ANTARCTIC SUBGLACIAL LAKE IDENTIFICATION

Anastasia Horne (Student)

COLORADO SCHOOL OF MINES - DSCI 575

Subglacial hydrology connects glacial and oceanic systems, modulates ice dynamics, and remains a major physical uncertainty in future ice-sheet projections. Subglacial water is often stored in subglacial lakes, some of which are active, episodically filling and draining on short timescales. Lake drain-fill cycles cause changes in water distribution, grounding-zone stability, freshwater flux, and nutrient and carbon export into the Southern Ocean and sub-ice-shelf cavities. Subglacial lakes can also contain thousands of cubic meters of water (Livingstone et al., 2022). These subglacial lakes can burst causing sudden outburst floods that present a risk to existing downstream infrastructure. In 1892, during the dead of night, the Tete Rousse Glacier experienced such an event. The glacier released 200000 cubic meters of water from subglacial lakes, sending a rapid wall of water, sediment, and debris toward the town of Saint Gervais, ultimately claiming the lives of 200 people (NOVA). While subglacial lakes in Antarctica do not provide the same level of risk to human life as subglacial lakes in France, they are still impacted by subglacial bursts. These bursts can weaken the ice surrounding the lake as it has experienced a sudden draining event and it can also alter the behavior of the ice where the water finally 'burst' through, which can have impacts on glacier melting and speed. So, understanding where these active subglacial lakes are in order to collect more data on them is vital to understanding how ice sheets and glaciers change. Traditional methods for identifying subglacial lakes can be time-consuming and expensive, so this project proposes an alternative way to identify subglacial lakes using a supervised deep neural network.

This solution employs a supervised deep neural network to automatically identify active subglacial lakes in Antarctica. This approach leverages existing airborne ice penetrating radar surveys, satellite imagery and other datasets that serve multiple glaciological research purposes in order to create a comprehensive feature set for lake detection. This approach also maximizes cost-efficiency and reduces the need for dedicated data collection trips. The neural network utilizes a custom deep neural network architecture designed to widen the feature space to better learn complex patterns and then pruning the features to have a final classification. This approach provides rapid and large-scale detection to identify subglacial lakes, and by utilizing a neural network, we are also able to prioritize which lake candidates to study in the field based on the probability output from the neural network. The model provides validation statistics and uncertainty quantification to help the user assess prediction reliability. Additionally, it provides visualizations in order provide qualitative evidence of its success and usefulness. Overall, this solution significantly reduces the time and resources required for subglacial lake detection while maintaining high accuracy standards, enabling more comprehensive monitoring of these important glaciological features.

The solution relies on open access to high temporal resolution satellite data that is assumed to have adequate coverage of glacial regions containing subglacial lakes. The model assumes that the list of identified subglacial lakes is accurate and the list of candidate subglacial lakes is exhaustive to ensure that the model can properly train on the given dataset and identify clear margins between subglacial lakes and nonlakes. If multiple true lakes (identified or proposed)

were not excluded from the continent before nonlake points were chosen, the model could inaccurately attribute certain feature values to nonlake points, while in truth, they indicate a subglacial lake. The most important assumption the model makes is that the surface patterns some features represent either impact subglacial lake existence or are impacted by subglacial lake existence.

The datasets used in this model are large and thus the storage, processing, and use of them requires large computational resources, which are not accessible to everyone. Additionally, the model requires specialty data forms and dependencies such as geodataframes and PyTorch which some users might not have access to or are unfamiliar with using. There are some environmental constraints on the data. If atmospheric conditions were not ideal during airborne or satellite data collection, the resulting values risk being outliers and reduce the quality of our data. Additionally seasonal variations in ice conditions could cloud our model's judgement. For example, if the 'firn' value for lake 'a' was taken in the summer when ablation increases, but the 'firn' value for lake 'b' was taken in the winter when accumulation is dominant, then the resulting 'firn' values would not be comparable. Finally, some of these lakes exist in increasingly remote and dangerous locations, meaning there are limited ground-truth validation opportunities.

