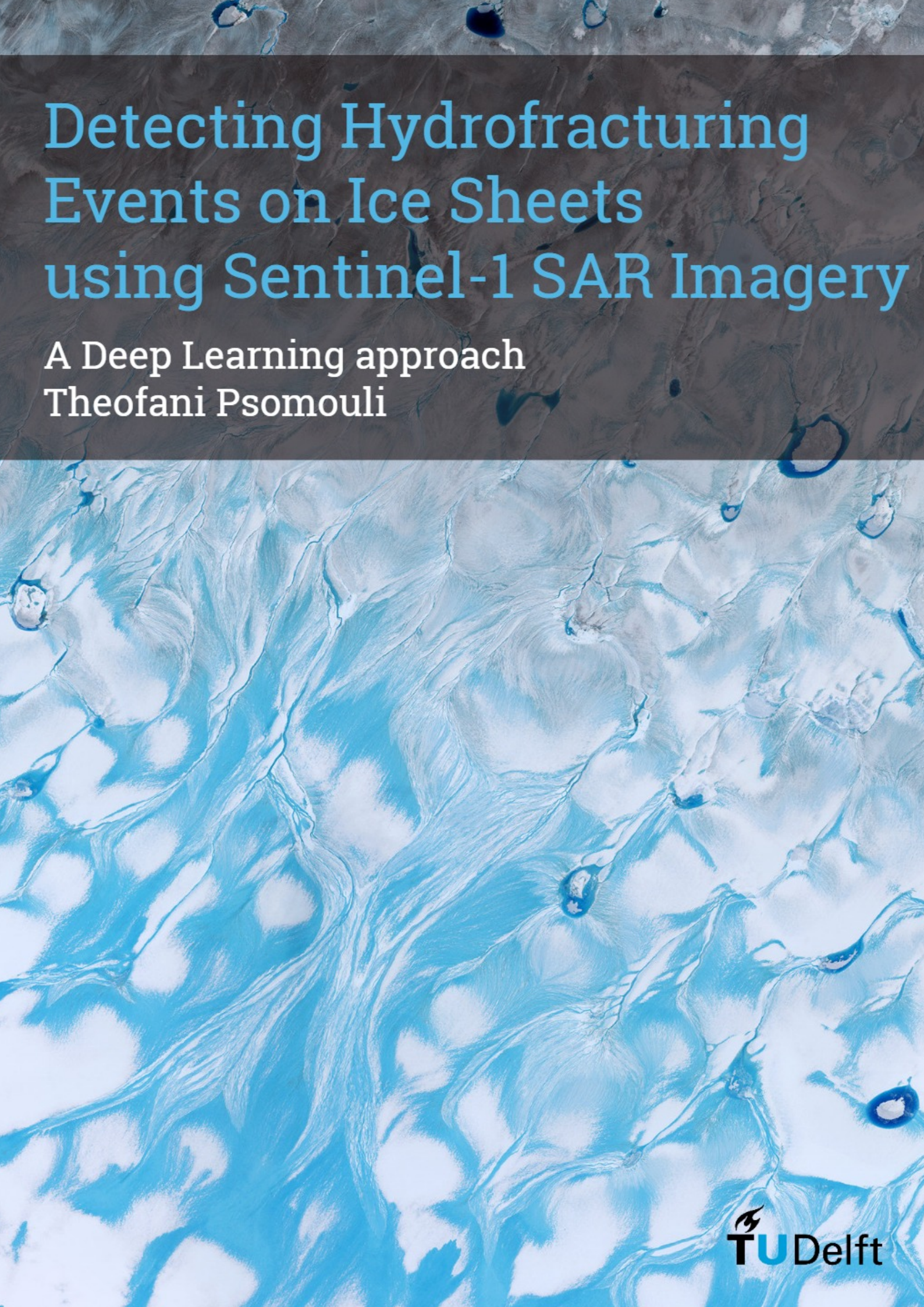
Detecting Hydrofracturing Events on Ice Sheets using Sentinel-1 SAR Imagery

A Deep Learning approach

Theofani Psomouli



TU Delft

Detecting Hydrofracturing Events on Ice Sheets Using Sentinel-1 SAR Imagery.

A Deep Learning approach

by

Theofani Psomouli

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday July 2, 2024 at 10:45 AM.

Thesis committee:	Dr. Ir. S.L.M. Lhermitte, TU Delft
	Ir. S. de Roda Husman, TU Delft
	Dr. Ir. B. Wouters, TU Delft
Project Duration:	February, 2024 - July, 2024
Faculty:	Faculty of Civil Engineering & Geosciences, TU Delft
Student number:	5833124

Cover: Meltwater and surface lakes on the Greenland ice sheet in August 2021. ESA, Sentinel-2

TU Delft

Abstract

Climate change, with global temperatures rising over the past decades, is a primary driver of sea level rise through the thermal expansion of seawater and the melting of the Antarctic and Greenland Ice Sheets (AIS and GIS). These ice sheets are crucial for predicting future sea level changes, as increased melting forms supraglacial lakes. These lakes can induce hydrofracture, leading to ice shelf instability and accelerated ice flow into the ocean, further elevating sea levels and affecting global climate systems.This study focuses on the AIS and GIS, emphasizing the development and application of a deep learning model to detect and classify the behavior of summer supraglacial lakes using Sentinel-1 SAR satellite data.

