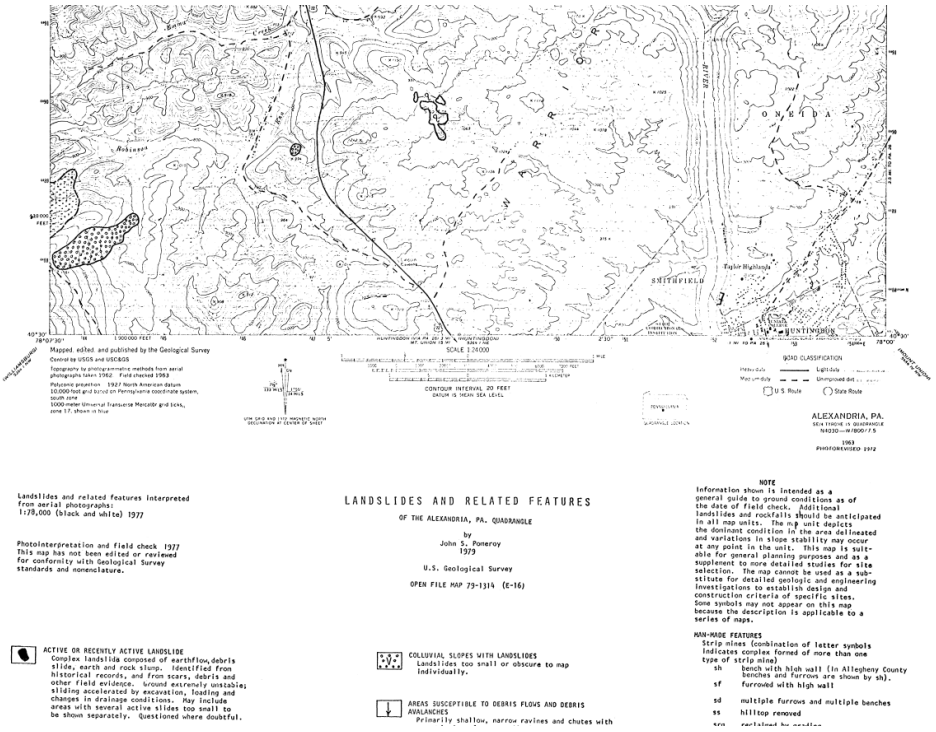


Figure 1. Landslides and related features map by John S. Pomeroy in 1979.

Landslides and landslide potential for southwestern Pennsylvania were then digitized from those USGS Topo sheets. The digitized map is openly available on the Pennsylvania Department of Conservation and Natural Resources website. The database is stored in the format of polygon shapefiles corresponding to landslide locations on the map, which is easy for further process and conversion. The digitally documented landslides in southwestern Pennsylvania are shown in Figure 2.

Figure 2. Documented active landslide polygons in southwestern Pennsylvania.

In this study, the digitized USGS landslide inventory was incorporated as the database for spatial analysis as it contains much more landslides than other available sources, and the polygon shapefile provides the possibility of implementing various methods for LSM compared to the point shapefile.

2.1.2 Data sources for spatiotemporal analysis

Temporal analysis requires landslide events with accurate event dates so that time-varying precipitation data can be incorporated. In the present study, landslide data mainly comes from three sources: USGS Landslide Inventory, NASA Cooperative Open Online Landslide Repository (COOLR), and PennDOT District 12 Slide Database. The first two data sources are available from relevant official websites and are open for the public to download. District 12 Slide Database was provided by PennDOT district engineers, and the database contains comprehensive and detailed information of each landslide recorded by the district engineers. The compiled database contains landslides mainly in Pennsylvania; however, considering similar terrain and climate conditions (i.e., precipitation), landslide data in adjacent areas is also included to extend the database. Hence, the landslide database for spatiotemporal analysis covers Pennsylvania, northern West Virginia, eastern Ohio, and New Jersey.

USGS Landslide Inventory is a web-based interactive map with a consistent set of landslide data. The searchable map includes contributions from many local, state, and federal agencies and provides links to the original digital inventory files for further information (Eric et al. 2022). Considering that landslide inventories are typically collected and maintained by different agencies and institutions, usually within specific jurisdictional boundaries and often with varied objectives and information attributes or even in disparate formats, USGS collaborated with state geological surveys and other federal agencies and released these data to provide an openly accessible, centralized map of existing information on landslide occurrence across the entire U.S. (Eric et al. 2022). Given the wide range of landslide information sources in this data compilation, an attribute is provided to assess the relative confidence in the characterization of the location of each landslide. Confidence (1): possible landslide in the area; Confidence (2): probable landslide in the area; Confidence (3): likely landslide at or near this location; Confidence (5): confident consequential landslide at this location; and Confidence (8): high confidence in extent or nature of landslide (Eric et al. 2022). Since the current USGS landslide inventory is not comprehensive, as further mapping is still needed in many parts of the country, periodic updates of the database are planned as new or improved data become available. To compile a reliable landslide database in Pennsylvania, only landslide data points with Confidence (5) and (8) are collected and added to our database in this study. Figure 3 shows landslide data points in Pennsylvania and adjacent states displayed on USGS Landslide Inventory online map.

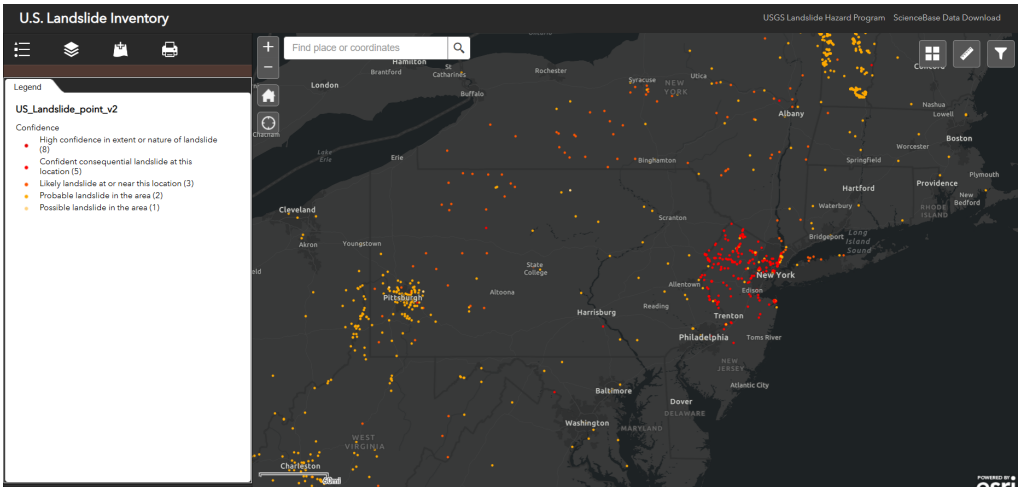


Figure 3. USGS landslide inventory online map.

The NASA Cooperative Open Online Landslide Repository (COOLR) project provides an open platform where scientists and citizen scientists around the world can share landslide reports to guide awareness of landslide hazards for improving scientific modeling and emergency response. Landslides can be submitted to the Landslide Reporter web application or directly to the NASA landslides project team. All the data submitted is made available on the data portal Landslide Viewer, which shows referenced and imported landslide inventories from all over the world (Kirschbaum et al. 2010 and 2015)

As an important part of the NASA landslide project, COOLR is a worldwide inventory of landslide events. COOLR currently includes three data sources: NASA's Global Landslide Catalog (GLC), Landslide Reporter Catalog (LRC), and other collated landslide inventories. NASA GLC currently contains more than 11,500 reports on landslides, debris flows, rock avalanches, etc. around the world, in which reports of landslides are found primarily in online media, including news articles and other databases. However, the compilation of GLC has been a manual and very time-consuming process that is hard to maintain individually; hence, LRC and collated landslides from other institutions are needed to provide an open platform for the global citizen science community to add reports to expand and fill in the gaps in current data (Kirschbaum et al. 2010 and 2015).

In this study, some landslide data points were collected from the data portal NASA Landslide Viewer. The data points in Pennsylvania from Landslide Viewer are shown in Figure 4. The number within each yellow circle in the Viewer represents the number of landslide events recorded in that area, which can be zoomed in for accurately locating the landslides within the area.

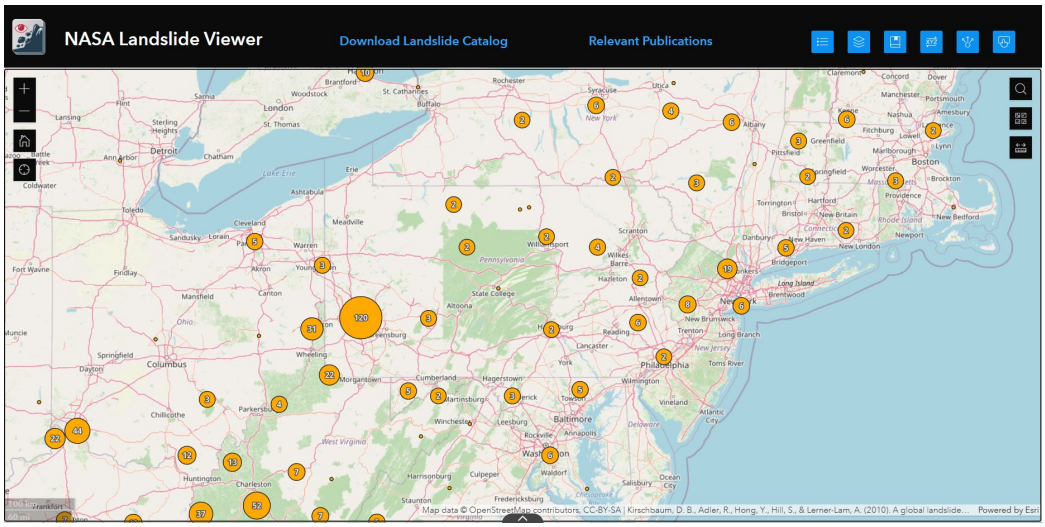


Figure 4. NASA Landslide Viewer.

PennDOT District 12 is responsible for the state-maintained transportation network in the region of several counties, including Fayette, Greene, Washington, and Westmoreland. PennDOT District 12 Slide Database contains pertinent information on identified landslide sites, including the accurate location of the landslide, failure type, identified event date, and other important information. There are 212 events in total in the database, among which there are 134 slides based on the failure type. Since not all slide events have identified event dates, 50 events with event dates were collected and added to the database. Figure 5 shows portions of the District 12 Slide Database in Excel format.

District 12 Slide Database e-IQ Map 12-1: 24 12-2: 75 12-4: 86 12-5: 27 Road Closures: 13 Lane Closures:

Print Date: December 24, 2022

Sort by Rating | Sort by SR | Sort by ADT

		Site Identification														Base Point Criteria			
		County	SR	Segment	Offset	Municipality	Failure Type	Date Identified	Failure Length (ft)	Failure Width (ft)	Failure Depth "Height" (ft)	Approx. Excavation (yd³)	Rip-Rap Estimate (CY)	Last Field Review	Year Repaired	Roadway Classification	Roadway Impact	ADT	Detour Length
97	92	Greene	4023	0010	1435	Morris Twp.	Slide - Shallow		50	10	5	93		12/9/2021		7 - Local	Depression	15	5-15 Miles
98	93	Greene	4029	0030	3121	Washington Twp.	Slide - Shallow		25	10	5	46		12/9/2021		6 - Minor Collector	Depression	461	5-15 Miles
99	94	Greene	4029	0040	0000	Washington Twp.						0		12/9/2021		6 - Minor Collector	Shoulder Only	158	5-15 Miles
100	95	Greene	4033	0040	3800	Washington Twp.		12/9/2021				0		12/9/2021		7 - Local	Shoulder Only	135	5-15 Miles
101	96	Greene	4033	0050	1400	Washington Twp.	Drainage Malfunction					0		12/9/2021		7 - Local	Depression	135	5-15 Miles
102	97	Greene	4033	0070	1400	Washington Twp.	Slide - Shallow					0		12/9/2021		7 - Local	Right-of-Way Only	236	5-15 Miles
103	98	Greene	4035	0020	1520	Franklin Twp.		12/9/2021				0		12/9/2021		7 - Local	Depression	204	5-15 Miles
104	99	Greene	4035	0030	0800	Franklin Twp.	Stream Erosion					0	5,075	12/9/2021		7 - Local	Depression	204	5-15 Miles
105	100	Washington	0018	0330	2050	North Franklin Twp.						0		3/15/2022		4 - Minor Arterial	Depression	7,256	<5 Miles
106	101	Washington	0040	0810	0000	West Brownsville		4/21/2020				0		1/5/2022		3 - Other Principal Arterial	Shoulder Only	4,391	<5 Miles
107	102	Washington	0040	0821	1400	West Brownsville		4/21/2020				0		1/5/2022		3 - Other Principal Arterial	Shoulder Only	4,799	<5 Miles
108	103	Washington	0088	0340	0800	California	Slide - Shallow		100			0		1/5/2022		4 - Minor Arterial	Depression	1,732	5-15 Miles
109	104	Washington	0136	0100	2133	South Strabane Twp.						0		2/23/2022		4 - Minor Arterial	Depression	8,376	5-15 Miles

Main Database | Scoring Criteria | Repaired Slides | Contract Candidates | ADT Data | RMS Admin All | RMS-Seg-#

Figure 5. PennDOT District 12 Slide Database.

2.2 Data acquisition and database establishment

2.2.1 Database for spatial analysis

Based on the digitized map of USGS Topo sheets, the database for spatial analysis covers the regions in southwestern Pennsylvania. The study area is shown in Figure 6, which consists of eight counties: Butler, Armstrong, Indian, Beaver, Allegheny, Westmoreland, Washington, and Greene.

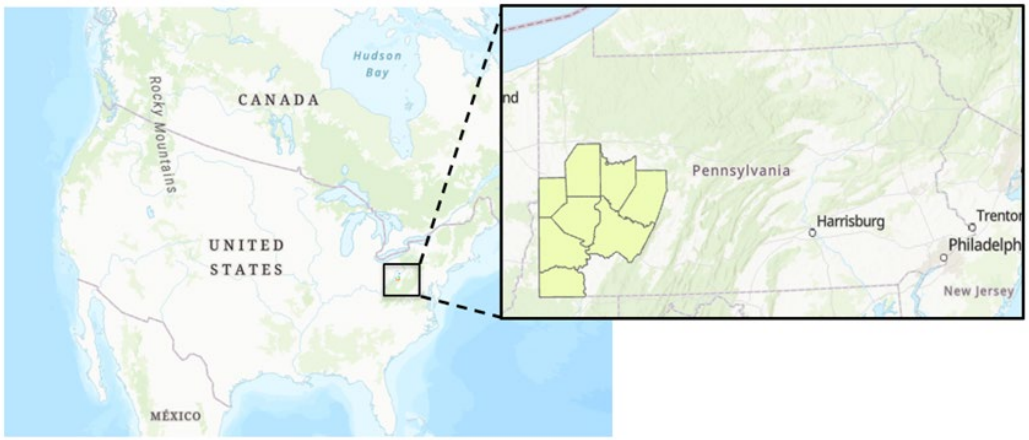


Figure 6. Study area in southwestern Pennsylvania.

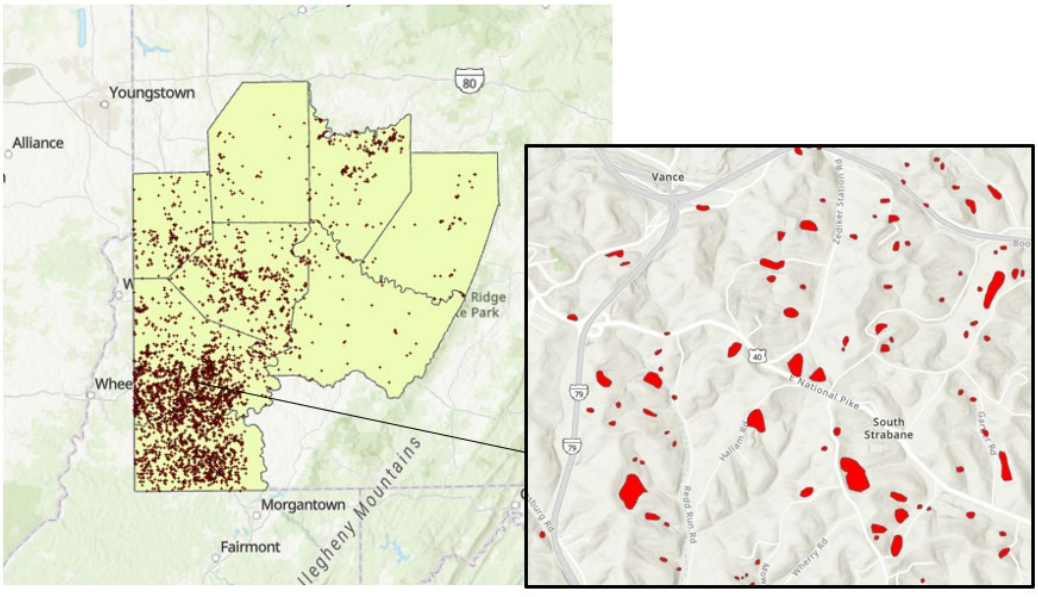


Figure 7. Landslide polygons in the study area for spatial analysis.

In total, there are 4,543 landslide events compiled in the database for spatial analysis, as Figure 7 shows. To display the data distribution clearly and handle relevant calculations and conversion conveniently, all data were compiled and stored in the form of point shapefiles on the ArcGIS platform, as shown in Figure 8. ArcGIS is a geographic information system (GIS) for working with maps and geographic information maintained by the Environmental Systems Research Institute (Esri). As a powerful geospatial software, ArcGIS offers various and comprehensive tools for users to view, edit, manage, and analyze geographic data. In the database, the original polygon shapefile of landslides downloaded from the database of the Pennsylvania Department of Conservation and Natural Resources was converted into a point shapefile. The red dots and marks on the map represent the recorded landslide events. The information attached to each landslide is in the format of standard tables in ArcGIS, as shown in Figure 8.

It is convenient to transfer all information in the database from ArcGIS to other platforms. For example, after data calculation and processing, to better visualize the relationship between the terrain and landslide distribution, the database could be transferred to Google Earth directly from ArcGIS, together with all pertinent information. Google Earth is a free desktop geographic information system with satellite imagery covering all of Earth's landmasses. The database can be stored in Google Earth as KML or KMZ file. A KML file stores geographic modeling information in the Keyhole Markup Language (KML), which is a geographic information systems data format. It includes placemarks, points, lines, polygons, and images. A KMZ file consists of a main KML file and supporting files that are packaged using a Zip utility into one unit called an archive. The KMZ file can then be stored and emailed as a single entity. Figure 9 shows the landslide spatial database transferred to Google Earth. For each landslide, the relevant information is attached and can be displayed when the mark is clicked, as shown in Figure 9.

In addition to being transferable to Google Earth, the database is also exported in the format of an Excel table. In practical engineering applications, tabular databases are often the most intuitive and convenient tools to handle a large amount of data. The Excel tabular database is shown in Figure 10. In this database, landslide location, data source, and pertinent geotechnical variables are recorded in rows and columns. The tabular database is also convenient to serve as ML input data.

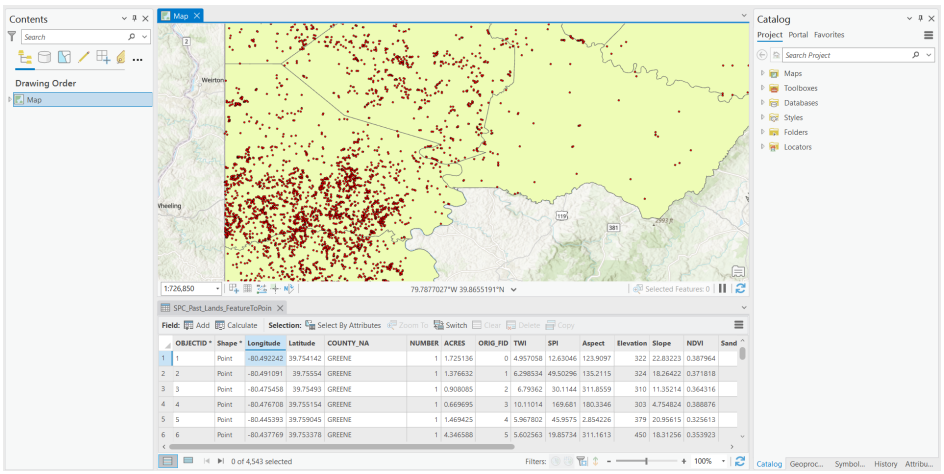


Figure 8. Database for spatial analysis on ArcGIS platform.

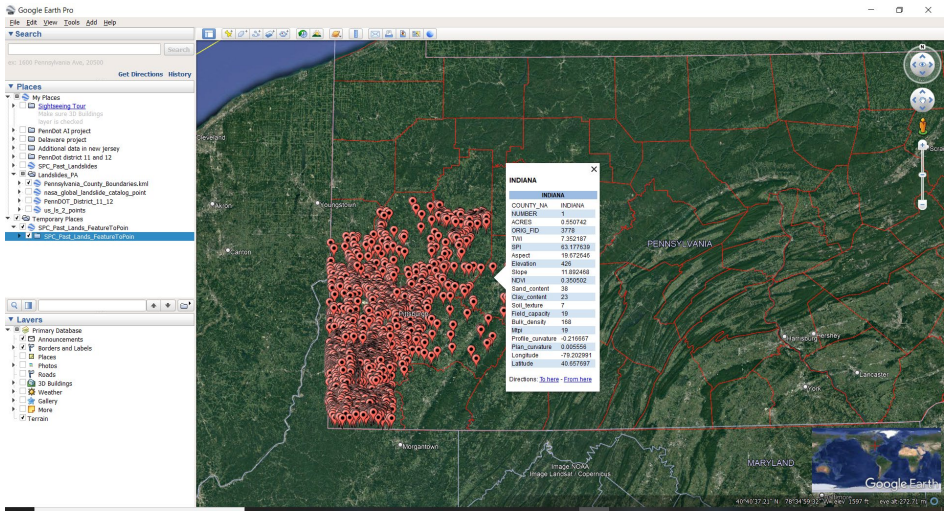


Figure 9. Database for spatial analysis on the Google Earth platform.

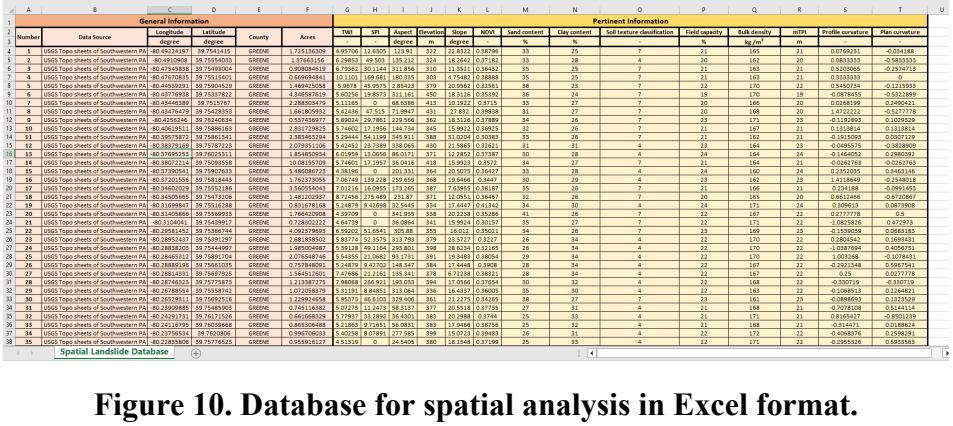


Figure 10. Database for spatial analysis in Excel format.

2.2.2 Database for spatiotemporal analysis

Based on the three data sources: USGS Landslide Inventory, NASA COOLR, and PennDOT District 12 Slide Database, a landslide database for spatiotemporal analysis was compiled with 387 landslide events in total. Since there is an overlap between the recorded data from USGS and NASA, the overlapped data were identified by manual comparison and were not included in the database. The number of landslides with their corresponding source and regional distribution is shown in Table 1.

Table 1. Landslides distribution with data source and region.

Number of landslides	Region	Data source
161	Pennsylvania	NASA COOLR
50	Pennsylvania	PennDOT District 12
55	West Virginia and Eastern Ohio	NASA COOLR
121	New Jersey	USGS Landslide Inventory

USGS Landslide Inventory and NASA COOLR catalog offer different formats of data to download as file geodatabase (.gdb), shapefiles (.shp), or comma-separated values (.csv). PennDOT District 12 provides a slide database in Excel format. The platforms for storage are the same as the database for spatial analysis, which are ArcGIS, Google Earth, and Excel tabular formats. The compiled database in ArcGIS, Google Earth, and Excel is shown in Figures 11-13.

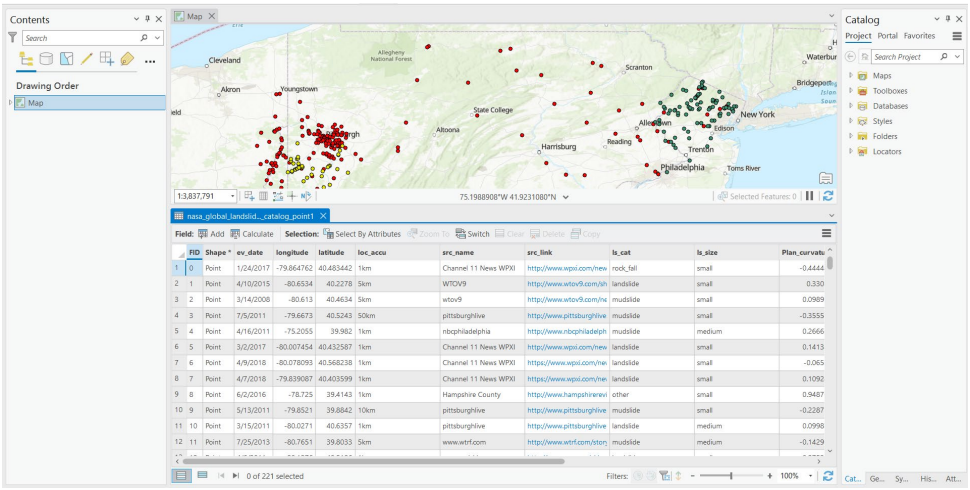
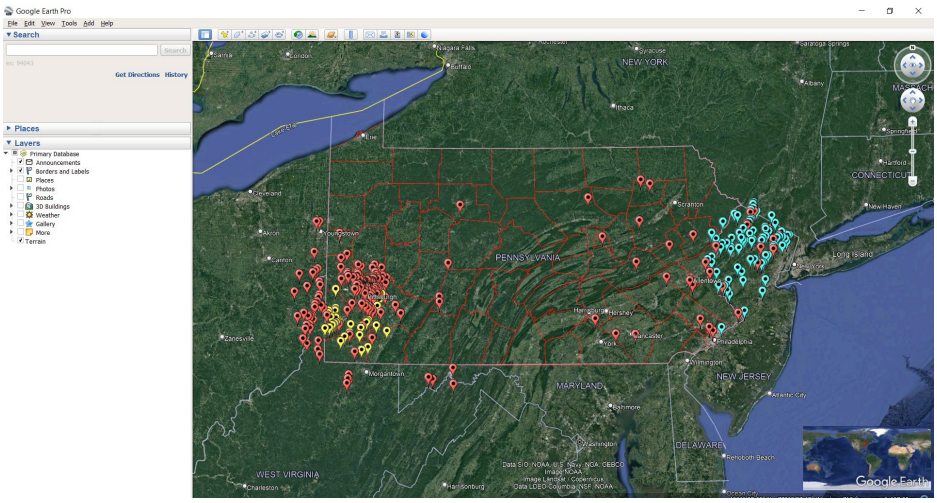
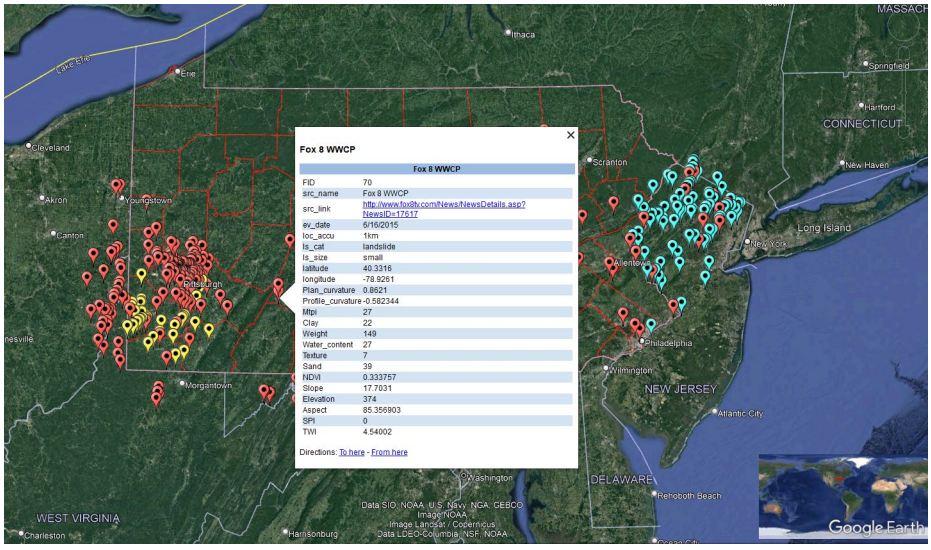


Figure 11. Spatiotemporal landslide database in ArcGIS.

