UMBC

THESIS APPROVAL SHEET

Title of Thesis: Immersive Visualizations and Quality Metrics for Polar Science

Name of Candidate: Naomi Tack ntack1@umbc.edu

Master of Science 2024

Graduate Program: Computer Science

Thesis and Abstract Approved:

DocuSigned by: Don Engel 39F989A7A17C496...	DocuSigned by: Rebecca Williams D1BEF1914CC8438...
Don Engel	Rebecca Williams
donengel@umbc.edu	rmwillia@umbc.edu
Assistant Professor	Assistant professor
Computer Science and Electrical Engineering	CSEE
5/1/2024 | 3:34 PM EDT	5/1/2024 | 6:46 PM EDT

NOTE: *The Approval Sheet with the original signature must accompany the thesis or dissertation. No terminal punctuation is to be used.

ABSTRACT

Title of dissertation: IMMERSIVE VISUALIZATIONS AND QUALITY METRICS FOR POLAR SCIENCE

Naomi Angela Tack, Masters of Science, 2024

Thesis directed by: Professor Don Engel
Department of Computer Science

Polar science studies many aspects of the north and south poles, and glaciers. In order to better understand the internal structure of ice sheets and how they are impacted by climate change, ice penetrating radar is used. This radar captures the layers inside the ice sheet which can then be used to understand layer behavior and flow inside the glacier. Layers can be picked out of these radargrams and then visualized to give an indication of the internal topology. We explore two areas in which these radargrams are used, namely analysis of machine-picked layers and immersive visualization of these layers. Analysis of machine picked layers is achieved through the development of quality metrics which are independent of number of layers. We present these metrics and analysis of their utility. Immersion is accomplished through WebXR, inspired by recent research in both polar science and other scientific fields. We present the system, its controls and limitations, and evaluations to be performed on this system.

IMMERSIVE VISUALIZATIONS AND QUALITY METRICS FOR POLAR SCIENCE

by

Naomi Angela Tack

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
Masters of Science
2024

Advisory Committee:
Professor Don Engel, Chair/Advisor
Professor Rebecca Williams, Co-Advisor
Professor Tejas Gokhale

Acknowledgments

I would like to thank God, first and foremost, for the ability and skill to complete this work.

I would like to thank my advisors and mentors Dr. Don Engel and Dr. Rebecca Williams. I would also like to thank my many professors and mentors who have helped developed my understanding and knowledge in computer science. I am also grateful to the members of the Pi^2 Lab who encouraged and fostered a community of collaboration and open exchange of ideas. In addition, I would like to thank iHARP for giving me to opportunity to research in the polar science space.

I would also like to thank my family and friends who have supported me throughout this research. I am eternally grateful for their support and encouragement throughout these last one and a half years.

Table of Contents

List of Tables

List of Figures

Chapter 1: INTRODUCTION

1.1 Polar Science

Polar science is a broad domain, considering many aspects of glaciers and their surrounding regions. This field aims to gauge the impact of climate change on the glaciers in present and the past. One technique to study this impact is the use of ice penetrating radar to capture radargrams of the internal structure of the ice sheets. These radargrams are collected from both ground and air sources, giving different resolutions. Once these radargrams are captured, they are then annotated with the layers inside the ice sheet which is then used to track layers throughout a glacier.

1.2 Data

The majority of data used in this work is provided by Center for Remote Sensing of Ice Sheets (CReSIS) [3]. Radargrams are collected from radar sounders attached to planes and flown over Greenland and Antarctica. These planes often cover areas of interest with many intersecting paths to better trace flow or identify objects in the ice sheet.

1.3 Quality Metrics

This tracing of layers is a highly time consuming process, requiring many hours of a glaciologist's time. To ease this burden of data production, AI models are being trained to automatically pick layers from unprocessed radargrams [4]. From our observed datasets, we note that the models seem to produced more annotations per radar slice than the corresponding expertly produced annotations, potentially indicating many false positives. It has been generally observed that model annotations better followed the underlying layer flow, rather than remaining accurate to a single layer. This leads to potential difficulty in measuring their accuracy since there is concern that the models objective function will bias toward higher layer count, rather than layer accuracy. We propose several quality metrics, which are independent of the number of layers and descriptive of the features of the layer placement. These metrics show potential in highlighting quality layer annotations, by allowing outside evaluation of automatic annotations. The value of these metrics is their independence from the source of the underlying data, focusing on the output alone as an indication of a 'good' layer annotation.

1.4 XR

Another use for radargrams is the visualization of fence diagrams, which enable glaciologists to develop an accurate 3D mental model of the glacier and identify 3D entities inside the glacier. Fence diagrams are constructed by joining and aligning intersecting radargrams along the corresponding collection path. Intersections allow for layer tracing and identification between non-intersecting slices. Traditionally these diagrams have been generated and displayed in Matlab on a 2D screen, reducing the dimensionality of the 3D data into a 2D projection. We hypothesize that this projection looses valuable spatial data and limits the hypothesis that can be derived. Research into displaying these diagrams in 3D has so far been limited to a specific VR headset due to the method of development. We propose a lightweight, hardware independent system in WebXR. We believe that this will lower the barrier to access and aid scientists in their spatial understanding of glaciers and hypothesis generation.

1.5 Outline of Thesis

We begin first by examining the quality metrics developed in Chapters 2 and 3. We then consider the WebXR contributions to the polar science domain in Chapters 4 and 5. Finally we conclude and offer suggestions for future work in these directions in Chapter 6.

Chapter 2: METRICS FOR THE QUALITY AND CONSISTENCY OF ICE LAYER ANNOTATIONS [1]

This work was published by IEEE, presented at IGARSS 2023 and is publicly available at: https://ieeexplore.ieee.org/document/10283420 .

Authors: Naomi Tack, Bayu Adhi Tama, Atefeh Jebeli, Vandana P. Janeja, Don Engel, Rebecca Williams

University of Maryland, Baltimore County (UMBC), Baltimore, Maryland, USA

2.1 Abstract

Ice layers in glaciers, such as those covering Greenland and Antarctica, are deformed over time. The deformations of these layers provide a record of climate history and are useful in predicting future ice flow and ice loss. Cross sectional images of the ice can be captured by airborne radar and layers in the images then annotated by glaciologists. Recent advances in semi-automated and automated annotation allow for significantly more annotations, but the validity of these annotations is difficult to determine because ground-truth (GT) data is scarce. In this paper, we (1) propose GT-dependent and GT-independent metrics for layer annotations and (2) present results from our implementation and initial testing of

GT-independent metrics, such as layer breakpoints, local layer density, spatial frequency, and layer orientation agreement.

2.2 Introduction

Englacial ice layers are deformed and influenced by flow fields, and as these layers are buried, the basal conditions are recorded as described by Holschuh et al. [5]. Thus, englacial ice layers can be used to infer climate history, glacial dynamics, and physical ice properties, among others [6]. Ice penetrating radar can detect these layers [7,8], which must then be annotated to further understand and describe the flow fields. Semi-automated methods are being developed [7] to reduce the annotation burden, but these methods require substantial ground truth validation in order to train a model.

Ground truth (GT) is defined as manual annotation by domain specialists, which is time-consuming and limited by the quality of the data, which often makes reliable manual annotation impossible. We have developed several quantitative metrics to characterize the quality and agreement of automatic englacial layer detection. Because GT is scarce, we separate our metrics into two groups: those that require GT and those that are separate from GT.

To aid in understanding these metrics, we produced a method of quantifying and visualizing the local metrics as quality maps. For example, quantification of layer density allows for the identification of image artifacts and glacial phenomena such as lakes or melt [9]. However, it also highlights the need for additional annotations of layers, enhanced image processing, and need for alternative partitioning for the training-testing-validation data incorporating layer density as a factor [9].

2.3 Contributions

We investigate and implement concepts from fingerprint identification/quality analysis [10] and multi-target tracking [11–13] to produce feature and quality maps and representative statistics for englacial layer segmentations that are outputs of an unsupervised algorithm [14], and surface & bedrock annotations output from a two-step deep neural network model [4]. These methods use airborne ice-penetrating radar data hosted by the Center for Remote Sensing of Ice Sheets (CReSIS) [3], collected during several overlapping flight paths.

We investigate an initial proof-of-concept metric suite and analysis framework upon which we will build in future work. We also provide recommendations for standardization and enhancement for metrics with high "labeling utility." To this end, we propose structured families of metrics that evaluate different aspects of englacial layer detection. We consider two conceptual groupings of metrics:

Metrics that require GT vs. do not require GT: When GT is not available or is expensive to curate, it is beneficial to consider quality metrics that don't require the use of reference data. The goal is not to completely negate the use of GT but to decrease the manual burden of annotating it by producing as many *a priori* automated annotations as possible. Attempting to measure layer fidelity/correctness or association errors necessarily require GT. Metrics that do

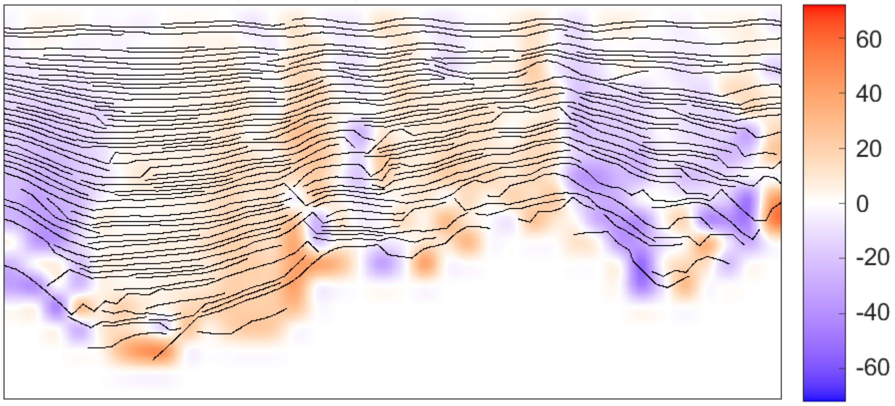

Figure 2.1: Local layer orientation map. Red indicates positive slope in degrees and blue indicates negative slope in degrees. These orientations are used to compute orientation agreement map, as shown in Fig. 2.2

not require GT can be assembled based on their utility as annotated GT labels for training supervised approaches.

