Leveraging Machine Learning for Wide-Scale Climate Change Informed Slope Stability Prediction in the United States

Alaguvalliappan Thiagara
Dept of Liberal Arts and Sciences
University of Florida
Gainesville, FL, 36111
athiagarajan@ufl.edu

Suguang Xiao
Dept of Civil and Env Eng
Clarkson University
Potsdam, NY, 13699
sxiao@clarkson.edu

***Abstract*— Machine learning has emerged as a valuable tool for assessing landslide susceptibility and slope stability, surpassing traditional statistical methods in accuracy and numerical simulations in terms of time-cost. Machine learning for slope stability has been hindered by the scarcity of real-world case studies, necessitating the use of simplified algorithms or simulated data for model training. To address this challenge, this project leverages NASA's landslide inventory, which encompasses nearly 4000 landslides in the United States to get labels for our training data. This data is augmented with SSURGO soil parameter data, USGS 1/3 arc-second elevation data, soil moisture data from SMAP, and precipitation data from Meteostat to get the parameters for our model. The random forest model used in this project can classify areas susceptible to landslides in a binary manner 90 percent of the time and the random forest helped evaluate which features were relevant in predicting which areas are susceptible to landslides. Our model outperforms LHASA 1.1 in the US in terms of evaluating the probabilistic susceptibility to landslides. Moreover, the model can be used to predict landslides with increased precipitation conditions. The preliminary results indicate that over 20% of currently stable sites become unstable with the predicted maximum daily precipitation in 2073.**

Keywords—Machine Learning, Slope Stability, Landslides, Random Forest, USGS, SSURGO, Meteostat, LHASA 1.1

I. Introduction

Landslides represent a critical issue with far-reaching consequences, impacting both infrastructure and the well-being of communities and natural ecosystems. Shuster (1978) concluded that landslides in the United States caused more than 1 billion 1978 dollars (4.6 billion 2023 dollars) in damages to property annually [1]. The potential devastation they bring underscores the urgent need to develop effective landslide forecasting methods. Accurate predictions play a pivotal role in implementing proactive measures to safeguard existing infrastructure and preserve the integrity of surrounding landscapes. Finding sites with low landslide susceptibility is also essential since it allows planners of future infrastructure projects to make well-informed choices. Landslide forecasting systems will become even more relevant in the future as extreme weather conditions rise owing to climate change. Despite the importance of wide-scale landslide forecasting tools, the tools available are limited in nature.

The traditional approaches for evaluating slope stability are Limit Equilibrium Method (LEM) and Finite Element Method (FEM) which both suffer from requiring parameters that are difficult to collect at scale. While machine learning (ML) approaches exist, they suffer from using either too general or too specific data. Some machine learning models, like LEM and FEM, use parameters that must be collected in laboratories for a given slope, such as cohesion and internal friction angle, which makes it more difficult to use these approaches at large scale. [2].

Moreover, other models (e.g., NASA's LHASA 1.1) that can be deployed at scale tend to use 1 arc-second elevation data which is imprecise and when accounting for precipitation, they tend to use precipitation data gathered from satellites which are far less accurate than precipitation data gathered by weather stations [3,4]. By using more precise data sources that are still available for the entirety of the United States, a widely deployable model can be created that is far more accurate than current wide-scale landslide prediction tools. With ensemble models, the model will be more interpretable than a deep-learning model allowing us to understand the influence of various parameters on determining landslide susceptibility in our model.

II. Methodology

A. Label Collection

Binary labels were created for the sake of simplicity and the ability to collect more data. The two labels are stable and unstable, with stable points assuming that there has never been a slope failure in the area and unstable points supposing that a landslide has occurred there.

Unstable cases were collected from the NASA Global Landslide Catalog (GLC) which is a repository where scientists can share landslide reports [5,6]. The quality of the data collected varies so data points where the locations of the landslides were exactly known were chosen. The total dataset had 39617 data points with 3917 occurring in the United States, of which 1293 data points had exact locations. The data was stored as a comma-separated values (CSV) file and the latitude and longitude information for each landslide as well as the date the landslide occurred were used to characterize the unstable points.

To characterize stable points, random locations were chosen in the United States that are at least 50 miles apart from any landslide points in the GLC, and for the date of the stable points, a random date between 2010 and 2022 was chosen since 97.6% of points in the US with precise locations in the GLC occurred during this period.

B. *Parameter Collection*

When looking for predictors for landslides, the three types of parameters to consider are parameters relating to slope geometry, slope materials, and external conditions.