In Figure 11, red dots represent landslide data from the NASA COOLR catalog, green dots represent landslide data from USGS Landslide Inventory, and yellow dots represent landslides provided by PennDOT District 12. Each landslide point in the database contains detailed information about the location of the landslide with the XY coordinates of longitude and latitude, identified event date, data source, and other pertinent geotechnical information.



(a)



(b)

Figure 12. Spatiotemporal landslide database in Google Earth: (a) display of data points; (b) information attached to a data point.

Figure 12(a) shows the landslide spatiotemporal database transferred to Google Earth. The marks with different colors on the map represent landslide events from different sources: red dots represent landslide data from the NASA COOLR catalog, blue dots represent landslide data from USGS Landslide Inventory, and yellow dots represent landslides provided by PennDOT District 12. The red lines on the map show the boundaries of different counties of Pennsylvania, providing a clear relative geographic location of landslides. For each landslide, the relevant information is attached and can be displayed when the mark is clicked, as shown in Figure 12(b).

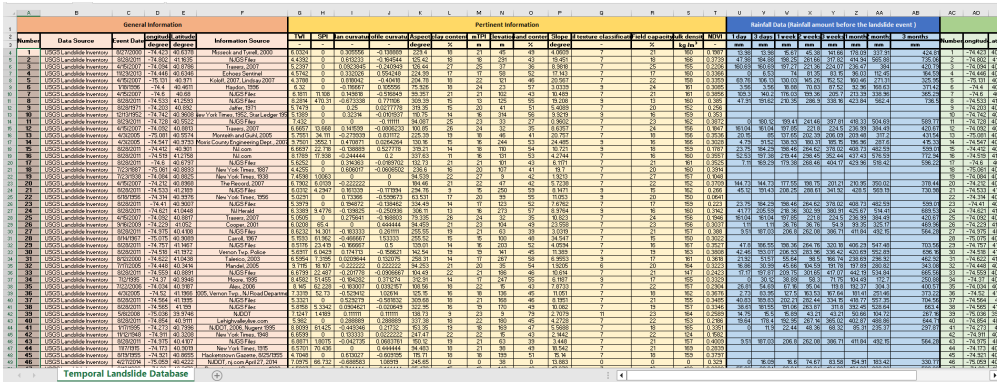


Figure 13. Spatiotemporal landslide database in Excel format.

The Excel tabular database for spatiotemporal analysis is shown in Figure 13. In this database, landslide location, event date, data source, and pertinent geotechnical variables are recorded in rows and columns.

Every landslide event also includes a pre-event image and a post-event image, together with the event date and longitude and latitude coordinates of the event location. The satellite images of landslides are from Google Earth with a Landsat database map. The NASA/USGS Landsat Program provides the longest continuous space-based record of Earth's land in existence. Landsat data provides information essential for making informed decisions about Earth's resources and environment (Tucker et al. 2004). With the function of historical images in Google Earth, pre-event and post-event images of landslides are determined manually based on their location and date. An example of event images is shown in Figure 14. The example landslide event date is 09/25/2018; the longitude and latitude coordinates of the event location are (-80.05269922, 40.53872962). The pre-event image was recorded on 04/2014 while the post-event image was

recorded on 09/2019. The images show that the landslide caused damage to the road. It should be noted that some landslides are not evident in images, and there are some cases where event images could not be obtained for landslides. It is because those historical images were obtained only at certain points in time, during which small landslides may have been repaired completely. In addition, the landslide location has an uncertainty of up to several kilometers; hence, it is challenging to locate every landslide accurately from Google Earth; for those landslides with an old event date, there was no satellite serving to get clear photography of the earth at the time. Links to these images are provided in the Excel database, and these links would work as long as the event images are stored in the same folder as the Excel database.

(a)

(b)

Figure 14. Example of satellite images of landslide (#123) in the database for spatiotemporal analysis: (a) pre-event image; (b) post-event image.

3 Collection of Pertinent Information

Pertinent information regarding the causative factors for the landslides is collected. The causative factors are related to geology, geomorphology, hydrology, land cover, seismicity, manmade activities, etc. (Raghuvanshi et al. 2014; Anbalagan 1992). ML algorithms are used to find underlying relationships between landslide occurrence and these causative factors. However, determining the exact number and type of causative factors to be incorporated in ML models is one of the most critical and challenging tasks in LSM, and many researchers differ on what causative factors should be included in LSM (Kavzoglu et al. 2019). Thus, there is no universally agreed selection of landslide causative factors; in general, these factors can be broadly categorized in Table 2.

Table 2. Category of landslide causative factors (from Moziihrii et al. 2022)

Type of factors	Causative factors
Topography	Slope, aspect, elevation, plan curvature, profile curvature, and sediment transport index
Hydrology	Rainfall, solar radiation, stream power index, topographic wetness index (TWI), distance to rivers, and density of the river
Geological	Lithology, distance to faults, and density of fault
Land use/cover	Land Use and Land Cover (LULC) and normalized difference vegetation index (NDVI)
Man-made	Distance to roads and road density

In this study, fourteen landslide causative factors are chosen for LSM, as Table 3 shows. The values of these causative factors can be downloaded as GeoTIFF files from Google Earth Engine. Google Earth Engine combines a multi-petabyte catalog of satellite imagery and geospatial datasets with planetary-scale analysis capabilities. GeoTIFF is based on the TIFF format and is used as an interchange format for georeferenced raster imagery. GeoTIFF is widely used in NASA earth science data systems.

Table 3. Causative factors used for landslide susceptibility mapping.

Number	Causative factor	Number	Causative factor
1	Elevation	8	Stream power index (SPI)
2	Slope	9	Normalized difference vegetation index (NDVI)
3	Aspect	10	Sand content
4	Multi-scale topographic position index (mTPI)	11	Clay content
5	Profile curvature	12	Bulk density
6	Plan curvature	13	Texture classification
7	Topographic wetness index (TWI)	14	Field capacity

Based on various geospatial datasets in Google Earth Engine, the data of these causative factors can be obtained using codes in the Google Earth Engine code editor. The interface of the Google Earth Engine code editor is shown in Figure 15.

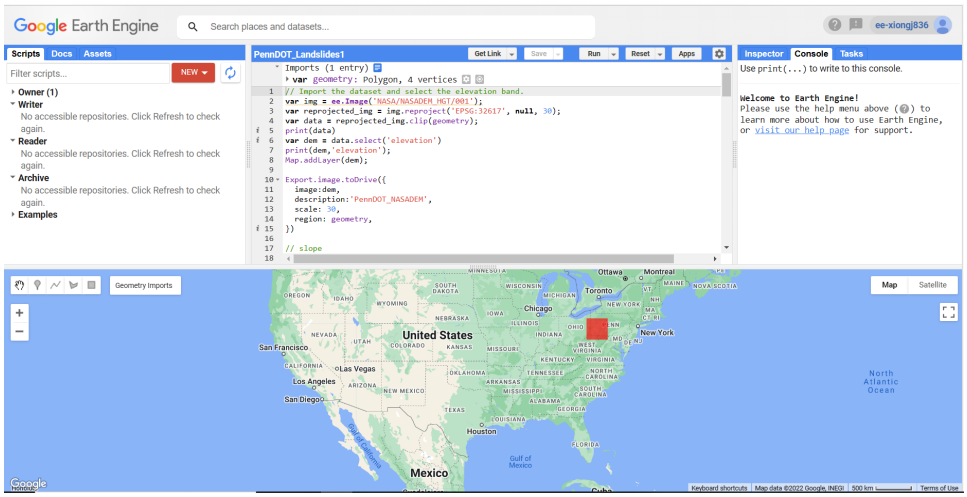


Figure 15. Interface of Google Earth Engine code editor.

The most fundamental terrain data are the elevation, slope, and aspect data, which are extracted from the NASA Digital Elevation Models (NASADEM). A Digital Elevation Model (DEM) is a representation of elevation data to represent the bare ground topographic surface of the Earth, excluding trees, buildings, and any other surface objects. Elevation can be obtained directly from the DEM dataset, while slope and aspect can be calculated based on DEM. NASADEM is a modernization of the DEM and associated products generated from the Shuttle Radar Topography Mission (SRTM) data and has an effective ground resolution of 30 m. Hence, the resolution of the factors derived from DEM is 30 m. The description of NASADEM from Google Earth Engine is shown in Figure 16.

The elevation, slope, and aspect raster data were downloaded for entire Pennsylvania, covering all landslide data points in the databases for spatial and spatiotemporal analyses. Since the area is too large to make a one-time download, the area was divided into three parts for data download separately, and all landslide data points are covered by the three segments. For example, the elevation data for the whole area is shown in Figure 17. For demonstration purposes, the contributing data of the landslide dataset for spatial analysis is shown next (within eight counties in southwestern Pennsylvania); the corresponding data of the landslide dataset for spatiotemporal analysis is similar and hence not shown herein. The raster data of elevation, slope, and aspect in the study area for spatial analysis are shown in Figures 18, 19, and 20, respectively.

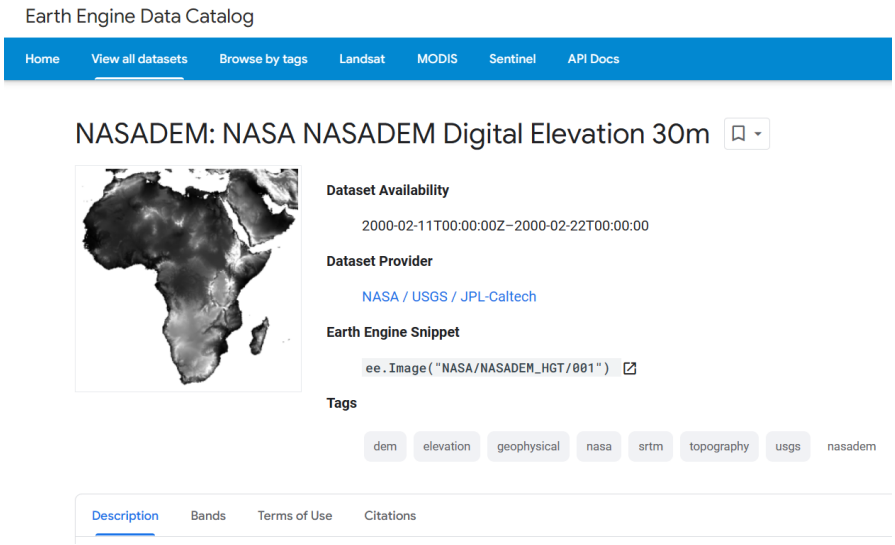


Figure 16. Dataset of NASADEM in Google Earth Engine.

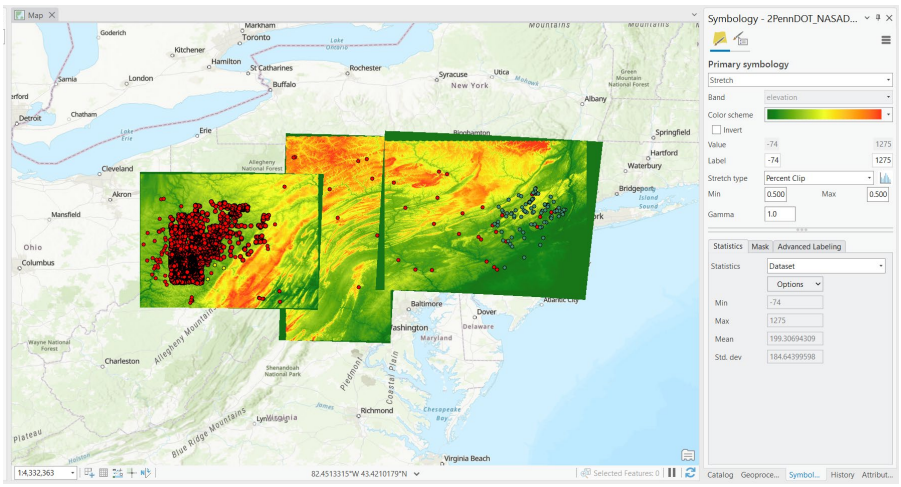


Figure 17. Elevation raster data for the whole study area.

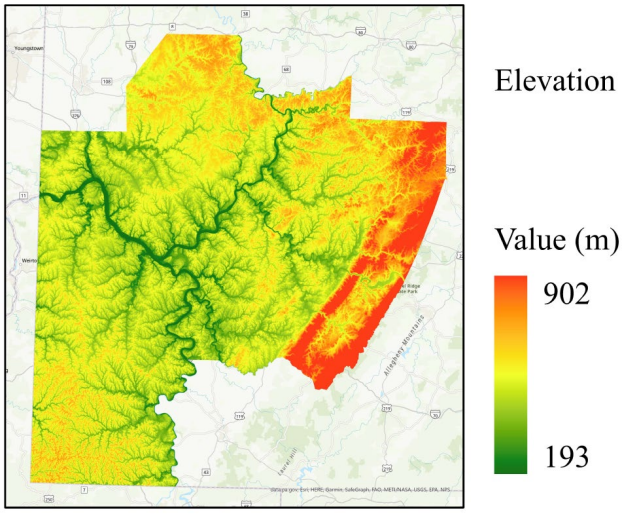


Figure 18. Elevation raster data for the study area for spatial analysis.

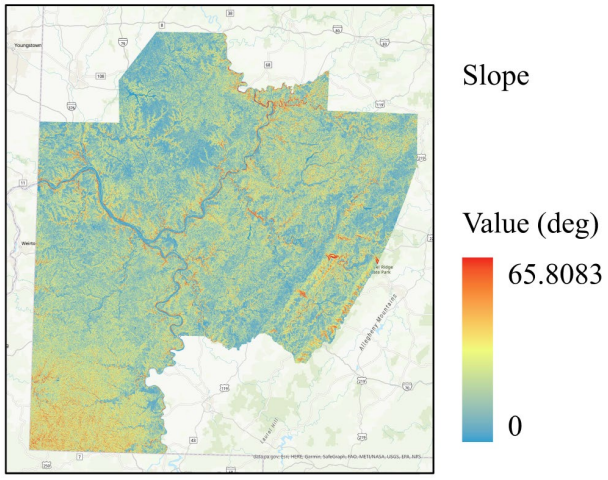


Figure 19. Slope raster data for the study area for spatial analysis.

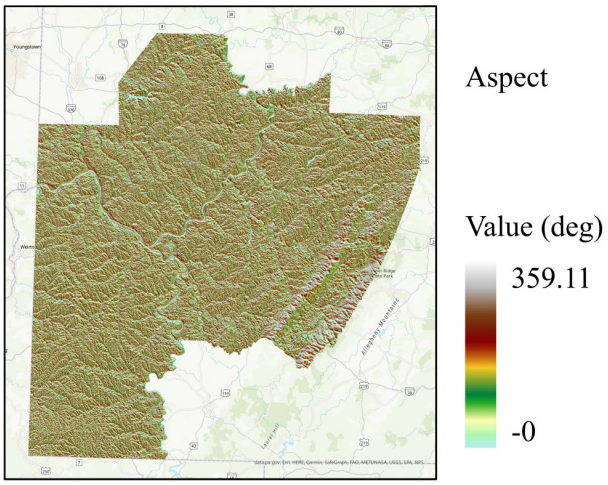


Figure 20. Aspect raster data for the study area for spatial analysis.

Topographic Position Index (mTPI) is relative elevation data and is calculated using elevation data for each location subtracted by the mean elevation within a neighborhood; hence, mTPI can distinguish ridges from valley forms. mTPI in the study area of the landslide database for spatial analysis is shown in Figure 21.

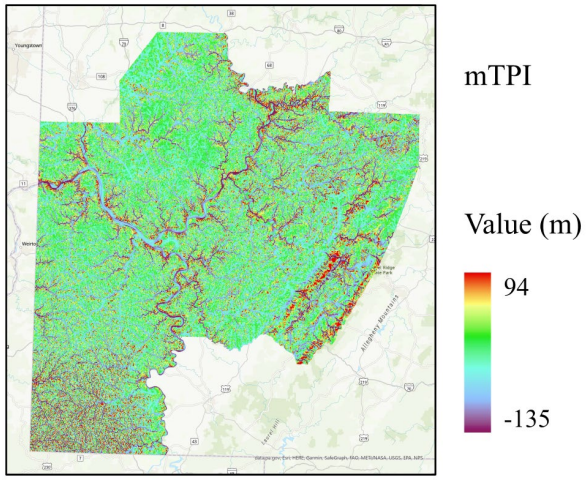


Figure 21. mTPI raster data for the study area for spatial analysis.

Profile and plan curvatures are calculated from the bathymetry surface for each raster cell using the ArcGIS Spatial Analyst "Curvature" Tool based on DEM data. They reflect the rate of change of curvature in different directions. The profile curvature is parallel to the direction of the maximum slope, while the plan curvature is perpendicular to the direction of the maximum slope. As important topographical factors, they contribute to the occurrence of landslides. Profile and plan curvatures in the study area for spatial analysis are shown in Figure 22 and Figure 23, respectively.

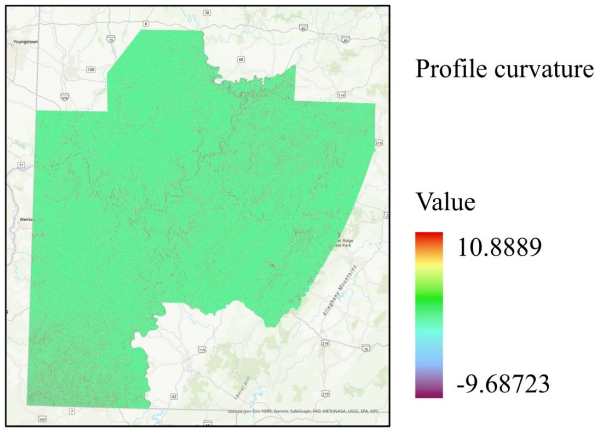


Figure 22. Profile curvature raster data for the study area for spatial analysis.

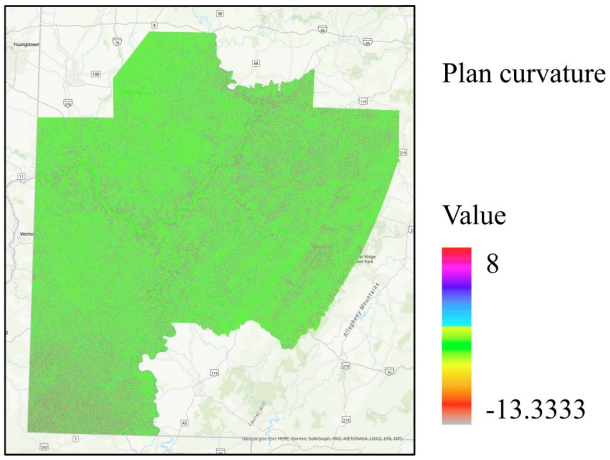


Figure 23. Plan curvature raster data for the study area for spatial analysis.

Topographic Wetness Index (TWI) is a significant causative factor that measures the degree of water accumulation at a location. Stream Power Index (SPI) is a measure of the erosive power of flowing water. They are calculated based on flow accumulation and the slope angle of each location. Flow accumulation is calculated using ArcGIS Spatial Analyst "Hydrology" Tool based on DEM data. TWI and SPI in the study area for spatial analysis are shown in Figures 24 and 25, respectively.

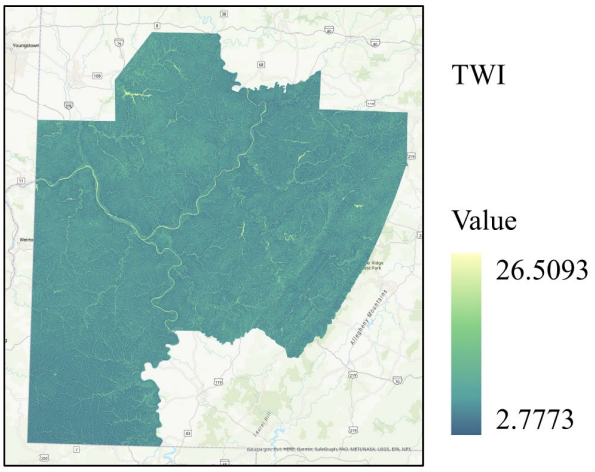


Figure 24. TWI raster data for the study area for spatial analysis.

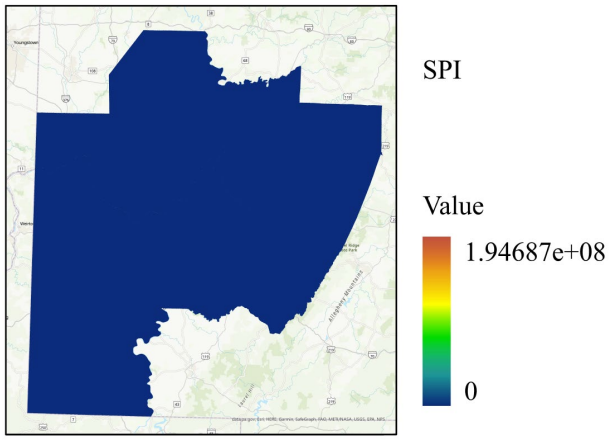


Figure 25. SPI raster data for the study area for spatial analysis.

Normalized Difference Vegetation Index (NDVI) is an index that researchers commonly use in remote sensing as it quantifies vegetation by measuring the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs). NDVI data were downloaded through Landsat 8 Collection 1 Tier 1 dataset, as shown in Figure 26. These composites are created from all the scenes in each 8-day period, beginning from the first day of the year and continuing to the 360th day of the year. The average NDVI value was calculated and downloaded through the code editor. NDVI in the study area for spatial analysis is shown in Figure 27.

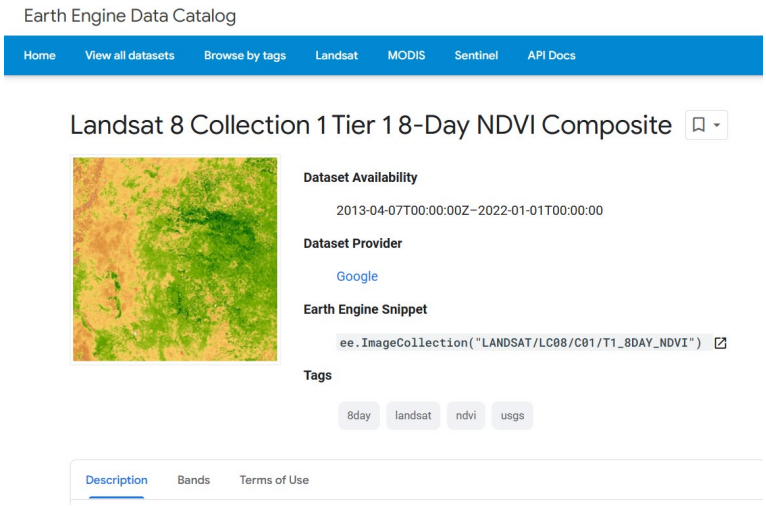


Figure 26. Dataset of Landsat 8 Collection 1 Tier 1 8-Day NDVI Composite in Google Earth Engine.

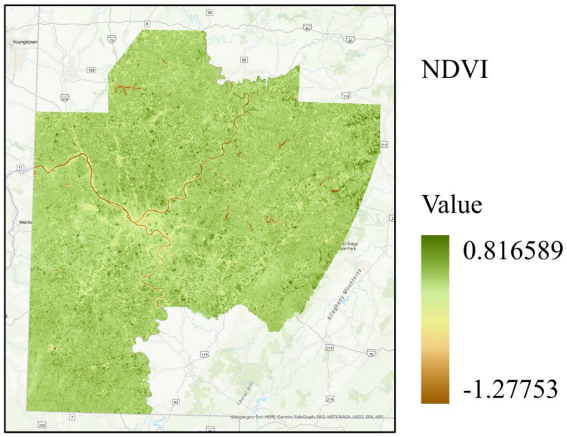


Figure 27. NDVI raster data for the study area for spatial analysis.

Sand content, clay content, soil bulk density, soil texture classification, and field capacity represent soil properties. Soil property data was obtained from the dataset of OpenLandMap in Google Earth Engine, as shown in Figure 28, and they have a spatial resolution of 250 m. The band downloaded in this study is b100, which represents the soil properties at 100-cm depth. Sand and clay contents and the bulk density of soil play important roles in the occurrence of landslides by affecting the slope weight and shear strength. Soil texture is a classification instrument used to determine soil classes based on their physical texture. The United States Department of Agriculture (USDA) soil taxonomy uses 12 textural classes to classify soil, considering the percentages of sand, silt, and clay in the soil (Figure 29). Since soil texture classification is a categorical variable, there are integer values from 1 to 12 in OpenLandMap, with each value representing a classification of soil as shown in Table 4. Field capacity is the amount of water content held in the soil after excess water has drained away and the rate of downward movement has decreased; the field capacity used in this study is soil water content for 33 kPa suctions at 100-cm depth. The raster data of soil properties in the study area are shown in Figures 30 through 34.

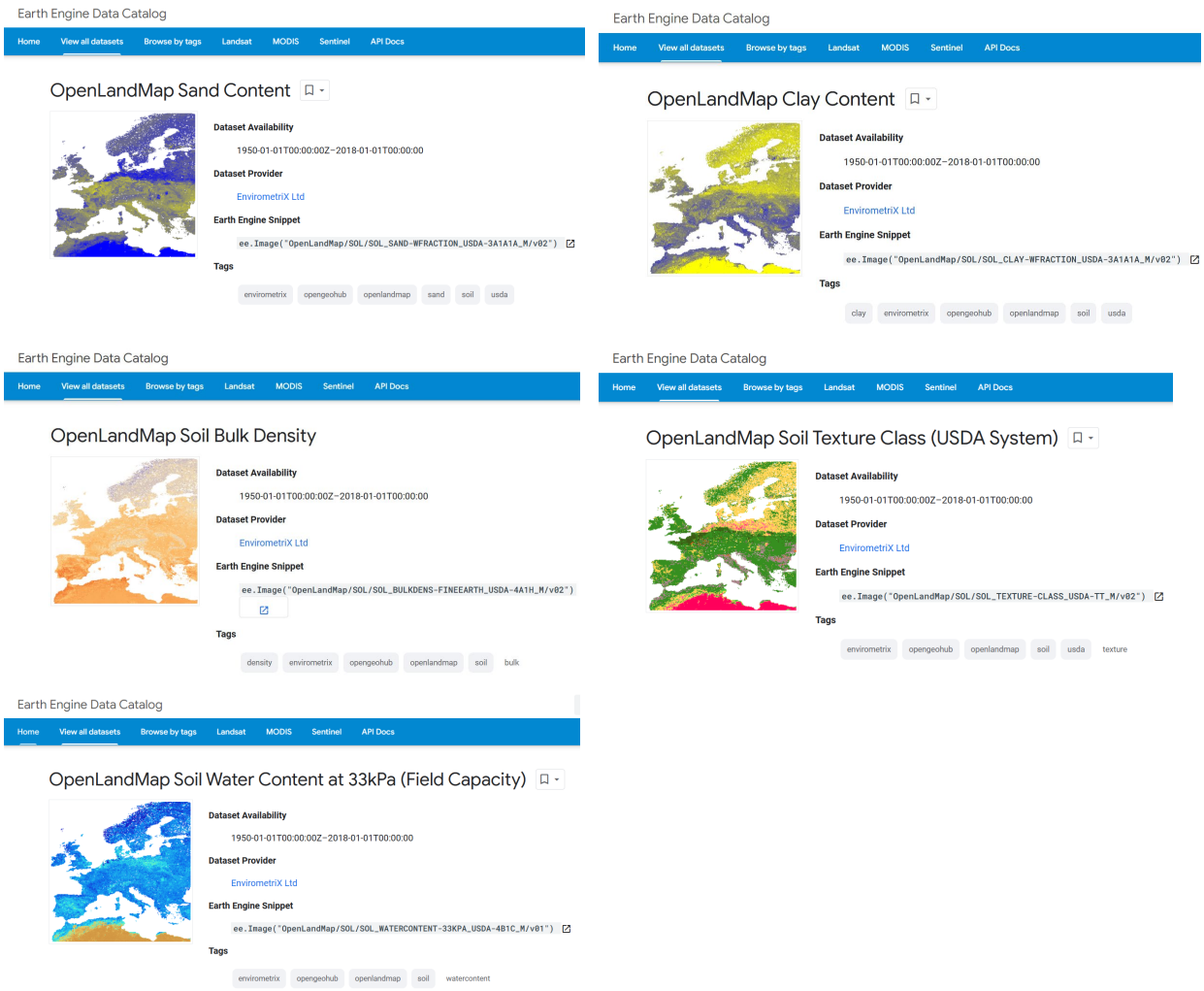


Figure 28. Datasets of OpenLandMap in Google Earth Engine.