The methodology involved normalizing SAR imagery data to enhance data consistency, training a deep learning model using a U-Net architecture on a labeled dataset of Greenland lakes for semantic segmentation, and evaluating its performance on both Greenland and Antarctic datasets using metrics such as accuracy, precision, recall, F1-score, and SSIM. The model is trained to distinguish between draining and refreezing lakes based on backscatter intensity patterns captured in the satellite images. Furthermore, a sensitivity analysis is conducted by creating ten different perturbations of the testing dataset, which included variations in intensity, rotation, and zoom levels. The trained model is then applied to Antarctic data to create an Antarctic-wide map of lake behavior.

The deep learning model exhibited high performance, achieving an accuracy of 90.6%, precision of 90.9%, recall of 90.6%, and an F1-score of 90.0%. It showed high classification accuracy for non-lake pixels (97.5%) and draining lake pixels (84.4%), but lower accuracy for refreezing lake pixels (55.4%) due to class imbalance. Sensitivity analysis revealed optimal performance on the 'zoomed out' dataset with an overall accuracy of 92.9%. Applying the model to Antarctic data successfully identified regions of draining and refreezing lakes, providing a starting point for monitoring ice sheet dynamics and their implications for climate change.

This study underscores the potential of deep learning models to enhance supraglacial lake monitoring, contributing to a better understanding of ice sheet stability and the impacts of climate change. Future work should address class imbalance and explore further model optimizations to improve classification accuracy across both lake types.

Preface

Working on this thesis has been a remarkable journey, filled with exciting and frustrating experiences. Over these six months, I have learned a great deal about cryosphere dynamics, SAR data, and deep learning. It is bittersweet to conclude this chapter of my life, which has been filled with wonderful memories.

I wish to express my heartfelt thanks to Sophie, my daily supervisor, for her unwavering support over these past six months. Working on a project that aligns with my interests has been incredibly fulfilling, and our frequent meetings, often more than once a week, have been invaluable. Your constructive feedback, endless patience, and the freedom you granted me to explore different paths have been crucial to my progress. I loved working with you and hope we can collaborate again in the future.

I am also grateful to Stef, my second supervisor, for offering his insights and valued input throughout this process. Your feedback has been instrumental in clarifying several key aspects of my research.

Additionally, I would like to thank Bert from my thesis committee for his involvement in assessing the project. Your feedback has been essential in shaping this work.

Lastly, a big thank you to my family and friends for sticking by me through these challenging yet precious two years. Your support and encouragement have meant the world to me.

Theofani Psomouli
Delft, June 2024

Contents

Nomenclature

List of Abbreviations

Abbreviation	Definition
AIS	Antarctic Ice Sheet
AP	Antarctic Peninsula
CNN	Convolutional Neural Network
EO	Earth Observation
ESA	European Space Agency
EW	Extra-Wide swath
GHG	Greenhouse Gas
GIS	Greenland Ice Sheet
GEE	Google Earth Engine
GeoTIFF	Geographical Earth Orbit Tagged Image File Format
GRD	Ground Range Detected
IPCC	Intergovermental Panel on Climate Change
IW	Interferometric Wide Swath
MAR	Modèle Atmosphérique Régional
MODIS	Moderate Resolution Imaging Spectroradiometer
NDWice	Normalized Difference Water Index adapted for Ice
QGIS	Quantum Geographic Information System
RADAR	Radio Detection and Ranging
ROI	Region Of Interest
SAR	Synthetic Aperture Radar
S1	Sentinel-1
SLs	Summer Supraglacial Lakes
SM	Stripmap
SMB	Surface Mass Balance
SSIM	Structural Similarity Index Measure
WM	Wave Mode

1 Introduction

1.1. Background

Over the past decades, the temperature across the world has been increasing, with global surface temperature reaching 1.5°C above pre-industrial levels in 2011–2020. One of the major impacts of climate change is global sea level rise as it is unavoidable for centuries to millennia due to the continuous warming of the deep ocean and ice sheet melting (Lee and Romero 2023). Sea level rise is primarily driven by thermal expansion of seawater, melting glaciers and ice caps, and the loss of ice from the ice sheets (Church and White 2011). In this research, the focus is on the two ice sheets, the Antarctic Ice Sheet (AIS) and the Greenland Ice Sheet (GIS).