Local vs global metrics: Quality measures at a local level preserve spatial/regional information, whereas global measures assign the whole image a single value, which enables ranking and comparing results directly. Metrics can be computed at the layer level to pinpoint instance-level anomalies and trends or can be windowed and/or gridded to provide full coverage of the image. Local metrics can be assembled into feature heatmaps, quality maps, and histograms. In contrast, global metrics are reported on a per-image basis, typically aggregated from local metrics and then aggregated into a single quality score. The advantages of global metrics include the ability to provide a single number to rank and compare models/algorithms. However, global quality scores can cause aggregation effects that are opaque and may not reflect outputs that are "mostly correct" or "good enough."

Quality values and maps typically involve determining a threshold of acceptance, so we present both the raw local/global feature values and the computed quality values. The visualization goals are to output a variety of quality maps/feature maps wherein anomalous regions appear salient and to display drill-down information such as local and global histogram values. Future work will streamline and extend the visualization for enhanced quality exploration, including in virtual reality [2].

Unsupervised and supervised performances are dependent on both radar image quality and annotation quality. In a two-stage framework using an unsupervised model to produce labels for a supervised approach [15], the quality of the detected/segmented layers affects the supervised model's performance, such that errors are propagated. Therefore, in this work we focus on detected layer quality; future work may incorporate radar image quality measures as well.

For the GT-independent sub-family, we compute and visualize quality maps and local histograms representing layer density/spacing (Figure 2.1), layer orientation agreement with neighbors (Figure 2.2), local frequency components, and minutiae detection (breakpoints, branch points, corners). The goal of computing these local metrics is to accumulate a quality feature vector in order to compute a quality score that accurately represents "label utility" of the detected layer mask.

2.4 Data

We use layers generated by [4, 14] on the ice-penetrating radar data hosted by the Center for Remote Sensing of Ice Sheets (CReSIS) [3]. The radar images were collected during flight, with several intersecting flight paths.

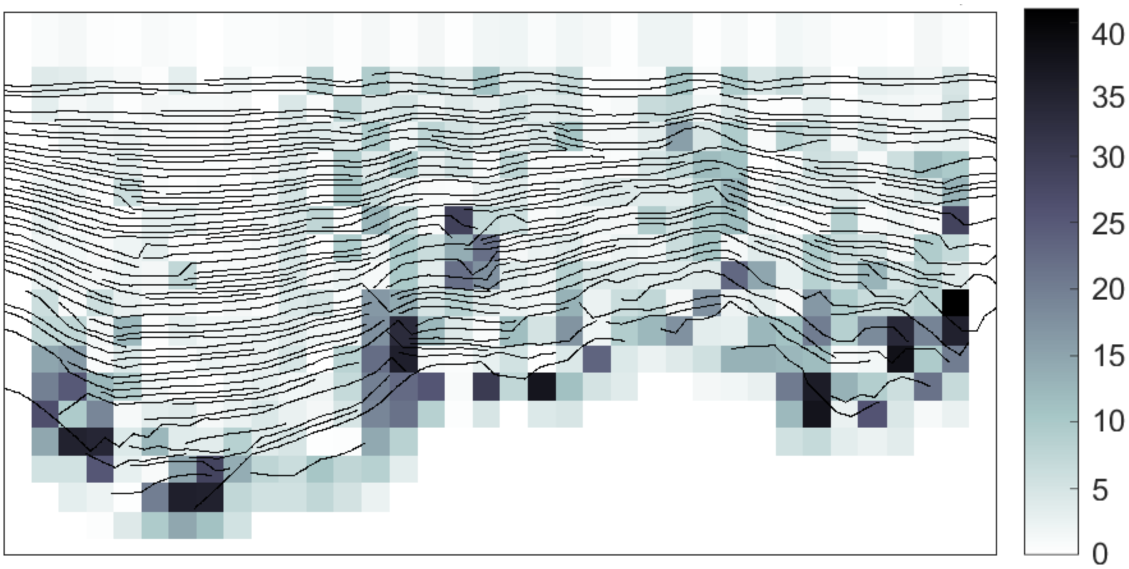

Figure 2.2: Quality map of local layer agreement with 8-connected neighbors. Darker areas indicate orientation disagreement, while light areas indicate agreement between neighboring patches. Fairly continuous areas with zero slope have high agreement, while areas near the bedrock with discontinuities or drastic direction changes have low agreement.

Because we want to reduce dependence on GT for evaluation while still enabling useful evaluation assessments, we focus on evaluation/quality metrics that can be computed without GT. For the purposes of this paper, we consider “GT” to include layer instance ids, layer pixel locations, and any image artifacts or ice anomalies. Radar parameters and flight paths are considered supporting metadata available for computing both GT-dependent and GT-independent layer quality metrics.

2.5 Results

We compute local and global metrics and visualize them using feature maps and quality maps, and accumulate a histogram and quality feature vector that can

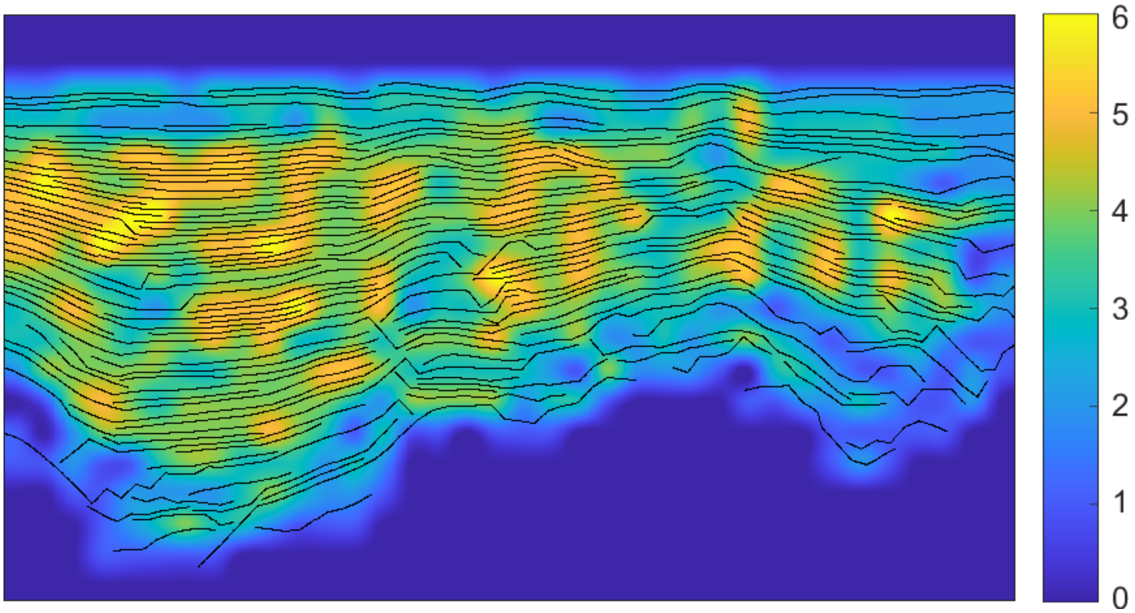

Figure 2.3: Local layer density map, where bright yellow areas indicate areas with high layer density, and dark blue represents areas with few or no layers. Layer density maps can be used to identify areas with artifacts and/or incomplete annotation, or to cue in on interesting glaciological morphology (e.g. ice lenses, crevasses, melt ponds, etc.)

be used to determine layer quality effects on the supervised stage of Jebeli *et al.* [15].

2.5.1 Ground-truth Independent Metrics

Our first metric suite encompasses global and local measures that can be computed without requiring accurate GT, as defined in Section 4.3. In the GT-independent sub-family, we further discuss the computed quality maps and features as discussed in Section 2.3. The advantages of these measures stem from the retention of spatial information that can be used to produce a map of high vs. low-quality areas based on each metric. We use the aforementioned average layer density to compute an appropriate window size for measuring orientations.

Layer breakpoints - Layer breakpoints are identified when a layer ends before the end of the image. Layers are generally continuous throughout the radar images, and identifying the "dropped" layers can help identify either layers that have actually "dropped" from the image or if they are a product of the layer detection model.

Local layer density - The local layer density is calculated using a sliding window average with a 50% overlap (Fig. 2.3). By counting the number of connected components per window, we can identify where dense layers areas are calculated in the AI approach. These metrics may prove useful when evaluating the overall performance of the approach for annotation as they may indicate areas where layers are easier to automatically annotate and areas the model fails to capture the complexity of the layers.