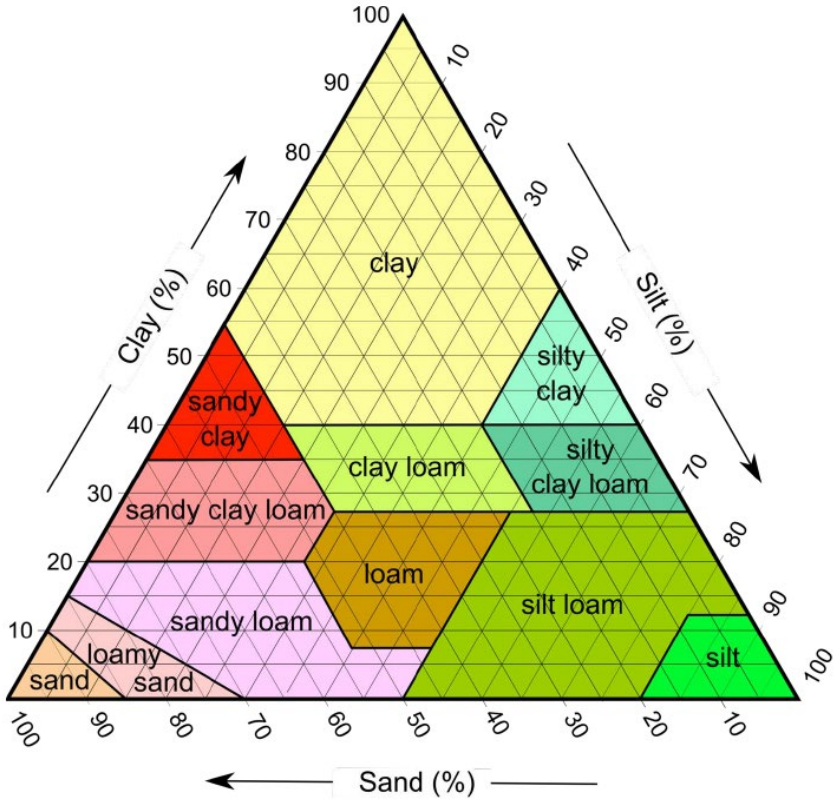


Figure 29. USDA soil textural classification triangle.

Table 4. Soil texture value and soil classification.

Soil texture value in OpenLandMap	Soil classification
1	Clay
2	Silty clay
3	Sandy clay
4	Clay loam
5	Silty clay loam
6	Sandy clay loam
7	Loam
8	Silt loam
9	Sandy loam
10	Silt
11	Loamy sand
12	Sand

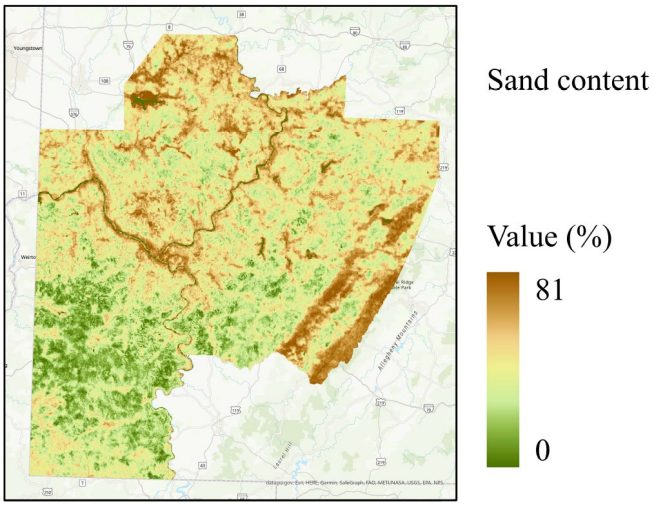


Figure 30. Sand content raster data for the study area for spatial analysis.

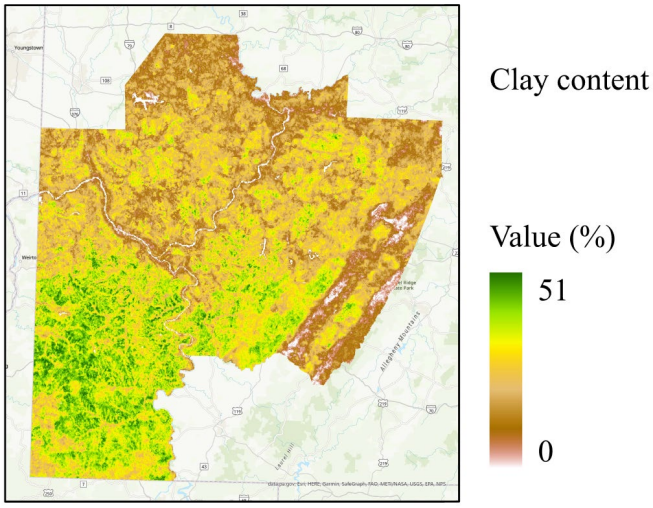


Figure 31. Clay content raster data for the study area for spatial analysis.

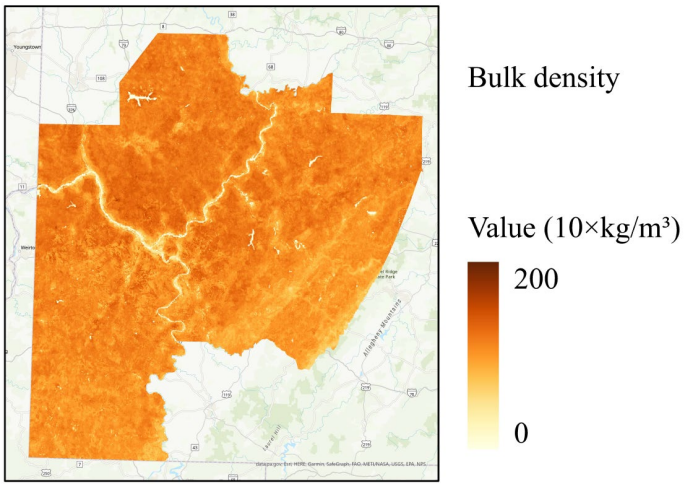


Figure 32. Bulk density raster data for the study area for spatial analysis.

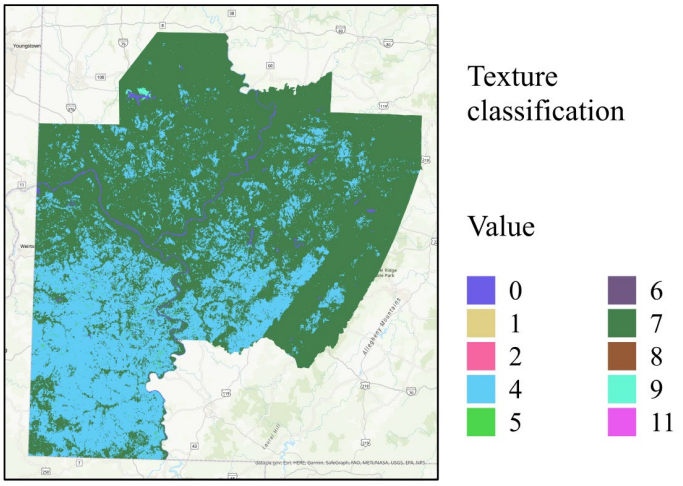


Figure 33. Soil texture classification raster data for the study area for spatial analysis.

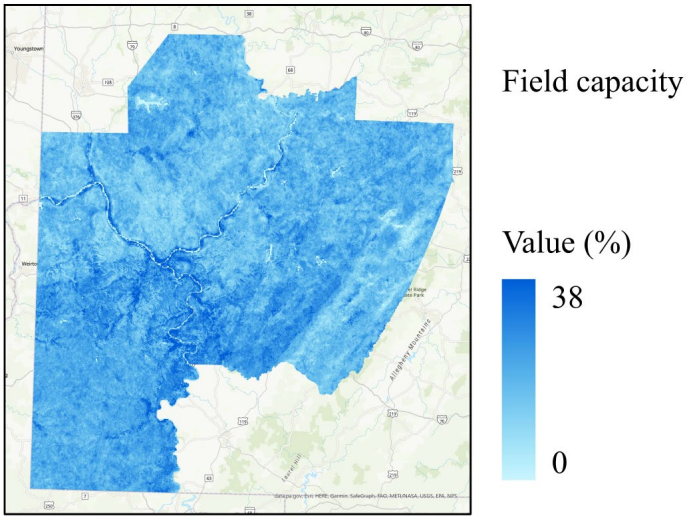


Figure 34. Field capacity raster data for the study area for spatial analysis.

The values of these soil properties are based on ML predictions from a global compilation of soil profiles and samples. OpenLandMap uses a compilation of published point data from various national and international soil point data providers to develop ML models for mapping soil properties and classes. The most important sources of training points include:

- USDA National Cooperative Soil Characterization Database,
- Africa Soil Profiles Database,
- LUCAS Soil database,
- Repositório Brasileiro Livre para Dados Abertos do Solo (FEBR),
- Sistema de Información de Suelos de Latinoamérica y el Caribe (SISLAC),
- The Northern Circumpolar Soil Carbon Database (NCSCD),
- Dokuchaev Soil Science Institute / Ministry of Agriculture of Russia (soil profiles for Russia),
- WHRC global mangrove soil carbon dataset,
- Local data sets such as Silva et al. (2019).

Additional points, if not available through these databases, have also been imported from the WoSIS Soil Profile Database (Batjes et al. 2017). Predictions are based on 3D ML ensemble models estimated using the SuperLearner and Caret packages (Hengl and MacMillan 2019). Data import, overlay, and model fitting to produce predictions of soil properties and classes are

explained by Hengl and MacMillan (2019). The general workflow for the generation of soil properties and classes using ML is shown in Figure 35.

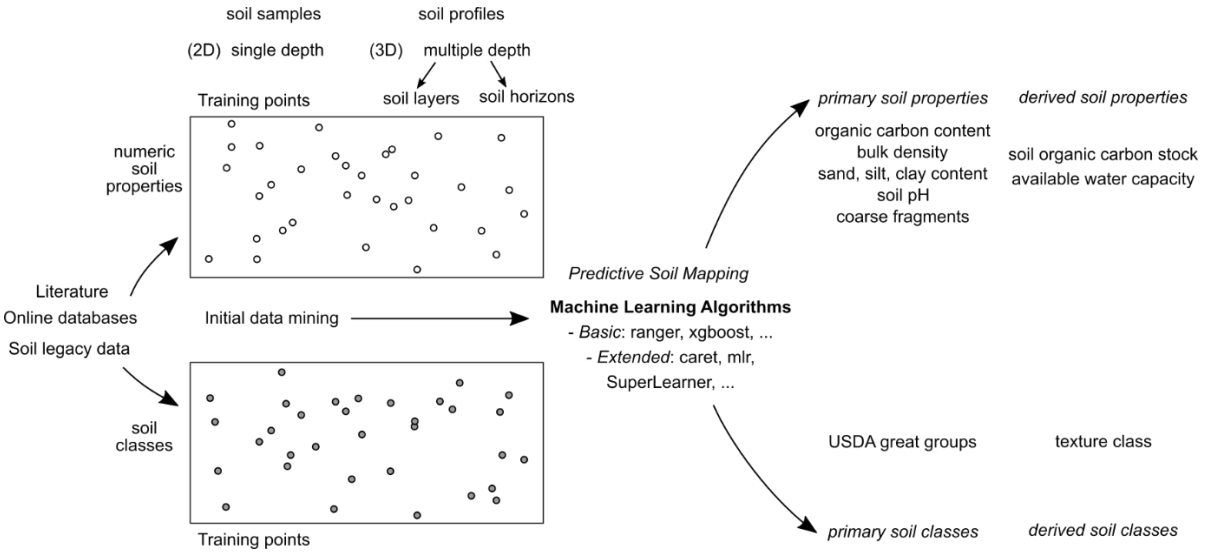


Figure 35. General workflow for generation of soil properties and classes using ML (from Hengl and MacMillan 2019).

In the present study, these causative factors of soil properties were downloaded or generated through Google Earth Engine and ArcGIS, and they were prepared as different raster layers in ArcGIS. The values of every factor are extracted for each landslide point in the landslide databases for spatial and spatiotemporal analyses. They are recorded as pertinent information together with the location, event date, and other information to develop the final landslide database in different platforms, as shown in Figures 8 through 13 in Section 2. It is noted that some landslide events have missing values of soil properties in the databases (almost all the values of soil properties for those cases are 0). Those missing values are null values based on ML prediction by OpenLandMap; hence, they will not be used as input in the ML of this study.

In addition to these fourteen contributing factors, rainfall data was also collected for the database to conduct spatiotemporal analysis. Daily rainfall amount data was downloaded from NASA Daily Surface Weather and Climatological Summaries (Daymet). Daymet provides long-term, continuous, gridded estimates of daily weather and climatology variables by interpolating and extrapolating ground-based observations through statistical modeling techniques. Through the

Single Pixel Extraction Tool of Daymet, daily rainfall amount data before each landslide event can be downloaded using the event date and longitude and latitude coordinates of the landslide location as input. An example of downloaded rainfall data is shown in Figure 36.

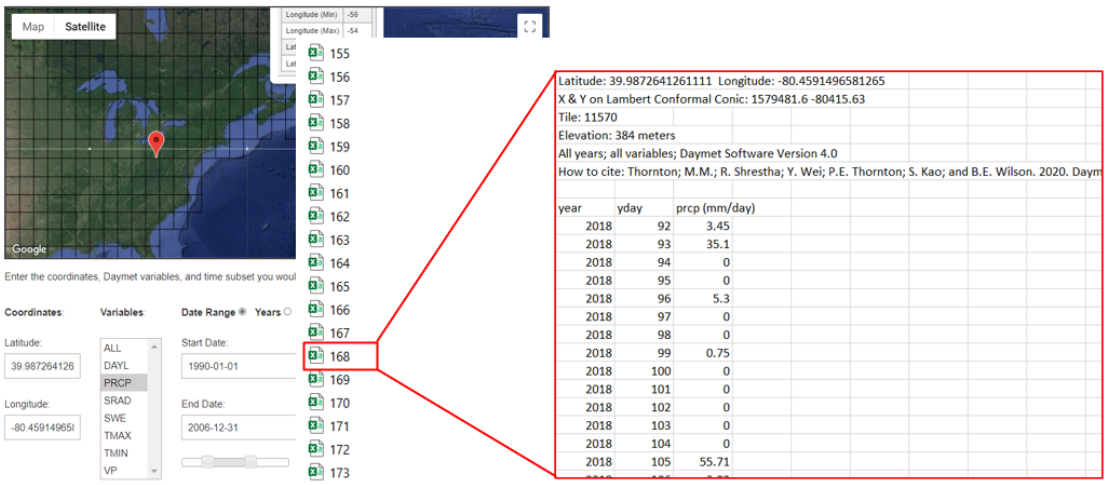


Figure 36. An example of rainfall amount data downloaded from Daymet.

4 Landslide Susceptibility Assessment

Landslide susceptibility is a measurement of the occurrence probability of landslides under certain geo-environmental conditions. Landslide susceptibility assessment is considered an important and effective approach for assessing landslide risk. The methods and techniques of landslide susceptibility mapping as well as the comparisons between them, have been widely studied (Yesilnacar and Topal 2005; Chacón et al. 2006; Lee et al. 2007; Yilmaz 2010; Akgun 2012; Youssef 2015; Wang et al. 2016).

The methods of landslide susceptibility assessment can be broadly categorized into three fundamental types: knowledge-guided methods, data-driven methods, and physics-based methods. Knowledge-guided methods are qualitative methods by weighting and ranking related landslide causative factors based on the knowledge of experts. Data-driven methods evaluate landslide risk by analyzing quantitative relationships between related geo-environmental characteristics and landslide occurrence. Physics-based methods predict landslide susceptibility based on the mechanisms and processes that control slope failures. All three types of methods have advantages and drawbacks. For knowledge-guided methods, subjectivity would be inevitably involved, while

physics-based methods are usually only effective for specific conditions. Generally, large scales of landslide susceptibility mapping involve more complex variables and larger amounts of data. Since data-driven methods benefit from ample and high-quality data, they are applied for landslide susceptibility mapping in the present study.

Data-driven methods broadly include statistical methods and ML methods. Statistical methods commonly used include the frequency ratio method (Lee and Talib 2005; Kannan et al. 2013), the weight of evidence method (Lee and Choi 2004; Thiery et al. 2007), the fuzzy logic method (Ilanloo 2011), information value method (Sarkar et al. 2013), and geographic information system (GIS) matrix method (Irigaray et al. 2007). ML methods commonly used include logistic regression (Ayalew and Yamagishi 2005; Wang et al. 2013), artificial neural network (Ermini et al. 2005; Tsangaratos and Benardos 2014), support vector machine (Yao et al. 2008), and random forest (Youssef et al. 2016). Among statistical methods, the frequency ratio method is one of the most popular and can outperform other methods, as shown in several case studies (Guo et al. 2015; Ding et al. 2016). In the present study, the frequency ratio method and several ML methods are applied to predict landslide susceptibility in southwestern regions of Pennsylvania, and an LSM is generated in the ArcGIS environment to serve as a tool for landslide risk assessment.

5 Frequency Ratio Method for LSM

To predict landslides, it is generally assumed that landslide occurrence is determined by related causative factors, and future landslides will occur under similar conditions. On this basis, the frequency ratio method analyzes the quantitative relationships between landslide inventory and related causative factors using probabilistic approaches. In the present study, the frequency ratio method is used to evaluate the level of correlation between the distribution of landslides and causative factors in the study area. The database for spatial analysis is used for the frequency ratio method, which contains 4,543 landslides in the format of the polygon as shown in Figure 7. Since the frequency ratio method was used to generate a preliminary LSM, only eight causative factors were selected for the analysis, which are shown in Table 5. As this study progressed, six additional factors were considered, and fourteen factors in total were included in ML methods to generate the final LSM, which is presented in Section 6.

Table 5. Causative factors used in frequency ratio method.

Number	Causative factor	Number	Causative factor
1	Elevation	5	Stream power index (SPI)
2	Slope	6	Normalized difference vegetation index (NDVI)
3	Aspect	7	Sand content
4	Topographic wetness index (TWI)	8	Clay content

5.1 Framework of frequency ratio method

In the case of landslide occurrence,the landslide-occurrence event is denoted by L, and a given factor's attribute is denoted by F. Given that the factor F is subdivided into n classes, the frequency ratio (FR) for the ith class of factor F (F_i) can be written as:

$$FR_i = \frac{\mathrm{P}(L_i)}{\mathrm{P}(F_i)} = \frac{\text{the frequency of landslides in the } F_i \text{ area}}{\text{the frequency of the } F_i \text{ area}}$$

$$= \frac{\text{the area of landslides in the } F_i \text{ area / the area of landslides in the study area}}{\text{the area of the } F_i \text{ area / the area of the study area}} \qquad \text{Eq. (1)}$$

A frequency ratio FR_i larger than 1 indicates that the frequency of landslides in the F_i area is larger than the frequency of the F_i area and further indicates that the ith class of factor F has a positive contribution to landslide occurrence. On the contrary, a frequency ratio FR_i smaller than 1 indicates that the ith class of factor F doesn't favor landslide occurrence.

To better demonstrate the statistical meaning of frequency ratio, Eq. (1) can be transformed as:

$$FR_i = \frac{\text{the area of landslides in the } F_i \text{ area / the area of the } F_i \text{ area}}{\text{the area of landslides in the study area / the area of the study area}}$$

$$= \frac{\text{the probability of landslides in the } F_i \text{ area}}{\text{the probability of landslides in the study area}} = \frac{\mathrm{P}(L|F_i)}{\mathrm{P}(L)} \qquad \text{Eq. (2)}$$

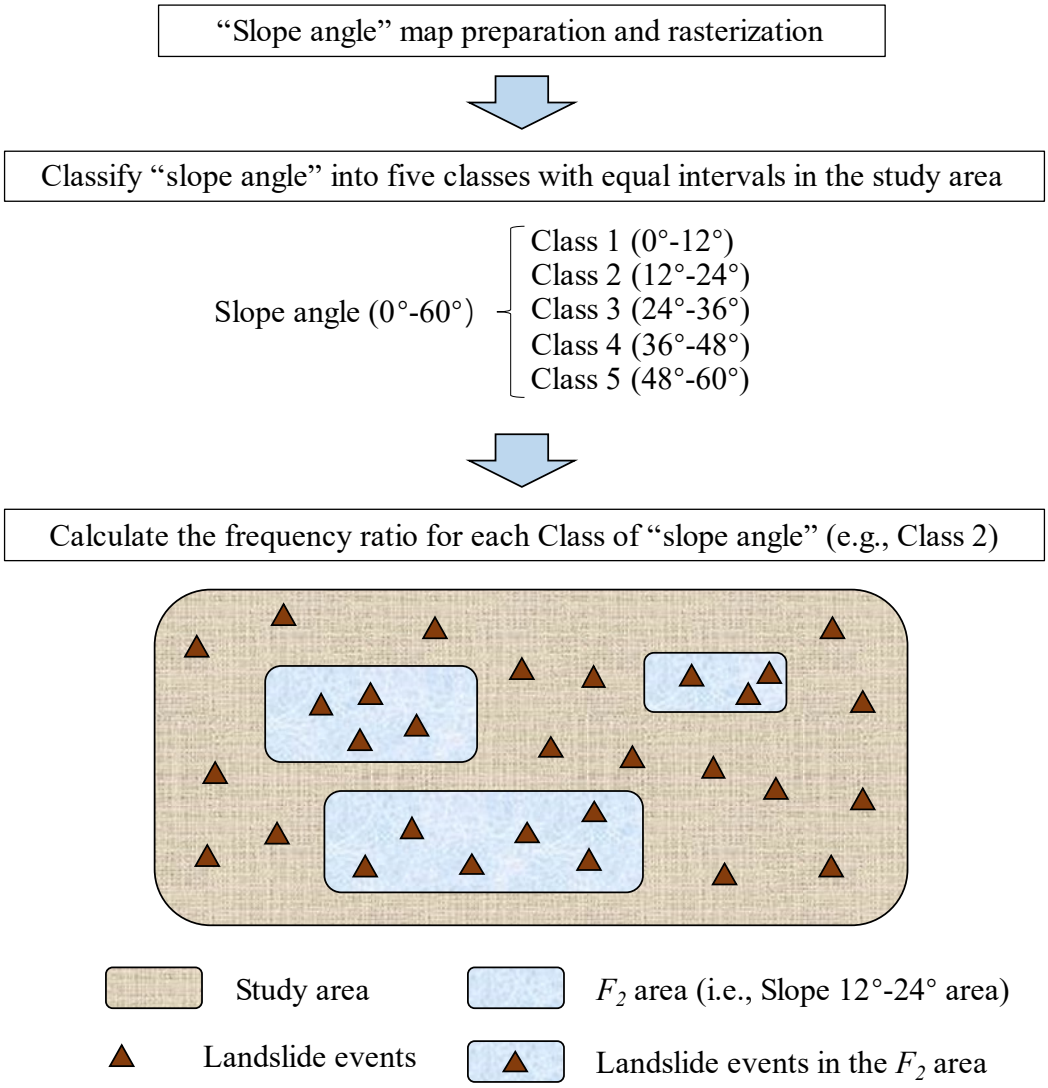
Since the probability of landslides in the study area $P(L)$ is predetermined by the landslide inventory, the frequency ratio (FR_i) is determined by the probability of landslides in the F_i area $P(L|F_i)$, which is the conditional probability of L given F_i. Hence, a higher conditional probability $P(L|F_i)$ means that there is a higher probability of landslides occurrence in the ith class of factor F (F_i), and that would be reflected with a higher frequency ratio (FR_i). When the frequency ratio value is greater than one, it indicates a strong correlation between the factor's class and landslide occurrence, while a value smaller than one indicates a weak correlation.

Take slope angle, one of the causative factors, as an example to demonstrate the calculation process of the frequency ratio. As Figure 37 shows, the first step after preparation and rasterization of data is to classify slope map into five classes with equal intervals in the study area. Eq. (1) or Eq. (2) can then be used to calculate the frequency ratio (FR) for each class of the slope map.

Considering different landslide causative factors $F^{(j)}$ (j=1, 2, …, m), their frequency ratio with n class $FR_i^{(j)}$ (i=1, 2, …, n; j=1, 2, …, m) can be calculated according to Eq. (2). That means if the class of $F^{(j)}$ at a particular location is $F_i^{(j)}$, the frequency ratio of this factor at this location will be $FR_i^{(j)}$. Therefore, a landslide susceptibility index (LSI) was introduced for any given location, which is the summation of the frequency ratios of different causative factors at this location (Lee and Pradhan 2007):

$$\mathrm{LSI} = \sum_{j=1}^{m} FR^{(j)} \qquad \text{Eq. (3)}$$

A high LSI value indicates a high risk of landslide occurrence at the location. Hence, LSI can also be represented as the landslide hazard index (LHI) in the study of landslide susceptibility mapping (Pradhan and Lee 2009). To obtain the final landslide susceptibility map, the LSI map is then reclassified into several classes with equal intervals to distinguish the areas with different landslide susceptibility.



$$FR_2 = \frac{\text{the area of landslides in the } F_2 \text{ area /the area of landslides in the study area}}{\text{the area of the } F_2 \text{ area / the area of the study area}} \quad \text{(Eq. 1)}$$

Or

$$FR_2 = \frac{\text{the area of landslides in the } F_2 \text{ area / the area of the } F_2 \text{ area}}{\text{the area of landslides in the study area / the area of the study area}} \quad \text{(Eq. 2)}$$

Figure 37. Calculation process of the frequency ratio for slope angle.

In the present study, there are six essential steps in applying the frequency ratio method, which are shown in Table 6.

Table 6. Steps in applying the frequency ratio method.

Steps	Description
1	Preparation of landslide inventory map and landslide causative factors map.
2	Rasterization of all maps including landslide inventory and causative factors.
3	Classifying all causative factor maps.
4	Calculating the frequency ratio of each causative factor.
5	Summing up all frequency ratio maps to obtain the landslide susceptibility index.
6	Reclassifying the landslide susceptibility index into different classes to generate landslide susceptibility map.

The framework of the frequency ratio method applied in this study is shown in Figure 38.

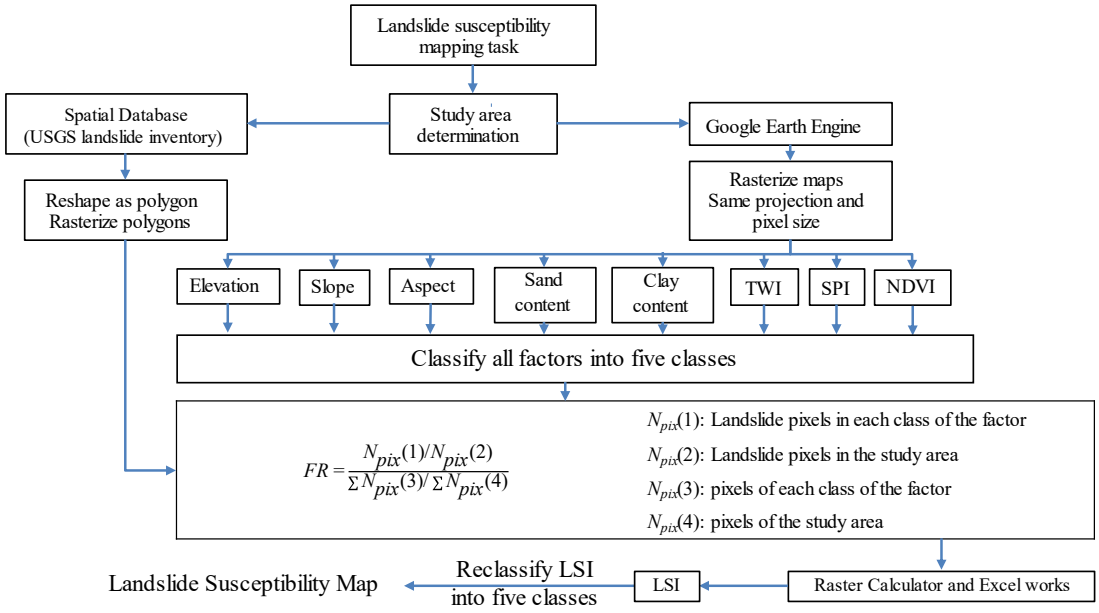


Figure 38. Flowchart of the frequency ratio method applied in this study.