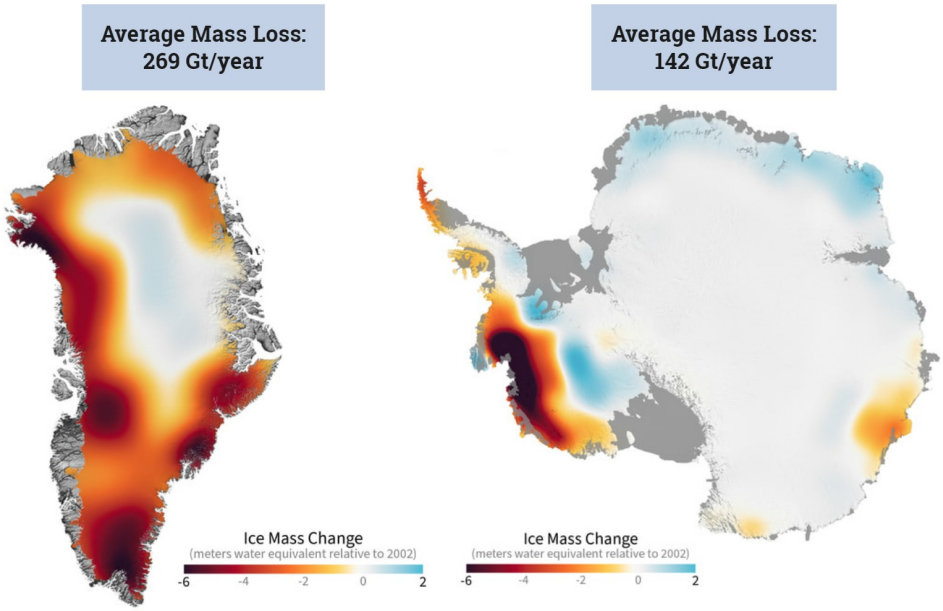


Figure 1.1: Image created from GRACE (2002-2017) and GRACE-FO (2018 -) data, shows ice changes in Arctic and Antarctic ice mass since 2002. Orange and red indicate areas that experience ice mass loss, while blue indicates areas that gained ice mass. White indicates areas where there has been little to no change in ice mass since 2002. Lastly, gray indicates floating ice shelves whose mass change GRACE and GRACE-FO do not measure (NASA and JPL/Caltech 2023).

The AIS is the largest single mass of ice on Earth, covering 14 million square kilometers and containing about 90% of the world's ice. The significance of the AIS extends beyond its size as it is considered a key factor when it comes to global sea level rise, climate regulation, and overall scientific research. The AIS has a sea level rise equivalent of approximately 58 meters, which means that if the whole ice sheet would melt it would lead to a sea level rise of approximately 58 meters (Fretwell et al. 2013). Out of the 14 million square kilometers, more than 1.5 million are covered by ice shelves which is comparable to the size of the GIS (E. Rignot et al. 2013). Ice shelves are comprised of thick mass of floating ice that is attached to land ice (*National Snow and Ice Data Center* 2023). One of the most important

factors for climate change evolution is ice shelf stability as it affects its contribution to global sea level rise. For Greenland, the sea level rise equivalent is approximately 8 meters (**Fretwell et al. 2013**).

Altimetry and gravimetry measurements have shown that AIS and GIS have been losing mass over the last two decades. As seen in Figure 1.1, between 2002 and 2023, Greenland experienced an average loss of 269 Gigatons of ice per year due to surface melting and iceberg calving. This caused global sea level to rise by 0.8 mm per year. When looking at the figure, it can be seen that higher elevation areas (center Greenland) have experienced little to no change. On the contrary, lower-elevation areas experienced over six meters of ice mass loss (expressed in water equivalent height: dark red) over 21 years. The largest mass loss is focused along the West Greenland coast. Furthermore, for the same years, Antarctica experienced an average loss of 142 Gigatons of ice per year. This caused global sea level to rise by 0.4 mm per year. The AIS is overall shrinking but there are variations between East and West Antarctica, with some regions experiencing large losses while others experience small mass gains. East Antarctica experienced moderate amounts of mass gain due to snow accumulation and did not experience severe ice loss, while most of the loss was concentrated south of the Antarctic Peninsula in West Antarctica (**NASA and JPL/Caltech 2023**). According to the latest IPCC AR6 report, global mean sea level will rise between 0.18 (SSP1-1.9) and 0.23 m (SSP5-8.5) by 2050, and between 0.38 (SSP1-1.9) and 0.77 m (SSP5-8.5) by 2100 (**Fox-Kemper et al. 2021**).

A study by **Gilbert and Kittel (2021)** examines the potential impact of sustained atmospheric warming on Antarctic ice shelves using the Modèle Atmosphérique Régional (MAR) to assess the effects of warming scenarios of 1.5°C, 2°C, and 4°C above pre-industrial temperatures on the surface mass balance (SMB). By using melt and runoff as indicators of ice shelf stability, they find that several ice shelves (Larsen C, Wilkins, Pine Island, and Shackleton) are vulnerable to collapse at 4°C of warming. Furthermore, **Lai et al. (2020)** associated the vulnerability of ice shelves to hydrofracture with the accumulation of meltwater and the presence of damage. Nonetheless, the term ”damage” can be interpreted in many different ways and it remains uncertain which specific types of damage and to what degree they contribute to vulnerability.

As the atmospheric temperature increases, if there is sufficient surface melt and low firn air content, surface melt can accumulate and start ponding (lake formation). This process can contribute to ice shelf collapse as it usually occurs after a relatively warm summer with increased surface melting (**Davies 2020**). As a consequence, surface lakes form on the ice sheets which highly affect its future behavior, contribution to sea level rise, stability, and overall implications to climate change (**Alley et al. 2018**). One of the mechanisms affecting ice shelf stability and facilitated by the increase in temperature is hydrofracture which is an erosion process in which existing damage propagates under the load of water through an ice shelf. More specifically, hydrofracture occurs when water infiltrates and fills crevasses, and if the pressure increase becomes high enough the ice can be compromised as illustrated in Figure 1.2. In order for the fracture to propagate, pressure needs to keep increasing which happens as more water flows in from the reservoir which is the ponding of meltwater on the surface of the ice shelf. With many closely spaced fractures, part of the ice may break off causing the ice shelf to collapse. On the other hand, lakes that do not drain tend to refreeze after the summer. Rapid drainages, meaning that as the weight of the water widens the crack, it creates a connection to the underlying ocean leading to the water being drained rapidly are considered an indicator of hydrofracture (**J. Sommer and Izeboud. 2023**).

