Using Machine Learning for Alpine Land Classification in Mountainous Terrain

Joe Tayabji

Department of Geography, University of Utah

GEOG 6160 - Spatial Modeling and Geocomputation

Simon Brewer, Moongi Choi

May 3, 2024

Abstract

Alpine terrain is an important research topic for scientists in various fields, including those at the Betty Ford Alpine Gardens (Vail, CO). Defining and mapping what constitutes alpine terrain is a study that has come a long way, but there is no definitive source. This study aims to apply machine learning pixel classification techniques to remotely sensed imagery and digital elevation models. Literature review finds that contemporary definitions have shifted more toward bioclimatic variables. In an effort to bring modern solutions to more traditional thinking, Landsat 9 OLI-2 images were combined with DEMs to train random forest models as well as neural networks. In spite of high AUC scores, the models generated did not provide satisfactory predictions. However, their current state is one that can be built upon with proper resources.

Introduction

Alpine flora has long been of interest to botanists, ecologists, and biologists. In the early 20th century, alpine species and genera diversity were being analyzed in a fairly robust and statistical method (Jaccard, 1912). However, the study of mountains is made relatively more difficult by how loosely they are defined (Gerrard, 1990). There is a general understanding of their definition, which is aided by basic human cognition and intuition. Many think of steep terrain and hills that rise for hundreds of meters. Several attempts to provide definitive databases of mountains, and more specifically alpine terrain, have not resulted in universal acceptance (Körner et al., 2011). A rough definition of alpine can be given as areas in mountains where a combination of temperature, growing season, elevation, and latitude (in addition to several other variables) contribute to an environment where trees cannot grow (Paulsen & Körner, 2014).

The Betty Ford Alpine Gardens (2024), located in Vail, Colorado, is committed to conservation and education surrounding flora found in the alpine. They have a team of botanists and biologists who spend much of their time hiking in search of different species of alpine flora. They document their findings using global navigation navigation system (GNSS) devices so they can return to

Figure 1: Columbine (Aquilegia). Source: Todd Pierce, Betty Ford Alpine Gardens

monitor any changes. Their work is made difficult by a combination of short summers (when snow has melted) and a vast alpine area. Modern machine learning methods offer a chance to aid the Betty Ford Alpine Gardens, and all alpine researchers, in understanding specifically where they should and should not look search.

Literature Review

The Global Mountain Biodiversity Assessment (GMBA & University of Bern, 2024) is a group aimed at the conservation and sustainability of mountain biodiversity. They have applied the term "ruggedness, " which relates to local elevation changes, to define mountains. Their definition ignores absolute elevation (e.g. meters above sea level), which was used by Kapos et al. (2000) and is another commonly accepted framework. The GMBA definition has been translated into a publicly available map and data, including ESRI format shapefiles.

The GMBA has also adopted Paulsen and Körner's (2014) research to define alpine. This definition effectively focuses on temperature and growing season, or bioclimatic belts. This definition was taken even further by Testolin et al. (2020). They used 19 climatic variables in a principal component analysis in conjunction with a clustering algorithm to create what is arguably the most robust map of alpine areas. This map excludes Antarctica but is otherwise worldwide.

Testolin et al.'s methods have yet to capture everything, which is more than reasonable given the scope of the effort. For instance, some areas in New Mexico are missing, as well as the Tarryall Mountains in Colorado, which are near South Park and Pikes Peak. The Tarryall Mountains remain an area that the Betty Ford Alpine Gardens visit.