5.2 Results of the frequency ratio method

Since landslide polygons and all causative factors are rasterized in ArcGIS and the size of every pixel is 30 m×30 m, the area of landslides and factors is represented by the number of pixels, which can be calculated using tools readily available in ArcGIS. To calculate the frequency ratio, an Excel table (see Table 7) was constructed for each landslide causative factor, and Eq. (1) was used for the calculation of the frequency ratio. First, the area ratios for landslide occurrence within each factor's class to the total landslide area in the whole study area were calculated, then the area ratio for each class of factors to the total study area was calculated. Finally, the frequency ratios for each class of factors were calculated by dividing the landslide occurrence ratio by the factors' class ratio.

Table 7. Frequency ratios of factors to landslide occurrences.

Factor	Class	Class pixels	Class pixels (%)	Landslide pixels	Landslide pixels (%)	FR
Elevation	193.0 ~ 334.8 m	4411804	25.73	10048	30.68	1.19
	334.8 ~ 476.6 m	11415708	66.58	22667	69.21	1.04
	476.6 ~ 618.4 m	1082468	6.31	32	0.10	0.02
	618.4 ~ 760.2 m	137557	0.80	3	0.01	0.01
	760.2 ~ 902.0 m	98537	0.57	0	0	0
Slope	0 ~4.65°	5090411	29.69	1497	4.57	0.15
	4.65 ~ 8.77°	5591795	32.61	4753	14.51	0.45
	8.77 ~ 13.68°	3971571	23.16	10899	33.28	1.44
	13.68 ~ 20.13°	1976044	11.52	11195	34.18	2.97
	20.13 ~ 65.81°	516253	3.01	4406	13.45	4.47
Aspect	0 ~ 65°	3227426	18.82	7243	22.12	1.17
	65 ~ 138°	3458279	20.17	8113	24.77	1.23
	138 ~ 210°	3505869	20.45	5024	15.34	0.75

	210 ~ 280°	3646002	21.26	5576	17.03	0.80
	280 ~ 360°	3308498	19.30	6794	20.75	1.08
Sand content	0 ~ 16.2 %	157620	0.92	109	0.33	0.36
	16.2 ~ 32.4 %	6967438	40.63	22331	68.30	1.68
	32.4 ~ 48.6 %	9908434	57.79	10226	31.28	0.54
	48.6 ~ 64.8 %	112599	0.66	30	0.09	0.14
	64.8 ~ 81 %	534	0	0	0	0
Clay content	0 ~ 10.2 %	155445	0.91	106	0.32	0.36
	10.2 ~ 20.4 %	2296563	13.39	2027	6.20	0.46
	20.4 ~ 30.6 %	12239139	71.38	16859	51.56	0.72
	30.6 ~ 40.8 %	2455116	14.32	13701	41.90	2.93
	40.8 ~ 51 %	362	0.002	3	0.01	4.35
TWI	2.78 ~ 6.22	7043427	41.08	16862	51.49	1.25
	6.22 ~ 7.90	6569230	38.31	11465	35.01	0.91
	7.90 ~ 10.50	2369547	13.82	3257	9.95	0.72
	10.50 ~ 14.13	902037	5.26	967	2.95	0.56
	14.13 ~ 26.51	261833	1.53	199	0.61	0.40
SPI	0 ~ 3.9e7	5265707	30.75	5289	16.16	0.53
	3.9 ~ 7.8e7	1623307	9.48	1015	3.10	0.33
	7.8 ~ 11.7e7	9848630	57.51	25818	78.86	1.37
	11.7 ~ 15.6e7	368886	2.15	612	1.87	0.87
	15.6 ~ 19.5e7	17892	0.10	3	0.01	0.09
NDVI	-1.28 ~ -0.86	157152	0.92	30	0.09	0.10
	-0.86 ~ -0.44	745431	4.35	521	1.59	0.37
	-0.44 ~-0.02	3679572	21.46	6561	20.02	0.93

	-0.02 ~ 0.40	10564152	61.61	21812	66.55	1.08
	0.40 ~ 0.82	2000203	11.67	3853	11.76	1.01

Table 7 shows the relationships between eight causative factors and landslide occurrence. According to the statistical meaning of frequency ratio, a ratio larger than 1 suggests a strong relationship between the factor and landslide occurrence. For example, for slope angles below 8.77°, the frequency ratios are smaller than 1, indicating a low probability of landslide occurrence; while the frequency ratios of slopes larger than 8.77° are larger than 1, indicating a high probability of landslide. After calculating frequency ratios for all causative factors, LSI was calculated using Eq. (3) to get a summation of each factor's frequency ratio. The LSI values were then classified into different levels of landslide susceptibility zones using equal breaks in the ArcGIS tool.

The landslide susceptibility map using frequency ratios for the selected causative factors is generated in ArcGIS, as shown in Figure 39. The susceptibility map is classified into five classes based on LSI with equal intervals, representing very low, low, moderate, high, and very high risks corresponding to different landslide occurrence probabilities. The correlation between the susceptibility classes and the probability of landslide occurrence is shown in Table 8.

Table 8. Correlation between susceptibility classes and probability of landslide.

Susceptibility classes	Probability of landslide
Very low	0% - 20%
Low	20% - 40%
Moderate	40% - 60%
High	60% - 80%
Very high	80% - 100%

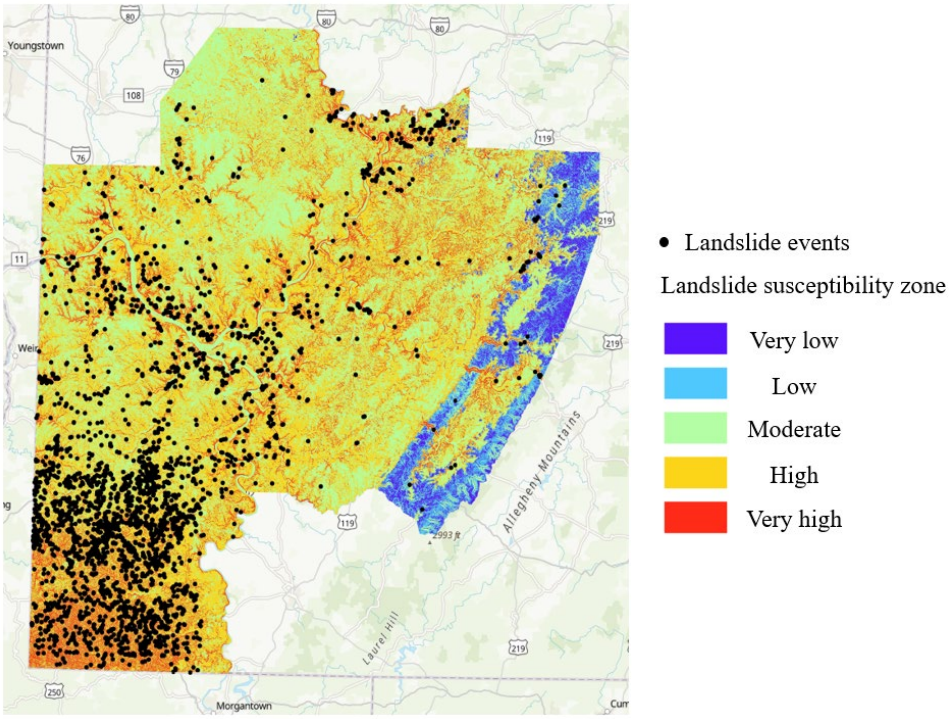


Figure 39. LSM using the frequency ratio method.

It is shown that the susceptibility map derived using the frequency ratio method corresponds to the location of actual landslide occurrences very well. The red zones, which represent a very high risk of landslide occurrence, follow a similar distribution as the recorded landslides in the study area. Through frequency ratio analysis, a preliminary landslide susceptibility map was created, and the level of correlations between the selected causative factors and landslide occurrence was verified. The preliminary susceptibility map provides a basis for landslide susceptibility mapping using ML methods, the power of which has been shown in many studies in recent years.

6 ML Methods for LSM

In recent years, the big data era has brought enormous benefits to society and different sectors. Big data refers to the vast amount of data that can be obtained or generated with advanced techniques. ML techniques, as an important subfield of AI, have achieved considerable success in various domains, from computer science to commercial fields. With the power of different mathematical algorithms and statistical techniques, ML can extract information from tremendous

data and imitate the way that humans learn to deal with different tasks. In addition to traditional fields related to data and computer systems, ML is becoming more significant in the fields of natural science and engineering.

In geotechnical engineering, the technical explosion of ML has promoted its application to landslide susceptibility mapping as well. For example, Reichenbach et al. (2018) reviewed published works on various aspects of landslide susceptibility mapping with ML; based on the comparisons of modeling approaches and model evaluation criteria, they provided recommendations for the preparation, selection, and evaluation of ML models for landslide susceptibility mapping. Merghadi et al. (2020) summarized the popular ML techniques available for landslide susceptibility mapping and highlighted the advantages and disadvantages of each model with a case study in Algeria. Naemitabar et al. (2021) analyzed the selection of effective landslide causative factors and compared four ML algorithms for landslide susceptibility mapping. Moziihrii et al. (2022) conducted a comprehensive literature survey and showed the current trend of landslide susceptibility mapping using ML techniques. The survey analyzed published works, including the studies of ML models, landslide causative factors, study location, datasets, evaluation methods, and model performance.

Previous studies indicate that ML-based methods are effective for assessing complex relationships between landslide occurrence and causative factors. Typically, the capability to deal with vast amounts of data makes ML techniques suitable for the tasks associated with regional landslide susceptibility assessment compared to other physical methods.

6.1 ML algorithms

ML algorithms are mathematical models that allow people to explore, analyze, and find meaning in complex data sets. Each algorithm is implemented as a set of unambiguous step-by-step instructions that a program can follow to achieve a particular goal. For ML tasks, the general goal is to discover or establish patterns that people can use to make predictions on quantities or categories. Hence, the two most fundamental and important types of algorithms in ML are regression and classification. Regression algorithms predict a continuous value based on the input data. For example, if the target variable that needs to be predicted is a quantity like income, scores, height, or weight, regression models can be very effective. Classification algorithms predict

discrete class labels based on the input data. For example, image recognition systems can automatically classify different items in one image. Email spam detection systems can classify an email as spam or non-spam based on the content of the mail. Given a person's symptoms, an ML-based disease diagnosis can classify the person as suffering from a disease or not. For regression and classification algorithms, there are many conventional and advanced algorithms born with the development of computer technology.

In the present study, landslide prediction is treated as a binary classification problem. The objective is to predict whether a landslide will occur in a specific location based on input data of landslide causative factors. Four ML models are used in the study, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machines (GBM). These models are commonly used in geotechnical fields and were chosen due to their applicability, small biases, and reasonable results in previous studies. For example, Ayalew et al. (2005) applied LR to predict landslide occurrence and distribution in the Kakuda-Yahiko mountains in central Japan. Ballabio et al. (2012) analyzed the application of SVM to landslide susceptibility mapping in the Staffora river basin in Italy. Zhang et al. (2023) compared the performance of RF and GBM applied to landslide susceptibility mapping in Fengjie County in southwestern China.

LR accomplishes binary classification tasks by predicting the probability of an event; it analyzes the relationship between independent variables and classifies data into binary classes by using a sigmoid function to map probabilities. SVM is based on the principle that minimizes errors associated with the training dataset and maximizes the generalization of the model (Vapnik 1995). The main idea behind SVM is to find a hyperplane that maximally separates the different classes in the training data. RF is an ensemble learning algorithm based on the decision tree (DT) algorithm. It solves the problem by collecting the results from different DT models built randomly. The algorithm selects a feature subset of examples to develop different DTs during model training. After the creation of DTs, each DT makes a separate prediction, and RF considers the mean of those separate predictions to make the final prediction (Ho 1995). GBM is also one of the ensemble learning algorithms where multiple weak models are created first and then are combined to yield better performance (Natekin and Knoll 2013). Boosting works as it reduces error with each additional weak learner into a strong learner sequentially to correct its predecessor. The above four

models are widely used for classification problems. According to the different mechanisms behind ML algorithms, the performance of an ML model would vary when applied to different datasets and tasks. Therefore, no model is always the best. In the present study, all four models are established, and the performance is compared based on different evaluation methods.

6.2 Landslide database for ML

The database compiled for ML contains 4,543 landslides from USGS Topography sheets. Different than the eight causative factors consisered in the frequency ratio method, fourteen landslide causative factors (see Table 3) are considered in building the ML model for LSM. To reduce the bias caused by a high concentration of landslides in the southwestern region of the study area, 3000 landslides in the study area are randomly chosen as the database for ML. For binary classification problems, ML algorithms require both positive and negative samples with features so that the model can be trained to distinguish the pattern of features for different classes. Therefore, 3000 non-landslides with the same number as landslides are sampled in the study area, as shown in Figure 40. Non-landslides are randomly sampled in the study area outside circular buffers set around landslides. The buffer ensures that non-landslide will not be sampled within 500 m around a landslide to avoid the possibility of coinciding with landslide locations in sampling.

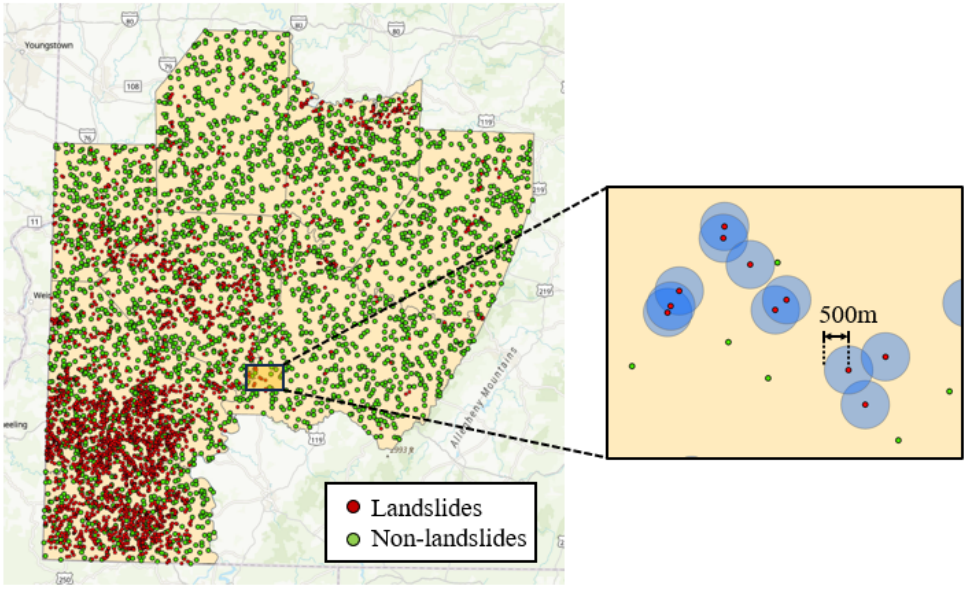


Figure 40. Landslide and non-landslide samples in the study area.

Positive and negative samples are usually represented by labels 1 and 0, respectively, for ML tasks; hence, the database compiled for landslide classification contains 3000 positive landslide samples with a label of 1 and 3000 non-landslide samples with a label of 0. Both positive and negative samples are associated with the fourteen landslide causative factors. A portion of the database is shown in Figure 41.

TWI	SPI	Plan_curvature	Profile_curvature	Aspect	Clay_content	mTPI	Elevation	Sand_content	Slope	Soil_texture	Field_capacity	Bulk_density	NDVI	label
-	-	-	-	degree	%	m	m	%	degree	-	%	10 × kg/m^3	-	
4.99238	12.1911	0.045045	1.3783799	100.107	21	17	246	44	22.1158	7	26	120	0.291498	1
7.107127	0	-0.0277778	-0.0277778	315.655	21	-1	266	39	1.350655	7	25	129	0.3255065	0
5.68224	6.10066	0.225951	-0.107383	145.597	22	-25	238	43	11.4949	7	24	130	0.237815	1
6.3491	31.3498	-0.0666667	0.933333	297.216	21	-44	232	42	14.6414	7	24	130	0.20215	1
6.94507	17.2207	-0.18994	0.0322823	126.207	23	-18	256	39	8.16667	7	22	134	0.347043	1
5.720965	0	0.1666667	-0.0555556	59.77807	20	-14	222	44	5.557273	7	31	136	0.2183223	0
11.0021	0	0	0	0	22	-14	216	41	0	7	33	141	-0.1006884	0
6.322749	3.201004	0.0433604	-0.2899729	219.4045	23	-14	223	46	6.090526	7	28	141	0.0875986	0
12.10071	0	0	0	0	19	-11	235	45	0	7	32	141	0.0325174	0
5.767116	5.601785	0.168	0.5013334	100.9125	19	-10	236	45	10.57699	7	32	141	0.0569549	0
6.665237	0	0.1111111	-0.2222222	27.29059	19	-19	217	40	2.132382	7	27	142	0.0422743	0
6.665237	0	0.3555556	0.0222222	153.9813	19	-19	217	40	2.132382	7	27	142	0.258859	0
6.326734	0	0.0777778	-0.1444445	162.1451	24	-11	234	38	3.013085	7	30	143	0.0629181	0
7.6985	403.962	-0.180719	0.70817	328.09	19	-29	238	48	23.479	7	26	144	0.232276	1
6.912302	64.26605	0.1063554	-0.6714224	274.2828	23	8	264	36	14.99095	7	23	144	0.2427415	0
8.721562	15.84019	-0.1277778	0.0944444	252.3127	21	-27	228	45	3.022497	7	32	144	0.2011265	0
7.446209	21.88198	-0.2180294	0.5597484	254.7515	25	-4	235	42	6.931355	7	28	144	0.0900959	0
7.109727	14.58812	0.0649895	0.1761006	254.7895	20	-27	230	45	6.931419	7	27	144	0.0912877	0
6.205274	3.603761	-0.1111111	-0.3333333	214.4067	18	19	281	43	6.849968	7	22	145	0.2929403	0

Figure 41. Input database compiled for ML.

When building an ML model, the model is first trained on a training dataset. The trained model is then evaluated on whether it can perform well on unseen data. Hence, it is a common approach to split the original dataset into training and testing subsets to check if the ML model performs well on data that it has not seen. A split ratio of 80%/20% is applied in this study, which means 80% of the original data is chosen for model training and the remaining 20% is used to check the performance of the model applied to unseen data.

However, the drawback of using only one split of training and testing sets is that the ML model performance can vary greatly depending on which samples were used in the training and testing sets. One way to avoid this problem is to build a model several times using different training and testing sets each time, then calculate the performance to be the average of all test results. This general method is known as cross-validation. Through cross-validation, the original dataset is divided into k folds, k-1 folds are used as the training set, and the remaining one fold is used as the testing set. The process is repeated k times so that each fold is used as a testing fold, and the final performance of the model is evaluated by the average performance for each testing fold. In

this study, five-fold cross-validation, corresponding to an 80%/20% data split, was used. Figure 42 illustrates the cross-validation procedure in this study.

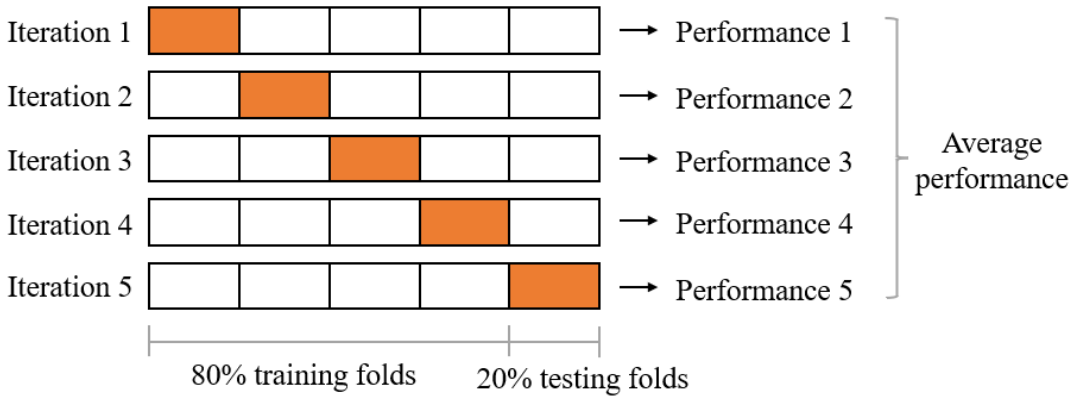


Figure 42. Five-fold cross-validation procedure in ML.

6.3 Evaluation methods

For classification problems, there are many evaluation methods that can be adopted to assess the performance of ML models. In the present study, Accuracy, Precision, Recall, F1 score, and AUC score are used for the evaluation of ML algorithms in landslide binary classification.

Accuracy is the fraction of all predictions that the model gets right. Formally, Accuracy has the following definition:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \qquad \text{Eq. (4)}$$

For binary classification problems, Accuracy can also be calculated in terms of positives and negatives as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Eq. (5)}$$

where TP=True positives, TN=True negatives, FP=False positives, and FN=False negatives. These positive and negative values can be determined by the confusion matrix (Figures 43 through 46), which is used for evaluating the performance of a classification model as it compares the actual

target values with those predicted by the ML models. Accuracy, as shown in Figure 43, summarizes the performance of a classification model as the number of correct predictions divided by the total number of predictions, so it reflects the overall performance of the model.

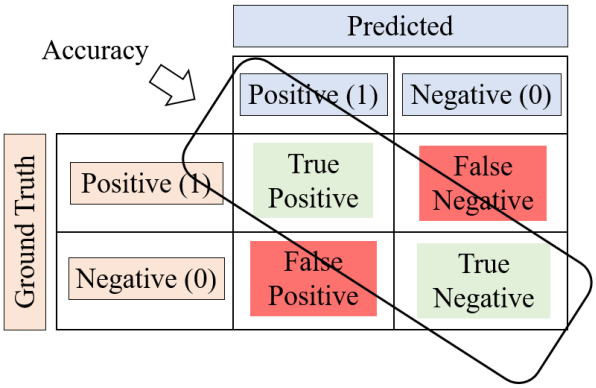


Figure 43. Confusion matrix for binary classification and definition of Accuracy.

Precision is a measure of correctness that is achieved in positive prediction. It tells how many predictions are actually positive out of the total positive predicted. Precision is defined as the ratio of the total number of correctly classified positive classes divided by the total number of predicted positive classes (see Figure 44).

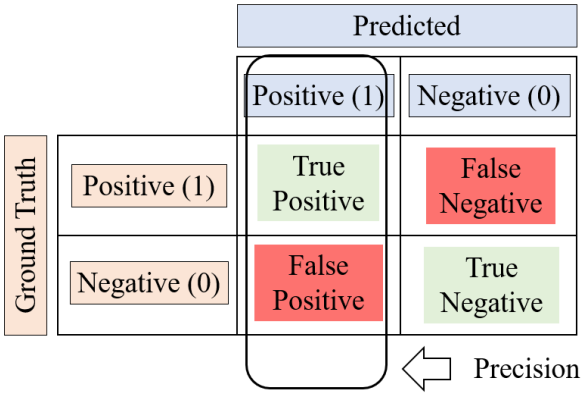


Figure 44. Confusion matrix for binary classification and definition of Precision.

Hence, Precision is calculated as:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Eq. (6)}$$

Precision reflects the reliability of the prediction model. If a model always predicts the occurrence of landslides, it might produce many false positives, making the model have a low Precision value and thus unreliable for users. Although a model with low Precision will be unlikely to miss positive events, the model is unreliable and will cause a waste of time, money, manpower, and public anxiety in the case of landslide predictions.

Recall, on the other hand, is a measure of actual observations that are predicted correctly, i.e., how many observations of positive class are actually predicted as positive. It is also known as Sensitivity. It is defined as the ratio of the total number of correctly classified positive classes divided by the total number of positive classes (see Figure 45).

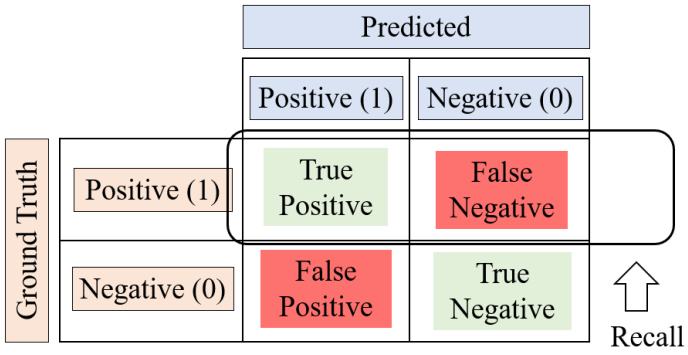


Figure 45. Confusion matrix for binary classification and definition of Recall.

Recall is calculated as:

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{Eq. (7)}$$

Recall considers all the positive samples and how many of them are identified correctly. Recall is important to applications in the context of medical diagnostics or severe hazards because missing positive class will come with serious consequences.

However, Precision and Recall are often in a collision, which means improving one score can come at the cost of decreasing the other. Given that there is a trade-off between Precision and Recall, the F1 score is introduced as the harmonic mean of Precision and Recall, representing a balance between them. F1 score is a number between 0 and 1, where 0 is the worst possible score (i.e., the model predicts all observations incorrectly) and 1 is a perfect score (i.e., the model predicts all observations correctly). F1 score is calculated as:

$$F1 = 2\times\frac{(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})} \qquad \text{Eq. (8)}$$

For a binary classification problem, many classification algorithms like LR use probability threshold to distribute samples into classes. In most cases, the probability threshold defaults to 0.5. That means the algorithm classifies a sample as positive if the probability of being positive is above 0.5 and negative if the probability is below 0.5. However, the probability thresholds are not necessarily 0.5 in many cases like medical diagnostics or severe hazards, where it is more rational to choose a low probability threshold to prevent any chance of the positives being misclassified. A receiver operating characteristic curve (ROC curve) is a curve showing the performance of the classification model for all probability thresholds. This curve plots two parameters: true positive rate (TPR) and false positive rate (FPR), as shown in Figure 46. TPR is the same as Recall, while FPR is defined as:

$$\text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \qquad \text{Eq. (9)}$$

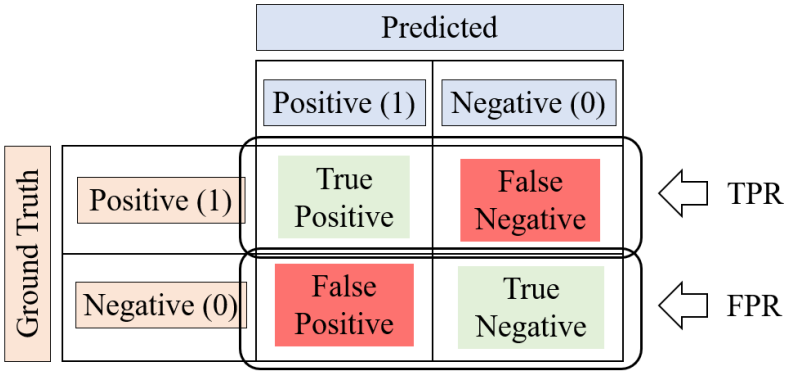


Figure 46. Confusion matrix for binary classification and definition of TPR and FPR.

As Figure 47 shows, the ROC curve provides a simple way to demonstrate the model performance under different probability thresholds. Specifically, it summarizes all the confusion matrices that each threshold produces without having to sort through the confusion matrices; each point in the ROC curve represents a relationship between TPR and FPR under one probability threshold used for classification. AUC score is the area under the ROC curve, and it helps decide which classification model is better by comparing the area under different ROC curves of the models. As Figure 47 shows, a perfect classifier yields an AUC score of 1, whereas a random classifier yields an AUC score of 0.5. Hence, the characteristic of assembled evaluation makes AUC score an important metric for the evaluation of model performance; a higher AUC score suggests a superior overall performance of a classification model.