Spatial frequency - The identified layers were modeled mathematically using cubic splines to generate their equations and interpolate the normals along evenly spaced intervals. By testing the intersections of the normal with the nearest layer above, we then calculate the Euclidean distance between layers. This allows us to generate mean distances across several columns of the image and collapse these representative distances into means for the column. These were compared to the 2D FFT spatial frequency map computed, as shown in Fig. 2.4. Areas with low spatial frequency can be seen near incomplete layers.

Orientation and orientation agreement - Fig. 2.2 shows the orientation agreement quality map. By calculating the layer normals we can understand the overall orientation of the layers (shown in Fig. 2.1) and allows us to take a sliding window average and see how orientations agree with the neighboring layers. This could indicate the flow of the ice or bias in the model to produce layers that are

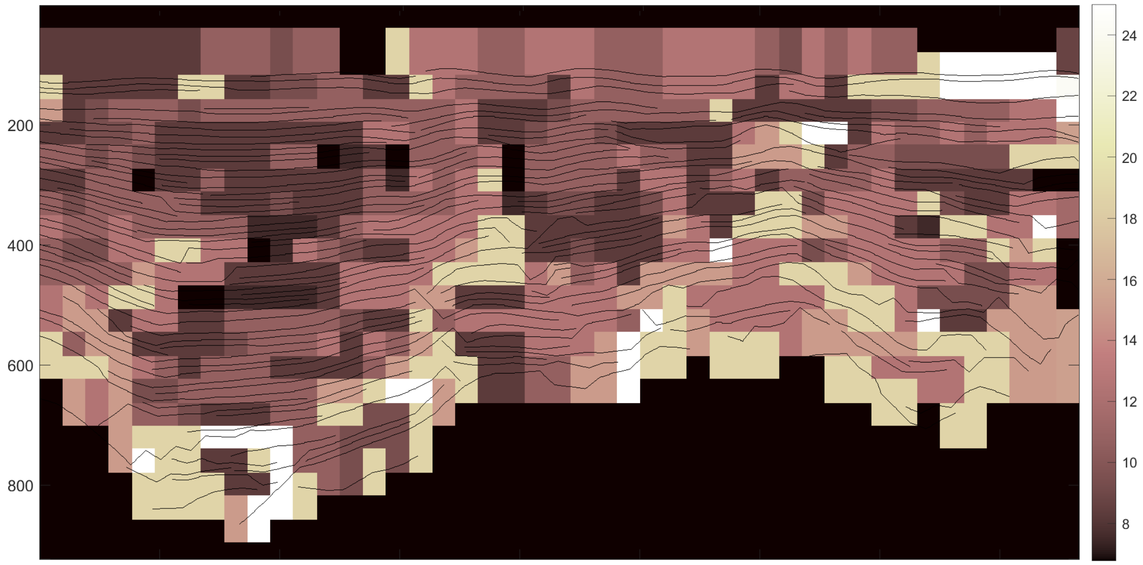

Figure 2.4: Local spatial frequency map. Bright areas correspond to lower spatial frequency (units of px/cycle), darker areas indicate higher spatial frequency (i.e. decreased distance between layers). The maximum frequency recorded is the half the window size (in this case, 30 px/cycle for a window size of 75 px).

"going in the same direction."

2.6 Future Work

In this paper, we focus mainly on ground-truth independent metrics to avoid over-reliance on the availability and quality of hand-picked/annotated layers. In our ongoing work, we are planning to also compute several GT-dependent metrics, from both the local and global families of metrics. These include inter-annotator agreement for GT (per available annotated layer) and layer completeness for GT (per image).

For both the GT dependent and independent metric sets, we plan to assemble a quality vector and examine the effects of both high and low-quality layers (i.e., their predictive utility) using the U-net framework in development by [4].

Our current work aims to establish a standardized framework for evaluating the performance of models and techniques that enable englacial layer detection, localization, and association. In the near-term, we will also perform a sensitivity analysis of our quality metrics and supervised algorithm performance. We will compare these metrics with domain-expert quality assessment and investigate feature importance using dimensionality reduction and other techniques.

Longer-term, we intend to explore the agreement of annotations in multiple images captured in proximity to each other, such as images captured by parallel flights of aircraft as well as images captured by intersecting flight paths.

Our ultimate goal is to develop and implement a standardized evaluation, visualization, and annotation tool for ice layer detection. This will help users employ the quality metrics to aid in correction/new annotation, provide quality assurance and quality control (QA/QC) for hand-picked layers, and could ultimately support training a supervised algorithm to correct them automatically.

Chapter 3: QUALITY METRIC EVALUATION OF AUTO-ANNOTATION APPROACHES FOR POLAR SCIENCE

The majority of the work was done in Fall 2023 for CMSC 491 - Computer Vision with Professor Tejas Gokhale. Since the submission of this paper, additional metrics have been considered from the results, and these are documented and discussed in Section A

Authors: Naomi Tack, Luke Zimmerman, Naren Sivakumar

University of Maryland, Baltimore County, Baltimore, Maryland, USA

3.1 Abstract

The melting of Earth's glaciers is a major contributing factor in sea level rise and understanding the flow and structure of these ice sheets is of high importance to the polar science community. Ice penetrating radar captures the layers of ice sheets and annotations provide a better understand ice dynamics. This is very time intensive process, requiring domain knowledge. Recently there have been efforts to produce automatic annotations of the ice sheet. However, these auto annotations often jump between neighboring layers or completely miss the underlying structure, producing erroneous results. In this paper we examine several models for scoring

these annotations, based on ground truth agnostic metrics. Our goal is provide a method of scoring automatic annotation to ensure quality annotations.

3.2 Introduction

Earth's glaciers contribute significantly to sea level rise and understanding ice flow and structure is important to the polar science community to forecast future developments [5]. Annotations of the ice sheet layers preformed by glaciologists aid in this understanding but in a very time intensive manner. This highlights the need for automatic annotations of ice sheet layers and [15] have taken up the challenge. In [1] we explored some metrics which could further evaluate the model in [15] and perhaps provide feedback to help refine produced annotations.

In this paper we aim to continue the work by training several models with the metrics from [1] to analyze the quality of auto annotated ice sheets, with the aim to provide an outside evaluation to [15]. This work is in association with iHARP and focuses on aiding the work of glaciologists by supporting the development of automatic annotation of ice sheets.

We continue by discussing the related work in Section 3.3. Section 3.4 details data processing and augmentations. We highlight the results found in Section 3.5, limitations in Section 3.6 and individual contributions in Section 3.7. We conclude in Section 3.8.

3.3 Related Work

In the field of biometrics, the quality of fingerprint images has been successfully scored using a Random Forest (RF) model [10]. This approach also utilized quality metrics, derived from the fingerprint images, as the input to the random forest. Our dataset is similar in that it is composed primarily of lines following a pattern, so we hypothesis that RF will be a good fit for our data.

Another model, developed specifically to handle binary classification of images, is AlexNET [16]. AlexNET has been extensively used to classify images where there are very few data samples per class, as in [17]. There is precedent of deep learning models being used to determine the calving margins of glaciers in Greenland and Antarctica, and in debris-covered glacier mapping [18] [19]. Due to the size of our dataset we believe that a CNN, such as AlexNET, would be too complex so we consider a traditional neural network.

3.4 Methods

We propose segmentation and perturbation of our datasets following the standard methods for data augmentation. We also highlight the three models we will construct and discuss their uses.

3.4.1 Segmentation & Augmentation

Due to the time intensive work of annotating ice sheet layers there is only one dataset available for expertly annotated radar images [?]. This dataset contains approximately 2300 annotated radar slices. At the time of data collection only 46 were available for download in the format provided by Nick Holschuh. Due to this small section of the dataset, each radar slice was segmented into non overlapping 250 × 250 pixel areas. Of these areas only segments where layers were annotated are retained and small segments are padded. Each sections quality metrics are then computed, resulting in 995 segmented portions for input [1].

Each segment needs a label to train the classification and regression models described in Section 3.4.2. A range of 1 to 5, 5 being the best, was chosen to represent the agreement of a segments layers with the underlying radargram. Due to time constraints, only 743 segments were successfully labeled. Given that the dataset was annotated by experts, the majority of labels were closer to 5. We discuss some of the limitations of the labeling process and distribution in Section 3.6.

In an attempt to offset the minimal data available, we also performed data augmentation on the labeled data. These augmentation were limited to rotation with a maximum angle of 20°, and horizontal flips. These specific augmentations were chosen to mimic the expected trends of layers in glaciers. Vertical slopes are not found in glacier formations and very steep slopes are rare, with the exception of near ocean glaciers. Labels for these augmented segments are taken to be the same as the original segments. Due to time constraints the metrics for each segment could not be computed so these data will be incorporated in future iterations.

We provided two different representations from our datasets to the models. The first representation, *LayerData*, took a binary representation of the annotated layers as an input to the model and the mean and standard deviation for each metric. This representation has 743 labeled instances. The second representation, *MetricData*, took the entirety of the quality metrics and omitted the annotated layers. This representation only has 635 labeled representations due to more stringent size requirements. Neither representation contained the raw radar data so as not to bloat the feature space, this is discussed more in Section 3.6. We compare these data representations effects in Section 3.5.

3.4.2 Models

Several different machine learning models were selected as potential baselines for quality prediction, including both regression and classification models. Expected classification for an annotation is in the range 1 to 5, but intermediate values are possible so regression is also considered.