Slope geometry refers to the shape of the slope, including the slope angle, slope length, slope aspect, and slope curvature. The parameter chosen to represent the slope geometry in our model is slope angle because a higher slope angle increases the downward force of gravity, making a slope far more susceptible to landslides. The highest resolution digital elevation model (DEM) in the United States is the United States Geological Survey (USGS) 10-meter elevation dataset and by taking a gradient of the elevation dataset at the data point using the python library richDEM, a value for the slope angle of the 10-meter cell where the data point is located was obtained [7,8]. RichDEM calculates the slope “using a central difference estimation of a surface fitted to the focal cell and its neighbours” [9]. From this, the slope angle is chosen in the direction of the highest value.

Slope materials refer to the properties and the composition of the materials that compose the slope. Rockfalls were removed from the dataset, so we’ve data is gathered on soil parameters to account for the slope materials. Data was collected from the SSURGO National Cooperative Soil Survey [10]. The data is split into map units, and to collect point data within the map unit, values are collected in soil pits which are generally between 153 to 203 cm deep and 1.5 to 2 m wide [11]. As for the specific choice of soil parameters, initially, a wide range of available soil parameters were included in our dataset and machine learning models and domain expertise decided which soil parameters were relevant in our model.

External conditions are often the triggers for landslides and can include seismic activity, human activity, and precipitation. Precipitation was the first external condition considered in our model and to collect precipitation data, Meteostat was used which has an Application Programming Interface in Python that combines bulk data from different weather stations [12]. Through Meteostat, the maximum daily precipitation in the location for the past 7 days before the event date was queried to see if there was any extreme precipitation in the preceding week that could have triggered a landslide and the number of hours/days with heavy rain 365 days before the event date was queried to get the number of heavy rainfall events in the past year [13,14]. Of the 1293 exact landslides in the United States, 270 non-rockfall locations had all the parameters needed so 270 stable cases were generated to have a dataset of 540 entries.

C. *Random Forest*

The machine learning model used for our model was a random forest. The benefits of random forests are that they are an interpretable machine learning model that is resistant to overfitting while having strong accuracy with relatively small datasets. Random forests are an ensemble machine learning model that works by averaging the results of decision trees with bootstrap where each decision tree is given a random subset of features. Decision trees are a supervised learning algorithm that can be used for classification tasks [15]. In classification tasks, the trees follow a hierarchical structure that originates from a root node and depending on the values of the feature at each node, a path is followed through internal nodes until a leaf node is reached. The leaf node represents the outcomes, in this case, stable or unstable [15].

III. Results

A. *Feature Importances*

Looking at feature importance allows us to see which features are relevant to our predictions. By keeping the top 6 features in our model, (the slope angle collected from taking the gradient of the USGS 10-meter elevation dataset, the deepest horizon layer of the slope profile, the slope of the soil map unit collected by SSURGO, the maximum daily precipitation 7 days preceding the event date collected using Meteostat, the Bulk Density of the soil map unit, and the amount of Organic Material in the soil map unit) the model maintains the same accuracy as the model trained with all the features. Gini importance, the metric used to evaluate feature importance, is defined as the total decrease in node impurity weighted by the proportion of samples reaching the node, and then averaged over all the trees in the ensemble [16].

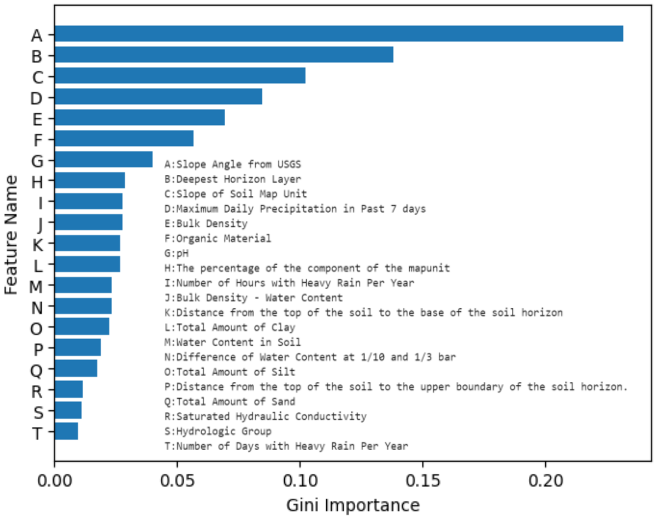
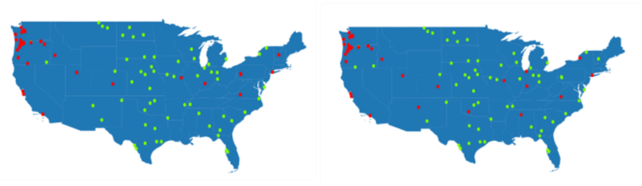


Fig 2: Gini Importances for Random Forest Model

B. *Test Set Performance*

Because our dataset is relatively small, with an 80/20 train test split where 108 samples are used for testing and 432 samples are used for training, the accuracy of our model ranges between 87 percent and 94 percent accuracy depending on the choice of test set. To better understand the general performance of our model, cross-validation was used. Cross-validation is a means for evaluating the model by averaging the test performance over different train/test splits. The 10-fold cross-validation accuracy of our model was 90.18 percent where the model had a slight bias towards false positives.