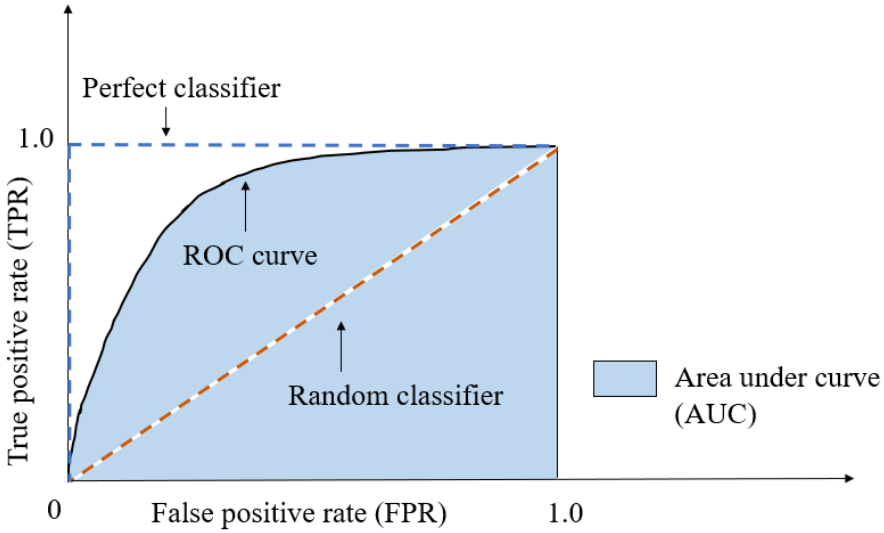


Figure 47. Illustration of ROC curve and AUC score.

6.4 ML results and LSM

Four ML algorithms (LR, SVM, RF, and GBM) were used for landslide susceptibility mapping using the database containing 3000 landslide points (positive samples) and 3000 non-landslide points (negative samples). Table 9 shows the model performance of the four algorithms, and it is found that GBM has the best classification performance compared to the other three algorithms. Among the four models, the AUC value of the GBM algorithm is 0.872, which indicates very good performance in classifying landslides and non-landslides under different probability thresholds. Therefore, the trained GBM model is used to predict the probability of landslide occurrence for

the whole study area. Specifically, fourteen landslide causative factors are attached to every pixel in the study area, and the trained model is applied to predict the probability of landslide for each pixel.

Table 9. Model performance using ML methods for LSM.

Model	Accuracy	Precision	Recall	F1	AUC
LR	0.773	0.761	0.796	0.778	0.851
SVM	0.782	0.758	0.827	0.791	0.855
RF	0.797	0.769	0.848	0.807	0.869
GBM	0.803	0.775	0.855	0.813	0.872
Avg.	0.789	0.766	0.832	0.797	0.862

A landslide susceptibility map for the study area, as shown in Figure 48, is generated using the GBM model. According to the probability of landslide occurrence from 0 to 1, five susceptibility zones are classified with an equal interval of probability. The relationship between susceptibility classes and the probability of landslide occurrence is shown in Table 8. By comparing the susceptibility maps generated by the frequency ratio method and the GBM model (see Figure 49), it is found that the model based on the frequency ratio method overestimates landslide risk in many areas. In contrast, the model based on ML shows a better performance in matching the spatial distribution of actual landslide events. For example, the landslide susceptibility map based on the frequency ratio method shows that there are large areas with a high risk of landslides (yellow background) but few actual landslide occurrences. On the other hand, the landslide susceptibility map based on the ML method shows a better consistency between landslide occurrence and susceptibility. Most landslide data points are distributed within the areas of high and very high risk (orange and red backgrounds in the map using the ML method), which indicates a more accurate predictive model.

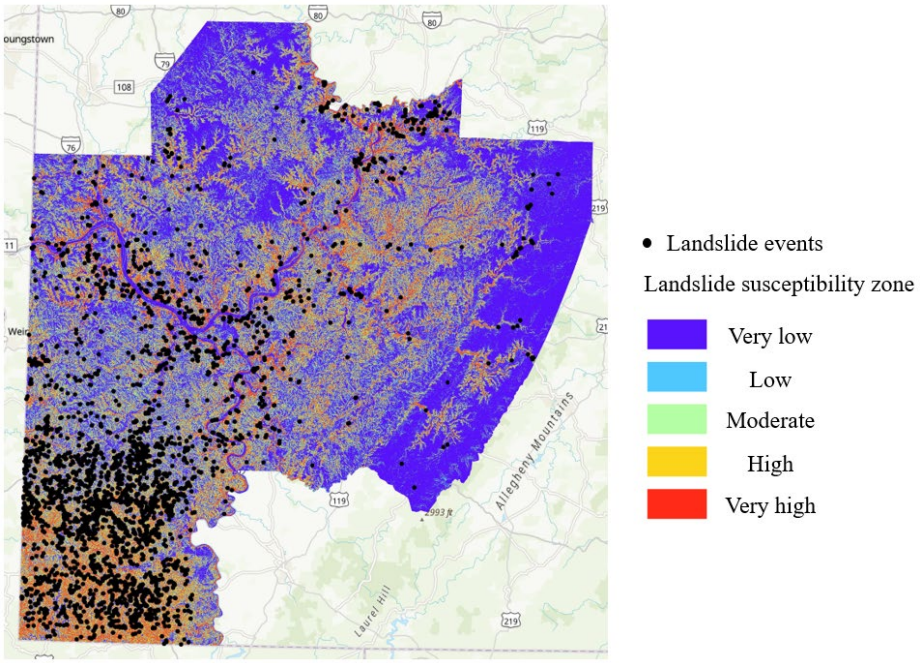


Figure 48. Landslide susceptibility map of the study area using ML method.

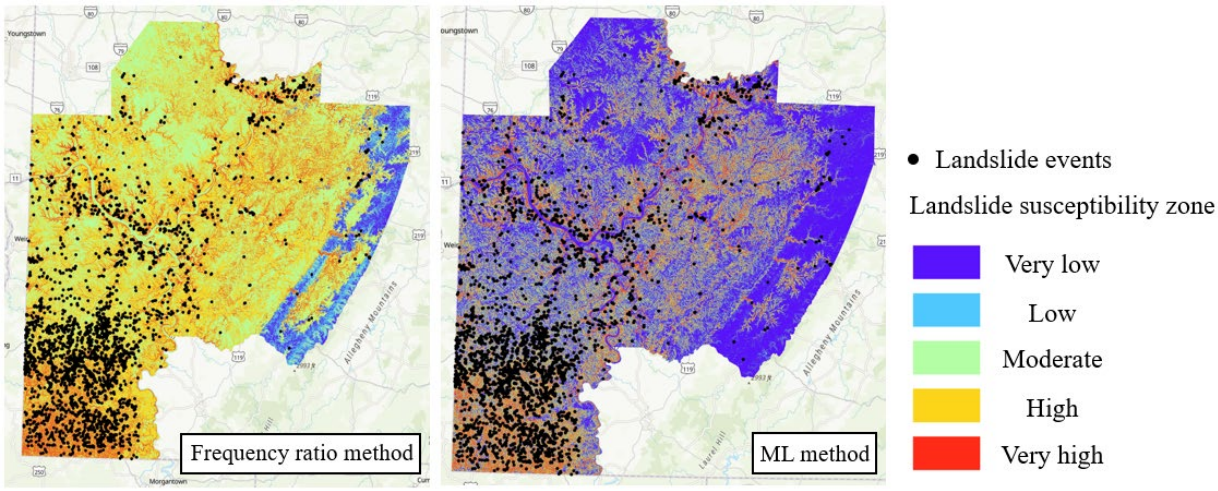


Figure 49. Comparison of landslide susceptibility maps using frequency ratio and ML methods.

6.5 Model explainability

ML models are often referred to as black boxes, and a significant downside of using ML method is losing the ability to quickly interpret the results and explain the relationships between causative factors and predicted outcomes. However, in scientific domains, it is necessary to understand how models learn the problem so that it is possible to combine the ML algorithms and physical mechanism from domain knowledge to analyze the reliability and generalizability of the model. Specifically, it is important to understand the main factors that affect the output of the model and the importance ranking of all input factors. Hence, explainable ML techniques are needed to unravel some of these aspects.

One of these techniques is the SHAP (SHapley Additive exPlanations) method, which is used to explain how much each feature has contributed to the output and thus allows local and global analysis for the dataset. In this study, the SHAP plot for landslide susceptibility mapping based on the GBM model is shown in Figure 50.

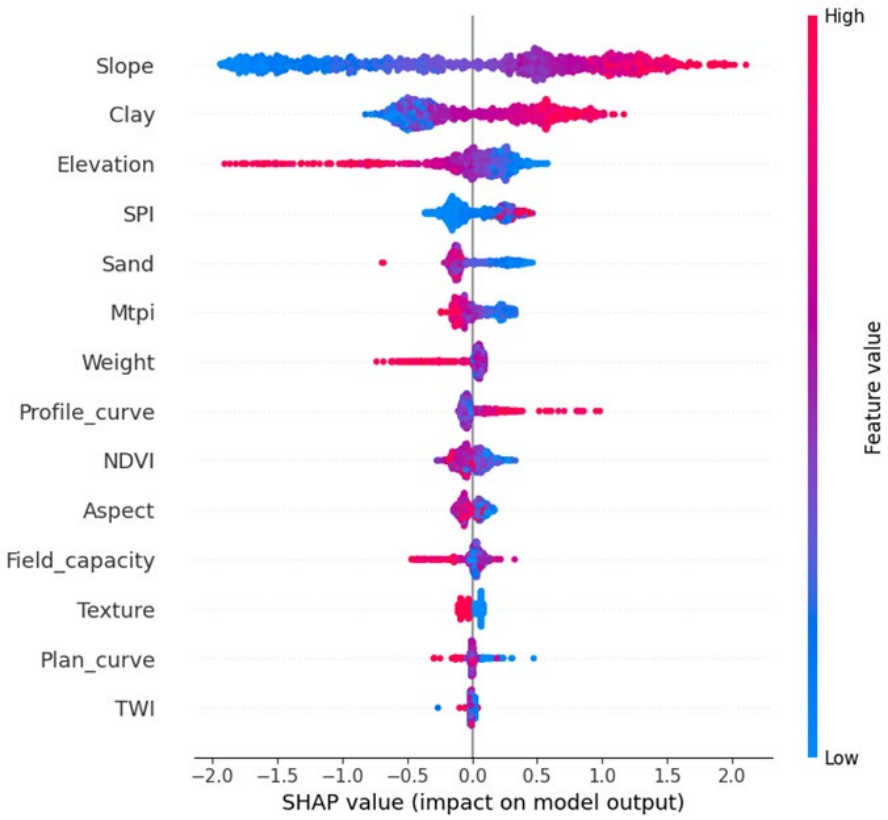


Figure 50. SHAP plot for GBM model of LSM.

SHAP plot is based on cooperative game theory and used to increase the transparency and interpretability of ML models. SHAP plot shows how much a single feature affects the prediction. In Figure 50, the sequence of the left list represents the importance ranking of all factors. For example, the slope makes the most significant contribution to the output, and the SHAP value increases as the slope value increases. That means a higher slope value steers the ML model toward a more positive prediction (i.e., a higher probability of landslide). In the SHAP plot, it is found that among all factors, slope, clay content, elevation, SPI, weight (bulk density), and profile curvature have a relatively high contribution to the prediction of a landslide occurrence. According to the physical model of landslide, slope angle directly influences the shear stress of soil along potential failure surfaces of a slope, and that is the reason for its top ranking of importance to the model prediction. Clay content in the slope material also has a great impact on the shear strength and other characteristics of soil that contribute to the occurrence of landslides. Higher elevation values have a high negative contribution, which is because in high altitude areas, clay content decreases and rock becomes the main geological material; as a result, the probability of landslides decreases as elevation increases. Compared to the plan curvature, profile curvature contributes more to the model prediction, which is reasonable based on the definition of plan and profile curvatures from a geometric perspective. The typical geometry of profile curvature makes it have a more significant impact on landslide occurrence.

7 Spatiotemporal Analysis for LSM

Conventional landslide susceptibility mapping focuses on the prediction of landslide spatial distribution based on static causative factors (e.g., topographic factors), which only vary in space. Positive labels (landslides) and negative labels (non-landslides) are sampled in different locations and spatial causative factors are obtained for each sample. ML models are then trained as classifiers to predict the probability of landslides in space based on all samples and the corresponding causative factors. By extracting the information in spatial factors, ML models learn the relationships between landslide occurrence and spatial features; thus, the spatial distribution of landslide risk can be predicted. However, pure spatial features cannot fully explain the timing of landslide occurrence at a given location. To further predict landslides both in spatial and temporal scales, spatiotemporal landslide susceptibility mapping is conducted in the present study.

7.1 Landslide database for spatiotemporal ML

Landslide inventories used for temporal analysis have been compiled and discussed in Section 2. Given that most landslides in Pennsylvania are concentrated in the southwestern regions of Pennsylvania, PennDOT Districts 11 and 12 were selected for the study. Integrating temporal analysis requires the creation of a database of landslide events that includes accurate landslide event dates. In this case, as the initial available landslide data set (from PennDOT Districts 11 and 12) containing the date of the landslide was limited, data from adjacent areas with accurate landslide event dates were included in the database. As a result, northern West Virginia and eastern Ohio were incorporated into the study area. In the present study, the landslide inventories are based on NASA Cooperative Open Online Landslide Repository (COOLR) project and PennDOT Districts 11 and 12 Slide Database. There are 223 landslide data points in the study area, as shown in Figure 51, where red dots represent 173 recorded landslides from NASA COOLR project and yellow dots represent 50 recorded landslides from PennDOT Districts 11 and 12. All 223 landslides are provided with accurate event dates.

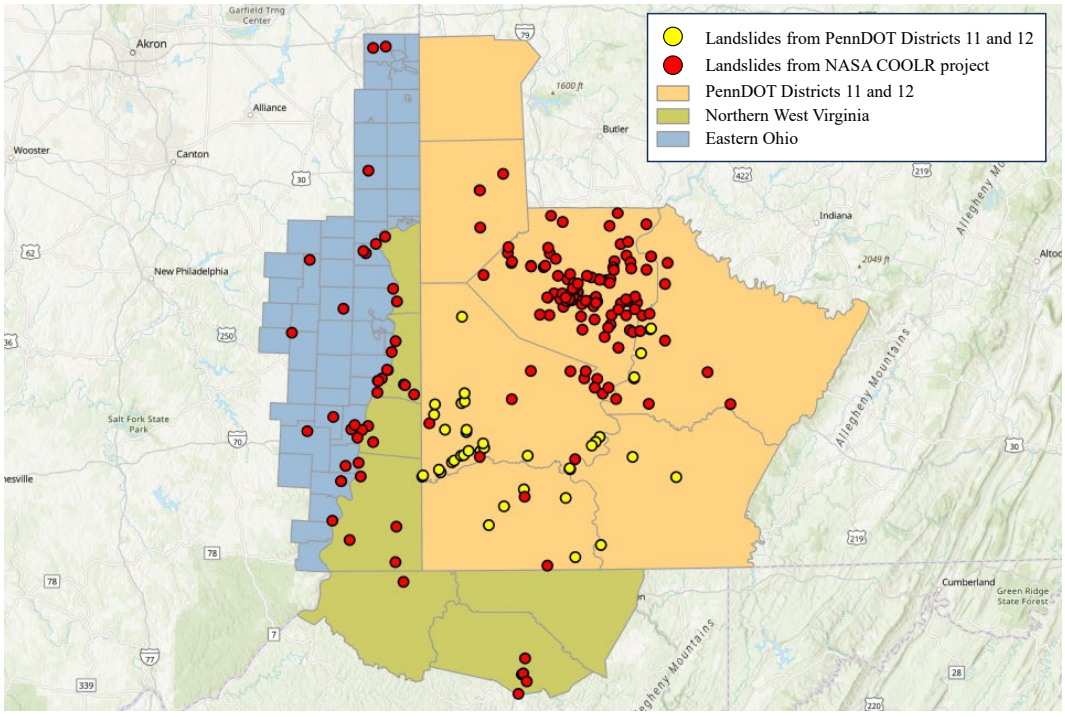


Figure 51. Landslide distribution and study area for spatiotemporal analysis.

7.2 Spatiotemporal causative factors

In addition to spatial topographic factors, previous studies have shown that landslide occurrence is closely related to rainfall, which is time-varying and can explain the timing of landslide occurrence at a given location. Hence, to train ML models with the ability to predict landslides on the temporal scale, rainfall factors should be included as additional features. For spatiotemporal prediction of landslides, the same fourteen topographic factors are kept to represent spatial features, and eight additional rainfall factors are included to represent temporal features in the database. These eight rainfall factors are cumulative precipitation 1 day, 3 days, 1 week, 2 weeks, 3 weeks, 1 month, 2 months, and 3 months preceding landslide events, which are downloaded from Daymet (https://daymet.ornl.gov/). A portion of the spatiotemporal database is shown in Figure 52.

Bulk_density	NDVI	1 day	3 days	7 days	14 days	21 days	30 days	60 days	90 days	label
$10 \times kg/m^3$	-	mm	mm	mm	mm	mm	mm	mm	mm	-
120	0.291498	0	1.24	18.7	25.5	26.01	54.08	103.91	195.58	1
130	0.20215	0	0	66.49	89.82	106.97	127.47	160.85	333.75	1
130	0.237815	0	0	34.97	74.11	99.24	113.36	199.71	295.17	1
134	0.347043	0	8.31	12.97	62.12	83.38	100.26	210.34	271.81	1
144	0.232276	0	0	17.44	18.76	24.42	46.02	101.93	189.61	1
146	0.270945	0	0	12	21.32	66.53	80.51	219.16	306.29	1
148	0.331658	11.18	35.08	43.73	76.79	98.24	98.24	218.32	313.24	1
149	0.317233	3.6	41.54	59.36	59.36	70.87	98.33	200.54	319.94	1
149	0.359738	44.08	99.38	102.82	128.18	161.32	228.83	324.12	446.3	1
150	0.219423	23.91	33.28	70.33	75.55	123.72	149.6	228.68	295.47	1
151	0.30739	3.41	35.47	47.36	93.24	98.04	126.85	295.91	362.22	1
151	0.30739	5.8	5.8	5.8	22.37	26.95	47.81	153.92	245.55	1
151	0.338981	9.21	23.95	47.02	72.72	269.88	275.79	406.77	489.92	1
152	0.278281	0	13.5	36.62	50.27	63.07	72.07	196.79	302.45	1
152	0.265096	0	12.55	12.55	34.09	42.6	116.67	201.24	245.83	1
152	0.294835	27.33	27.33	43.96	115.95	126.68	179.95	285.9	469.84	1
153	0.128275	3.83	31.15	33.14	46.96	68.48	87.07	203.12	227.12	1

Figure 52. Input database compiled for spatiotemporal prediction.

By introducing cumulative precipitation in different periods as input factors, ML algorithms can find relationships between precipitation and the probability of landslide occurrence. Previous studies have shown that landslides triggered by a storm are not necessarily just caused by that specific storm alone. Soil saturation is an important triggering mechanism for landslides, and in addition to being affected by current rainfall, soil saturation is also related to previous cumulative

precipitation. For example, a landslide can occur with little precipitation if the soil is already in a high-water-content condition due to previous precipitation events. Therefore, it is necessary to include cumulative precipitation in different periods preceding landslide event dates. These additional rainfall factors also make it possible for ML models to predict the timing of landslide occurrence at a given location because when static topographic factors are the same, different rainfall factors can provide extra temporal information for the algorithms to classify landslides and non-landslides. From this perspective, the database for spatiotemporal prediction should contain landslides and non-landslides samples both in spatial and temporal scales.

7.3 Spatiotemporal sampling methods

To incorporate spatial and temporal information into the database, non-landslides need to be sampled both in space and time. Figure 53 shows a demonstration of the sampling method used for spatiotemporal analysis. For each landslide event/sample, one non-landslide is randomly sampled within a ring-shaped zone (buffer), between 0.5 km and 1.5 km, from the landslide location, and it is assigned the same date as that of the landslide event. Buffer-controlled sampling creates non-landslides that have different spatial features from landslides. The size of the ring-shaped buffer zone accounts for the typical landslide sizes in the study area; the minimum distance of 0.5 km ensures that non-landslides randomly sampled in space will not coincide with the landslides, while the maximum distance of 1.5 km restricts the sampling range of non-landslides. This restriction reduces the possibility that non-landslides are sampled in areas like rivers, buildings, or roads that are not comparable with landslide areas. After spatial sampling, temporal sampling is conducted both for the landslide samples and the non-landslide samples. Different landslide window periods are used for temporal sampling in this study; the window period of 1 year is illustrated in Figure 53 as an example. The window period of 1 year assumes that there was no landslide 1 year prior to the landslide event date. This assumption is based on the typical period and frequency of landslide investigation in the study area. In addition, it is assumed that landslides are regularly reported in the study area. Finally, for each landslide event in the database (red solid circle in Figure 53), there are three corresponding non-landslide samples (green solid circles in Figure 53) in both spatial and temporal scales as shown in Figure 53. Every sample (positive or negative) is attached with fourteen topographic factors and eight rainfall factors, forming the spatiotemporal database for training ML models.

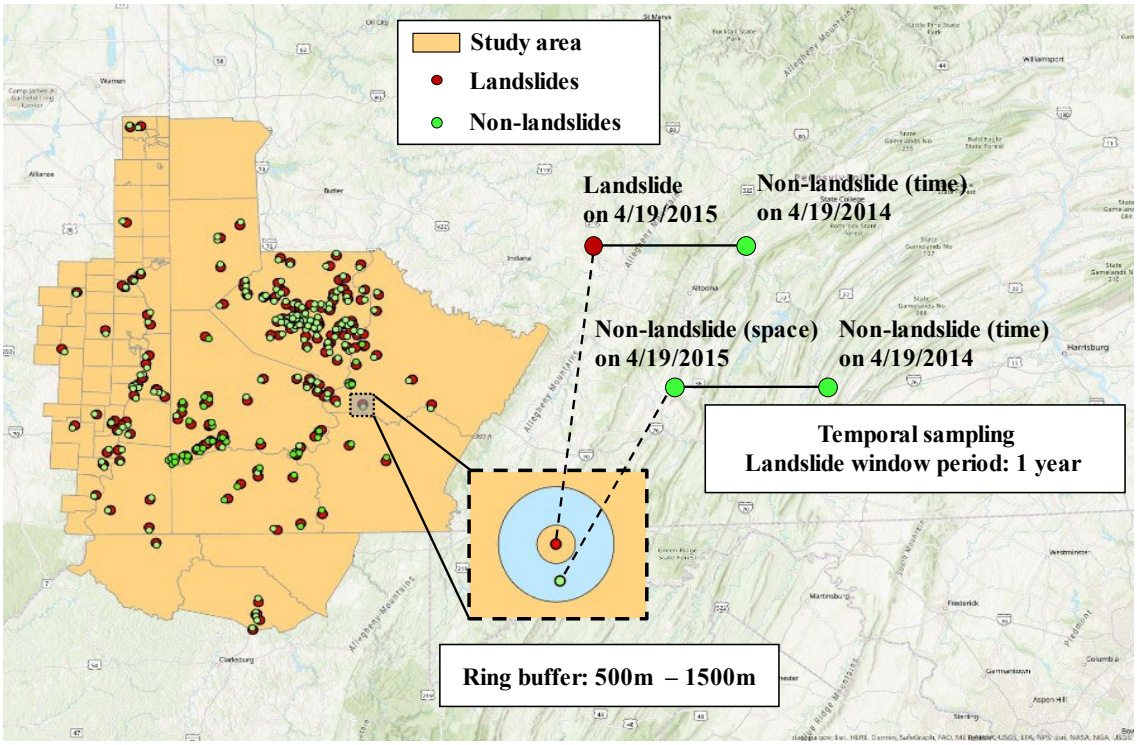


Figure 53. Sampling approach for spatiotemporal analysis (landslide window period: 1 year).

In the sampling approach shown in Figure 53, the ratio of positive to negative samples is 1:3, which makes the spatiotemporal dataset imbalanced. The imbalanced dataset can cause poor classification performance of minority classes. If the number of majority samples is much greater than that of minority samples, ML models tend to always predict the outcome as the majority class. Given that the general mechanism of ML algorithms is to minimize the error between predicted results and the ground truth, a poor classification of minority samples can still achieve very high overall accuracy in the model. Therefore, conventional ML models may be unreliable when dealing with imbalanced datasets.

In general, there are three approaches to tackling imbalanced datasets. The first one is applying a weighted cross-entropy loss function for ML algorithms. This method addresses the structures and mechanisms of ML algorithms. By modifying the loss function, the ML model pays more attention to the minority class while conducting classification tasks. However, this method can only increase the Recall of the model with fewer minority samples being misclassified. However,

the Precision will decrease significantly, and the model is unreliable. The second approach is over-sampling, through which new minority samples will be created using statistical theories to achieve a balanced dataset. However, as a product of natural factors, the characteristics of landslides are complex. New landslide samples created based on data science will not be interpretable. In the present study, the third method, which is under-sampling, is adopted to tackle the imbalanced spatiotemporal dataset. A demonstration of the under-sampling method is shown in Figure 54.

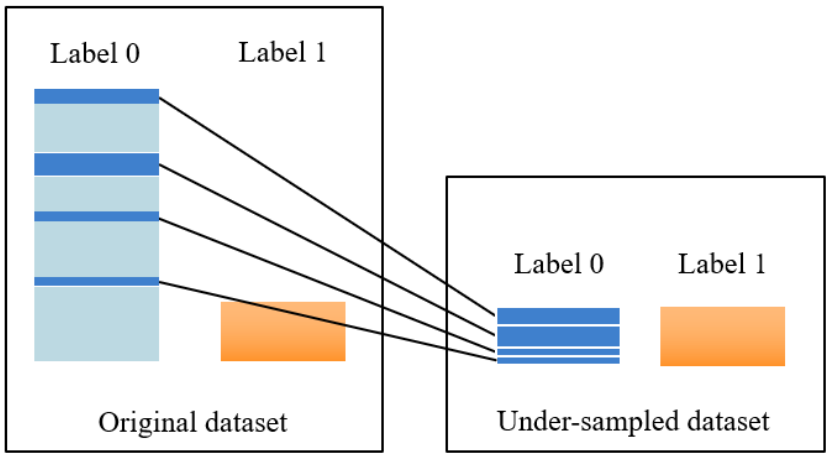
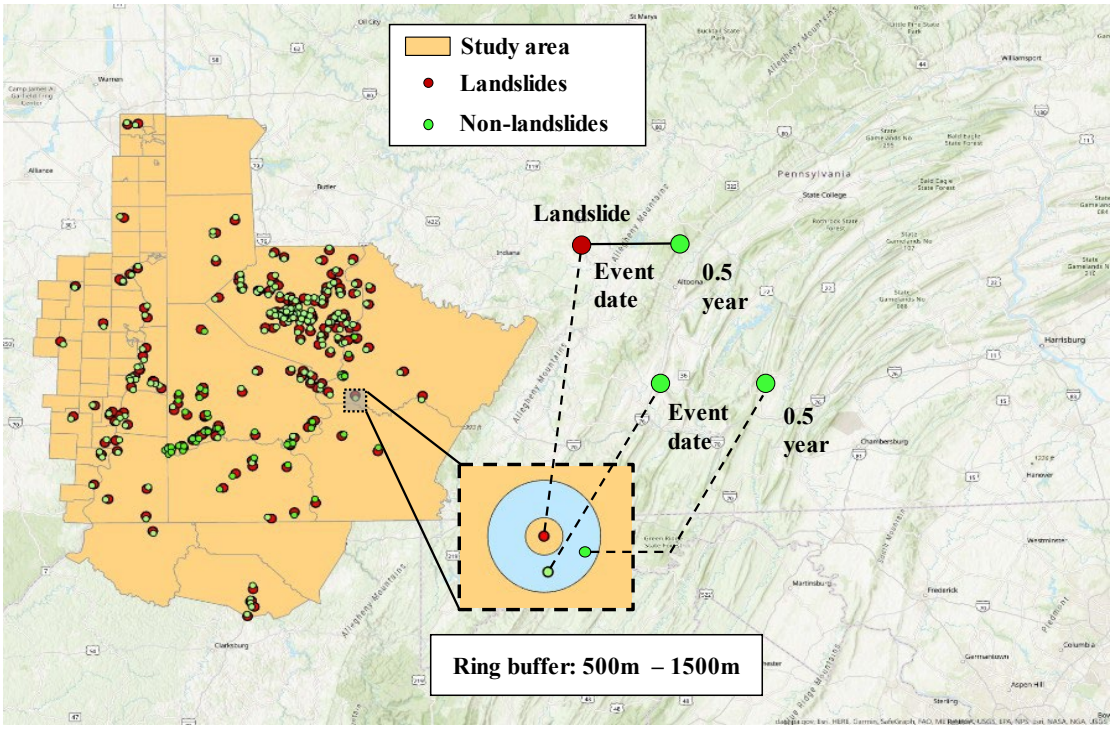


Figure 54. Under-sampling method for imbalanced dataset.