The results of my model have limited regional ethical or societal implications because the region of interest is extremely remote and is solely inhabited by researchers and government personnel. However, knowing were these subglacial lakes are can help us learn more about climate change. If the glaciers in Antarctica and Greenland were to all melt, sea level would rise by approximately 210 feet. This would cause a large increase in coastal erosion and an increase in storm surges, as warming air and water temperatures contribute to more frequent storms. So understanding any and all causes of glacial melt is very important. The faster glaciers flow/move the more frictional heat increases and this leads to more ice melt. The flow of glaciers is extremely dependent on their subglacial hydrology, including subglacial lakes. So, if we know exactly where subglacial lakes are, and what geophysical characteristics and features represent them, we can better track them and potentially visit them and collect other valuable data. Additionally, because places like Antarctica are so cold they are able to preserve information about the past very well. Subglacial lakes contain information on ancient climate, and ancient microorganisms who have been able to survive for thousands of years (Goeller et al., 2016). Similar to how ice cores provide information on how greenhouse gasses have changed throughout the past, we can look at the chemical composition of the water in these lakes to gain vital information on atmospheric conditions. The results of my model will have a larger impacts on the glaciological and scientific communities. I am able to show that machine learning is over 80% accurate at identifying subglacial lakes, suggesting that machine learning can be used to model and solve other complex subglacial processes. The model provides sufficient enough identification, such that a glaciologist can (in a reasonable amount of time) examine false positives and negatives to determine what features of some subglacial lakes are unique and region specific.

The model I have designed utilizes 16 different features related to Antarctic processes and characteristics to detect subglacial lakes (Appendix A). Additionally, it utilizes a list of 131 identified lakes (Siegfried & Fricker, 2017) and a list of 3008 lake candidates (Sauthoff, 2022). We can use the geometry values in the lake and lake candidate lists to select 262 nonlake points on the Antarctic continent that exclude lakes and candidates. We can combine the identified lake and nonlake datasets to form our final target dataset. In order to assign feature values to lake points we must compare target geometries (locations) to locations where features are available. The geographically closest point of features to a target object will become the features for that specific object. Once we have all feature and target datasets read into the model and concatenated together, we are able to process them and prepare for model training. Looking at the feature values, each individual feature has a different range of values (see Table 1).

Feature	Minimum Value	Maximum Value
Firn	0	38.0924
Ice Surface Elevation	68.6411	3967.4246
Ice Thickness	0	4126.224

Table 1. Range for a subset of features

In order to limit feature bias we must fit a scaler to all features. However, as previously mentioned, due to atmospheric conditions and other sensor issues, some of our features may contain outliers. The scaler function used in this model will take extreme values and convert them to NaNs. After scaling the features, it is necessary to either remove rows that contain NaN values or to replace the NaNs with a global mean. This model opts for the latter because there is a limited number of identified lakes, to remove ones where one feature is missing would severely decrease the robustness and applicability of the model. The specific neural network used within this problem does not allow for categorical variables, so it was necessary to convert our target data into a dummy variable using the following mapping scheme: {Nonlake: 0, Lake: 1}. Now that the data has been scaled and processed, we are able to build our model.