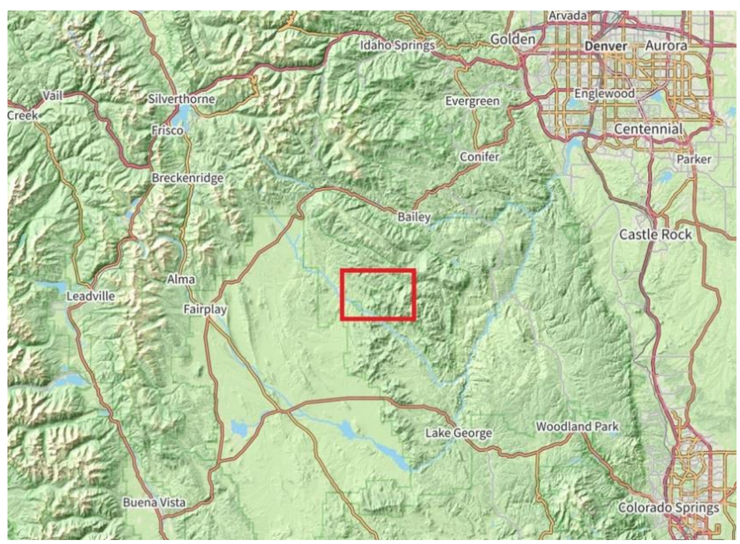

Figure 2: The location of the Tarryall Mountains is denoted by the red box. Source: OpenStreetMap

Machine learning has successfully been used to perform pixel classification on land use land cover (LULC), and has proven especially useful when considering urbanization

(Talukdar et al., 2020). Random forest, support vector machine, and artificial neural network methods used with Landsat 8 imagery were more accurate than other methods in Talukdar et al.'s LULC study along an area of the river Ganga, between Rajmahal and Farakka in India.

Methodology

Methods undertaken in this study were supplemental to methods used by Emily Griffoul, Conservation Scientist at Betty Ford Alpine Gardens, and Dara Seidl, Associate Professor of GIS at Colorado Mountain College. Griffoul used ESRI ArcGIS Pro in an attempt to combine in situ observations with the GMBA mountain inventory. Ultimately, the accuracy was not satisfactory and the time cost was expensive. Seidl has been rather successful using remote sensing and digital elevation models (DEMs) in conjunction with Google Earth Engine (GEE). Ultimately, a Landsat 9 OLI-2 scene with seven bands was combined with a DEM in a specific study area, mostly in Colorado. This can all be retrieved easily in GEE, which also provides random forest methods. The first difficulty Seidl encountered was in trying to combine multiple areas into one model. The second was the limitation in GEE's methods and only being able to use random forest models.

In this study, a similar approach to Seidl's was taken, but using random forests and neural networks in R. Landsat 9 OLI-2 imagery was retrieved for summer months (mid-July to mid-September, 2018 to 2024) in an attempt to minimize the effects of snow on reflectance. Additionally, cloud coverage was limited to 4%. This data was sourced from the USGS, as were the DEMs. Three study areas were chosen in an attempt to represent some level of terrestrial diversity. The first area included parts of the Canadian Rockies, Calgary, Alberta, and the plains to its south. The second included parts of Denver, Colorado, nearby cities, various mountain ranges in the Colorado Rockies, and high-elevation valleys. The final area included the White Mountains in New Hampshire and southwestern Maine.

Band	Description
1	Visible Coastal Aerosol (0.43 - 0.45 μm)
2	Visible Blue (0.450 - 0.51 μm)
3	Visible Green (0.53 - 0.59 μm)
4	Red (0.64 - 0.67 μm)
5	Near-Infrared (0.85 - 0.88 μm)
6	SWIR 1(1.57 - 1.65 μm)
7	SWIR 2 (2.11 - 2.29 μm)

Figure 3: Landsat 9 OLI-2 Bands. Source: USGS

Using QGIS, several polygons were drawn by hand in each area to represent alpine and non-alpine land cover, resulting in six total shapefiles. These polygons were created by referencing Google Maps Satellite imagery, as was done in Seidl's GEE methods. These were loaded in Rstudio and 999 random samples were taken from each of the six shapefiles (5,994 points total). Then, three different raster stacks were made, one for each area. Each stack had eight bands; the first seven were the Landsat 9 OLI-2 bands, and the eighth was a DEM. A data frame was then created where the raster information was extracted at all 5,994 locations and was combined with coordinates and a factor to indicate whether or not the point was in an alpine area. A task was created using this data as the backend, and having a target of the presence/absence factor. All of this data was scaled on a range of 0 to 1 based on the minimum and maximum values in each column. Then, n/a values were removed, resulting in 5,889 points.