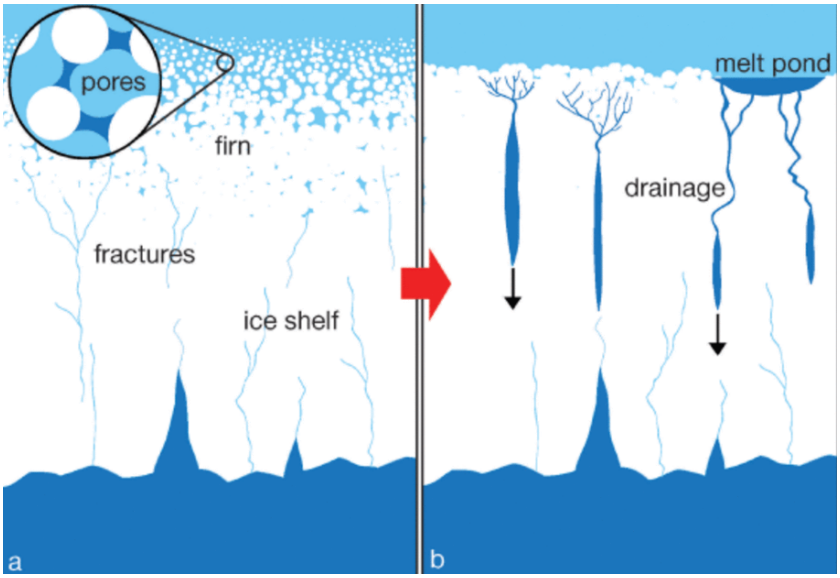


Figure 1.2: Conceptual illustration of firn air depletion and its consequences for ice-shelf hydrology and stability. (a) An ice shelf covered by a firn layer containing sufficient air. The inset shows meltwater being stored in the pore space of the firn. (b) An ice shelf with a depleted firn layer. Due to the absence of pore space, meltwater forms ponds that drain into fractures. Alternatively, water is routed to the fractures efficiently as shown in the leftmost fractures (**Kuipers Munneke et al. 2014**).

The above-mentioned process likely caused the complete or partial collapse of several ice shelves on the Antarctic Peninsula (AP). However, not all ice shelves are vulnerable to hydrofracture. More specifically, in regions that are characterized by high snow accumulation or permeable, porous firn, any meltwater produced in the summer months infiltrates the upper firn layer and refreezes. In order to prevent hydrofracture, the firn layer below the meltwater lakes must be adequately saturated with refrozen meltwater to prevent efficient downward percolation. Both modeling studies and observations confirm that ice shelves that collapsed on the AP in the past had very minimal firn air thickness prior to their disintegration which indicates that they were preconditioned for the hydrofracture process to take place (**Alley et al. 2018**). Ice shelves play a crucial role in sea level rise because they respond to rising temperatures more quickly than ice sheets or glaciers, which highlights why ice shelf collapse poses a problem. While their collapse alone does not directly contribute to sea level rise, it triggers an acceleration of glaciers that feed into them. As these glaciers accelerate and flow into the ocean, they contribute to the overall sea level rise (**Scambos et al. 2004**).

1.2. Current Methods and Limitations

Currently, in order to detect hydrofracture events in Antarctica research focuses on the use of optical and Synthetic Aperture Radar (SAR) imagery. To detect hydrofracture events at the Amery Ice Shelf, **Trusel et al. (2022)** employed a combination of optical and radar imagery and tidal data analysis. They monitored a supraglacial lake from 2014 to 2020 using high-resolution satellite images, which allowed them to observe changes in the lake's surface and identify post-drainage features, characteristic of rapid, vertical drainage. By correlating the timing of these drainage events with periods of high daily tidal amplitudes, they found that the flexing of the ice shelf due to tidal forces played a key role in inducing hydrofractures. This approach provided a comprehensive understanding of the mechanisms behind hydrofracture events in this region (**Trusel, Pan, and Moussavi 2022**).

Given the challenges posed by optical data (cloud cover, seasonal darkness, high albedo, low sun angle), SAR can be used complementary. The use of SAR offers a way to bypass these limitations since it is not affected by cloud cover or the lack of sunlight. Such research has been carried out to detect hydrofracture events in Greenland using a statistical method by **Benedek and Willis (2021)**.

The above-mentioned study looked at drainages using Sentinel-1 data, by using a statistical automated method. In order to confirm that a significant sudden increase in mean backscatter reflects a change in a specific lake rather than an artifact, the mean backscatter change of each lake is compared to that

of all other lakes in the same scene across consecutive image pairs. For selected lakes, backscatter frequency distributions were examined and found to be nearly normally distributed, ensuring that lake medians and means were similar. The z score of backscatter change for each lake is calculated relative to all lakes within the study site, using a threshold of +1.5 to identify lakes with greater-than-average increases. To ensure these increases in backscatter are sustained and not isolated occurrences, filters were applied to check for reversals in the subsequent three images (within 48 days). This approach is limited by the criteria used to determine lake drainage events as they are conservative and may have missed drainage events (false negatives) rather than identifying events that were not real (false positives) (**Benedek and Willis 2021**). Therefore, while this approach provides valuable insights into lake drainage events, it is important to take these limitations into account when focusing on the approach in this thesis.

Lastly, **Miles et al. (2017)** also conducted a study that looked at surface lakes in Greenland (draining and refreezing) using Sentinel-1 (EW - HH, HV polarizations) data following a threshold-based approach to define lake areas by examining the bimodal distribution of backscatter values in the masked Sentinel-1 images. Next, they selected the lowest point between the peaks as the backscatter threshold value to define the lake area for each particular image. This approach might be limited if applied to Antarctica due to the fact that they used coarser resolution (25 - 40 m) for Sentinel-1 data and in general Antarctic lakes are smaller and might not be detected when having coarser resolution.