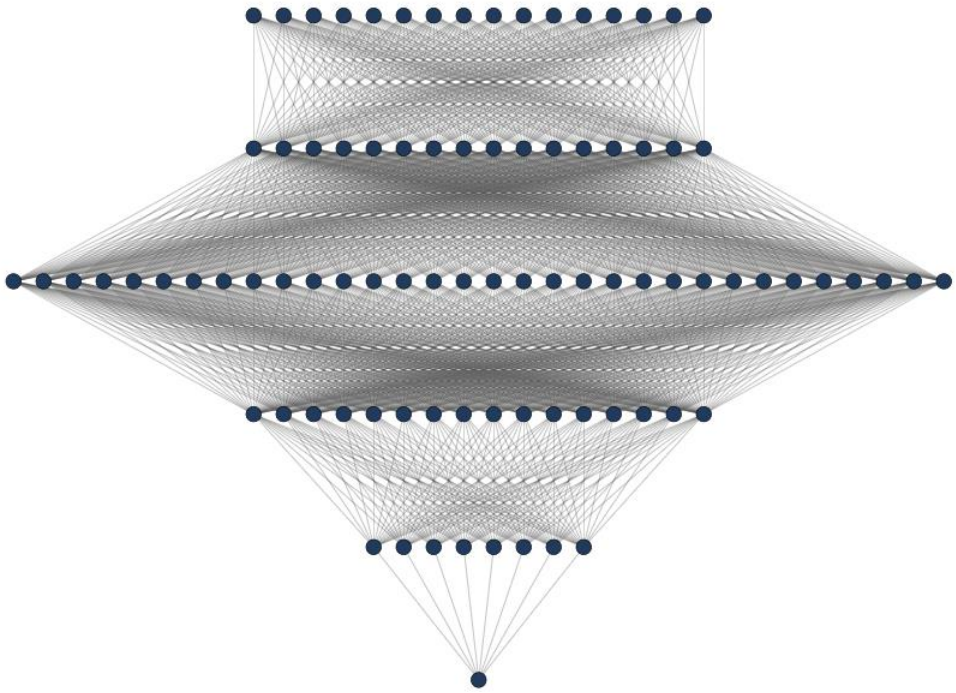

Figure 1. Neural Network Diagram

The specific neural network used in the problem contains 5 hidden layers (Figure 1). The input layer takes in all 16 features and then applies a linear PyTorch layer to them (hidden layer 1), while maintaining the size of our feature space. The next hidden layer applies another linear layer and widens the feature space to 32 features, and then in the third hidden layer we go back down to 16. The fourth hidden layer prunes our feature space and shrinks it to 8 features, and our final hidden layer reduces the feature space to a single logit. Finally, before we output this logit, we apply a rectified linear unit (ReLU) layer which applies a threshold operation to our final output transforming any negative values to zero.

There are two parameters needed to train and test a neural network that depend on the model itself: learning rate and epochs (training iterations). To determine the proper value for both, the neural network was first trained with 70% of the data and validated against the remaining 30%. The result of the training and testing accuracy over each epoch can be seen in the figure below.

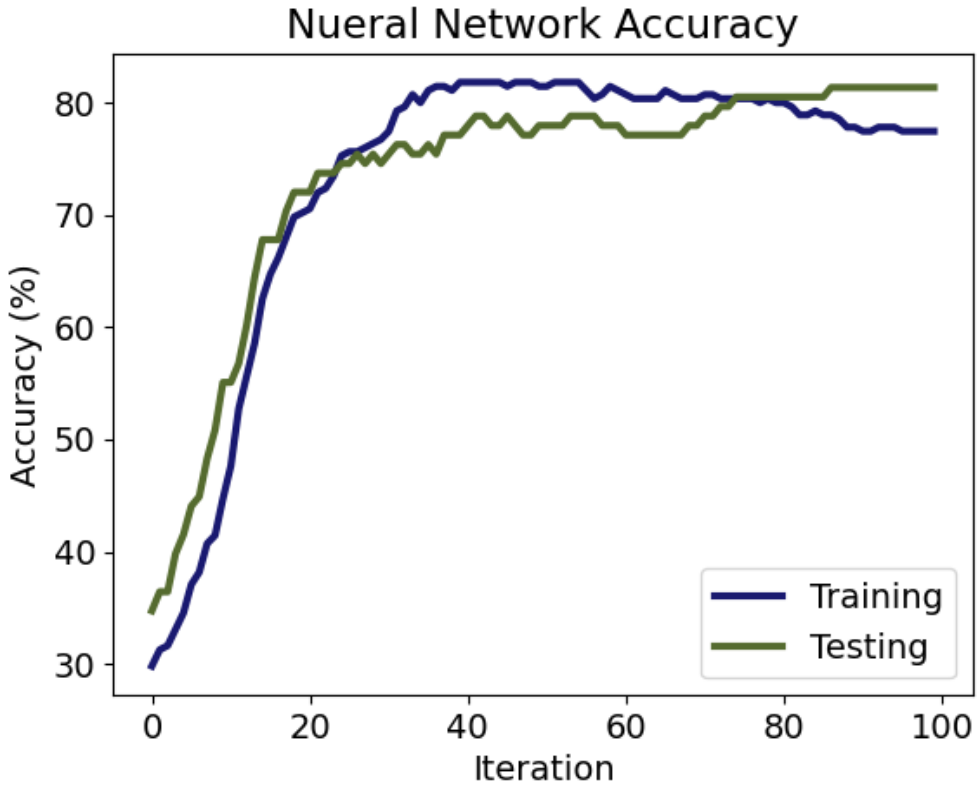

Figure 2 Test and Train Accuracy per Epoch

We notice that at approximately forty epochs both the training and testing accuracy begin to level off. If we were to run the code multiple times, this 'leveling' would be observed between epochs 35-60, so in to remain conservative we will run future simulations for 50 epochs. Additionally, in Figure 2 we notice that during some epochs the test accuracy (green line) is greater than the train accuracy (blue line) this is not expected and could be signs of an underfitting model, or it could be a result of the distribution of lakes versus nonlakes in our training and test datasets.