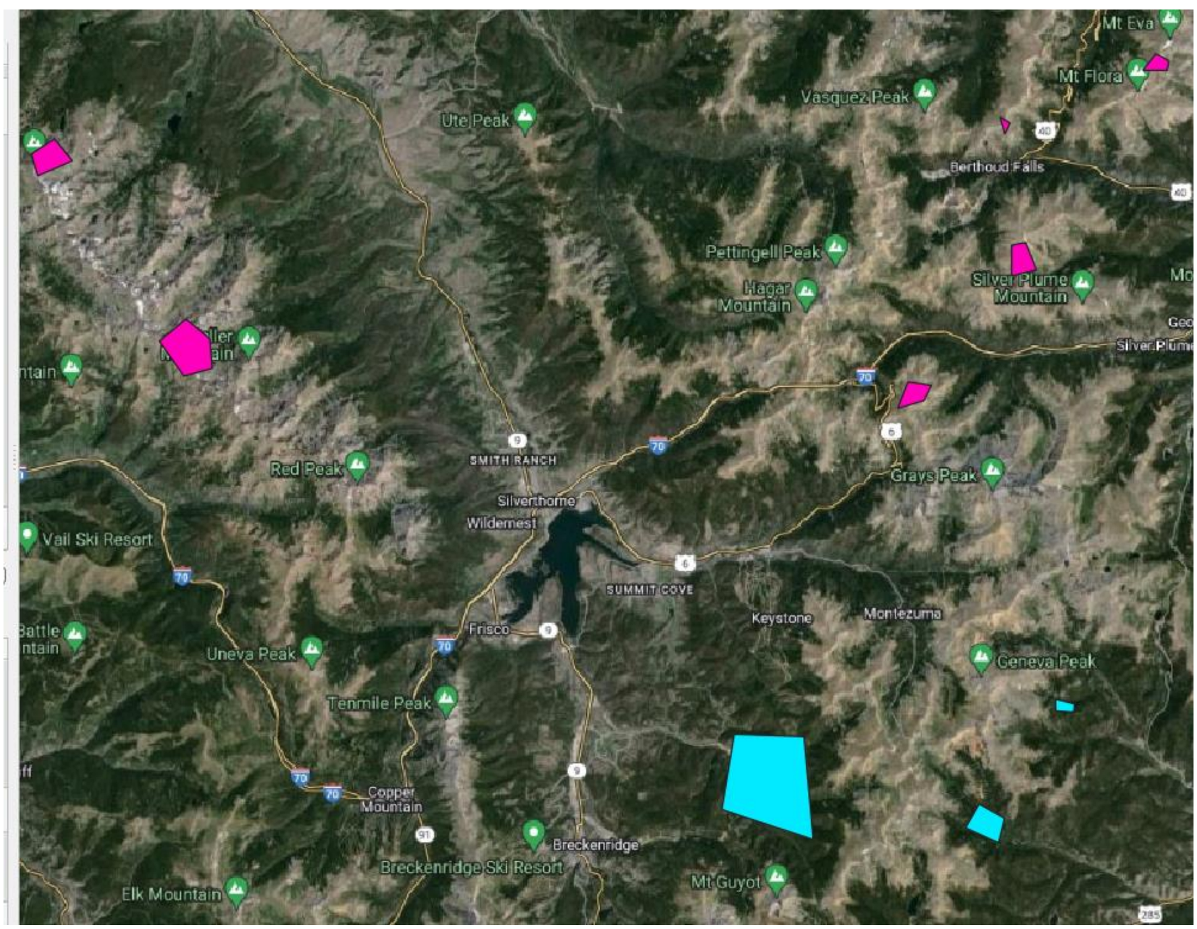

Figure 4: A QGIS screenshot showing a small subset of the polygons used for training data samples. Here, magenta polygons occur in alpine environments, while the cyan ones do not.