1.3. Research Questions

This thesis will focus on a combined approach, that uses a deep learning Convolutional Neural Network (CNN) and more specifically a U-Net in order to distinguish between draining and refreezing events on ice sheets by leveraging known hydrofracture and refreezing events **(de Roda Husman, unpublished)**. Furthermore, ground truth data will be used to complement the Sentinel-1 time series that will be used to train, validate, and test the model. The ground truth data will be used as a label for each lake (time series). Next, ten perturbed datasets will be created from the test dataset to perform a sensitivity analysis on when and why the model performs well. Those are split into temporal and spatial perturbations in order to conclude how model performance is affected. Lastly, the trained model will be applied to Antarctica where no reliable ground truth data was available. Following from the Introduction, the research questions of this thesis are listed below:

How can we develop a deep learning model to classify draining and refreezing lakes in Greenland using Sentinel-1 data? **Approach:** This question will be answered by creating a workflow that includes the exporting and preprocessing of the Sentinel-1 dataset by leveraging known hydrofracture and refreezing events on the Greenland Ice Sheet. Next, an attention U-Net will be trained, validated, and tested to conclude how such a workflow can be efficiently implemented.

How do spatio-temporal patterns and changes affect model performance? **Approach:** This question will be answered by performing a sensitivity analysis and more specifically by creating two temporal and eight spatial perturbations of the testing dataset to investigate how those affect the model performance.

How can this method be improved? **Approach:** By interpreting the results of the approach, several aspects that would improve the overall results will be revealed. Furthermore, the sensitivity analysis on the ten perturbed datasets will provide a better understanding of how the model works.

Can this model be generalized to other polar regions? **Approach:** This question will be answered by applying the model to the Antarctic Ice Sheet and reviewing the model's performance and ability to generalize.

What are the advantages of this approach? **Approach:** This question will be answered by comparing what is currently happening in this field of research and by elaborating on the choices that can improve the methods already employed.

What are the limitations of this approach? **Approach:** This question will be answered by addressing the limitations of certain choices along the chosen workflow.

2
Methodology

2.1. Approach Overview

This study follows an approach that starts with the downloading and preprocessing of Greenland Sentinel-1 SAR time series for lakes identified by optical imagery as seen in Figure 2.1, then follows normalization, the split of the data into training, validation, and testing and lastly a U-Net is employed. Furthermore, ten perturbations of the testing dataset are created in order to evaluate model performance. Lastly, a Sentinel-1 dataset is downloaded and processed for Antarctica. The already trained model is then applied to the Antarctica dataset.

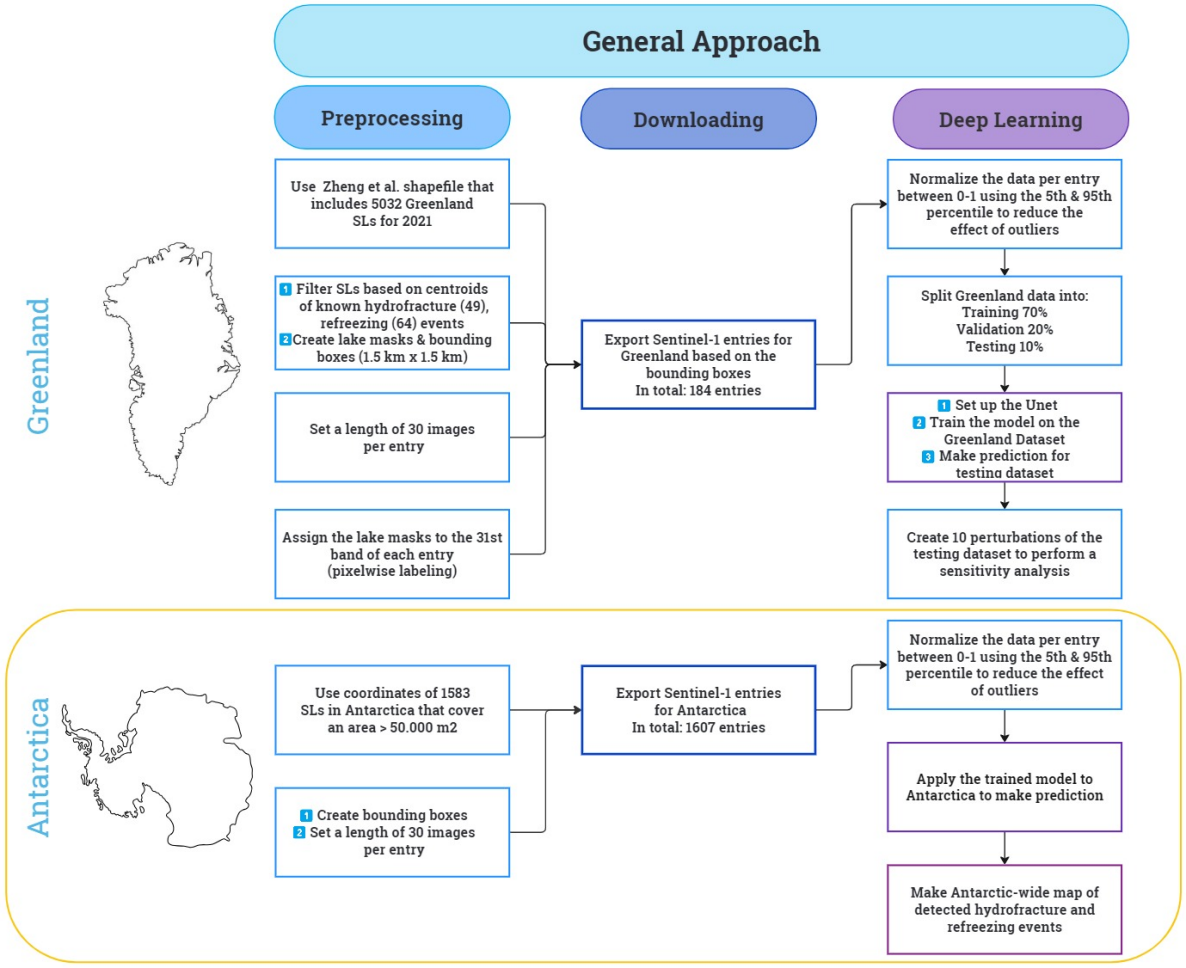


Figure 2.1: Schematic overview of the methodology of this thesis. The yellow box indicates the methodology for Antarctica.