In order to test the robustness of our model, we applied a 5-fold cross validation approach. This approach split the data into 5 distinct folds, where each fold served as a validation set once, while the remaining data served as the training set. After training each fold, it was necessary to reset the weights of the neural network to prevent overfitting and ensure independent evaluation. The results of each fold and each epoch within a fold training were tracked and recorded, leading to the following results.

Confusion Matrix		Predicted Label	
		Lake	Nonlake
True Label	Lake	22.4 ± 4.08	3.8 ± 2.23
	Nonlake	9.4 ± 1.86	43 ± 3.16

Table 2. Confusion Matrix

	Precision	Recall	F1-score
Lake	0.70 ± 0.05	0.85 ± 0.10	0.77 ± 0.06
Nonlake	0.92 ± 0.04	0.82 ± 0.03	0.87 ± 0.01

Table 3. Classification Report

Over the 5-folds we found the confidence interval for both training and validation accuracy to be **0.8671 ± 0.0225** and **0.8321 ± 0.0217**, respectively. The model has demonstrated acceptable discriminative capability, achieving a mean validation accuracy of at least 83% across all folds, with notably low variance (± 2%). This performance is particularly impressive given the inherent complexity of subglacial lake detection and the challenging nature of remote sensing data. Examining Tables 2 and 3 we notice that the true positive lakes had the most variability, this is not a desired outcome because we prefer our model to most accurately identify lakes. However, we also notice that the average recall score for lakes is higher than that of nonlakes. Additionally, from the F1-score of lakes, we can tell that our model does not do an effective job of balancing precision and recall. Overall, from the lake classification report our model is falsely identifying too many non-lake features as lakes. However, by applying this model we could significantly reduce the number of lake candidates and then examine them individually to determine which are the most likely lakes.

Subglacial lakes, hidden beneath the thick Antarctic ice sheet, play a crucial role in understanding the dynamics of the Antarctic ice sheet and potential impacts its melting has on global sea-level rise. Traditional methods for identifying subglacial lakes involve geophysical surveys which can be time-consuming, expensive, and in some cases impossible. This project presented an alternative approach to efficiently and accurately identify subglacial lakes using a neural network machine learning algorithm. The model performed well overall, however after further investigation it is clear that model falsely identified nonlake points as lakes at a high rate.

References

"Descent into the Ice". NOVA. https://www.youtube.com/watch?v=Dt7KZanolcI.

Goeller S, Steinhage D, Thoma M, Grosfeld K. Assessing the subglacial lake coverage of Antarctica. *Annals of Glaciology*. 2016;57(72):109-117. doi:10.1017/aog.2016.23

Sauthoff, Wilson. "Multi-mission altimetry of variable Antarctic active subglacial lake geometries and causal inference networks to inform geostatistics-based subglacial water modeling". 2022.

Morlighem, M. (2022). MEaSUREs BedMachine Antarctica, Version 3 [Data Set]. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/FPSU0V1MWUB6. Date Accessed 04-09-2024.

Mouginot, J., B. Scheuchl, and E. Rignot. (2017). MEaSUREs Annual Antarctic Ice Velocity Maps, Version 1 [Data Set]. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/9T4EPQXTJYW9. Date Accessed 04-09-2024.

Li, L., Aitken, A.R.A., Lindsay, M.D. et al. Sedimentary basins reduce stability of Antarctic ice streams through groundwater feedbacks. Nat. Geosci. 15, 645–650 (2022). https://doi.org/10.1038/s41561-022-00992-5

Siegfried MR, Fricker HA. Thirteen years of subglacial lake activity in Antarctica from multi-mission satellite altimetry. Annals of Glaciology. 2018;59(76pt1):42-55. doi:10.1017/aog.2017.36

Livingstone, S.J., Li, Y., Rutishauser, A. et al. Subglacial lakes and their changing role in a warming climate. Nat Rev Earth Environ 3, 106–124 (2022). https://doi.org/10.1038/s43017-021-00246-9

Appendix A – List of Features

From BedMachine Antarctica 3: Firn, Surface Elevation, Ice Thickness, and Bed Elevation. Hydroptotential was a calculated byproduct.

From Antarctic Ice Velocity: Ice Surface Velocity

From Li et al.: Rebound Bed Topography, Gravity Variation ANTGG BG HP30 STD30, Gravity ANTGG BG, Gravity ANTGG FA', Curie Depth, Crust Thickness ANTCRUST, Heat Flux, Gravity ANTGG Iso Airy, Mean Curvature, and Basal Friction. Further details on these features can be found in the Li supplementary information.