GeoClimb: Adventure Is Out There!

Ian Christie, Ankur Varma

February 18, 2025

1 Introduction

The climbing industry relies on exploration, but manually searching for high-quality climbing areas is both costly and time-consuming. By developing a model that predicts rock climbing locations and generalizing it on a global scale, we aim to significantly reduce the search space for exploration. We hypothesize that a combination of satellite imagery [1], elevation data [2], and lithologic information [3] provides sufficient signals to predict rock climbing potential in a given area. To train this model, we will use Mountain Project data [4] from the United States as labels. Because we cannot assume we have a complete label-set (ie. all climbing areas in the US have been explored), we will apply Presence Unlabeled (PU) training and evaluation techniques to our model.

Research Question: Can we build a model that uses lithology, elevation data, and satellite imagery to predict areas with rock climbing potential and generalize to unexplored regions?

2 Related Work

Rock Information: Extensive research has been conducted in identifying potential climbing areas on a local scale. For instance, [5] determine route quality from high-resolution photogrammetry, and [6] establish a Route Stability Index as a safety metric using high-resolution photogrammetry and geological information collected from field surveys. However, this information is costly to produce and not available on a global scale, therefore little work has been done to identify climbing areas more broadly. [7] uses satellite imagery to identify potential bouldering areas, but not rope climbing areas. We hypothesize that lithology—the rock composition of an area—can serve as an indicator for route quality and safety. Macrostrat [8] is a platform that aggregates and distributes geological data on a global scale, providing APIs for acquiring lithology information worldwide.

Species Distribution Models (SDMs): SDMs [9, 10] have been successfully employed to predict the distribution of various species based on ecological covariates. These models typically rely on climatic variables, soil characteristics, vegetation indices, land use/land cover, measures of human influence, and remote sensing imagery. The techniques used by these models are directly analogous to those we will employ in our model development. By extending this approach to rock climbing, we hypothesize that integrating geologic covariates (topographic and lithological features) alongside satellite imagery will improve prediction accuracy.

Positive and Unlabeled (PU) Learning: [11] introduce the basic concept of binary classification and its challenges when only positive and unlabeled data are available, as in areas such as medical diagnosis. In their work, they assume that all unlabeled examples are negative. They also discuss the F1 score and its limitations in PU settings, presenting an alternative performance criterion—based on recall r—that can be estimated from PU data.

The spatio-temporal prior introduced in [12] can be combined with any image classifier to improve performance on fine-grained image classification tasks. This approach has potential for use in other domains involving presence-only data and contextual information. The authors utilize readily available geographic location and capture time metadata to estimate the probability of object category occurrence based on location and time. They generate "pseudo-negatives" through random sampling to represent likely absences, prioritize confirmed presence data, and downweight incorrect predictions.

3 Current Results

In this work we introduce the *GeoClimb Dataset*, a novel dataset containing Sentinel-2 RGB imagery, Digital Elevation Model (DEM) data, and lithologic information for 12,742 climbing locations across the continental United States. The dataset also includes the same information for 12,692 unlabeled locations randomly distributed across the continental US. The data point, d, is represented a tuple $d = \{l, k, s, d, r\}$ where:

- l is the latitude and longitude of the location

- k is a boolean indicator denoting if the location is a known (1) or unknown (0) climbing area
- s is the Sentinel-2 image represented as $\mathbb{R}^{\sim 50 \times \sim 50 \times 3}$
- d is the DEM information represented as $\mathbb{R}^{\sim 17 \times \sim 17 \times 1}$
- r is the rock type information provided as a JSON blob

Note that the exact height and width of the sentinel-2 and DEM data is variable. All code can be found at our github and data can be downloaded from our google drive.

3.1 Labeled Climbing Locations

We acquired individual climbing route information from kaggle. This information was scraped in 2020 from The Mountain Project, a popular crowdsourced application where climbers provide structured geotagged route information. From this dataset we are able to distill the climbing routes into 12,776 unique climbing areas shown in FIG 1 (left). For this project, we focus on roped climbing (traditional or sport) and do not do any filtering of the dataset based on ratings or route length (see extra analysis for details).