2.2. Sentinel-1 Data

The Sentinel-1 mission consists of two polar-orbiting satellites performing C-band SAR imaging and were launched in April 2014 (Sentinel-1A) and April 2016 (Sentinel-1B). The Sentinel-1 satellites acquire data in single and dual polarization with a revisit time of six days at the equator. For Greenland and Antarctica, the revisit time can be shorter due to overlap of the satellite orbits in polar regions (**Torres et al. 2012**).

The C-band SAR operates in four different acquisition modes: Stripmap (SM), Interferometric Wide swath (IW), Extra-Wide swath (EW), and Wave mode (WV). The images used for this study were acquired in IW mode which is predominantly used over land areas. The IW products have a 250 km swath at 5 m by 20 m resolution (**Torres et al. 2012**). The Sentinel-1 mission distributes data at three levels of processing; Raw Level- 0, processed Level-1, and Level-2.

The GRD (Ground Range Detected Georefrenced) Level-1 product was used for this study. GRD data products come in 10, 25, or 40 meters resolution and for this study, the 10 m resolution was used. Furthermore, GRD data products have been detected, multi-looked, and projected to ground range using an Earth ellipsoid model (WGS84). GRD images include only intensity/amplitude information, while the phase information is removed (**Potin 2013**).

Sentinel-1 GRD data is available in logarithmic ('COPERNICUS/S1_GRD' - dB) and linear scale ('COPERNICUS/S1_GRD_FLOAT' - unitless). For this thesis, the choice of using the linear scale was made since it represents data in its raw form and therefore the values are directly proportional to the radar backscatter strength. Linear scale data is directly related to the physical properties of the observed surface. In this case, using linear scale data ensures that when normalizing the Sentinel-1 data, the values will still reflect the actual variations in backscatter. In addition, CNNs (e.g. U-Net) rely on the spatial and intensity patterns within the data to learn features and linear scale provides a true representation of these patterns which is key for effective feature extraction and accurate segmentation. Both the dataset created for Greenland and Antarctica include data from a variety of different orbits. For Greenland, orbits 74, 83, 90, and 127 were used, and for Antarctica orbits 3, 38, 69, 85, 134, and 169.

2.3. Locating lakes using optical data

For semantic segmentation tasks such as the one in this thesis, ground truth data is necessary. In order to acquire ground truth data and more specifically create a labeled dataset for the summer lakes of interest to train the U-Net, optical data was utilized. Since the model is trained only on Greenland data, the Antarctica dataset does not include ground truth data.

2.3.1. Regions of Interest

For this thesis, and the training of the model, the focus lies in Greenland due to its' important link to sea level rise as mentioned in Chapter 1 and the availability of several known refreezing and draining events.

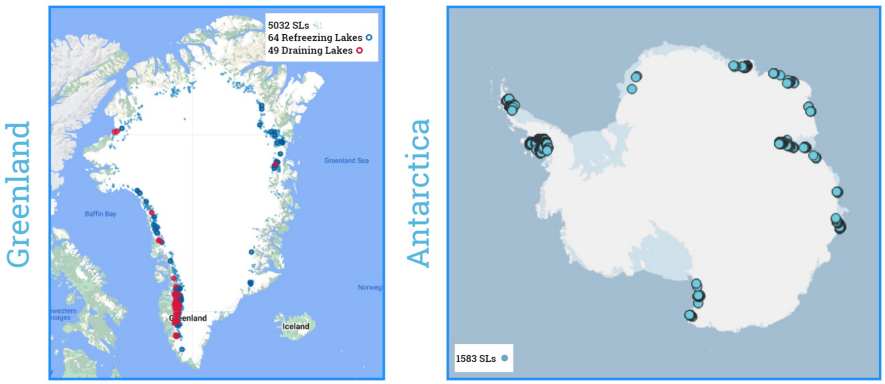


Figure 2.2: Depiction of the study areas. Greenland is on the left with 5032 summer supraglacial lakes for melt season 2021. Draining lakes (49) are shown with red circles while refreezing lakes (64) are shown with blue (created with GEE). Antarctica is on the right, depicting with blue circles the 1583 SLs for the austral summer of 2020-2021 (created using QGIS).

Those events were used to create the labeled dataset. More specifically, the centroid coordinates of 64 refreezing and 49 draining lakes were available along the GIS as seen in Figure 2.2. The behavior of those lakes was determined by visual inspection using high-resolution optical imagery from the Sentinel-2 and Landsat 8 missions **(de Roda Husman, unpublished)**.