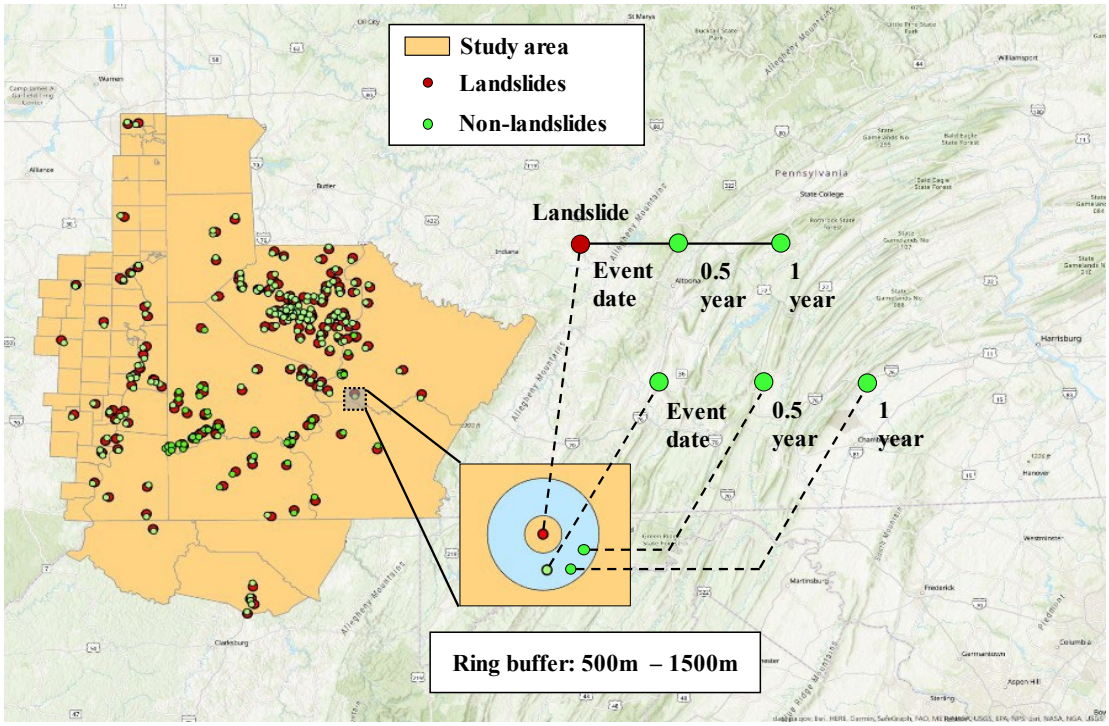
The under-sampling method randomly selects negative samples with the same number as the positive samples to form a balanced database for model training. Since under-sampled Label 0s are randomly selected from the original Label 0s, they are assumed to be able to represent the distribution of the original Label 0s. In the following section, different ways of including more negative samples in the original dataset to form different spatiotemporal datasets are discussed, and the performance of ML models based on these datasets is compared and analyzed.

7.4 Spatiotemporal ML with different spatiotemporal datasets

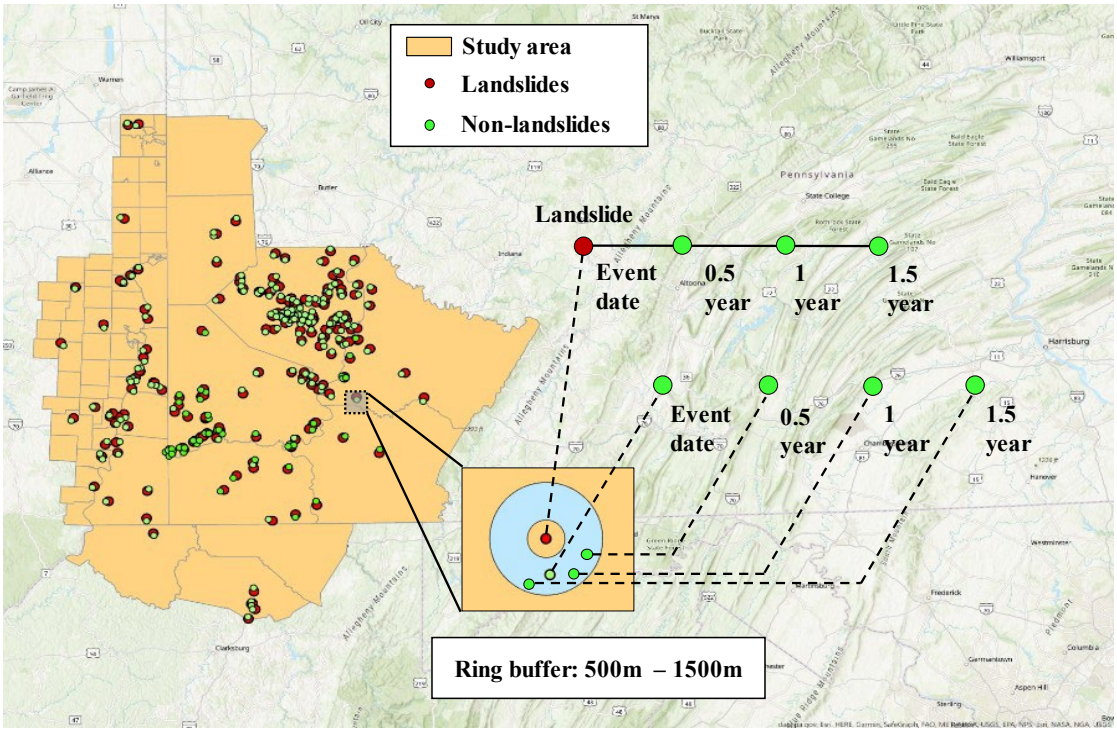
Using different landslide window periods, more non-landslide samples are added to the spatiotemporal dataset through sampling in space and time scales. Figure 55 shows spatiotemporal datasets constructed by including non-landslides with different locations and window periods.



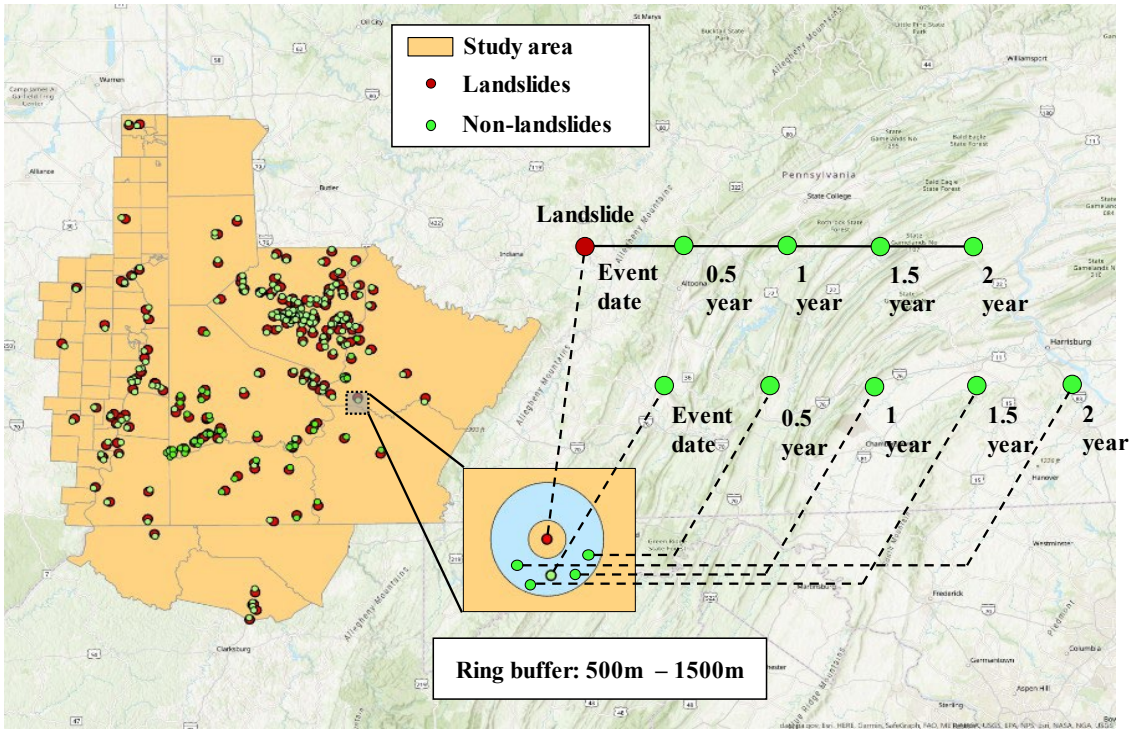
(a) Spatiotemporal dataset 1



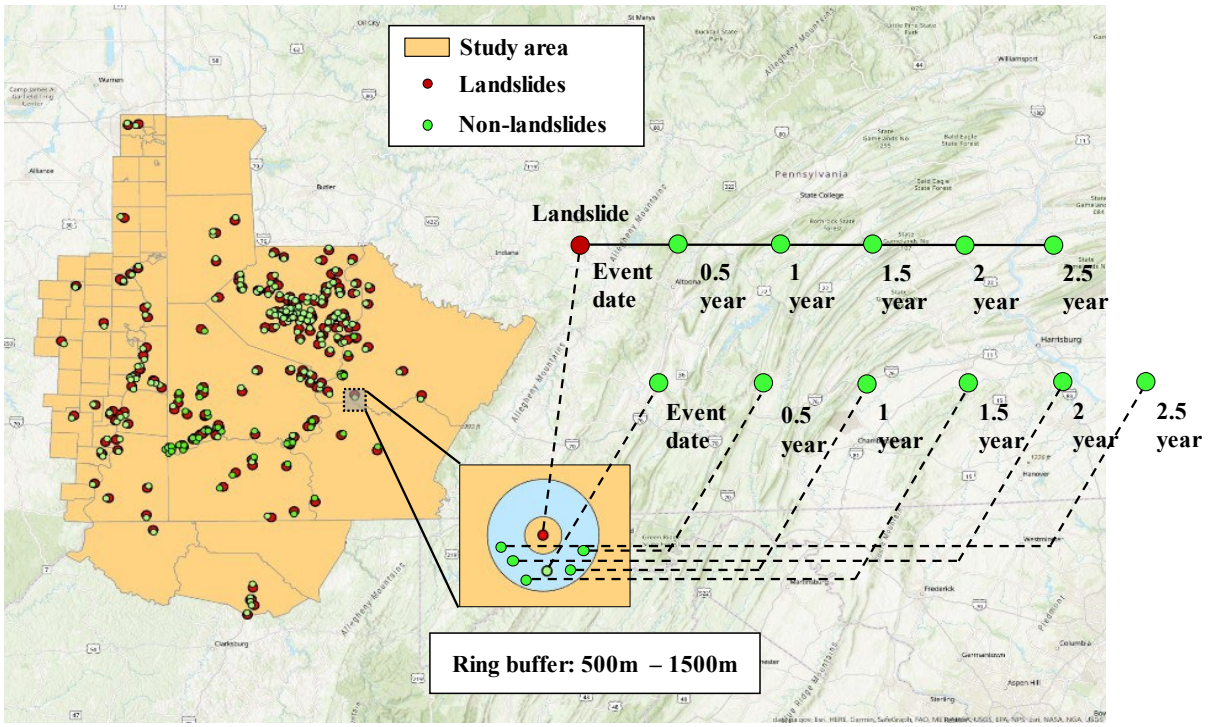
(b) Spatiotemporal dataset 2



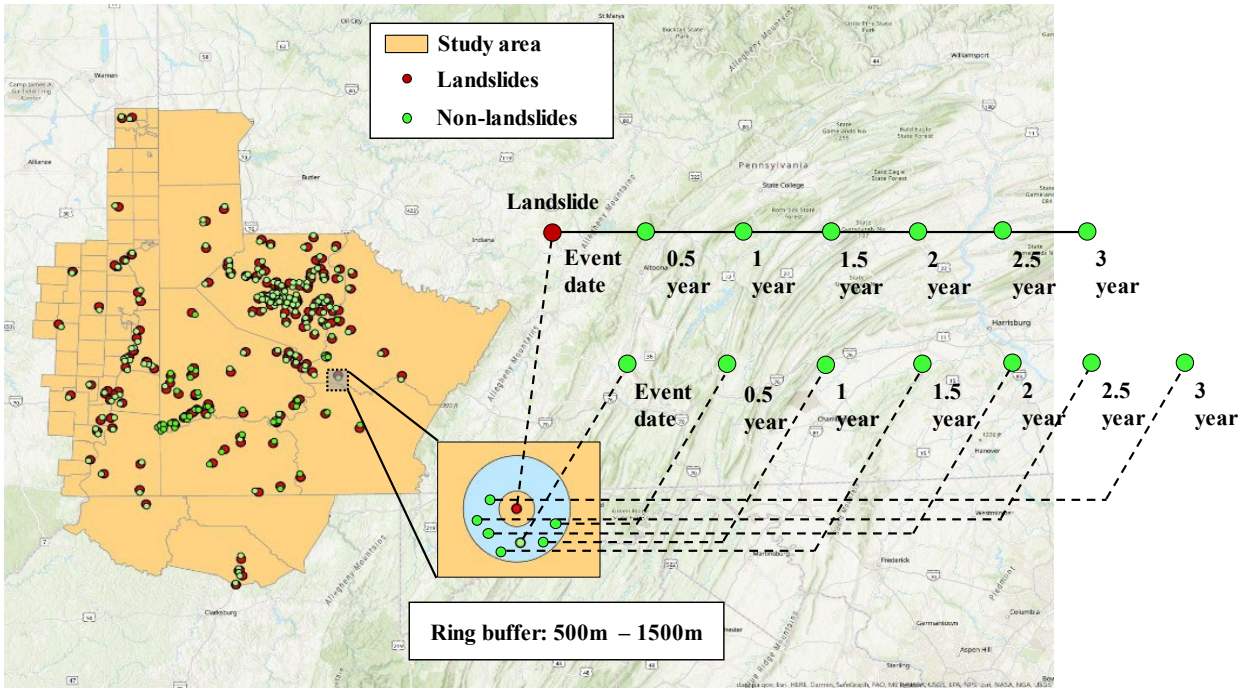
(c) Spatiotemporal dataset 3



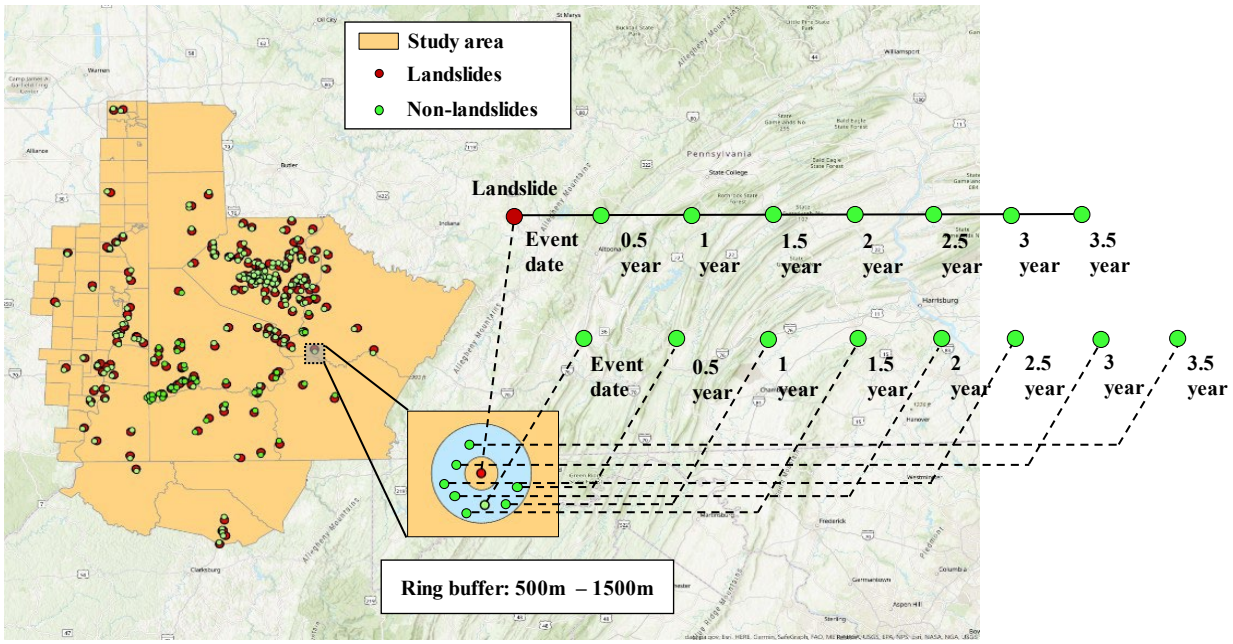
(d) Spatiotemporal dataset 4



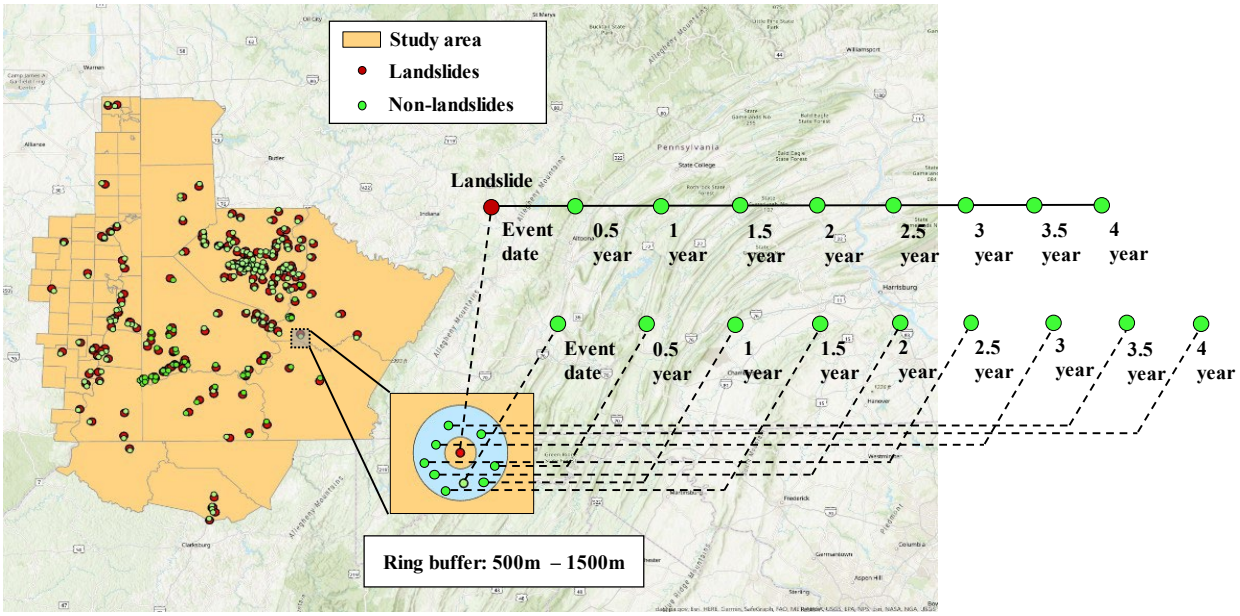
(e) Spatiotemporal dataset 5



(f) Spatiotemporal dataset 6



(g) Spatiotemporal dataset 7



(h) Spatiotemporal dataset 8

Figure 55. Spatiotemporal datasets with different non-landslide locations and landslide periods.

From Figure 55(a) to 55(h), more non-landslides are sampled within the buffer zone in space for each landslide event; in the time scale, non-landslides are sampled again for both the landslide and non-landslide locations. The number of temporal non-landslides increases as more landslide window periods are included. In the spatiotemporal dataset 8, for example, eight landslide window periods are used from 0.5 year to 4 years; for each landslide point (red solid circle in the figure), it is assumed that there is no landslide occurrence 0.5 year, 1 year, 1.5 years, 2 years, 2.5 years, 3 years, 3.5 years, and 4 years prior to the event date at the same location. For the nine spatial non-landslides in the buffer zone (green solid circles within the buffer zone in the figure), each of them is sampled with one window period. By doing this, both spatial and temporal information are introduced into the dataset, and the amount of information increases as more spatial and temporal non-landslides are sampled.

For each spatiotemporal dataset, four ML algorithms are used to conduct binary classification. The results show that the RF model outperforms the other three algorithms for all datasets, and the RF model performance for each dataset is shown in Table 10.

Table 10. RF model performance for spatiotemporal landslide susceptibility mapping.

Dataset	Accuracy	Precision	Recall	F1 score	AUC score
1	0.71	0.72	0.69	0.71	0.77
2	0.72	0.73	0.69	0.71	0.79
3	0.75	0.77	0.70	0.73	0.81
4	0.76	0.78	0.72	0.75	0.83
5	0.77	0.80	0.70	0.75	0.84
6	0.78	0.79	0.74	0.76	0.85
7	0.79	0.81	0.72	0.77	0.86
8	0.78	0.79	0.76	0.77	0.86

The results in Table 10 indicate that model performance improves as more non-landslides are included in the spatiotemporal datasets 1 to 8. The AUC score reaches the optimum of 0.86 in the spatiotemporal dataset 7, where eight non-landslides are sampled in space and seven landslide window periods are considered. The model performance doesn't meaningfully improve in dataset 8. From datasets 1 to 7, more non-landslide samples containing different spatial and temporal information have been added to the dataset for the training of ML models. As more information is provided for training, the ML models learn more representative and generalizable relationships from the dataset, and the model performance is steadily improved. However, such improvement will be limited as the model's performance reaches an optimum. Although more spatial and temporal information is introduced to the dataset as more non-landslides are sampled, the under-sampling method randomly selects negative samples with the same number as positive samples, which means there are always 223 non-landslides being under-sampled due to the fact that there are 223 landslides in the database. Consequently, although more spatiotemporal non-landslides are included, the weights of newly added non-landslides are diluted by under-sampling, and the effective spatial and temporal information in the dataset is saturated. Therefore, as the results of datasets 7 and 8 show, the model performance reaches an optimal level and then is kept at that level even with more non-landslide samples. The model trained by dataset 8 is used as the optimal model to conduct spatiotemporal landslide susceptibility mapping.

The feature importance of the model from SHAP plot is shown in Figure 56. It is shown that by incorporating topographic and rainfall factors into ML, the model predicts landslide occurrence from both spatial and temporal perspectives. Figure 56 shows that the cumulative precipitation 7 days, instead of 1 day, preceding the landslide event is the most important factor in landslide occurrence, which is reasonable since the 7-day cumulative precipitation reflects the increase in soil saturation caused by rainfall and accounts for the lag effect. Besides precipitation, the topographic factors of elevation and slope also contribute significantly to the model outcome according to the SHAP plot.

Figure 56 shows that cumulative precipitation of various periods has a great effect on the potential for landslides; this is consistent with geotechnical engineering experience that poor drainage and lack of drainage maintenance during and after rainfall events contribute to landslide occurrence. This consistency suggests good predictive capability and explainability of the

developed spatiotemporal ML model. However, it is cautioned that the disproportionate effect of precipitation in Figure 56 may be due to the fact that only rainfall-induced landslides are included for model training (i.e., the model output is just being consistent with the model input).

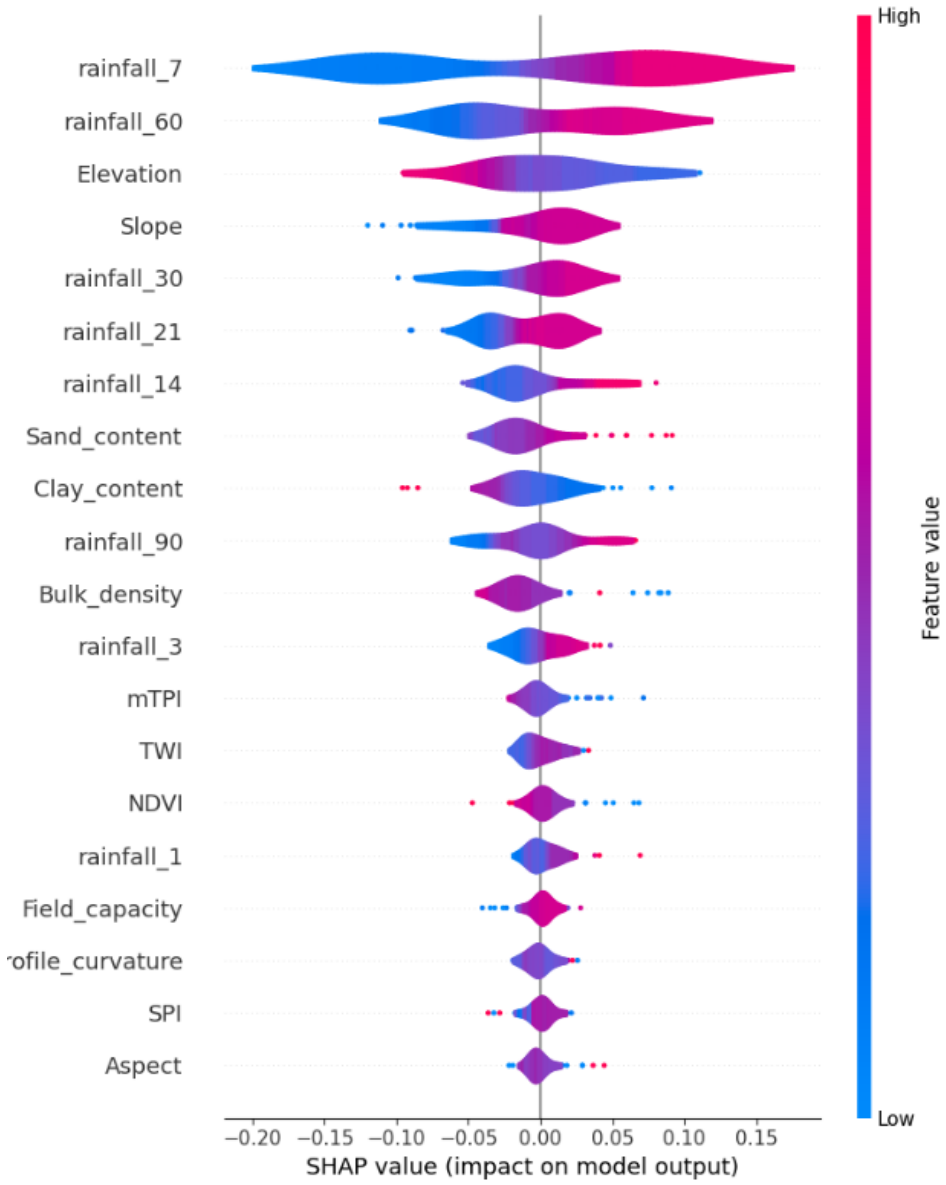


Figure 56. SHAP plot for the ML model in spatiotemporal landslide susceptibility mapping based on spatiotemporal dataset 8.

7.5 Spatiotemporal LSM

7.5.1 Pure spatial susceptibility map

Before generating a spatiotemporal landslide susceptibility map, a pure spatial susceptibility mapping is conducted as the baseline for comparison purposes. The framework of pure spatial susceptibility mapping is the same as that described in Section 6, except that a different landslide database is used. An illustration of the database for pure spatial analysis is shown in Figure 57. In the spatiotemporal dataset 8, the temporal/precipitation information can be disabled by removing the eight rainfall factors; as such, the dataset can be used to train ML models for pure landslide susceptibility mapping. The landslide susceptibility map generated is shown in Figure 58, where the black solid circles are the 223 landslide data points, and the AUC score of the ML model is 0.77. A comparison of the pure spatial landslide susceptibility map and the previous landslide susceptibility map using the much larger landslide database (without event dates) is shown in Figure 59, where only PennDOT Districts 11 and 12 are shown.