Random Forest showed to correct classification of fingerprints based on quality metric in [10]. We considered several configurations of random forest by varying the number of trees, the number of samples per node split, and the maximum depth. XGBoost is a modification of RF that uses a regularizing gradient boosting framework, handling large amounts of data efficiently by using parallel tree boosting.

This model provides an option to balance weights to handle unbalanced classes, so was tried. This parameter, scale_pos_weight, was tuned to see if its effect helped the unbalanced data.

Support Vector Machines are another common choice for regression and classification models. This was implemented as a regressive model with several kernels, including linear, radial basis function (RBF), and cubic polynomial. We also created three basic neural networks with varying layers. Each network trained for 10 epochs with batch sizes of 32 images.

Each models was trained with an 80/20 split of the data. Evaluation is provided for each model in terms of Accuracy, Macro Recall, Macro Precision, Macro F1, MAE, and MSE. Regressive models are evaluated by bucketing output values into five classes, by rounding to the nearest class.

3.5 Results

We consider first the likelihood of always selecting the most frequent label from our datasets. We find that the accuracy of this approach is 70.12%, highlighting the skewed nature of our labels. We consider mitigation in Section 3.6. We detail the findings and observations found through these initial models in the following sections. Tables 3.1-3.6 show an overview of each model set.

3.5.1 Random Forest

We see that overall the Random Forest model approach preformed the best out of the considered models. It's highest accuracy score is 76.51% when trained and evaluated on the *LayerData*, which is slightly better than random selection (Table 3.1). Interestingly this approach degraded when given the *MetricData* representation but was still better than random at an accuracy of 74.02% (Table 3.2).

Model Type	Accuracy	Macro Recall	Macro Precision	Macro F1 score	Mean Absolute Error	Mean Squared Error
500 Trees, 10 Split	0.758	0.346	0.306	0.323	0.335	0.604
1000 Trees, 10 Split	0.765	0.348	0.305	0.324	0.322	0.577
1000 Trees, 20 Split	0.758	0.346	0.308	0.325	0.348	0.644
10000 Trees, 10 Split, 5000 depth	*0.765*	0.348	0.305	0.324	0.322	0.577

Table 3.1: Metrics for the Random Forest Model with *LayerData* inputs. We prefer the 1000 Tree, 10 split over the 10000 Tree, 10 split, 5000 depth due to the difference in training time. We note that this model preforms slightly better than random guessing when predicting labels.

3.5.2 SVR and XGBoost

For SVM we found that normalizing the label range to $0 - 1$ produced the better results than $1 - 5$, or $0 - 4$. Overall, SVM preformed slightly worse than the XGBoost (Tables 3.3, 3.3). For the *LayerData* input the best preforming SVM and XGBoost both had an accuracy of 67%. The best XGboost had better mean absolute

Model Type	Accuracy	Macro Recall	Macro Precision	Macro F1 score	Mean Absolute Error	Mean Squared Error
500 Trees, 10 Split	0.716	0.235	0.382	0.239	0.440	0.944
1000 Trees, 10 Split	0.708	0.226	0.348	0.225	0.448	0.952
1000 Trees, 20 Split	0.677	0.193	0.166	0.175	0.503	1.102
10000 Trees 10 Split, 5000 depth	0.716	0.235	0.378	0.239	0.433	0.921
100 Trees 5 Split	0.740	0.242	0.416	0.243	0.417	0.952

Table 3.2: Metrics for the Random Forest Model with *MetricData* inputs. We see the simplest forest of 100 Trees, 5 split preformed the best of these metrics. There is a significant drop in performance when compared to Figure 3.1, with the best model from this data representation preforming slightly below the worst model above.

and mean square error compared to the best SVM, at 0.416 and 0.354 to 0.738 and 1.254, respectively. However, neither of these model types preformed better than random selection. Similar to the Random Forest models, we also observed that the models preformed better with *LayerData* as compared to the *MetricData*, which only attained an accuracy of 60%.

3.5.3 Neural Network

Unexpectedly, the considered neural network architectures significantly underpreformed. With *LayerData* and *MetricData* achieving an accuracy of 22% and 41%, respectively (Tables 3.5, 3.6). The *MetricData* did significantly improve the accuracy of the neural networks, indicating that perhaps additional data is needed

Model Type	Accuracy	Macro Recall	Macro Precision	Macro F1 score	Mean Absolute Error	Mean Squared Error
SVM: k=linear	0.24	0.08	0.12	0.12	1.175	2.130
SVM: k=RBF	0.42	0.17	0.16	0.16	0.767	1.151
SVM: k=poly_3	0.67	0.16	0.11	0.13	0.738	1.254
XGB: spw=0.5	0.62	0.42	0.46	0.44	0.421	0.337
XGB: spw=5	0.67	0.45	0.46	0.45	0.416	0.354
XGB: spw=30	0.65	0.44	0.47	0.38	0.388	0.324

Table 3.3: Metrics for the SVM and XGBoost models with *LayerData* inputs. Of the SVM the one with a 3^{rd} degree polynomial kernel preformed the best but it preformed slightly worse than the XGBoost model trained with a scale_pos_weight of 5 which had a better mean absolute and mean squared error. Both models though preformed worse than guessing the most frequent label.

Model Type	Accuracy	Macro Recall	Macro Precision	Macro F1 score	Mean Absolute Error	Mean Squared Error
SVM: k=linear	0.18	0.09	0.05	0.06	1.599	4.435
SVM: k=RBF	0.47	0.19	0.17	0.16	0.710	0.985
SVM: k=poly_3	0.54	0.17	0.15	0.15	0.712	1.107
XGB: spw=0.5	0.59	0.22	0.19	0.20	0.677	1.145
XGB: spw=5	0.60	0.24	0.21	0.21	0.673	1.163
XGB: spw=30	0.54	0.23	0.20	0.19	0.715	1.210

Table 3.4: Metrics for the SVM and XGBoost models with *MetricData* inputs. Of the SVM the one with a 3^{rd} degree polynomial kernel preformed the best but it preformed worse than all of the XGBoost models. The best XGBoost also ended up being the one with a scale_pos_weight of 5 similar to the results from training on *LayerData*.

to sufficiently train these architectures, or a larger data size.

Model Type	Accuracy	Macro Recall	Macro Precision	Macro F1 score	Mean Absolute Error	Mean Squared Error
3 Layer Model	0.14	0.21	0.42	0.13	1.3	2.6
4 Layer Model	0.21	0.21	0.24	0.19	1.2	2.2
5 Layer Model	0.22	0.23	0.21	0.19	1.1	2.0

Table 3.5: Metrics for the Neural Network model with *LayerData* inputs. We prefer the 5 layer network due to better accuracy and overall metric values.

Model Type	Accuracy	Macro Recall	Macro Precision	Macro F1 score	Mean Absolute Error	Mean Squared Error
3 Layer	0.32	0.12	0.11	0.09	0.86	1.29
4 Layer	0.41	0.12	0.19	0.14	0.79	1.12
5 Layer	0.37	0.17	0.19	0.12	0.80	1.07

Table 3.6: Metrics for the Neural Network model with *MetricData* inputs. We prefer the 4 layer network due to better accuracy.

3.6 Limitations

There are several limitations in this projects approach. Including, but not limited to data processing, data representations, and distribution bias.

Data collection presentations some interesting challenges. Our dataset has a clearly skewed distribution toward higher labels given that the slices were pulled directly from the expertly annotated sets. In addition, given size of our scale, there is a high level of subjective interpretation between what is considered a 3 or 4, or between a 4 and a 5. In order to attempt to mitigate this variability in inter-labeler agreement, only one author labeled data. There is a clear difference between a 3 and a 5, so future work may consider a smaller scale of $1 - 3$. Existing labels could be used by splitting labels 2 and 4 with a certain probability into the neighboring labels. This could reduce the need to re-label data but could introduce more variety into the labels so some correction is recommended.

In addition, there are many more slices for segmentation, augmentation and labeling. This represents a significant issue as data is available but requires significant processing and labeling for use. There is also a significant lack of auto-annotated data available, with only 8 slices currently available. These have not been segmented or labeled, however, their inclusion is important for increasing our models performance.

We also note this approach does not consider the underlying radar data and only considers what [1] call “metrics that don’t require GT ”. Labeling of annotated layers is done by comparing the layers to the observed layers in the underlying radar image, introducing some connection to the underlying GT. Future approaches should consider metrics that include the GT or more domain specific considerations.

Additional concerns to the approach are not noted here, as they address some of the larger issue in the domain of computer vision which are beyond the scope of this project. These issues contain but are not limited to layer identification on an instance level, track association and continuity, and distinction of tracks.

3.7 Author Contribution

3.7.1 Naomi

I contributed to the project in the following manner. I was responsible for explaining the projects goals and background information for team member understanding. I also handled all data consolidation, format shifts, data labeling and metric calculations. I provided code to load different data representations and labels to feed into models. I implemented and evaluated the Random Forest models.

3.7.2 Luke

I contributed to the project in the following manner. I designed the image augmentation algorithm that was used to augment the radar images. I also implemented and evaluated the SVM and XGBoost models.

3.7.3 Naren

My contribution to this project consisted of training, evaluation and comparison of simple neural networks, with regression instead of classification. The process consisted of splitting the data into training and testing data, and then bucketing the results into a 1 to 5 range, since the regression produce finer grained results than necessary. The neural networks used were a modification of simple neural networks with increasingly complex architectures used.