2.3.2. Greenland lakes

In order to map out the lakes of interest to Sentinel-1 data a shapefile by (**Zheng et al. 2023**) that included 5032 summer supraglacial lakes (SLs) for the melt season of 2021 in Greenland was employed. To create this shapefile, the maximum summer lake extent was extracted from Sentinel-2 and Landsat 8 optical imagery. One limitation of this method is that there is one lake mask for the whole melt season, meaning that when creating our dataset we have the same lake mask for all images of one lake. The study also uses a threshold value of 0.5 for the modified Normalized Difference Water Index adapted for Ice (NDWice) (**Yang and Smith 2013**) to extract SLs without surrounding streams, which is a bit higher than thresholds used in order studies (e.g., 0.25) (**Dell et al. 2020**; **Yang, Smith, et al. 2021**). This might result in smaller lakes being excluded and an overall underestimation of the total number of lakes.

In Greenland, SLs occur in early summer, peak in July, and disappear as the temperature drops below the freezing point in autumn (**Zheng et al. 2023**; **Zheng 2023**). The shapefile was filtered by keeping only the lakes with known behavior (Subsection 2.3.1), meaning the 64 refreezing and 49 draining lakes. Furthermore, bounding boxes of 1.5 km x 1.5 km were created and placed over the centroids of the lakes to act as outlines for the exporting part.

The choice of 1.5 km by 1.5 km was made due to the fact that most lakes in Greenland do not cover an area bigger than 2.25 km^2 and are generally widespread from each other. The boxes are large enough to capture the lake and some of the surrounding area but small enough so that they do not capture neighboring lakes.

2.3.3. Antarctic lakes

For Antarctica, the approach was similar but the lake centroids were based on Sentinel-2 data and were acquired using the method of Moussavi (**Moussavi et al. 2020**), **(de Roda Husman, unpublished)** which is a threshold-based method for detecting lakes on the Antarctic ice shelves. Furthermore, only lakes $>$ 50.000 m^2 were selected and the centroid coordinates were kept. Lastly, the Sentinel-1 data were downloaded in a bounding box of 1.5 by 1.5 km around these centroids.

2.4. Downloading Sentinel-1 time series

The Sentinel-1 data is available through a variety of different platforms such as the Copernicus Open Access Hub and Google Earth Engine (GEE). An already existing script for exporting Sentinel-1 data via GEE was used and the following input parameters were defined. The first parameter to be defined was the date range of interest, stretching from the 1st of May to the 31st of October 2021 (extended arctic summer). Next, the region of interest which was defined by bounding boxes (1.5 km x 1.5 km) as mentioned in 2.3, corresponding to the locations of the refreezing and hydrofracturing known lakes. In addition, the Sentinel-1 GRD_FLOAT data product and HH polarization were selected. Furthermore, a length of 30 images per entry was defined after investigating how many images are available on average for each region within the date range of interest (melt season 2021). Within the script two Greenland-wide masks were created, one for all draining lakes and one for all refreezing lakes. Each time an entry was exported, the corresponding mask (draining or refreezing) was added as the last band (31st) to the entry, as seen in Figures 2.3 and 2.4.

For Antarctica the approach was similar but one main difference was the date range of interest which in the case of Antarctica stretches from the 1st of September to the 1st of March of 2021 in order to achieve the same 30-image length per entry. For this dataset labels were not created, but the trained model on Greenland will be applied to predict whether lakes are draining or refreezing. In addition, bounding boxes of 1.5 by 1.5 km around the centroids were used to download the Sentinel-1 data. In a follow-up study, it would be interesting to compare the predictions to other remote sensing products such as optical or altimetry data.

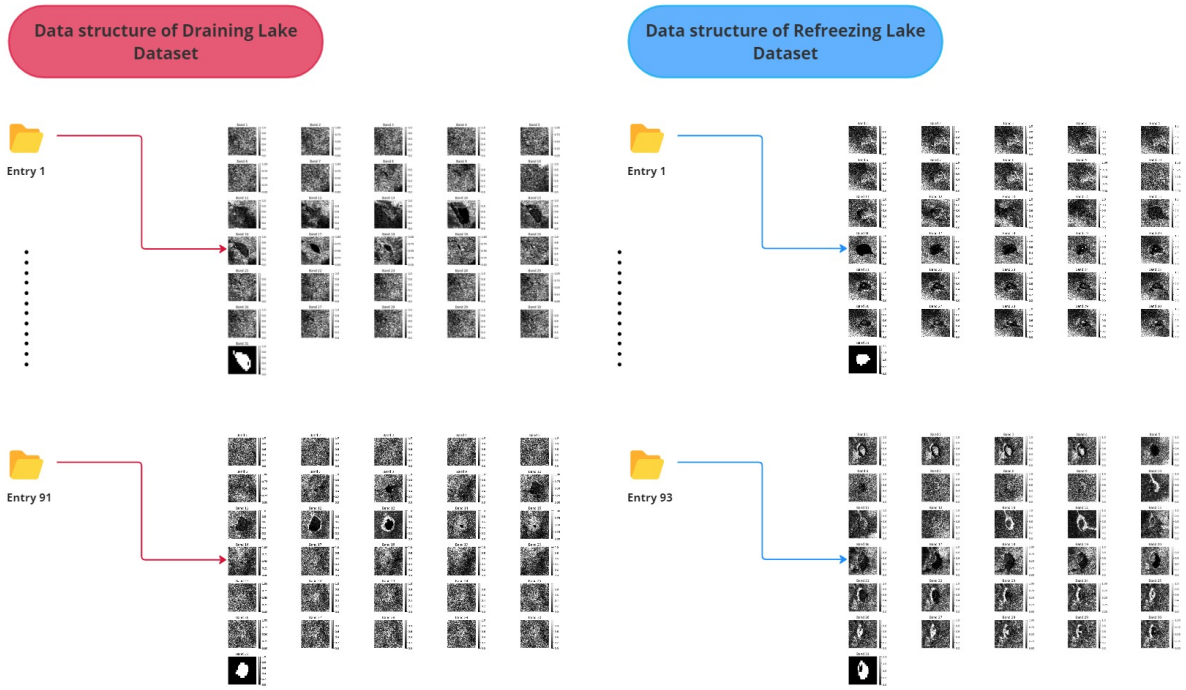


Figure 2.3: Diagram that shows the data structure of the draining and refreezing Greenland dataset.