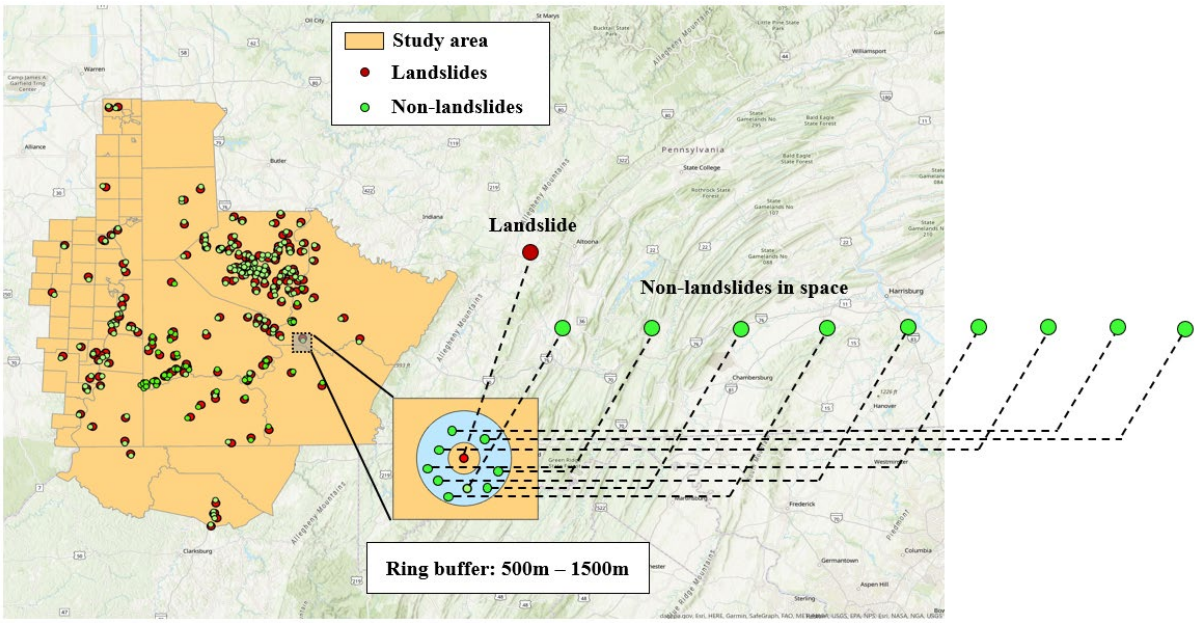


Figure 57. Spatiotemporal dataset 8 for pure spatial susceptibility mapping.

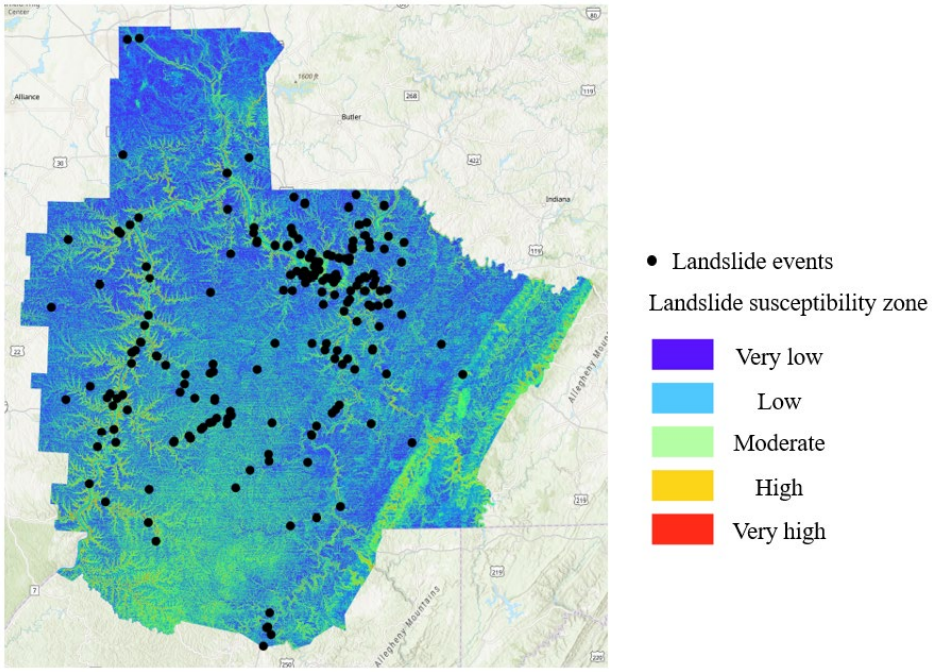


Figure 58. Pure spatial susceptibility map.

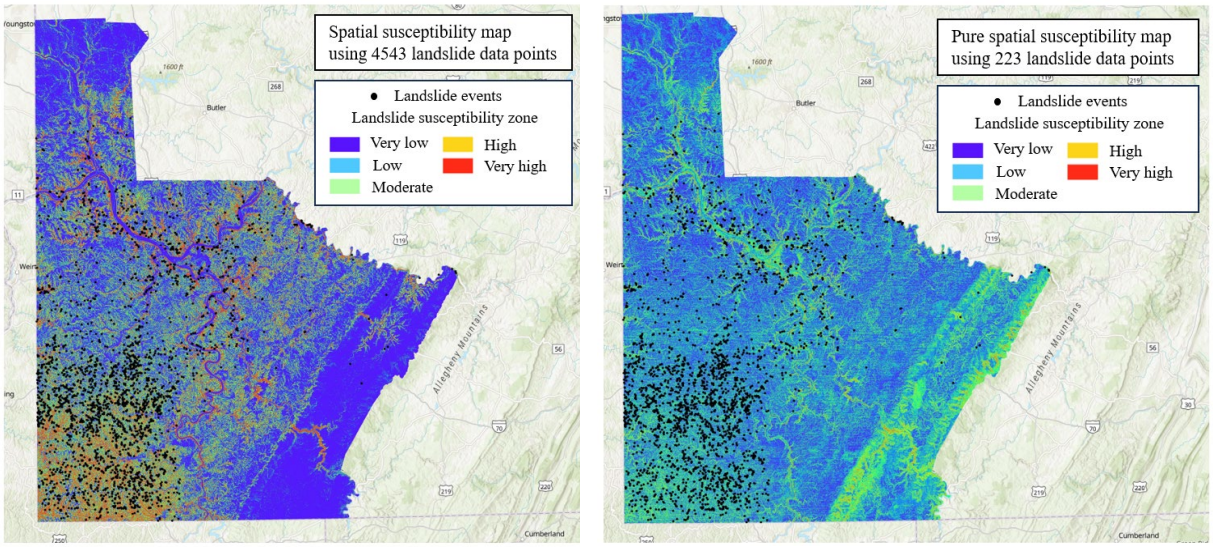


Figure 59. Comparison of pure spatial landslide susceptibility map using different databases.

In Figure 59, the black dots are 4,543 historical landslides (without event dates) as discussed in Section 6. The ML model trained using 4,543 landslide data points demonstrates superior

performance as the generated map can delineate areas with higher susceptibility. However, in the pure spatial susceptibility map based on the 223 landslide data points, there are large areas with a low probability of landslides (blue background) but numerous actual landslide occurrences. The AUC scores for the ML model trained using 4,543 landslide data points and the ML model trained using 223 landslide data points are 0.87 and 0.77, respectively, further confirming the superior performance of the former model. Figure 59 highlights the importance of the amount of landslide data points in conducting landslide susceptibility mapping.

7.5.2 Spatiotemporal susceptibility map

Based on the trained model using spatiotemporal dataset 8, the spatiotemporal landslide susceptibility map in the study area can be generated. Given that the rainfall factors are time-varying, the spatiotemporal susceptibility map also changes with time. PennDOT Districts 11 and 12 reported more than 12 landslide events on 02/15/2018 after a storm event. These reported events provide an opportunity to compare landslide susceptibility maps generated based on pure spatial mapping and spatiotemporal mapping to highlight the importance of incorporating precipitation data into susceptibility mapping. Figure 60 compares the pure spatial and temporal susceptibility maps for PennDOT Districts 11 and 12 and their surrounding areas for one rainfall event on 02/15/2018, where the reported landslide events are represented using red solid circles. Figure 60 shows that the spatial distribution of the reported landslide events on 02/15/2018 has a much better match with the spatiotemporal susceptibility map than with the pure spatial susceptibility map. Table 11 shows the predicted susceptibilities from the pure spatial ML model and spatiotemporal ML model for select 12 landslides reported on 02/15/2018. The pure spatial prediction shows a low probability (≤ 0.5) of landslide occurrence for some locations of the reported events; hence, considering only terrain factors, landslide susceptibility is significantly underestimated by the pure spatial susceptibility mapping. On the other hand, accounting for precipitation data for the event, the spatiotemporal ML model predicted much higher susceptibility for these locations for the single event, consistent with their actual occurrence.

Figure 60 shows that the spatiotemporal susceptibility map indicates a high probability of landslides in the places where landslides actually occurred, but also shows high probabilities of landslides in other places (yellow and red backgrounds) without reported landslide events. This

discrepancy may be due to a combination of the following reasons. First, there may be undiscovered or unreported landslide events in areas with high risk. Second, there is uncertainty in the output of the ML model due to noise (error or irrelevant information) in data, incomplete coverage of the domain, and imperfect models. In the present study, the location of landslide points has an uncertainty of 1-5 km according to the NASA COOLR project; therefore, the value of causative factors obtained at the location of landslides has an error (in the ideal situation, the causative factors should be obtained at the exact location of the landslide), which introduces noise into the training data for ML. Additionally, the small landslide database (223 data points) may result in the low generalizability of the model. Hence, the ML model has uncertainty when predicting data beyond the training samples. The model performance will improve as more landslide data points are included in the model training.

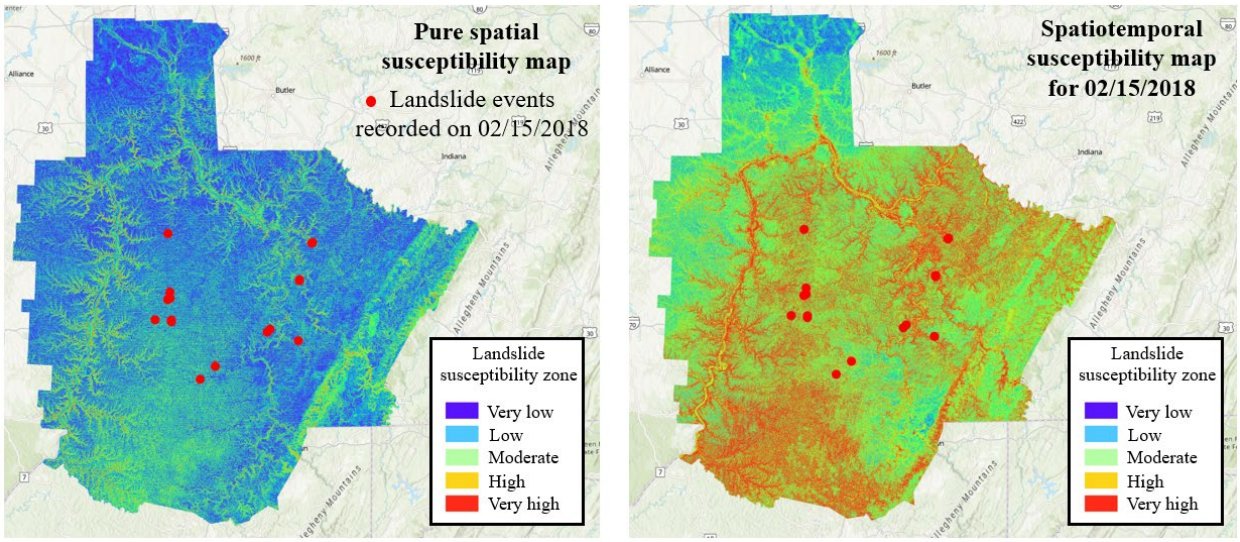


Figure 60. Comparison between pure spatial and spatiotemporal maps for 02/15/2018.

Table 11. Predicted susceptibilities from pure spatial ML model and spatiotemporal ML model for 12 landslides recorded on 02/15/2018.

Landslide Number	Latitude	Longitude	Susceptibility	
			Pure spatial ML Model	Spatiotemporal ML Model
1	-79.797	40.016	0.62	0.97
2	-80.238	39.890	0.74	0.86
3	-80.170	39.934	0.21	0.83
4	-79.923	40.057	0.85	0.99
5	-79.936	40.047	0.51	0.76
6	-80.438	40.093	0.51	0.86
7	-80.365	40.086	0.31	0.79
8	-80.364	40.092	0.58	0.87
9	-80.377	40.389	0.63	0.91
10	-80.380	40.162	0.74	0.88
11	-80.369	40.166	0.47	0.67
12	-80.370	40.188	0.59	0.82

To further explain the superior performance of the spatiotemporal susceptibility map, Figure 61 compares the pure spatial susceptibility map, 7-day cumulative precipitation map, and the spatiotemporal susceptibility map. The cumulative precipitation of 7 days prior to the event is taken as an example of the rainfall effect since the ML model has shown that this factor is the most important factor for spatiotemporal prediction (see Figure 56). Figure 61 shows that the spatiotemporal susceptibility map accounts for the combined effect of topographic factors, as demonstrated in Figure 61(a), and precipitation factors, as demonstrated in Figure 61(b). By incorporating rainfall factors, the proposed spatiotemporal susceptibility mapping approach is a useful tool to assist in predicting the occurrence and timing of potential landslides taking place due to precipitation events. The map highlights areas having very low to very high risk of landslide

susceptibility with precipitation, which may be used to establish a hierarchy and mitigate risk for slopes at "very high risk" for landslide susceptibility. The map may also be used for forecasting purposes. For example, it may be used as an aid for planning and programming purposes to address slopes with "very high" landslide susceptibility first. The map may be used in the event of incoming storms to target slopes with a very high risk of landslide susceptibility so that mitigation or preventative measures (such as temporary road closure) can be employed to ensure safe travel and minimize damage. In addition, the map may also help to target post-storm roadway/slope inspections to the most critical and high-risk locations first.

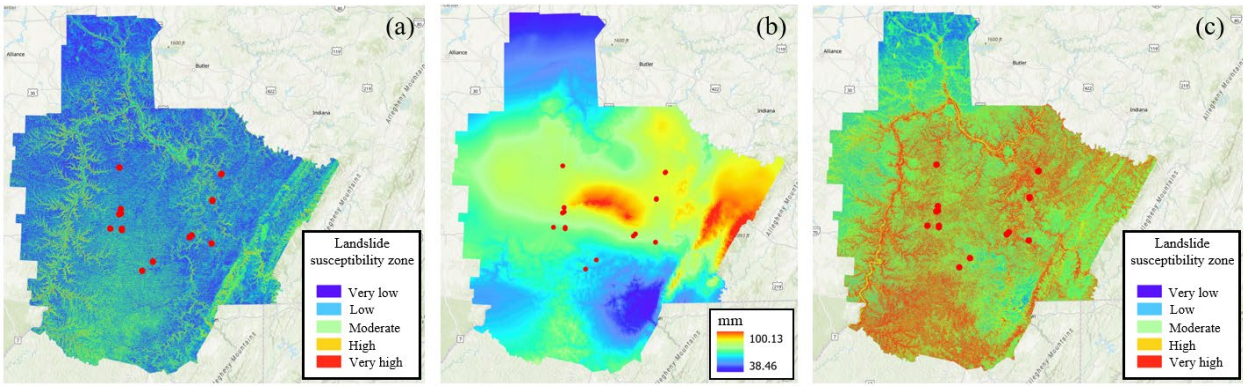


Figure 61. Comparison and interpretation of spatiotemporal susceptibility map: (a) pure spatial susceptibility map; (b) 7-day cumulative precipitation map for 02/15/2018; (c) spatiotemporal susceptibility map for 02/15/2018.

8 Conclusions and Limitations

8.1 Conclusions

To develop a warning system for rainfall-induced landslides, LSM is conducted using different ML techniques. Two landslide databases for spatial and spatiotemporal analyses of landslides in Pennsylvania and adjacent areas are compiled. Based on the digitized map of USGS Topo sheets, a spatial landslide database is compiled with 4,543 historical landslides in southwestern Pennsylvania. The database with pertinent information is stored on three platforms: ArcGIS, Google Earth, and Excel. The database for spatiotemporal analysis has 387 landslide events from three data sources: USGS Landslide Inventory, NASA COOLR, and PennDOT District 12 Slide Database. Due to similar terrain and climate conditions, landslide data in neighboring states is also

included to extend the database. The database is stored on the same three platforms. Based on landslide event dates, pre-event and post-event satellite images are collected and attached to each landslide in the database with hyperlinks. Fourteen causative factors of landslides are downloaded and calculated from Google Earth Engine and ArcGIS, and their values are extracted for every landslide data point in the database.

Based on the landslide databases compiled in the study, the landslide susceptibility assessment is conducted for PennDOT Districts 11 and 12 and adjacent areas, including northern West Virginia and eastern Ohio. Based on the spatial landslide database, the frequency ratio method is used to find the correlation between landslide occurrence and eight preliminarily selected causative factors, and a preliminary LSM is generated in terms of frequency ratios. ML methods are applied to conduct binary classification for landslide prediction with all fourteen landslide causative factors. The results show that the best model is GBM, which yields an AUC score of 0.87. The model is used to generate LSM showing the distribution of landslide probabilities. Based on the spatiotemporal landslide database, the spatiotemporal ML is conducted to predict landslide occurrence on both spatial and temporal scales. Eight additional rainfall factors are considered as temporal features, and non-landslides are sampled both in space and time to introduce spatial and temporal information into the datasets. Through the buffer-controlled sampling method and different landslide window periods, different spatiotemporal landslide datasets are constructed and the performance of ML models trained using these datasets is compared. The optimal model with an AUC score of 0.86 is used for susceptibility mapping. The spatiotemporal landslide susceptibility map is interpreted through comparisons with the pure spatial susceptibility map and the 7-day cumulative precipitation map. The results indicate that the spatiotemporal ML model can predict landslides, accounting for both spatial terrain factors and temporal rainfall factors, and the model outperforms pure spatial ML models with the same database size. Hence, the spatiotemporal LSM has the potential for applications in landslide hazard mitigation and forecasting.

The LSMs generated from this study highlight areas having very low to very high risk of landslide susceptibility with precipitation, which may be used to establish a hierarchy and mitigate risk for slopes at “very high risk” for landslide susceptibility. The maps may also be used for forecasting purposes. For example, they may be used as an aid for planning and programming

purposes to address slopes with “very high” landslide susceptibility first. The maps may be used in the event of incoming storms to target slopes with a very high risk of landslide susceptibility so that mitigation or preventative measures (such as temporary road closure) can be employed to ensure safe travel and minimize damage. In addition, the maps may also help to target post-storm roadway/slope inspections to the most critical and high-risk locations first.

8.2 Limitations

From the perspective of data science, the limitations of the current study are discussed below.

Uncertainty with the location and date of landslide events: The longitude and latitude coordinates of the landslide events have uncertainty. NASA COOLR catalog notes that there is a location uncertainly for every recorded landslide event. Most of the events have a location accuracy of 1 km or 5 km (Emberson et al. 2022), which means the landslides cannot be located accurately by longitude and latitude coordinates provided by NASA COOLR. For USGS Landslide Inventory, there is relative confidence in the characterization of the location of each landslide as discussed in Section 2.1.2. Although only landslides with Confidence (5) and (8) are chosen and included in the databases, the accurate locations of some landslides cannot be determined. In addition to the location, the event dates recorded in NASA and the USGS database are also estimated based on post-event on-site investigations. With the uncertainty in the location and event date, the contributing factors and rainfall data may have similar uncertainty.

Static causative factors: The fourteen causative factors are assumed to be static (i.e., they do not change with time). Since there was no advanced satellite imagery in the past, the true contributing factor data for old landslides cannot be obtained. The causative factors are assumed to be static; hence, modern satellite images representing the recent condition are assumed to represent the condition at the event. The difference between the real causative variables at the event time and those at the current time may cause errors in ML results.

Insufficient samples and variables in the current databases: From the perspective of data science, the size of the training dataset has a significant impact on the performance of ML. Ample training samples are required for ML algorithms to discover the true relationships between every variable and the target. The current databases contain 4,543 and 223 landslide data points for

spatial and spatiotemporal predictions, respectively, which may not be large enough for training an excellent ML model. In addition to the number of samples, the number of features (causative factors) is also important for model performance. The current databases contain fourteen spatial causative factors and eight precipitation factors, which may not be enough for explaining all landslide occurrences.

9 Recommendations and Instructions for Generating LSMs using the Developed ML Models

Based on the spatiotemporal landslide database, including 223 landslide data points with 14 topographic causative factors and eight rainfall factors, a Random Forest ML model was developed and trained to predict the probability of landslide occurrence both in space and time. The developed ML model and the relationship between causative factors (topographic and rainfall factors) and landslide occurrence are specific to the landslide database compiled and utilized for model training. If more landslides with event dates are reported in the study area, the landslide database can be updated and the ML model should be updated using the new training samples. Consequently, the relationship between causative factors and landslide occurrence should be updated acccordingly, which will result in different LSMs. Because rainfall factors are time-varying factors, the spatiotemporal LSM also changes with time: spatiotemporal LSM changes every day. Therefore, to generate a spatiotemporal LSM, it is required to provide the model with those causative factors at a specific date. In the present study, the fourteen topographic factors are assumed to be static, which means they will not change with time. Hence, the most important step in constructing a spatiotemporal LSM is to obtain rainfall factors for a specific date. If an incoming storm is forecasted, it is essential to incorporate the forecasted rainfall amounts into the rainfall factors to generate appropriate LSMs for the incoming storm.

The spatiotemporal LSM is not a single map as it dynamically varies with time based on precipitation data. The landslide database and ML codes can be stored in a cloud platform or a local platform. Procedures for generating the LSMs are presented below for each platform.

9.1. Cloud platform

9.1.1 Deliveries

Number	Items
1	Python code
2	Relevant data (landslide database in CSV and causative factors in TIF files)
3	ArcGIS template file

9.1.2 Requirements

Number	Items
1	An account for Google Drive login
2	A cloud platform for Python (Google Colab is recommended)

9.1.3 Instructions

To create spatiotemporal LSMs for different dates, the corresponding rainfall factors at these dates need to be downloaded and updated to the ML model. Hence, there will be frequent data savings and downloads. Using a Cloud platform to implement Python codes is a convenient and easy way for data import and export. The steps to generate an LSM are as follows.

Step 1: Google Drive login

A folder named "PennDOT_project" containing all relevant data, including the landslide database in Excel (e.g., Figure 13) and causative factors in TIF files (e.g., Figures 18 to 27 and 30 to 34), will be provided. Upload the folder to My Drive (see Figure 62). This folder contains the data of 14 topographic causative factors in the study area in the format of TIF files (see Figure 63). Since topographic factors will not change with time, they can be provided in advance and used continuously without needing periodic updates. The folder also contains the landslide database in CSV format for model training. It is noted that there is a subfolder named "rainfall_images" (see Figure 63), it will be created automatically as rainfall data is downloaded using the Python code, which will be explained in Step 3 in detail.

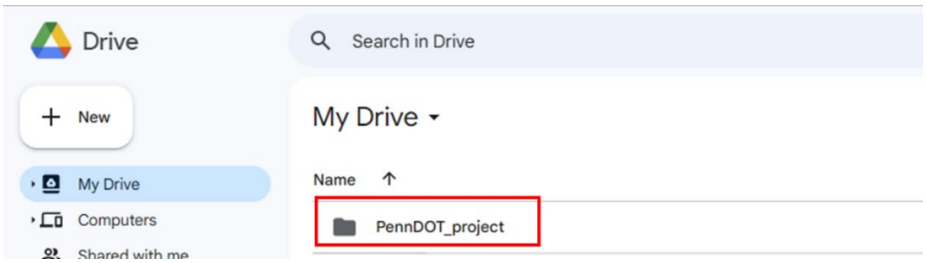


Figure 62. Upload the folder for relevant data.

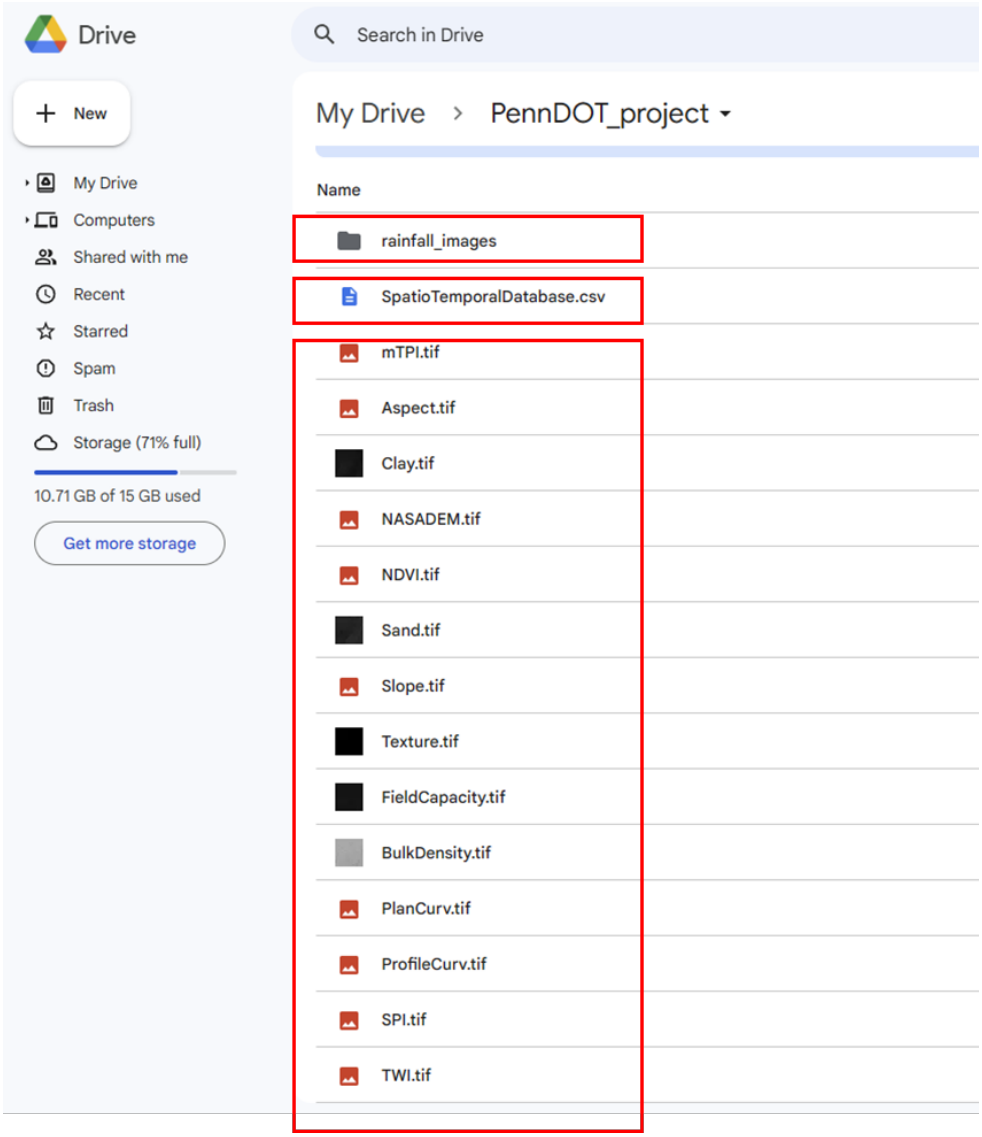


Figure 63. Contents in the folder of "PennDOT_project".

Step 2: Google Colab Setup

Google Colab is a cloud-based service that allows users to write and run Python code in the web browser. The Google Colab setup process can be completed with the following steps across all devices:

1). Visit the Google Colab page, which will direct the user to the Google Colaboratory Welcome Page.

2). Click the Sign in button on the right top (see Figure 64).

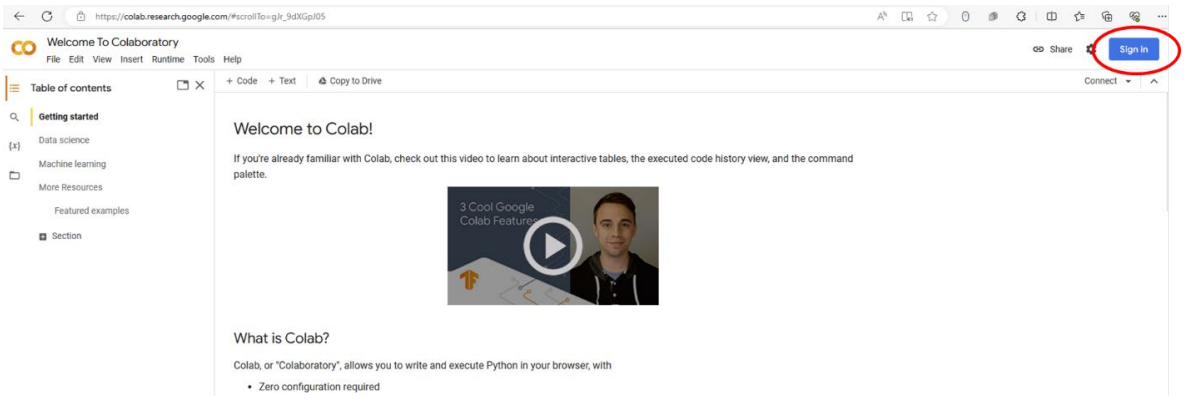


Figure 64. A Screenshot of Google Colab Welcome Notebook.