As part of the development leading to the methodology outlined above, smaller steps were, of course, taken. Initially, only a small subset of data in Colorado was used, and it was used to train a random forest model. With the inclusion of elevation and the limitation of physical area, a 0.99 AUC was achieved. However, when variable importance was examined, it was clear that elevation had a profound effect. Even with all three areas being used as training data, elevation was the most important variable in the random forest model. Ultimately, two random forest models were created that included all three areas: one with elevation and one without.

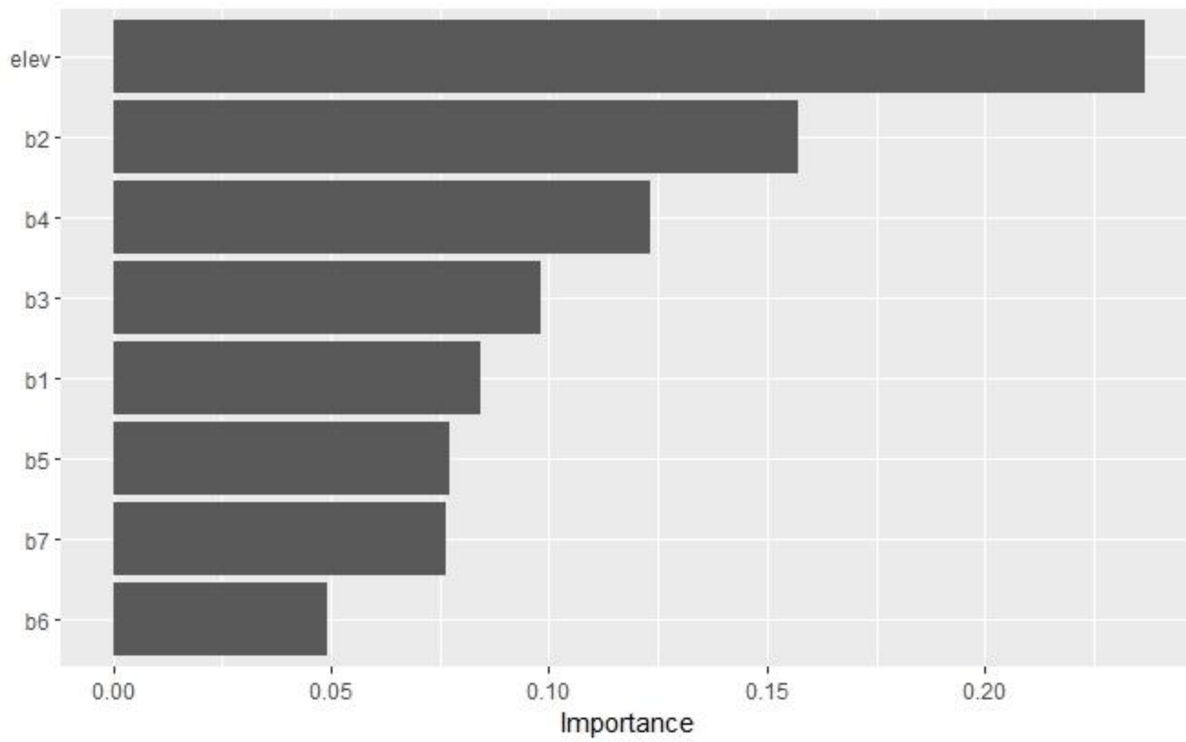

Figure 5: Variable importance plot when all three areas are used, and elevation is included.

Two neural network models were also created. Both of these had the following parameters:

```
lrn('classif.nnet', size = 25, decay = 1e-5, maxit = 500, trace = FALSE)
```

One of these models used only the seven Landsat 9 OLI-2 bands. The other model added elevation and latitude, both of which were also normalized. This study was undertaken to try and address any shortcomings of using only bioclimatic variables, which is why elevation was included. However, there is a clear link between the elevation of treeline and the latitude at the point being studied (Kapos et al., 2000), which is why latitude was also included. Attempts were made to tune hyperparameters for all four models, but none succeeded in improving the AUC. All models were resampled with an 80%/20% holdout strategy.