3.8 Conclusion

In this project we considered several models to evaluate the quality of annotations of ice sheet using several GT independent metrics. We found that *LayerData* showed slightly better performance for a Random Forest model, while *MetricData* improved performance of the SVM and neural networks. Initial findings indicate that these models are not much better than selecting the most frequent label. We believe that a better distribution of data would help mitigate these errors and increase performance. This project indicates that addition metrics should be considered to better rank annotations.

Chapter A: Confusion Matrices and Discussion

Here we present the confusion matrices from our best models trained on *MetricData* highlighted in the tables above. This representation of the models outputs lend significant insight into the limitations and pitfalls of this work. We omit *LayerData* as we wish to consider the developed quality metrics effects.

		Predicted Label				
		0	1	2	3	4
True Label	0	0	0	0	0	4
	1	0	0	0	0	1
	2	0	0	0	0	6
	3	1	0	0	2	24
	4	0	0	1	5	83

(a) RF: 100 Trees, 5 Split (3.2)

		Predicted Label				
		0	1	2	3	4
True Label	0	0	0	1	0	3
	1	0	0	0	1	0
	2	0	0	1	3	2
	3	0	0	0	10	17
	4	0	0	3	26	60

(b) XBG: spw = 5 (3.4)

		Predicted Label							
		0	1	2	3	4	5	6	7
True Label	0	0	0	0	4	0	0	0	0
	1	0	0	0	0	0	1	0	0
	2	0	0	0	3	3	0	0	0
	3	0	0	0	11	13	1	1	1
	4	0	0	1	28	44	14	1	1

(c) NN: 4 Layers (3.6)

Table A.1: Confusion Matrices showing the misclassifications by the best three models for *MetricData*. The off diagonals indicate where the models failed to distinguish the difference between labeled slices.

In Section 3.6 we briefly discussed the impact of labels ambiguity of individual data slices. These concerns are justified as the models confusion matrices showed significant misclassification between neighboring labels as seen in Table A.1. Both the XBG model (Table A.1b) and the 4-layer NN model (Table A.1c) predicted many label 4 slices as label 3. However the RF model (A.1a) predicted many label 3 slices as label 4.

Given the large bias in the data toward higher levels, no model was able to correctly identify labels 0 or 1. Adding additional slices with the lower labels would improve model performance for lower labels and may potentially help the model identify more distinguishing features between labels 3 and 4. However, with the current available data it is clear that the difference between labels 3 and 4 are largely indistinguishable. Binning or relabeling the slices is recommended for any future training of models.

In addition, we highly recommend increasing the amount of data, and a broader range on the quality of the data, if possible. Including slices from both auto-annotations and signal processing algorithms would give these models an increase range of labels, however would require labeling of these models performance by informed scientists. Incorporating additional metrics could also aid with the distinguishing of quality annotations, as you can extract additional representative statistics from the annotations. Finally, taking these model output for segments from the same radargram and averaging can provide an overall score for the annotated radargram. This approach allows for a fine grain evaluation of where the model is producing quality annotations as opposed to subpar annotations.

Chapter 4: DEVELOPMENT AND INITIAL TESTING OF XR-BASED FENCE DIAGRAMS FOR POLAR SCIENCE [2]

This work was published by IEEE, presented at IGARSS 2023 and is publicly available at: https://ieeexplore.ieee.org/document/10281776 .

Authors: Naomi Tack[1], Nicholas Holschuh[2], Sharad Sharma[3], Rebecca Williams[1], Don Engel[1]

[1]University of Maryland, Baltimore County, [2]Amherst College, [3]University of North Texas

4.1 Abstract

Earth's ice sheets are the largest contributor to sea level rise. For this reason, understanding the flow and topology of ice sheets is crucial for the development of accurate models and predictions. In order to aid in the generation of such models, ice penetrating radar is used to collect images of the ice sheet through both airborne and ground-based platforms. Glaciologists then take these images and visualize them in 3D fence diagrams on a flat 2D screen. We aim to consider the benefits that an XR

visualization of these diagrams may provide to enable better data comprehension, annotation, and collaborative work. In this paper, we discuss our initial development and evaluation of such an XR system.

4.2 Introduction

Earth's ice sheets are the largest source of uncertainty in models of future sea level rise [20]. While the systems governing ice loss are complex and have proven difficult to quantify, the topology of ice layers within the Greenland and Antarctic Ice Sheets contain information about the physical processes that govern ice dynamics. Layer shapes reflect spatial variability in the accumulation and flow of ice, and can therefore be used to constrain critical unknowns in ice flow models. Incorporating layer information into these models has the potential to reduce boundary condition uncertainty [21], and thereby to improve our projections of future ice flow behavior. Because ice flow (together with changes in surface mass balance) determines the ice sheet contribution to sea level rise [22,23], accurate predictions are a matter of great societal concern.

The topology of Greenland's ice layers is captured in ice-penetrating radar imagery, collected from both airborne and ground-based platforms. The current approach to making scientific use of these images requires interpretation by domain specialists and often starts with the annotation of layers within the ice. However, three-dimensional (3D) structures are difficult to explain from a single, planar cross-section of the ice. To build context, multiple radar images are often visualized together, in an effort to evaluate the continuity and spatial extent of measured features. To do that, radar imagery is commonly visualized in 3D space as fence diagrams (Fig 4.1).

Fence diagrams are typically rendered on flat computer screens, despite being 3D scenes. While users of these 2D interfaces (e.g., in Matlab) can pan and rotate the content, they lack the immersion afforded by a headset-based XR interface. In other scientific disciplines, XR has been shown to support scientific discovery by enhancing the ability of domain experts to understand their own data [24, 25], by empowering researchers to make spatial annotations more quickly and more accurately [26, 27], and by immersing collaborators in a shared XR visualization [28]. We anticipate that hypothesis generation by polar scientists will be improved in speed and quality with an immersive view. In this work, we describe our development and initial evaluation of an XR-based system for 3D fence diagrams of polar radar images. To ensure compatibility between headsets and simplicity for polar scientists who may be novice XR users, our visualizations are developed in WebXR, allowing polar scientists to load the fence diagrams with no software installation required in most commodity headsets.

4.3 Data

The Center for Remote Sensing of Ice Sheets (CReSIS) radar images are captured from overlapping flight paths in many areas of Greenland [3]. These cross-sections of the ice sheet show features which have arisen from the movement and formation of the layers over a period of more than 100,000 years. Historical events such as volcanic explosions, heavy precipitation, and melt are also recorded in layers within the sheet. By studying these cross-sections, scientists can determine flow and movement of ice over many years. For example, a set of three lines (the "three sisters") consistently appears on radar images taken across Greenland and the ice at these three layers is known to be $35,000$ to $55,000$ years old [29]. Deformations of these and other features have happened due to shifts in the ice in the time since then.

4.4 Spatial Positioning in XR

Due to the uneven altitude of the radar-equipped aircraft, the overall elevation varies across radar images. Therefore, we process each image to ensure a common elevation across each row of pixels [30]. We further introduce cropping to give a base elevation of $-2,000$ meters so that the bases of all the images align, and add padding so that each image is as tall as the highest calculated elevation. This ensures that the nth row of pixels in each image represents the same altitude as the nth row in the other images. Occasionally, there are slight differences in elevation within ± 5m. However, this error is almost negligible given the elevation range is in the thousands of meters.

Visually, these differences are occasionally noticeable but often the corresponding layers are easily identified. Additional information regarding the position of the radar slice, its time of collection, and its elevation data are collected and consolidated. These spatial data are then processed and transformed from latitude and longitude into the appropriate Cartesian coordinates needed to represent these images in XR. From this data we can determine the distance which the radar slice covers, its midpoint, and the direction along which the slice was captured. This directional information is vital to insuring the correct alignment in the visualization. Since each radar image has a unique identifier, which identifies its date of collection and radar parameters, we use this as a key when formatting into a JSON for the XR framework processing.

Two locations were chosen for initial diagrams as they provided both relatively simple crosshatch patterns as well as additional radar slices to incorporate once the orientation and direction of the images was determined. The first area is located in Northeastern Greenland (not pictured) and a small subset of 5 images were pulled from a much larger crosshatch pattern. Using the pre-processed and consolidated data, our framework is able to render the radar slices along the direction of collection with correct orientation and positioning between slices. In order to align the image with the correct orientation, we determined the relative position of the endpoints (East-West or North-South). By comparing the flight paths with the framework's positioning we were able to verify correct positioning.

Once this basis for aligning images was determined we moved on to another location in North Western Greenland which spans a larger area as depicted in Figure 4.2 and Figure 4.4. This allowed us to test the scalability of our design and ensure that we did not code to the specific case. The coordinate systems were different for the images and the XR world space, which cause the orientation to be mirrored.