3). Sign in with a Gmail account (see Figure 65).

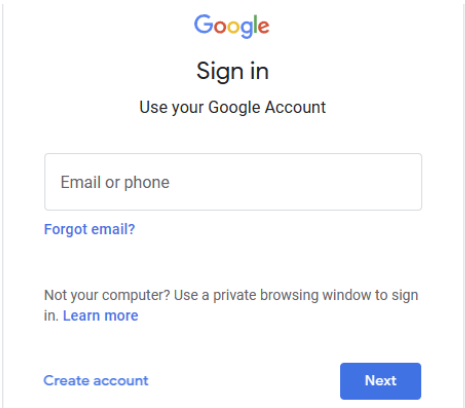


Figure 65. Google Sign-in Page.

4). Ready to use Google Colab.

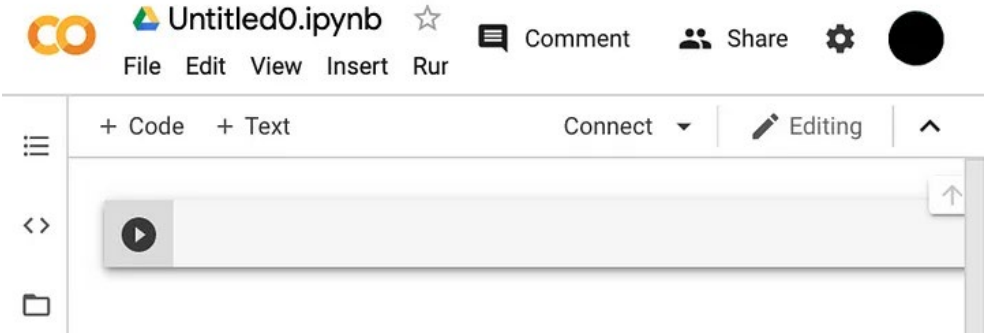


Figure 66. A Screenshot of an empty Google Colab Notebook.

Step 3: Python code implementation

A complete Python code will be provided for implementation. The code is divided into the following four parts to implement according to their functions.

Part 1. Package installation

Run the codes as Figure 67 shows; this will install the required packages.

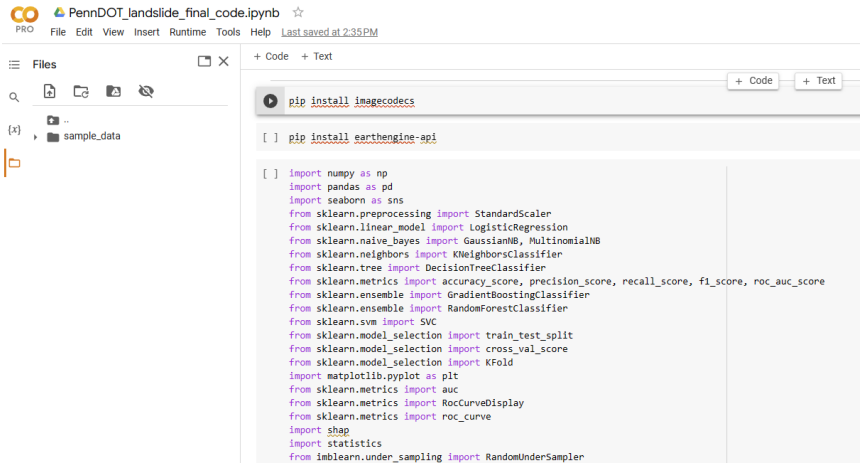


Figure 67. Python code in Part 1 for package installations.

Part 2. Link Google Drive and Python environment

Run the codes as Figure 68 shows, this will link the directories and data in Google Drive to Colab. Afterward, the relevant data provided in the folder "PennDOT_project" is available in Colab.

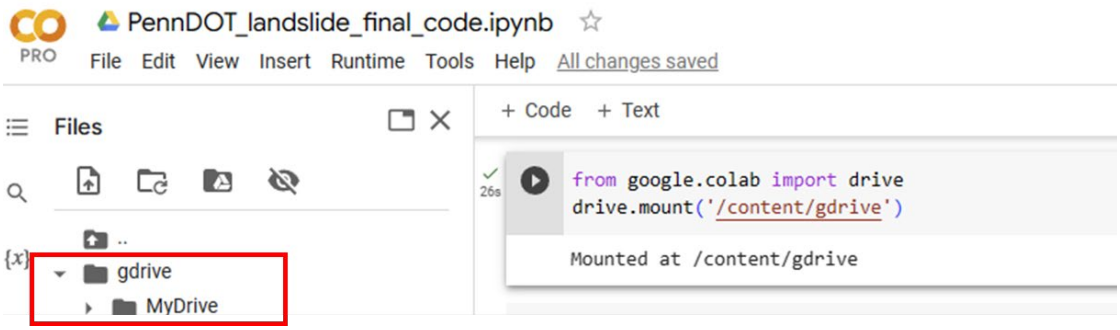


Figure 68. Python code in Part 2 for linking to Google Drive.

Part 3. Download rainfall data for a specific date

Run the codes as Figure 69 shows, open the following URL in a web browser, and follow the instructions to get a verification code for downloading data from Google Earth Engine.

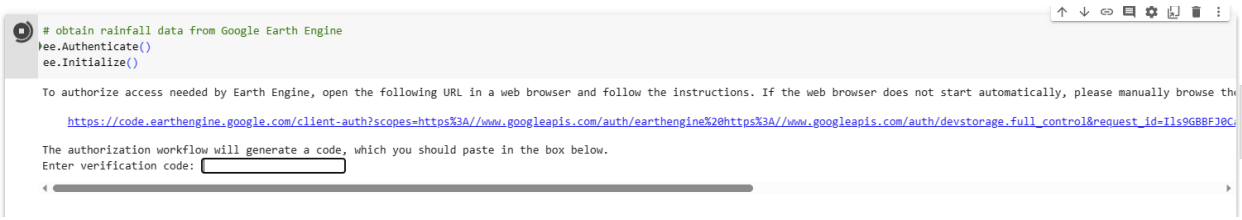


Figure 69. Python code in Part 3 for the authorization of accessing Google Earth Engine.

Run the codes as Figure 70 shows and enter a specific date in the box. The rainfall data for that date will be downloaded from Google Earth Engine and automatically saved in a new subfolder named "rainfall_images" in the folder "PennDOT_project".

Figure 70. Python code in Part 3 for downloading rainfall data from Google Earth Engine.

Part 4. Generate a landslide susceptibility map for the specific date

Run the remaining codes, then an LSM for the specific date entered above (2018-02-15 is entered as an example in this instruction) will be generated automatically, which is shown in Figure 71 (the map is on 2018-02-15). It is noted that the map generated in Python is within a larger rectangular area than the study area. The rectangular range is used for convenience for matrix operations in Python. To visualize the susceptibility map for the study area specifically, the following step in ArcGIS is required.

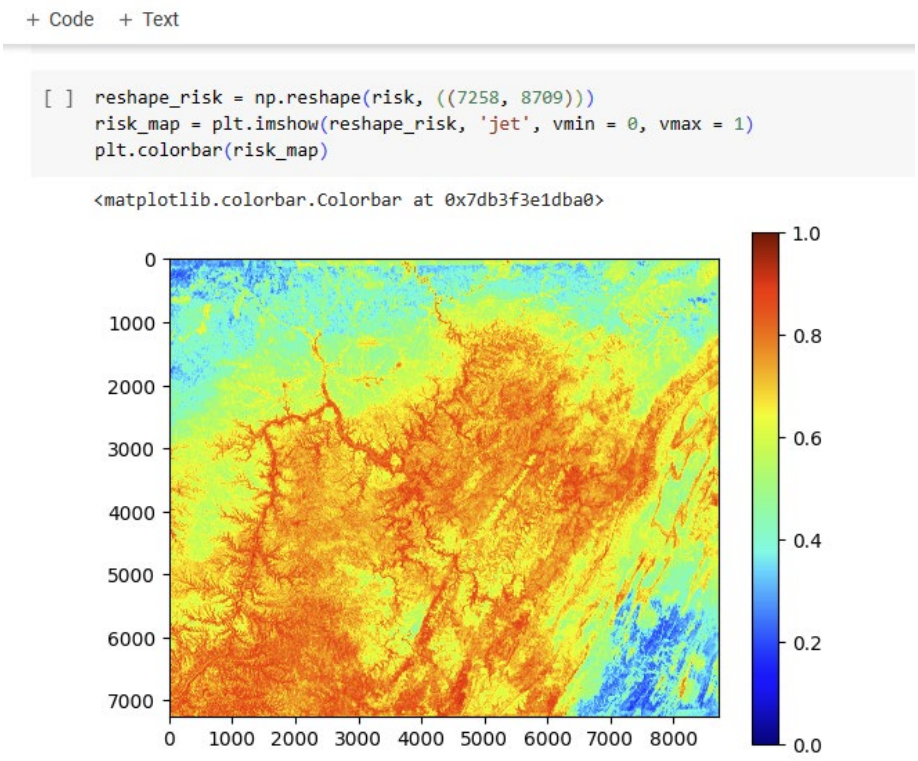


Figure 71. Python code in Part 4 for generating landslide susceptibility map.

Step 4: Map visualization in ArcGIS

Upload the LSM generated in Python to ArcGIS, as Figure 72 shows.

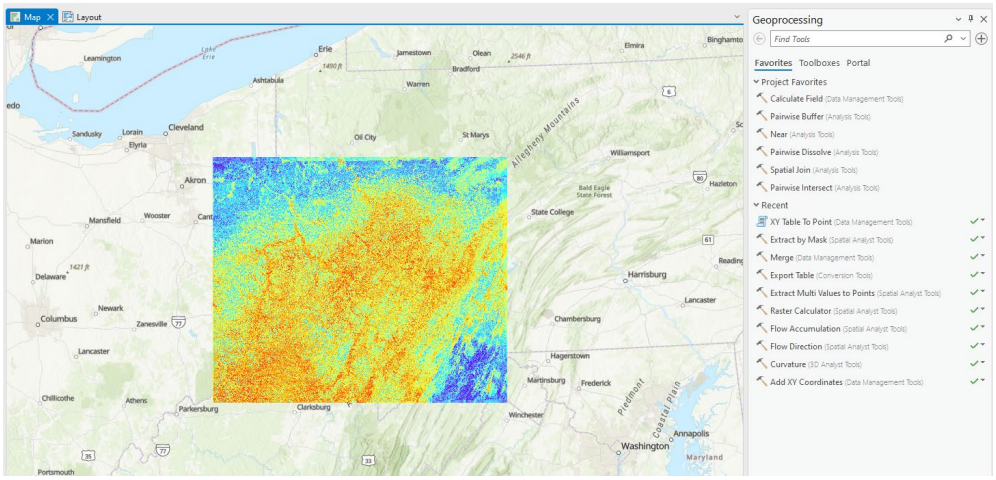


Figure 72. The landslide susceptibility map uploaded to ArcGIS.

An ArcGIS template file will be provided, where the boundary shape of the study area is included, as Figure 73 shows. To create the susceptibility map for the study area specifically, an Extraction Tool in ArcGIS can be used to clip the rectangular map (see Figures 73 and 74). Finally, to classify the probability of landslide occurrence into five classes with equal intervals (0%, 20%, 40%, 60%, 80%, 100%), edit the symbology for the map as Figure 75 shows. Click on any location on the map; the corresponding landslide probability will pop up, as Figure 75 shows.

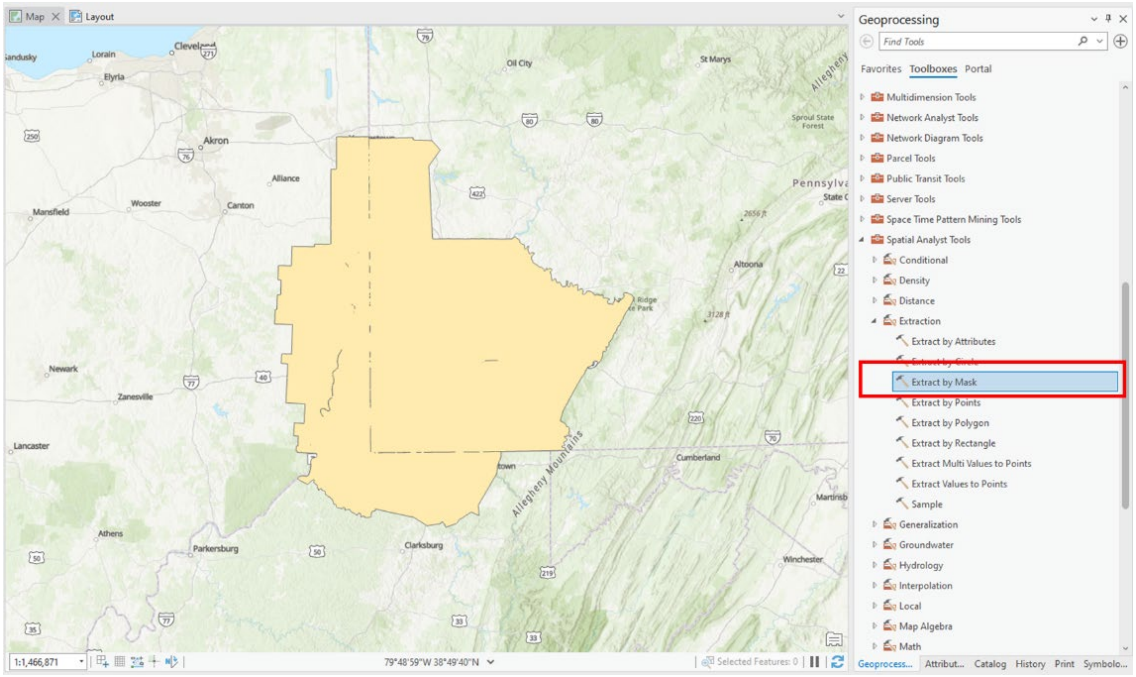


Figure 73. The boundary shape of the study area in ArcGIS.

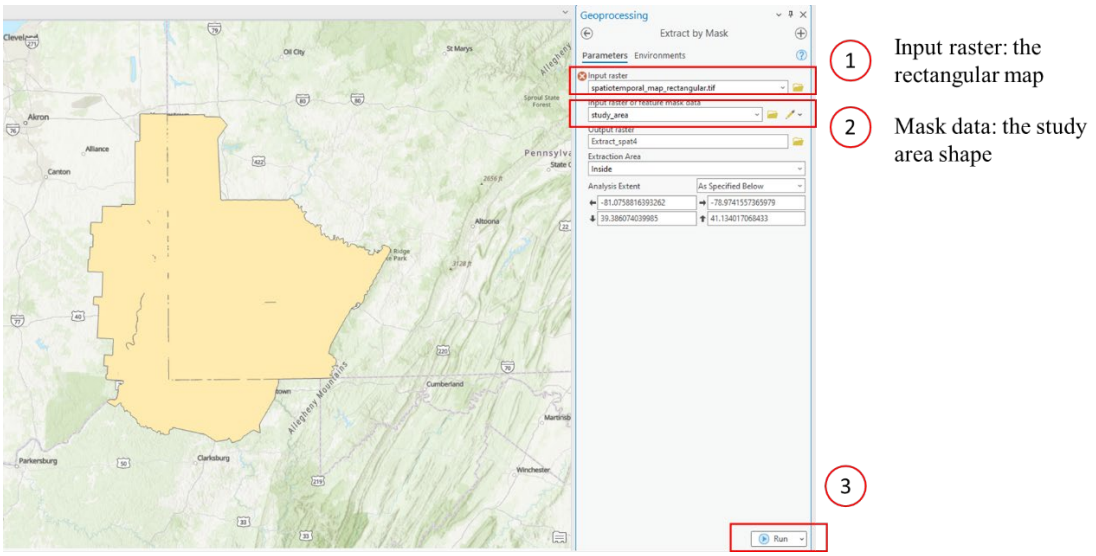


Figure 74. Steps for using Extraction Tool in ArcGIS to clip the rectangular map.

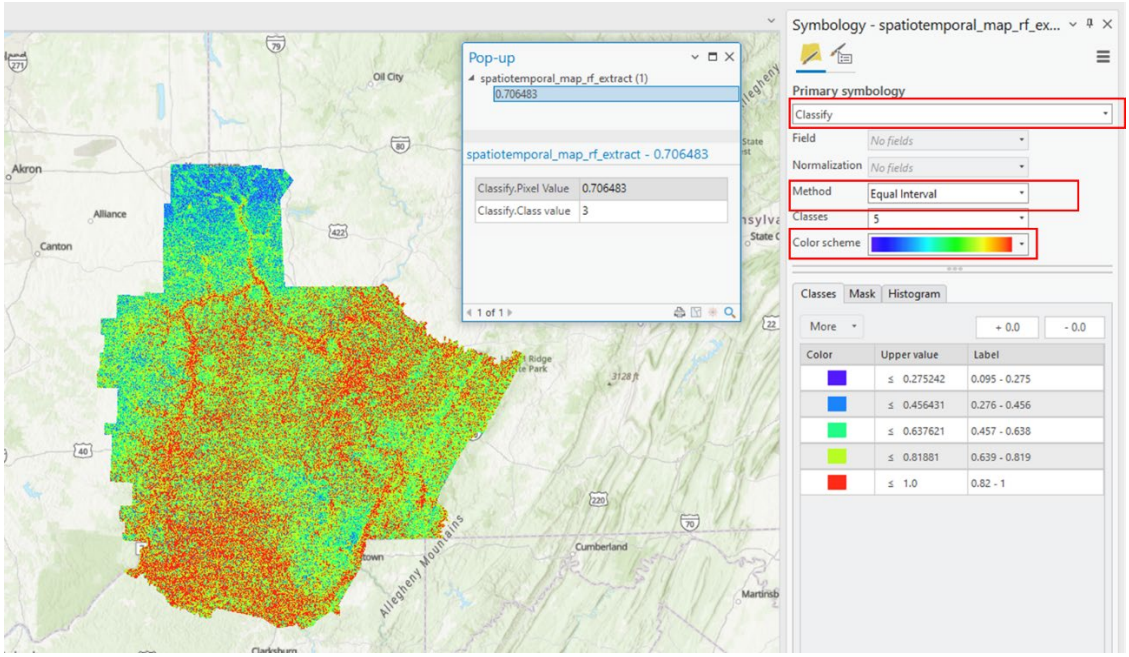


Figure 75. Classifying the probability of landslide occurrence in ArcGIS.

9.2. Local platform

9.2.1 Deliveries

Number	Items
1	Python code
2	Relevant data (landslide database in CSV and causative factors in TIF files)
3	ArcGIS template file

9.2.2 Requirements

Number	Items
1	A local platform for Python (Jupyter Notebook is recommended)

9.2.3 Instructions

The same items as those from the Cloud platform will be provided. The only difference between the local platform and the cloud platform is the location of directories for data storage. Instead of uploading the relevant data to Google Drive in Step 1, the user can upload the folder to the local Python environment directly, and the downloaded data (rainfall data) will be stored in the local directory. There is no need to link to Google Drive when implementing the code. The essential part of this approach is to install and set up a local environment for Python language correctly. Step 3 and Step 4 will be the same as those for the cloud platform.

References

Ado, M.; Amitab, K.; Maji, A.K.; Jasińska, E.; Gono, R.; Leonowicz, Z.; and Jasiński, M. (2022). "Landslide Susceptibility Mapping Using Machine Learning: A Literature Survey." *Remote Sens*, 2022, *14*, 3029.

Akgun, A. (2012). "A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio methods: a case study at İzmir, Turkey." *Landslides*, 9 (1), 93–106.

Anbalagan, R. (1992). "Landslide hazard evaluation and zonation mapping in mountainous terrain." *Engineering Geology*, 32(4): 269-277.

Ayalew, L., and Yamagishi, H. (2005). "The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan." *Geomorphology*, 65:15–31.

Ballabio, C., and Sterlacchini, S. (2012). "Support Vector Machines for Landslide Susceptibility Mapping: The Staffora River Basin Case Study, Italy." *Math Geosci*, 44, 47–70.

Batjes N.H., Ribeiro E., van Oostrum A., Leenaars J., Hengl T., and Mendes de Jesus J. (2017). "WoSIS - Providing standardized soil profile data for the world." *Earth System Science Data*, 9, 1-14.

Chacon, J., Irigaray, C., Fernandez, T., and El-Hamdouni, R. (2006). "Engineering geology maps: landslides and geographical information systems." *Bull. Eng. Geol. Environ*, 65, 341–411.

Cruden D.M. and Varnes D.J. (1996). "Landslide types and processes," Landslides—Investigation and mitigation: Transportation Research Board, Special report no. 247, A.K. Turner and R.L. Schuster (eds.), National Research Council, National Academy Press, Washington, D.C., pp 36–75.

Delano, H.L. and Wilshusen, J.P. (2001). "Landslides in Pennsylvania: Pennsylvania Geological Survey." *4th ser., Educational Series 9*, *34 p*.

Ding, Q. F., Chen, W., and Hong, H. Y. (2016). "Application of frequency ratio, weights of evidence and evidential belief function models in landslide susceptibility mapping." *Geocarto Int*.

Emberson, R., Kirschbaum, D., Amatya, P., Tanyas, H., and Marc, O. (2022). "Insights from the topographic characteristics of a large global catalog of rainfall-induced landslide event inventories." *Natural Hazards and Earth System Sciences*, 22, 1129-1149.

Eric, S.J., Stephen, L.S., and Benjamin, B.M. (2022). "Landslide Inventories across the United States version 2." *Geologic Hazards Science Center*, USGS, 10.5066/P9FZUX6N.

Ermini, L., Catani, F., and Casagli, N. (2005). "Artificial neural networks applied to landslide susceptibility assessment." Geomorphology 66:327–343.

Guo, C. B., Montgomery, D. R., Zhang, Y. S., Wang, K., and Yang, Z. H. (2015). "Quantitative assessment of landslide susceptibility along the Xianshuihe fault zone, Tibetan Plateau, China." *Geomorphology*, 248:93–110.

Hengl, T., and MacMillan, R.A., (2019). "Predictive Soil Mapping with R." *OpenGeoHub foundation*, Wageningen, the Netherlands, 340 pages. ISBN: 978-0-359-30635-0.

Hengl, T., de Jesus, J.M., Heuvelink, G.B., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B. and Guevara, M.A., (2017). "SoilGrids250m: Global gridded soil information based on machine learning." *PLoS one*, 12(2), p.e0169748.

Ho, T. K. (1995). "Random decision forests." *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, pp. 278-282 vol.1.

Ilanloo, M. (2011). "A comparative study of fuzzy logic approach for landslide susceptibility mapping using GIS: an experience of Karaj dam basin in Iran." *Procedia Soc Behav Sci*, 19:668–676.

Irigaray, C., Fernández, T., El-Hamdouni R., and Chacón, J. (2007). "Evaluation and validation of landslide-susceptibility maps obtained by a GIS matrix method: examples from the Betic Cordillera (southern Spain)." *Nat Hazards*, 41:61–79.

Kannan, M., Saranathan, E., and Anbalagan. R. (2013). "Landslide vulnerability mapping using frequency ratio model: a geospatial approach in Bodi-Bodimettu Ghat section, Theni district, Tamil Nadu, India." *Arab J Geosci*, 6:2901–2913.

Kavzoglu, T., Colkesen, I., and Sahin, E.K. (2019). "Machine learning techniques in landslide susceptibility mapping: A survey and a case study." *Landslides: Theory, Practice and Modelling*, 283-301.

Kirschbaum, D.B., Adler, R., Hong, Y., Hill, S., and Lerner-Lam, A. (2010). "A global landslide catalog for hazard applications: method, results, and limitations." *Natural Hazards*, 52(3), 561-575.

Kirschbaum, D.B., Stanley, T., and Zhou, Y. (2015). "Spatial and Temporal Analysis of a Global Landslide Catalog." *Geomorphology*, 10.1016/j.geomorph.2015.03.016.

Lee, S., and Choi, J. (2004). "Landslide susceptibility mapping using GIS and the weight-of evidence model." *Int J Geogr Inf Sci*, 18:789–814.

Lee, S., and Pradhan, B. (2007). "Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models." *Landslides*, 4 (1), 33-41.

Lee, S., and Talib, J. A. (2005). "Probabilistic landslide susceptibility and factor effect analysis." *Environ Geol*, 47:982–990.

Merghadi, A.; Yunus, A.P.; Dou, J.; Whiteley, J.; ThaiPham, B.; Bui, D.T.; Avtar, R.; and Abderrahmane, B. (2020). "Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance." *Earth-Sci Rev*, 207: 103225.

Moziihrii, A., Khwairakpam, A., Arnab, K.M., Elzbieta, J., Radomir, G., Zbigniew, L., and Michał, J. (2022). "Landslide Susceptibility Mapping Using Machine Learning: A Literature Survey." *Remote Sensing*, 14(13): 3029.

Naemitabar, M.; and Zanganeh Asadi, M. (2021). “Landslide zonation and assessment of Farizi watershed in northeastern Iran using data mining techniques.” *Nat. Hazards*, 108: 2423–2453.

Natekin, A., and Knoll, A. (2013). “Gradient boosting machines, a tutorial.” *Front Neurorobot*. doi:10.3389/fnbot.2013.00021.

Pomeroy, J.S. and William, E.D. (1979). “Landslides and related features, Pennsylvania - Pittsburgh 1°×2° sheet.” *U.S. Geological Survey*, 10.3133/ofr791314.

Pradhan, B., and Lee, S. (2009). “Landslide risk analysis using artificial neural network model focusing on different training sites.” *Int J PhysSci*, 4:001–015.

Raghuvanshi, T.K., Ibrahim, J., and Ayalew, D. (2014). “Slope stability susceptibility evaluation parameter (SSEP) rating scheme - an approach for landslide hazard zonation.” *Journal of African Earth Sciences*, 99: 595-612.

Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M., and Guzzetti, F. (2018). “A review of statistically-based landslide susceptibility models.” *Earth Sci Rev*, 180:60–91.

Sarkar, S., Roy, A. K., Martha, T. R. (2013). “Landslide susceptibility assessment using information value method in parts of the Darjeeling Himalayas.” *J Geol Soc India*, 82:351–362.

Silva, B.P.C., Silva, M.L.N., Avalos, F.A.P., de Menezes, M.D., and Curi, N. (2019). “Digital soil mapping including additional point sampling in Posses ecosystem services pilot watershed, southeastern Brazil.” *Scientific reports*, 9(1), 1-12.

Thiery, Y., Malet, J. P., Sterlacchini, S., Puissant, A., and Maquaire, O. (2007). “Landslide susceptibility assessment by bivariate methods at large scales: application to a complex mountainous environment.” *Geomorphology*, 92:38–59.

Tsangaratos, P., and Benardos, A. (2014). “Estimating landslide susceptibility through an artificial neural network classifier.” *Nat Hazards*, 74:1489–1516.

Tucker, C., Grant, D., and Dykstra, J. (2004). “NASA’s Global Orthorectified Landsat Data Set.” *Photogrammetric Engineering & Remote Sensing*, 70, 313-322.

Vapnik, V. N. (1995). “The nature of statistical learning theory.” *Springer*, New York.

Wang, L. J., Sawada, K., and Moriguchi, S. (2013). “Landslide susceptibility analysis with logistic regression model based on FCM sampling strategy.” *Comput Geosci*, 57:81–92.

Wang, Y., Song, C. Z., Lin, Q. G., and Li, J. (2016). “Occurrence probability assessment of earthquake-triggered landslides with Newmark displacement values and logistic regression: The Wenchuan earthquake, China.” *Geomorphology*, 258:108–119.

Yao, X., Tham, L. G., and Dai, F. C. (2008). “Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China.” *Geomorphology*, 101:572–582.

Yesilnacar, E., and Topal, T. (2005). “Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey).” *Eng. Geol*, 79 (3–4), 251–266.

Yilmaz, I. (2010). “Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine.” *Environ. Earth Sci*, 61, 821–836.

Youssef, A.M. (2015). “Landslide susceptibility delineation in the Ar-Rayth area, Jizan, Kingdom of Saudi Arabia, using analytical hierarchy process, frequency ratio, and logistic regression models.” *Environ Earth Sci*, 73, 8499–8518.

Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., and Al-Katheeri, M. M. (2016). “Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah basin, Asir region, Saudi Arabia.” *Landslides*, 13:839–856.

Zhang, W., He, Y., Wang, L., Liu, S., and Meng, X. (2023). “Landslide Susceptibility mapping using random forest and extreme gradient boosting: A case study of Fengjie, Chongqing.” *Geological Journal*, 58(6), 2372–2387.