Actual Values in Test Set

Predicted Values in Test Set

Fig 3: Graphical Comparison of Predicted Test Set Value to Actual Test Set Values

C. *Comparison to NASA LHASA 1.1*

Since landslides are underreported, points not in the landslide inventory aren't known to be stable. Because of this, to evaluate NASA's LHASA 1.1, the researchers found it more informative to use a graph of susceptibility to evaluate the model, rather than an Receiver Operating Characteristic (ROC) curve [4].

In their comparison, they compare landslide locations to all other locations whereas our comparison is to locations that are at least 50 miles away from landslide locations. In addition, LHASA 1.1 is a global model, but a comparison of our model to LHASA 1.1 gives us an idea of our model's performance. In order to give a probabilistic evaluation of our model, the number of decision trees in the random forest that predict that the slope is unstable is divided by the total number of decision trees.

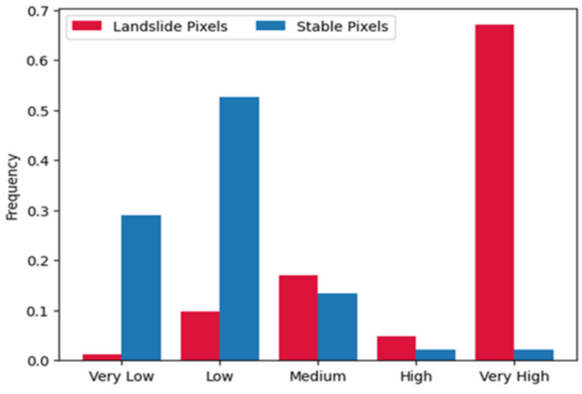


Fig 4: Distribution of susceptibility in our model for locations recorded in the GLC and locations at least 50 miles away that are in the test set.

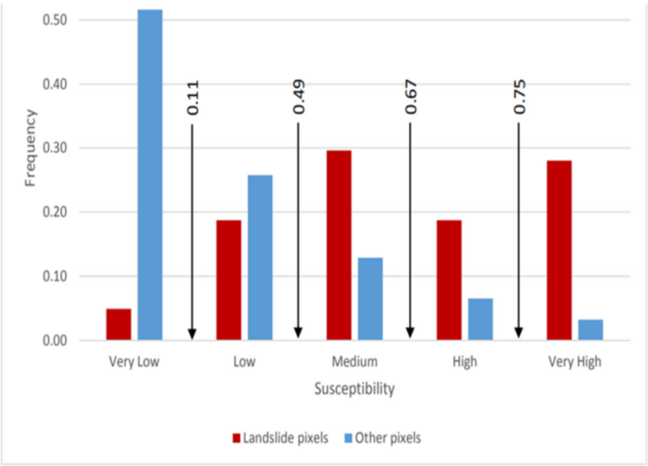


Fig 5: Distribution of susceptibility in NASA LHASA 1.1 for locations recorded in the GLC and other areas

D. *Prediction of Landslides under Increased Precipitation*

To provide a forecast of what precipitation conditions could look like in the future, the maximum daily precipitation per year from 1972 to 2022 is collected. The trend is used to forecast values for the maximum daily precipitation in the year 2073. When the values for the maximum daily precipitation for 2073 for each location were input into our model into the parameter representing the maximum daily precipitation 7 days before the event date, 21.6 percent of locations that our model currently classifies as stable were classified as unstable.

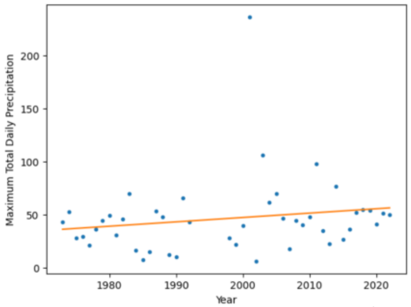


Fig 6: Maximum Daily Precipitation per Year from 1972-2022 for a random location in our dataset

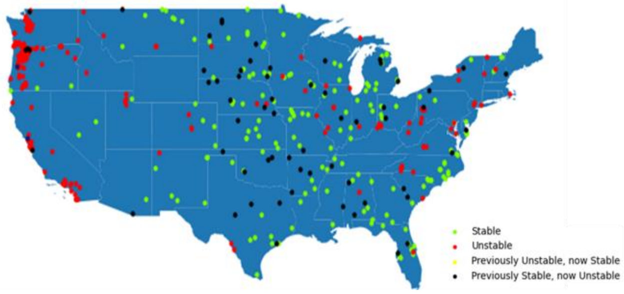

Stable
Unstable
Previously Unstable, now Stable
Previously Stable, now Unstable

Fig 7: Forecast of Stable and Unstable Locations with Forecasted Maximum Daily Precipitation in 2073