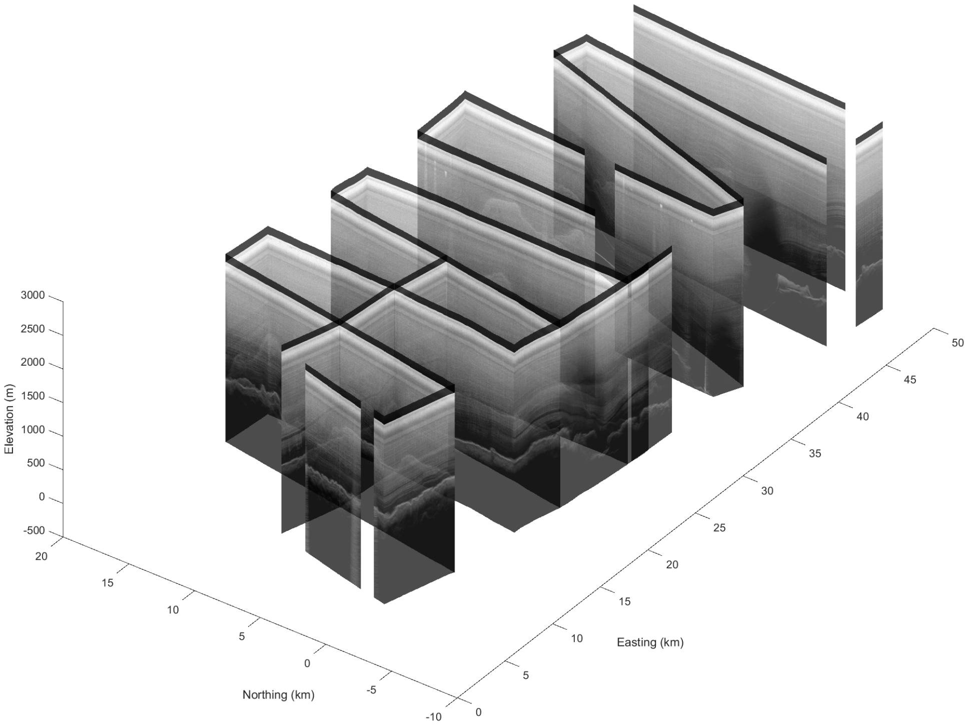

Figure 4.1: Fence diagram of radar data collected at Hercules Dome, Antarctica.

We adjusted the coordinates accordingly. Both regions also required curved paths for the images to lie along in order to accurately represent the flight paths and all local radar images. The initial development left these curved flight paths for future development as XR implementation of curved surfaces is significantly more complex than placement of the rectangles which result from straight portions of flight paths.

4.5 User Interface in XR

Initial development of the XR system began in BabylonJS. This provided the base for the code to generate the initial design and layout of the fence diagrams. This platform proved unsuitable as user controls and support were limited for the

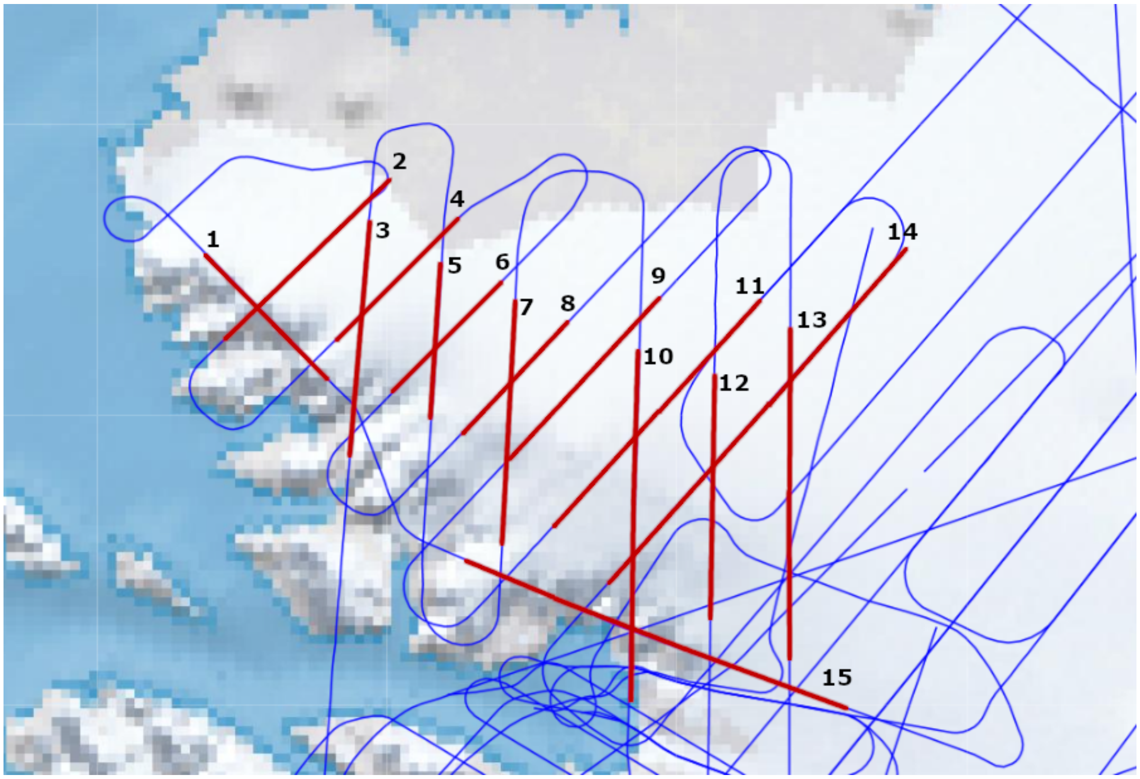

Figure 4.2: Northwest Greeland with CReSIS flight paths in blue and red. Ice penetrating radar from red paths is shown in XR in Fig. 4.4.

framework. By transitioning our development to A-Frame, we were able to leverage a wider range of community-developed tools for spatial navigation. We further developed controls which allow scientists to scale, rotate and move through the world with the handheld controllers. We programmed an interface which allows users to click-and-drag using a pair of controllers, fixing the controllers to two points in space and transforming the world as that pair of points is dragged in three dimensions. This functionality allows movement through the fence diagram in an intuitive way, while still allowing the user to move walk about the planes.

4.6 Future Work

Initial development of this system seems to suggest promising results for the utility of an XR visualization of fence diagrams. We hope to further develop this system to allow for annotation of the images in an exportable and collaborative manner. Developing controls for such layer annotation, combined with visualizations of automated annotation [2, 4], would further strengthen the understanding of ice sheet flow. We plan to allow XR-based glaciologists to select an area of interest from a floating 2D map of Greenland, with a 3D version of that scene automatically rendered in real time.

In addition to the capabilities of the software we also plan to develop a series of tasks and questions to best evaluate the efficacy of such a system. By including glaciologists in the development we are best able to tailor the system to best fit the needs and domain specific constrains. This also allows the evaluation metric to be more representative of the actual work being done by glaciologists.

These tasks and revisions aim to help answer our research questions:

RQ1 Are polar scientists able to navigate to areas of interest and develop hypotheses from an XR fence diagram visualization? How do they report the experience compared to doing similar hypothesis generation work compared to 2D fence diagrams?

RQ2 Given the size and quantity of radar images available, are there any hardware or software limitations which will be difficult to be overcome? What tech-

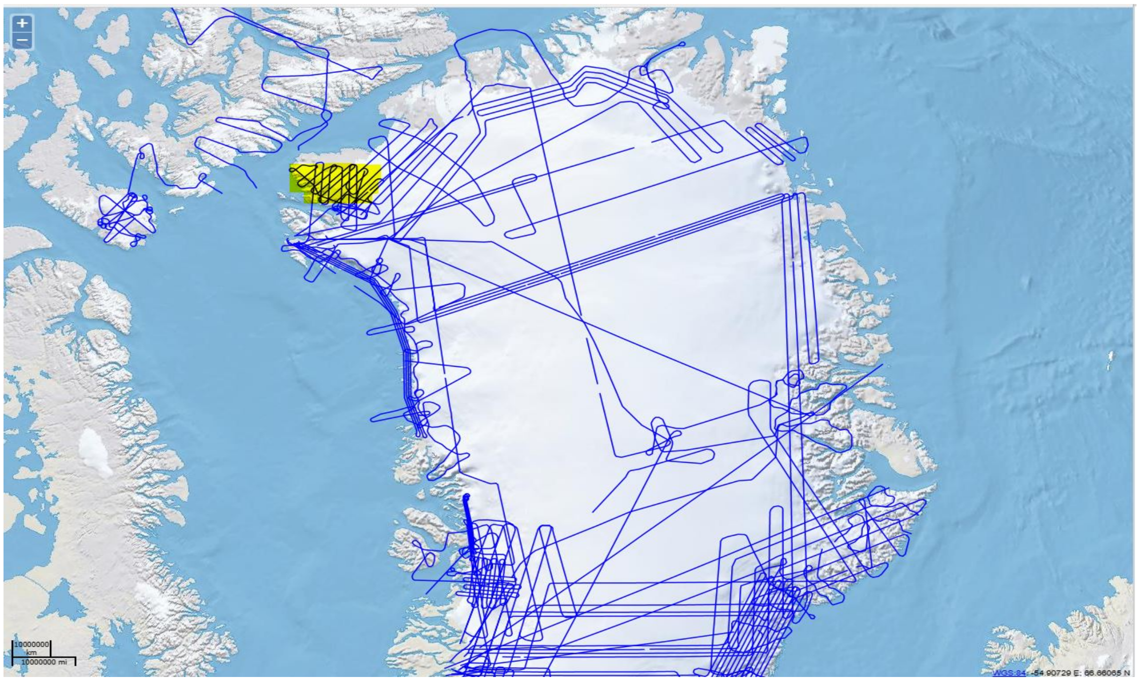

Figure 4.3: Map of Greenland with flight paths of CReSIS data in blue. Yellow area is shown in detail in Fig. 4.2.