Results

To reiterate, all the models presented used training data from all three areas combined into one data frame in an attempt to make a model that was generalizable to more study areas. The first model to be considered is the random forest model that only uses the seven Landsat 9 OLI-2 bands. This model had an AUC score of 0.97. Elevation, but not latitude, was then added to this model, and an AUC score of 0.998 was received. As Figure 5 shows, however, elevation's variable importance was considerable.

The neural network models resulted in similar AUC scores. When only the remote sensing imagery was used, the reported AUC was 0.979. When latitude and elevation were included, the reported AUC was 0.999. Figure 6 shows prediction rasters for the White Mountains and Maine. Figure 7 shows prediction rasters for an area near Leadville, Colorado. Finally, this model was used to make a prediction in Utah in an area including the Uintas and Wasatch. The results of this are shown in Figure 8.

The NN model that only uses the Landsat 9 OLI-2 bands suffers from error by commission. As shown in the images, it captures the actual treeline with a great amount of detail. However, it is erroneously including pixels that are clearly below treeline, as well as water. When elevation and latitude are added into the model, there is a much more definitive boundary between alpine and non-alpine. However, the image in Colorado appears to overstate the size of the alpine region, while the image in Utah does the opposite. While not shown in the image, these models unfortunately did not classify the Tarryall Mountains as alpine.