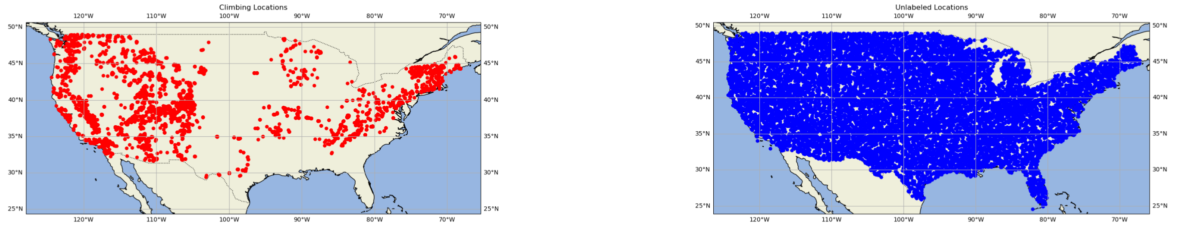


Figure 1: (left): Shows the spatial distribution of scraped climbing locations in the continental US. (right): Shows the spatial distribution of the generated unlabeled locations in the continental US. Note: these have the same number of points demonstrating the tight clustering of climbing locations.

3.2 Unlabeled Climbing Locations

To provide unlabeled data to our model we randomly sampled 12,776 points within the continental US that were more than 1 km away from any labeled location. The distribution of these can be seen in FIG 1 (right).

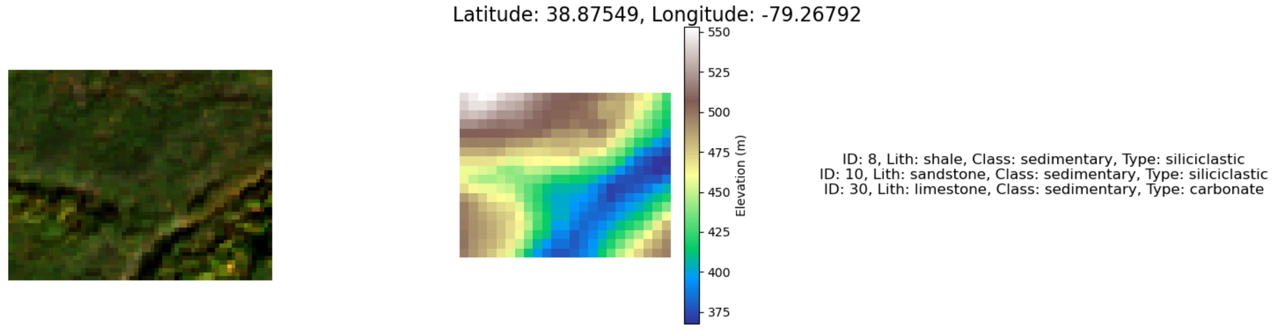

Latitude: 38.87549, Longitude: -79.26792
550
525
500
475
450
425
400
375
Elevation (m)
ID: 8, Lith: shale, Class: sedimentary, Type: siliciclastic
ID: 10, Lith: sandstone, Class: sedimentary, Type: siliciclastic
ID: 30, Lith: limestone, Class: sedimentary, Type: carbonate

Figure 2: Sample of data for one random climbing location. (left): Sentinel2 RGB images. (middle): Elevation data. (right): Lithology Data.

3.3 Sentinel2 Imagery

For all labeled and unlabeled locations, Sentinel-2 RGB bands were downloaded from Google Earth Engine from the COPERNICUS:S2_SR_HARMONIZED dataset. We acquired the latest 0.5 km square images centered around each location that had less than 10% cloud cover. The images have a spatial resolution of 10 meters and are approximately 50×50 pixels in size. Samples of the data can be viewed in FIG 2 & 3 (left).

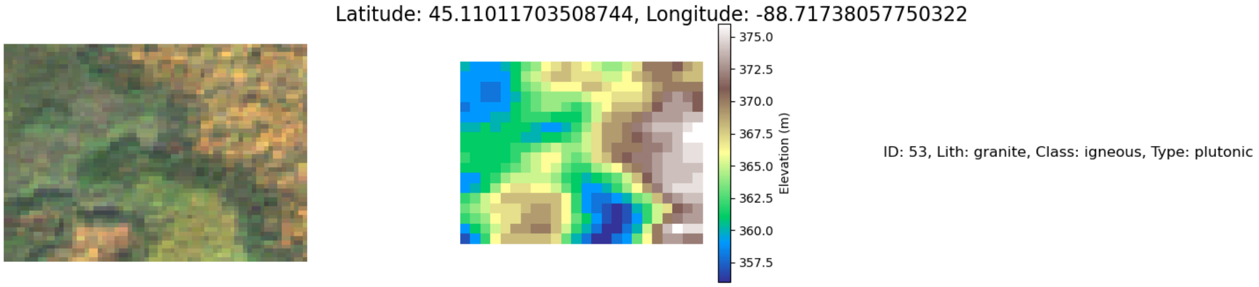


Figure 3: Sample of data for one unlabeled location. (left): Sentinel2 RGB images. (middle): Elevation data. (right): Lithology Data.