IV. Limitations and Future Work

Each slope angle pertains to the unit 10-meter cell at the specified latitude and longitude location listed in the GLC, so the slope angle isn't representative of the slope angle for the entire slope but for a 10-meter subset. The slope angle and the SSURGO soil parameters are assumed to be constant over time with the only temporal variable being precipitation. The model might struggle to distinguish the pivotal variable in determining landslide susceptibility in cases where variables are correlated which may have been why saturated hydraulic conductivity wasn't given high importance in our model. In the future, more precautions will be taken to account for correlation. More parameters (e.g., soil moisture, ground water, slope aspect, slope curvature) and other triggers (e.g., earthquakes) will be added to our model to provide more robust predictions. In the future, our data collection methods will be verified against data collected in landslide case sites.

V. Conclusions

The model in this study can evaluate which features were relevant in predicting landslide susceptibility. The random forest model was able to classify areas in the United States susceptible to landslides in a binary manner with a 10-fold cross validation accuracy of 90.18 percent. The model also outperforms LHASA 1.1 in probabilistic landslide susceptibility prediction. The model's responses to increased precipitation conditions were evaluated which will enable us to assess regions that might be vulnerable to landslides in the future due to the amplification of extreme precipitation events caused by climate change.

Acknowledgments

This work is supported by National Science Foundation under Grant Nos, OAC-2244049.

References

[1] Schuster, R. L., 1978, Introduction, Chapter 1, in 8~huster, R. L., and Krizek, R. J., eds., Landslides-Analysis and control: Washington, National Academy of Science, Transportation Research Board Special Report 176, p. 1-10.

[2] Guangjin Wang, Bing Zhao, Bisheng Wu, Chao Zhang, Wenlian Liu, Intelligent prediction of slope stability based on visual exploratory data analysis of 77 in situ cases, International Journal of Mining Science and Technology, Volume 33, Issue 1, 2023, Pages 47-59

[3] Matteo Gentilucci, Maurizio Barbieri, Gilberto Pambianchi, Reliability of the IMERG product through reference rain gauges in Central Italy, Atmospheric Research, Volume 278, 2022, 106340, ISSN 0169-8095, https://doi.org/10.1016/j.atmosres.2022.106340

[4] Stanley, T., and D. B. Kirschbaum. 2017. "A heuristic approach to global landslide susceptibility mapping." Natural Hazards, 1-20 1007/s11069-017-2757-y

[5] Kirschbaum, D.B., Stanley, T., & Zhou, Y. (2015). Spatial and temporal analysis of a global landslide catalog. Geomorphology, 249, 4-15. doi:1016/j.geomorph.2015.03.016

[6] Kirschbaum, D.B., Adler, R., Hong, Y., Hill, S., & Lerner-Lam, A. (2010). A global landslide catalog for hazard applications: method, results, and limitations. Natural Hazards, 52, 561-575. doi:1007/s11069-009-9401-4

[7] U.S. Geological Survey, 2022, 1/3rd arc-second Digital Elevation Models (DEMs) - USGS National Map 3DEP Downloadable Data Collection Geospatial_Data_Presentation_Form: raster digital data, accessed July-August 2023 at URL: https://data.usgs.gov/datacatalog/data/USGS:3a81321b-c153-416f-98b7-cc8e5f0e17c3

[8] Barnes, Richard. 2016. RichDEM: Terrain Analysis Software. http://github.com/r-barnes/richdem

[9] Barnes, Richard 2017. Terrain attributes¶. Terrain Attributes - RichDEM 0.0.03 documentation. Soil Survey Staff, Natural

[10] Resources Conservation Service, United States Department of Agriculture. Soil Survey Geographic (SSURGO) Database. Available online at https://sdmdataaccess.sc.egov.usda.gov. Accessed [month/day/year]

[11] McGuire, C. E. (n.d.). Collection and organization of soil data - ialcworld.org. https://www.ialcworld.org/conference/Pres-pdf/Mcgu1p1.pdf

[12] Lamprecht, C. (2022, August 17). Meteostat/meteostat-python: Access and analyze historical weather and climate data with python. GitHub. https://github.com/meteostat/meteostat-python

[13] Glossary of Meteorology (June 2000). "Rain". American Meteorological Society. Archived from the original on 25 July 2010. Retrieved 15 January 2010.

[14] Howard Perlman, U. (n.d.). Rainfall calculator, Metric Unitshow much water falls during a storm?. Rainfall calculator, metric-How much water falls during a storm? USGS Water Science School. https://water.usgs.gov/edu/activity-howmuchrain-metric.html

[15] What is a decision tree. IBM. (n.d.). https://www.ibm.com/topics/decision-trees

[16] Breiman, Friedman, "Classification and regression trees", 19