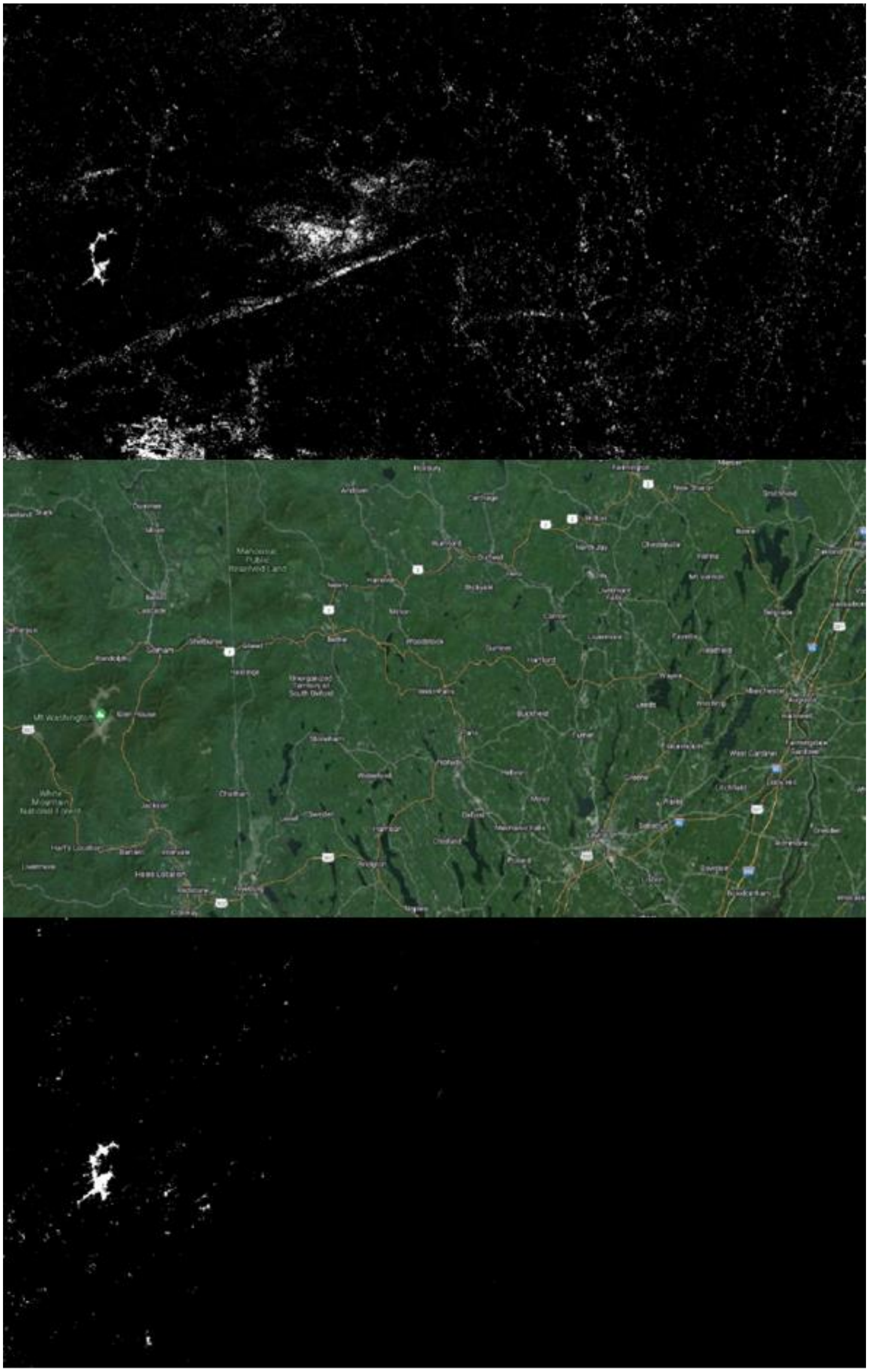

Figure 6: NN Predictions for the White Mountains. Top: Without elevation. Middle: Satellite imagery (source: Google). Bottom: With elevation and latitude