3.4 Digital Elevation Model (DEM)

We also used Google Earth Engine to download DEM information, acquiring a 0.5km square around each given location. The DEM data has a spatial resolution of 30 meters and are around 17x17 pixels. Although higher resolution DEM data exists for the US, we chose the USGS:SRTMGL1_003 dataset because it has global coverage, which may be critical if we want to apply this model outside of the US. Samples of the data can be viewed in FIG 2 & 3 (middle).

3.5 Lithology

The Macrostrat public API provides a tile server that allows users to retrieve lithology PNG images at a given (x, y) coordinate with a specified zoom level z. However, we opted not to use this data for several reasons. First, it was difficult to associate the zoom level with a 0.5 km square. Second, during manual testing, it was challenging to determine if the images were truly centered at the location of interest. Third, the tile server only provides color information and not actual lithology data.

Because of these limitations we chose to collect text data that contains rock classifications and richer textual descriptions of a location by scraping the macrostrat */map_query_v2* internal API. It takes a latitude, longitude, and zoom level and returns a JSON payload of lithology and other rock information. You can see a sample of the data we receive in FIG 2 & 3 (right).

4 Next Steps

We will develop and train a model with the architecture illustrated in FIG 4

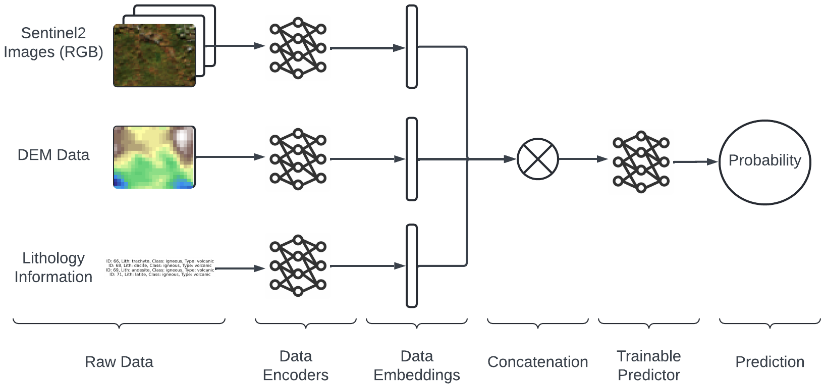


Figure 4: Proposed model architecture. The model is composed initial data encoders that will feed into a trainable CNN to produce a prediction probability.

4.1 Model Architecture

For each location l, the associated data will be converted into embeddings using data-specific encoders. For Sentinel-2 RGB images, we will use a pretrained ResNet50 model available from TorchGeo to generate image embeddings.

For lithology data, we will use a text encoder trained on scientific documents due to the specialized terminology; we have identified SciBERT as a suitable candidate. To our knowledge, no foundational pretrained DEM encoder exists. If none is found, we will consider training one using an unsupervised learning strategy; however, if this proves too challenging, we may defer the inclusion of DEM information to future work. The resulting embeddings from all data sources will be concatenated and used as input to a trainable model, which will output a predicted probability p of climbing occurrence at location l.

4.2 Model Training

The model will be trained using the Binary Cross-Entropy (BCE) loss function:

$$\text{BCE Loss} = -\left(y\log(\hat{y}) + (1-y)\log(1-\hat{y})\right),$$

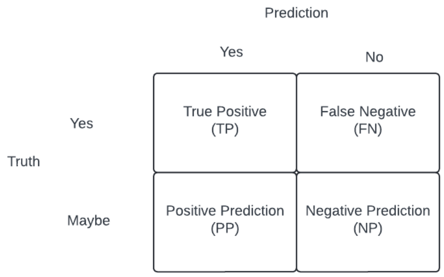
where y is the ground-truth probability and $\hat{y}$ is the predicted probability from the model.