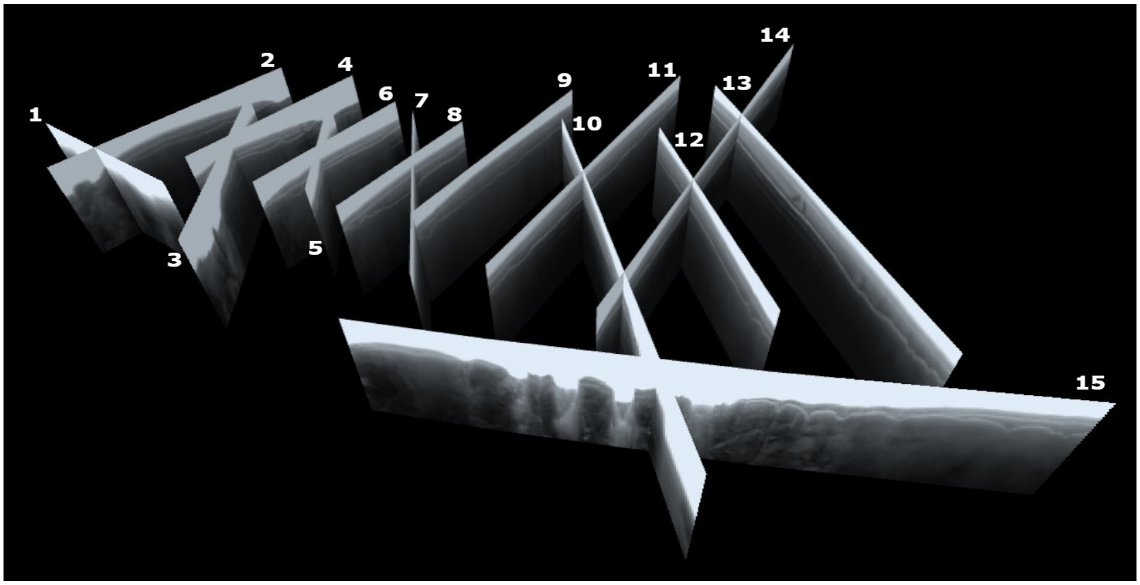

Figure 4.4: XR fence diagram of radar data collected in Fig. 4.3 and referenced to flight paths as shown in Fig. 4.2.

niques (e.g., dividing images into quad trees) are necessary to ensure smooth performance?

RQ3 Do polar scientists (our study participants) report any challenges or discomfort with using XR for Fence Diagrams which should be considered against or alongside any benefits?

While this initial work is focused purely on XR-based visualization without any capacity for annotation, it serves as a foundation on which to build future manual layer annotation tools in XR, as well as a path forward for more efficient human evaluation of automated layer annotations.

Chapter 5: EXTENDED CONTROLS AND RECOMMENDED EVALUATION OF WEBXR FENCE DIAGRAMS

Fence diagrams are a crucial part of developing glaciologist spatial understanding of an glacier. Often these are viewed on a 2D screen rather than the 3D space that they occupy, leading to a potential loss in understanding due to the dimensionality reduction. In our previous work we provided a basis for visualizing ice sheet data in 3D WebXR [2]. Here we discuss the increased functionality of the system as well as expand on methods to evaluate such a system. These evaluation are broad and provide general outlines for consideration rather than concrete steps. Our future work includes a small, well-defined, subset of these suggestions.

5.1 Extended Controls

In order to aid in the understanding of how layers flow throughout a larger region, the entirety of the cross-sections are displayed. However, when we wish to annotate a particular layer, this 3D structure creates occlusion and increases the cognitive load on the annotator. For this reason we wish to hide the supplemental slices and focus on a single slice. Following our previous work we chose the most intuitive way to select layers, point and click. We take advantage of the multiple buttons on the controllers to enforce intentionally in triggering this point and click, by using the farthest button, to ensure accidental selection or deselection is minimized. This choice allows nearer buttons to be used for more precise annotations.

When visualizing radargrams in Matlab, glaciologist change the opacity of the diagram to better understand the connection between layers appearing in intersecting slices. We have incorporated this utility into the WebXR controls allowing users to navigate through the scene while varying the level of opacity.

5.1.1 Recommend Future Controls

Annotations are a necessary part of this system and several methods exist for identifying picked layers. Recent development of painting games in VR provides many frameworks for which to allow users to freehand the layers. Initial consideration of this approach highlights that it is inefficient for developing complex multi-layer fence annotations. However, this could provide an intuitive and quick method for both the medium and cultivating 3D understanding. The major benefits to this approach are the wide body of relevant work in current XR work and the speed and isolation of desired layers through our the area of interest.

Alternatively line picking scripts are available in Matlab and in Python which can be incorporated. These scripts primarily identify the brightest horizon, e.g. surface and bedrock layers, through stepping through the gradient values. Integration of these scripts would give annotators a spring board from which to start, in addition to similarities between technologies they might have previously used.

However, consideration of both the time required to execute these scripts over large areas and selection of layers adjacent to the target layer must be considered when implementing this approach.

Line interpolation scripts seem to be a happy medium between the above methods. These Matlab scripts take several points along a layer and interpolate the layer using the underlying radargram. This approach is the most promising for long term annotation utility but also will require significant effort to integrate. Further consideration of the trade-offs between these approach's is recommended.

5.2 Proposed Future Evaluation Tasks

Evaluation of the the utility of this early stage system focuses on understanding the development of a mental 3D model of the fence diagram as a whole. We suggest simple tasks and a questionnaire to consider the effect of immersion in fence diagrams. These tasks should target several different use cases in-order to determine both strengths and weakness to inform future functionality. We comment on two proposed general tasks hypothesis generation and scientific communication, such as teaching or collaboration.

Immersing scientist in interactive visualizations has shown increased understanding of spatial data and increased annotation ability [24–28]. We hypothesis that immersing polar scientist in this 3D representation could allow a faster generation of hypothesis about 3D objects inside a glacier, layer connectivity or spatial structure of the glacier. Tasks such as layer picking across several intersecting radargrams and hypothesis about flow could be useful for this task.

Similarly we hypothesis that immersion of peer collaborators in a single scene will positively affect communication and exchange of ideas, as compared to the traditional 2D scenes. We suggest tasks which can be completed with both media and a shared questionnaire, adding qualitative analysis of all parties impressions of both the tool and its aid in the task. These tasks could include collaborative layer picking and object identification. Additionally, evaluation on the system as a teaching tool is recommended. Measuring students understanding of glaciers and layer identification when presented with a 2D vs 3D model would inform where each method is best used for transfer of knowledge.

Chapter 6: CONCLUSION

Throughout this work we have considered two related ways which computer science techniques can be applied to the study of polar science. Both computer vision and immersive visualization are tools that work towards developing a richer understanding of radargrams. We discuss both the contribution and directions for future research below.

6.1 Quality Metrics

6.1.1 Contributions

In order to independently evaluate layer picking models, we considered using derived properties of the layers themselves to provide a score for the 'goodness' of a picked layer. We provided in Section 2 and 3 several spatially informed quality metrics to describe slices of picked layers, and an initial evaluation of these metrics in their scoring ability. Our models indicated that our $0-5$ labels for 'bad' to 'good' was too broad and leaves room for future work, as discussed below. These models also indicate that these metrics can be used to identify disparate labels relatively well.

6.1.2 Future Work

The evaluation of quality metrics being measured has shown that a coarser grouping of ‘bad’ to ‘good’ is required to accurately evaluate the efficacy of these metrics. We recommend a $0 - 3$ or ‘good’, ‘average’ and ‘bad’ labels. With this coarser classification we anticipate that these quality metrics will provide a basis for evaluating picked layers, upon which future layer picking models can be further fine tuned. Future work in this area should consider a coarser classification and the expansion of this evaluation to a complete radar slice, rather than just segments of a slice.

Expanding the available metrics to include both additional local metrics and broader global metrics will also help in the classification of slices as whole. For ground truth, defined as expert opinion, dependant metrics we suggest a local metric to measure inter-annotator agreement for annotated layers. We also suggest a global layer completeness metric, which is a track association problem applied to annotated layers. Identifying if, or when, a disappearing layer reappears gives more context to the glaciologist about the structure of the glacier and maybe helpful to the models evaluation of a given annotation.

Radargrams can be captured by both ground based and airborne radar. This provides both short high detail slices and long lower detail slices respectively. The extent to which the quality of an underlying radargram impacts both the expert and automatically generated annotations is an interesting question. This radargram quality may impact the the downstream quality metrics described above. Determining a metric to encapsulate this difference in radargram quality could provide researchers with a clearer sense of what data can be accurately evaluated with these evaluation models.

6.2 WebXR

6.2.1 Contributions

We have developed and deployed a 3D WebXR fence diagram to immerse scientists in a hardware independent, open source, and portable manner. We have developed controls which allow for intuitive navigation and easy plane selection for considering single radar slices. This provides broader access to both scientists and students for visualizing radargrams, thereby lowering the barrier of access imposed by proprietary software and steep learning curves of such software.

6.2.2 Future Work

This visualization system serves as a base upon which annotation controls, layer tracing algorithms can be integrated. There are also several different user studies which are recommended to be performed to evaluate the usability and utility of the system, which have been detailed in Section 5.

6.3 Broader Impacts

We have considered some possibilities on how to use derived data from picked layers and how best to visualize this radar data to allow for easier annotation and spatial understanding of glaciers. The work presented above contributes toward advancing understanding in polar science and opens up interesting avenues for future research.