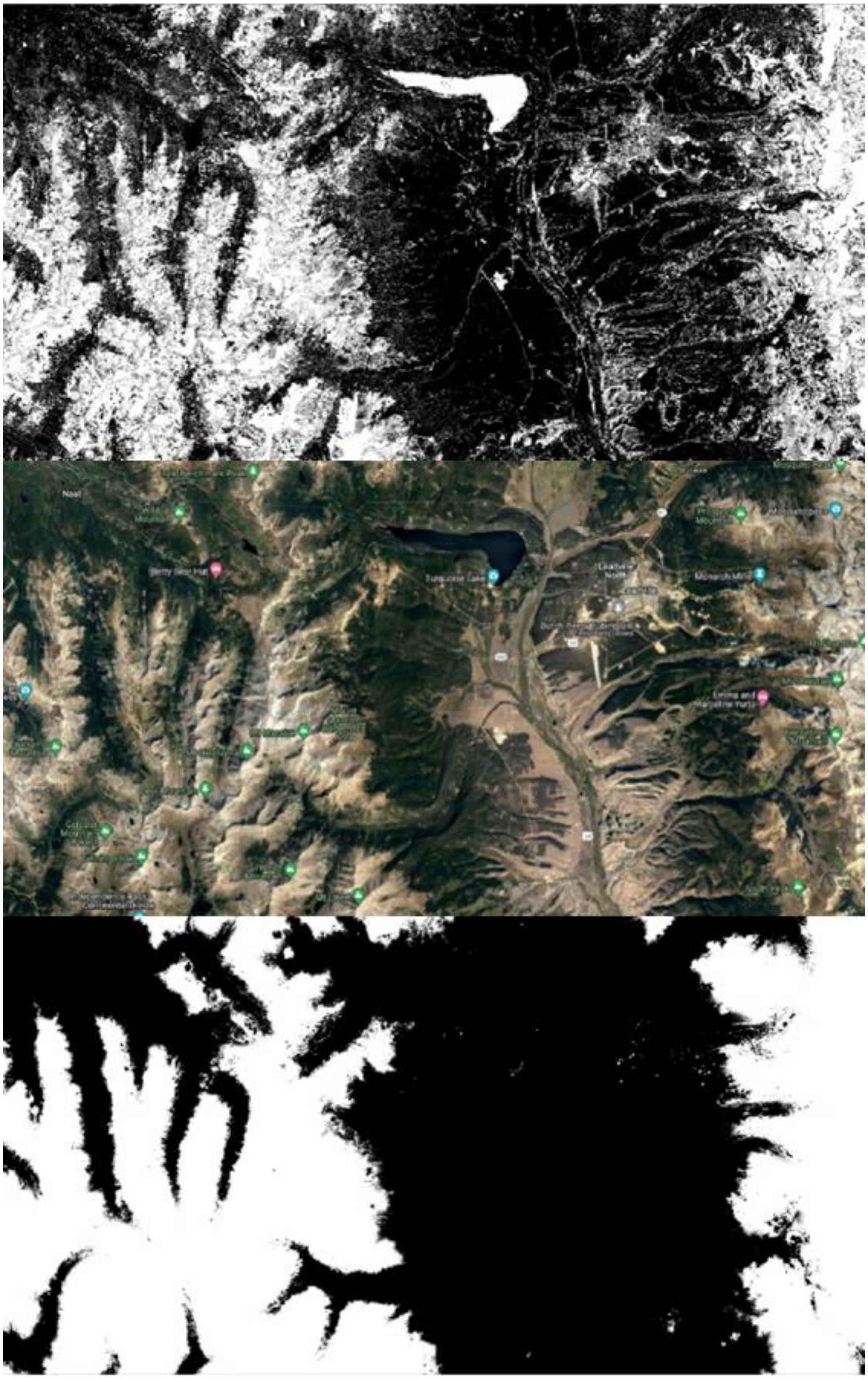

Figure 7: NN Predictions near Leadville, CO. Top: Without elevation. Middle: Satellite imagery (source: Google). Bottom: With elevation and latitude