For labeled locations, the ground-truth probability y will be set to 1. Establishing ”ground-truth” probabilities for unlabeled points presents a challenge, and we are exploring several options:

- **Assign $y = 0$ to Unlabeled Locations:** This approach assumes that all climbing locations in the United States have been discovered, treating all unlabeled locations as negatives.
- **Use a Fixed Probability Based on Spatial Metrics:** Assuming the United States is reasonably well-explored and the overall proportion of land with climbing routes remains relatively constant, we can partition the country into a grid of reasonable size. We then calculate the proportion of cells containing climbing routes and use this value as the ground-truth probability for unlabeled locations.
- **Distance-Based Probability Assignment:** Recognizing that climbing areas tend to cluster geographically, we can define the ground-truth probability for an unlabeled location as a function of its distance to the nearest known climbing area.

4.3 Model Evaluation

We will evaluate the model using the presence-only confusion matrix shown in FIG 5. We will utilize the corresponding Receiver Operating Characteristic (ROC) curve to establish a reasonable threshold probability. A good model should maximize the number of *True Positives*, minimize *False Negatives*, and maintain predictions at a reasonable proportion based on class priors.



		Prediction	
		Yes	No
Truth	Yes	True Positive (TP)	False Negative (FN)
	Maybe	Positive Prediction (PP)	Negative Prediction (NP)

Figure 5: Presence Only Confusion Matrix. Note that this differs from standard confusion matrices in that False Positives (FP) are not Predicted Postives (PP) and True Negatives (TN) are now Predicted Negatives (NP).

Based on the confusion matrix we can produce metrics for evaluating Presence Only Scenarios like below:

$$\text{Adjusted Precision} = \frac{TP}{TP + \alpha FP}$$

$$\text{Adjusted Recall} = \frac{TP}{TP + \beta FN}$$

where α and β are adjustment factors based on class prior probabilities. Determining appropriate values for α and β based on the data is a key challenge.

5 Requested Feedback

- Does this model architecture look reasonable?
- Any ideas for embedding the DEM data?
- Suggestions for determining ground-truth probabilities for unlabeled data.
- Other ideas for evaluating the model.

References

[1] https://developers.google.com/earth-engine/datasets/catalog/sentinel-2

[2] https://developers.google.com/earth-engine/datasets/tags/dem

[3] https://macrostrat.org/

[4] https://www.kaggle.com/datasets/pdegner/mountain-project-rotues-and-forums/discussion?sort=hotness

[5] Ruess, S., Paulus, G., & Anders, K.-H. (2022). The use of high-resolution photogrammetry for the survey and analysis of rock-climbing walls. *AGILE GIScience Series*, 3, Article 58. https://doi.org/10.5194/agile-giss-3-58-2022

[6] Beni, T., Gigli, G., Lombardi, L., Carlà, T., & Casagli, N. (2022). Route Stability Index (RSI): An index for the assessment of rockfall-related hazards in rock slopes equipped for sport climbing. *Geoheritage*, 14(3), Article 80. https://doi.org/10.1007/s12371-022-00715-7

[7] Finding boulders in satellite imagery using machine learning (AKA FART). (2022, August 12). *Mountain Project Forum*. Retrieved from https://www.mountainproject.com/forum/topic/122854457/finding-boulders-in-satellite-imagery-using-machine-learning-aka-fart

[8] Peters, S. E., & Husson, J. M. (2017). Macrostrat: A platform for geological data integration and deep-time Earth crust research. *Geochemistry, Geophysics, Geosystems*, 18(6), 2026–2037. https://doi.org/10.1029/2018GC007467

[9] Beery, S., Hughes, Z. J., Mallawaarachchi, V., Morrison, L. W., Oyler-McCance, S. J., Radosavljevic, A., & Franklin, J. (2021). Species distribution modeling for machine learning practitioners: A review. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 3069–3079). https://doi.org/10.1145/3460112.3471966

[10] Botella, C., Servajean, M., Bonnet, P., & Joly, A. (2019). Overview of GeoLifeCLEF 2019: Plant species prediction using environment and animal occurrences. In *Working Notes of CLEF 2019—Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings (Vol. 2380). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-2380/paper_238.pdf

[11] Bekker, J., & Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4), 719–760. https://doi.org/10.1007/s10994-020-05877-5

[12] Mac Aodha, O., Cole, E., & Perona, P. (2019). Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 9596–9606). https://doi.org/10.1109/ICCV.2019.00969

6 Extra Analysis

6.1 Rock Climbing

For this project, we focus on roped climbing (traditional or sport). We performed an analysis on the distribution of ratings and number of pitches in our dataset shown in FIG 6. Though the authors interests align with high-quality multipitch (longer than 35 meter) routes; however we decided not to filter on any metric such as quality or route length in order to provide a fuller dataset (3038 locations vs 12,776 locations). The trade off is that the model will learn lower quality and single pitch areas. Future work might involve stripping the dataset to only include high quality (2.5 stars or greater) multipitch routes and evaluating the differences between the models predictions.