Bibliography

[1] Naomi Tack, Bayu Adhi Tama, Atefeh Jebeli, Vandana Janeja, Don Engel, and Rebecca Williams. Metrics for the quality and consistency of ice layer annotations. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2023.

[2] Naomi Tack, Nicholas Holschuh, Sharad Sharma, Rebecca Williams, and Don Engel. Development and initial testing of XR-based fence diagrams for polar science. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2023.

[3] CReSIS. Radar depth sounder (RDS) data. http://data.cresis.ku.edu/, 2021. Online; accessed 31-May-2023.

[4] Atefeh Jebeli, Bayu Adhi Tama, Vandana Janeja, Nick Holschuh, Claire Jensen, Mathieu Morlighem, Joseph A. MacGregor, and Mark Fahnestock. TSSA: Two-step semi-supervised annotation for englacial radargrams on the greenland ice sheet. In *International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2023. In press.

[5] Nicholas Holschuh, Byron R. Parizek, Richard B. Alley, and Sridhar Anandakrishnan. Decoding ice sheet behavior using englacial layer slopes. *Geophysical Research Letters*, 44(11):5561–5570, 2017.

[6] Joseph A. MacGregor, William T. Colgan, Mark A. Fahnestock, Mathieu Morlighem, Ginny A. Catania, John D. Paden, and S. Prasad Gogineni. Holocene deceleration of the greenland ice sheet. *Science*, 351(6273):590–593, 2016.

[7] Joseph A. MacGregor, Mark A. Fahnestock, Ginny A. Catania, John D. Paden, S. Prasad Gogineni, S. Keith Young, Susan C. Rybarski, Alexandria N. Mabrey, Benjamin M. Wagman, and Mathieu Morlighem. Radiostratigraphy and age structure of the greenland ice sheet. *Journal of Geophysical Research: Earth Surface*, 120(2):212–241, 2015.

[8] Jason P. Briner, Joshua K. Cuzzone, Jessica A. Badgeley, Nicolás E. Young, Eric J. Steig, Mathieu Morlighem, Nicole-Jeanne Schlegel, Gregory J. Hakim, Joerg M. Schaefer, Jesse V. Johnson, Alia J. Lesnek, Elizabeth K. Thomas, Estelle Allan, Ole Bennike, Allison A. Cluett, Beata Csatho, Anne de Vernal, Jacob Downs, Eric Larour, and Sophie Nowicki. Rate of mass loss from the greenland ice sheet will exceed holocene values this century. *Nature*, 586(7827):70–74, Oct 2020.

[9] D. West, J. T. Harper, N. F. Humphrey, and W. T. Pfeffer. Measurement and Modeling of Firn Densification in the Percolation Zone of the Greenland Ice Sheet. In *AGU Fall Meeting Abstracts*, volume 2009, pages C31E–0477, December 2009.

[10] Elham Tabassi, Martin Olsen, Oliver Bausinger, Christoph Busch, Andrew Figlarz, Gregory Fiumara, Olaf Henniger, Johannes Merkle, Timo Ruhland, Christopher Schiel, and Michael Schwaiger. NIST fingerprint image quality 2, 2021-07-13 04:07:00 2021.

[11] Yan Song, Zheng Hu, Tiancheng Li, and Hongqi Fan. Performance evaluation metrics and approaches for target tracking: A survey. *Sensors*, 22(3):793, Jan 2022.

[12] Elisa Drelie Gelasca, Jiyun Byun, Boguslaw Obara, and B.S. Manjunath. Evaluation and benchmark for biological image segmentation. In *2008 15th IEEE International Conference on Image Processing*, pages 1816–1819, 2008.

[13] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008(1):246309, May 2008.

[14] Siting Xiong, Jan-Peter Muller, and Raquel Caro Carretero. A new method for automatically tracing englacial layers from MCoRDS data in NW Greenland. *Remote Sensing*, 10(1):43, 2017.

[15] Atefeh Jebeli, Bayu Adhi Tama, Vandana Janeja, Nicholas Holschuh, Claire Jensen, Mathieu Morlighem, Joseph A. MacGregor, and Mark Fahnestock. TSSA: Two-step semi-supervised annotation for englacial radargrams on the greenland ice sheet. submitted for publication.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[17] K Pravinkrishnan, N Sivakumar, P Balasundaram, and L Kalinathan. Classification of plant species using alexnet architecture. *Working Notes of CLEF*, 2022.

[18] Yara Mohajerani, Michael Wood, Isabella Velicogna, and Eric Rignot. Detection of glacier calving margins with convolutional neural networks: A case study. *Remote Sensing*, 11(1):74, 2019.

[19] Zhiyuan Xie, Umesh K. Haritashya, Vijayan K. Asari, Brennan W. Young, Michael P. Bishop, and Jeffrey S. Kargel. Glaciernet: A deep-learning approach for debris-covered glacier mapping. *IEEE Access*, 8:83495–83510, 2020.

[20] V. Masson-Delmotte, P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou. IPCC, 2021: Summary for Policymakers. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 3–32. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021.

[21] N. Holschuh, B. R. Parizek, R. B. Alley, and S. Anandakrishnan. Decoding ice sheet behavior using englacial layer slopes. *Geophysical Research Letters*, 44:5561–5570, 2017.

[22] Benjamin D. Hamlington, Alex S. Gardner, Erik Ivins, Jan T. M. Lenaerts, J. T. Reager, David S. Trossman, Edward D. Zaron, Surendra Adhikari, Anthony Arendt, Andy Aschwanden, Brian D. Beckley, David P. S. Bekaert, Geoffrey Blewitt, Lambert Caron, Don P. Chambers, Hrishikesh A. Chandanpurkar, Knut Christianson, Beata Csatho, Richard I. Cullather, Robert M. DeConto, John T. Fasullo, Thomas Frederikse, Jeffrey T. Freymueller, Daniel M. Gilford, Manuela Girotto, William C. Hammond, Regine Hock, Nicholas Holschuh, Robert E. Kopp, Felix Landerer, Eric Larour, Dimitris Menemenlis, Mark Merrifield, Jerry X. Mitrovica, R. Steven Nerem, Isabel J. Nias, Veronica Nieves, Sophie Nowicki, Kishore Pangaluru, Christopher G. Piecuch, Richard D. Ray, David R. Rounce, Nicole-Jeanne Schlegel, Hélène Seroussi, Manoochehr Shirzaei, William V. Sweet, Isabella Velicogna, Nadya Vinogradova, Thomas Wahl, David N. Wiese, and Michael J. Willis. Understanding of contemporary regional sea-level change and the implications for the future. *Reviews of Geophysics*, 58(3):e2019RG000672, 2020. e2019RG000672 2019RG000672.

[23] Ben Smith, Helen A. Fricker, Alex S. Gardner, Brooke Medley, Johan Nilsson, Fernando S. Paolo, Nicholas Holschuh, Susheel Adusumilli, Kelly Brunt, Bea Csatho, Kaitlin Harbeck, Thorsten Markus, Thomas Neumann, Matthew R. Siegfried, and H. Jay Zwally. Pervasive ice sheet mass loss reflects competing ocean and atmosphere processes. *Science*, 368(6496):1239–1242, 2020.

[24] Sebastian Pirch, Felix Müller, Eugenia Iofinova, Julia Pazmandi, Christiane V. R. Hütter, Martin Chiettini, Celine Sin, Kaan Boztug, Iana Podkosova, Hannes Kaufmann, and Jörg Menche. The vrnetzer platform enables interactive

network analysis in virtual reality. *Nature Communications*, 12(1):2432, Apr 2021.

[25] Kaur Kullman, Laurin Buchanan, Anita Komlodi, and Don Engel. Mental model mapping method for cybersecurity. In Abbas Moallem, editor, *HCI for Cybersecurity, Privacy and Trust*, pages 458–470, Cham, 2020. Springer International Publishing.

[26] Kaur Kullman and Don Engel. User interactions in virtual data explorer. In *Augmented Cognition: 16th International Conference, AC 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26 – July 1, 2022, Proceedings*, page 333–347, Berlin, Heidelberg, 2022. Springer-Verlag.

[27] Corentin Guérinot, Valentin Marcon, Charlotte Godard, Thomas Blanc, Hippolyte Verdier, Guillaume Planchon, Francesca Raimondi, Nathalie Boddaert, Mariana Alonso, Kurt Sailor, Pierre-Marie Lledo, Bassam Hajj, Mohamed El Beheiry, and Jean-Baptiste Masson. New approach to accelerated image annotation by leveraging virtual reality and cloud computing. *Frontiers in Bioinformatics*, 1, 2022.

[28] Torvald F Ask, Kaur Kullman, Stefan Sütterlin, Benjamin J Knox, Don Engel, and Ricardo G Lugo. A 3d mixed reality visualization of network topology and activity results in better dyadic cyber team communication and cyber situational awareness, Sep 2022.

[29] M. Fahnestock, W. Abdalati, S. Luo, and S. Gogineni. Internal layer tracing and age-depth-accumulation relationships for the northern greenland ice sheet. *Journal of Geophysical Research: Atmospheres*, 106(D24):33789–33797, 2001.

[30] Nicholas Holschuh. NDH_PythonTools. https://github.com/nholschuh/NDH_PythonTools, 2022. Online; accessed 31-May-2023.