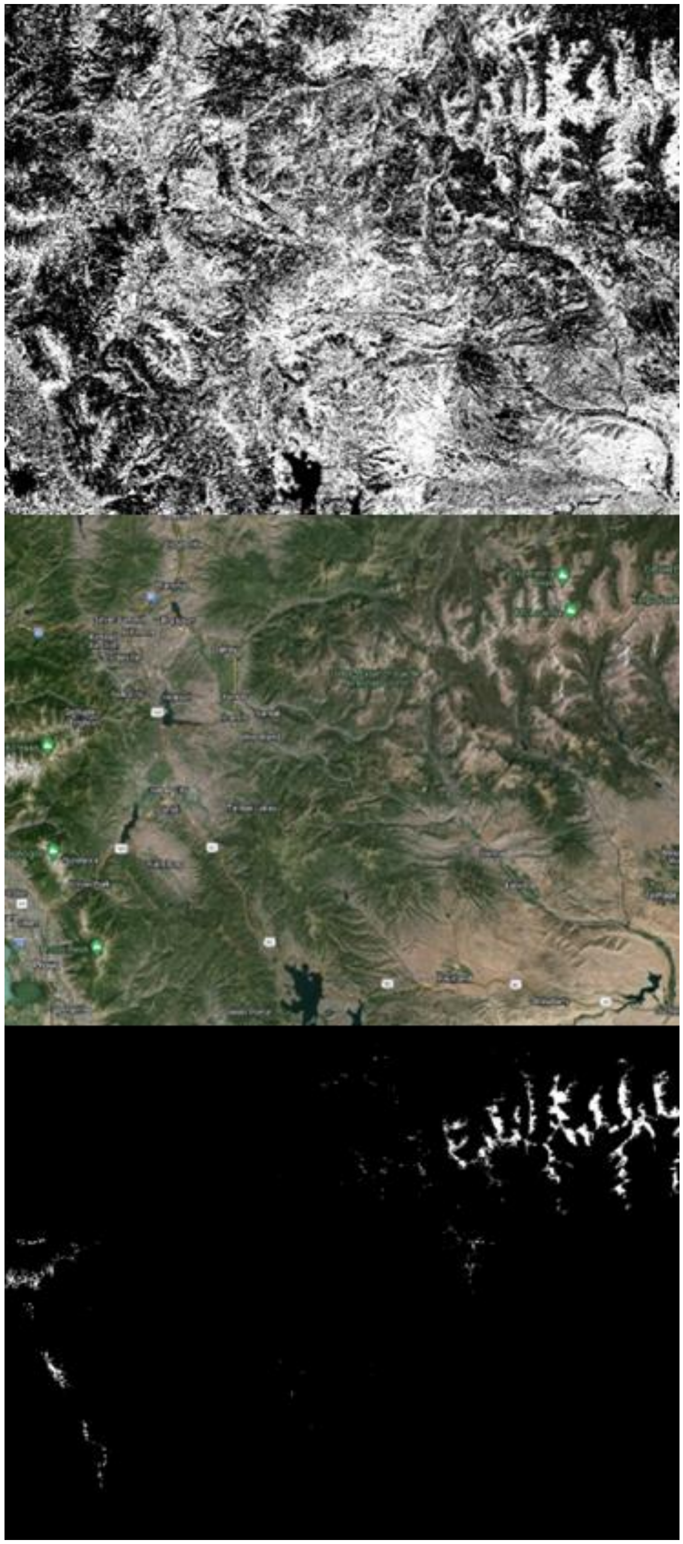

Figure 8: NN Predictions in Utah. Top: Without elevation. Middle: Satellite imagery (source: Google). Bottom: With elevation and latitude

Discussion and Conclusion

While predictions were made with some measure of accuracy in areas similar to where the training data came from, they were not made well in Utah. In its current state, this model cannot be generalized to a larger study area. Overfitting is a likely culprit, as evidenced by high AUC scores but unsatisfactory images. This makes sense, especially when using only three areas to make predictions in different landscapes. One solution would be to simply use less training data. The other would be to gather training data from even more diverse landscapes. Another issue to be determined is the proper treatment of elevation and latitude, as there is a clear association between them and they can lead to spatial autocorrelation. With that in mind, they are both undoubtedly important variables to be considered when attempting to define alpine.

While bioclimatic variables cannot exhaustively classify alpine, neither can spectral nor elevation and latitudinal variables. A practical next step in this work would be to combine these variables. While the definition of alpine remains ambiguous, it is clear that all of these variables can help explain different, however small, parts of the meaning.

An increasing number of variables brings up a final point for improvement: this work certainly requires a high level of computational power. While still manageable, the work was demanding on the consumer laptop used in this study. Using GEE offers a clear advantage as its processing power is among the greatest available to most users. If neural networks were desirable, high-performance computing, including parallel processing and vast storage, would be needed to make a larger classification.

With that in mind, it is clear that machine learning is a powerful and useful tool in LULC. In the specific case of alpine pixel classification, it is very promising. While the results found were not ideal, there is a clear path forward towards improving the models used.

References

Betty Ford Alpine Gardens. (2024b, April). *Conservation + Sustainability | Betty Ford Alpine Gardens Vail, Colorado*. https://bettyfordalpinegardens.org/

Gerrard, John. *Mountain environments: an examination of the physical geography of mountains*. MIT press, 1990.

GMBA & University of Bern. (2024). *Global Mountain Biodiversity Assessment (GMBA)*. Global Mountain Biodiversity Assessment (GMBA). https://www.gmba.unibe.ch/

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist*, *11*(2), 37-50.

Kapos, V., Rhind, J., Edwards, M., Price, M. F., & Ravilious, C. (2000). Developing a map of the world's mountain forests. In *Forests in sustainable mountain development: a state of knowledge report for 2000. Task Force on Forests in Sustainable Mountain Development.* (pp. 4-19). Wallingford UK: Cabi Publishing.

Körner, C., Paulsen, J., & Spehn, E. M. (2011). A definition of mountains and their bioclimatic belts for global comparisons of biodiversity data. *Alpine Botany*, *121*, 73-78.

Paulsen, J., & Körner, C. (2014). A climate-based model to predict potential treeline position around the globe. *Alpine Botany*, *124*, 1-12.

Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y. A., & Rahman, A. (2020). Land-use land-cover classification by machine learning classifiers for satellite observations—A review. *Remote Sensing*, *12*(7), 1135.

Testolin, R., Attorre, F., & Jiménez-Alfaro, B. (2020). Global distribution and bioclimatic characterization of alpine biomes. *Ecography*, *43*(6), 779-788.