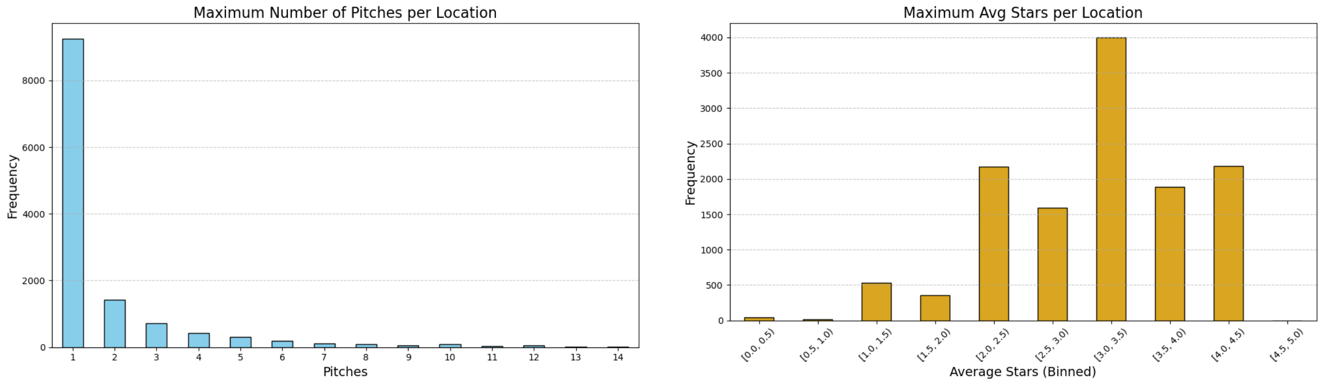


Figure 6: (left): the frequency of maximum number of pitches at a given location. There are 9248 locations that only contain single pitch climbing vs 3511 locations that contain multipitch climbs. (right): The distribution of the maximum user given rating for a given location. Most climbing locations have routes with decent quality (2 stars or above).

6.2 Lithology Analysis

There are 220 rock classifications that can be viewed at this link. The distributions of the rock types between labeled and unlabeled data is show in FIG 7. The distributions have a chi-squared distance of $3.87 * 10^{17}$ and a Pearson Correlation Coefficient of 0.60 indicating that these are highly dissimilar distributions. It validates our assumption that the model might be able to use lithology as a signal in prediction.

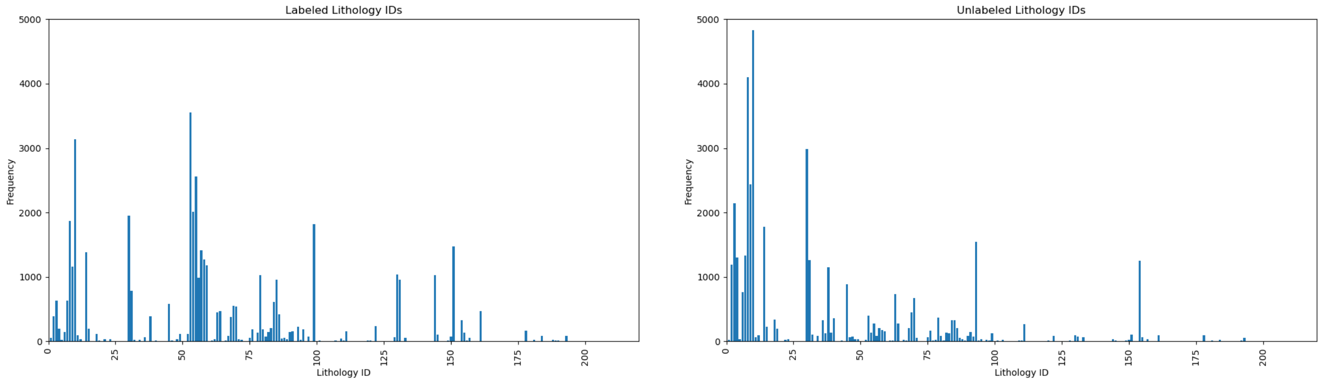


Figure 7: (left): The frequency distribution locations for lithology ids for the labeled climbing locations. (right): The frequency distribution locations for lithology ids for the unlabeled locations.