For Greenland, both the draining and the refreezing datasets follow the same structure. The draining lake dataset consists of 91 entries (i.e., lakes), and the refreezing lake dataset of 93 entries. Each entry corresponds to a draining or refreezing location/lake and it consists of 31 bands. The first 30 bands correspond to 'GeoTIFF' images within the defined date range and are of dimension 64 by 64 pixels (i.e., 1.5 km by 1.5 km), while the 31st band corresponds to the label of the same dimension. The pixel size is approximately 23.5 m by 23.5 m. The label band for the draining dataset is a binary image with 0 indicating 'no lake' pixels and 1 indicating 'draining lake' pixels. For the refreezing dataset, the label band consists of 0 for 'no lake' pixels and 2 for 'refreezing lake' pixels as seen in Figure 2.4.

For Antarctica, there is no label present and there are 1607 entries of length 30 within the dataset, based on the 1583 lake coordinates that were available for the austral summer of 2020-2021. The mismatch between the number of lakes and the number of entries is due to the fact that some lakes fall within multiple orbits.

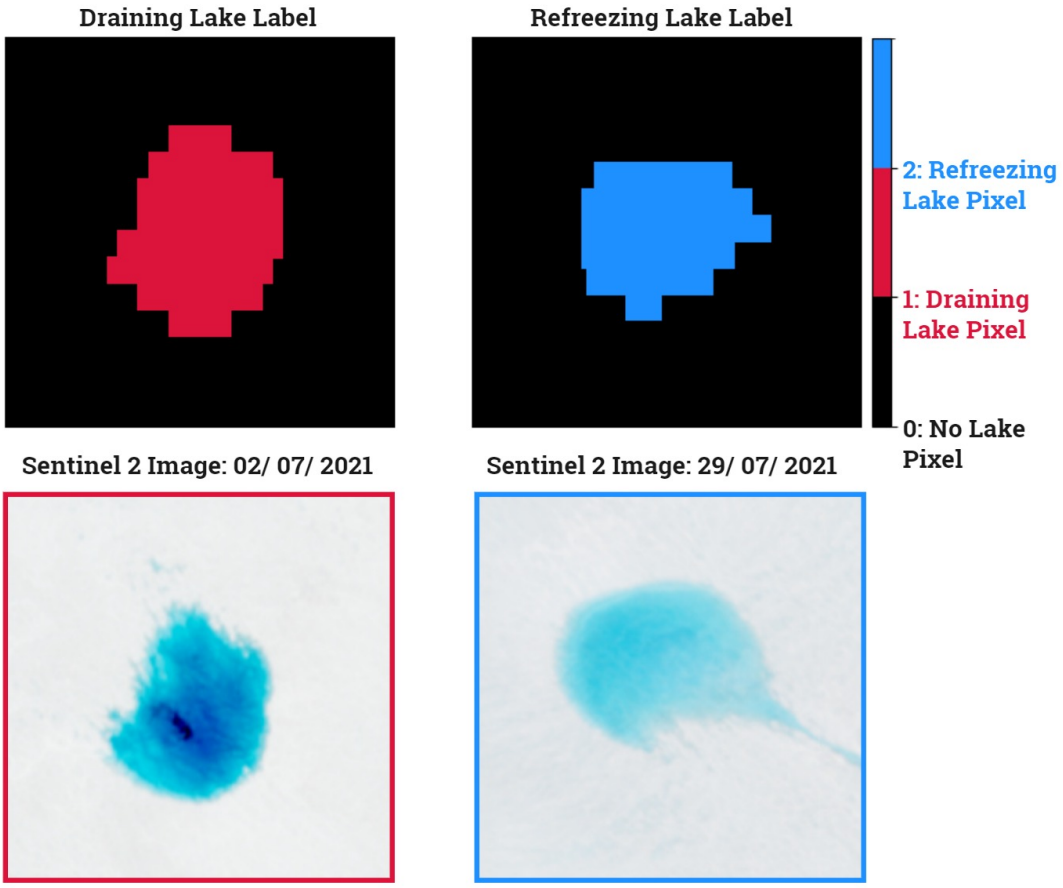


Figure 2.4: On the first row can be seen the 31st label band of example lake entry for a draining lake (left) and a refreezing lake (right). The label bands for draining lakes only include 0 (no lake pixel) and 1 (draining lake pixel) whereas the label bands for refreezing lakes only include 0 and 2 (refreezing lake pixel). On the second row can be seen Sentinel-2 images for the corresponding draining and refreezing lakes.

2.5. Normalization

For Greenland, the data is exported separately, meaning we have two datasets, one with images and labels from draining events and one from refreezing events. One key step in preprocessing data to apply a Deep Learning approach is normalization (Ali et al. 2014) which leads to input features having a similar scale that results in improved convergence speed. Normalization poses a key role in this study because data comes from different orbits (different viewing geometries), which influences the backscatter intensities. In this thesis, the data was normalized per entry between 0 and 1, using the 5th and 95th percentiles. The 5th percentile is the value below which 5% of the data falls and is used as a lower bound. Similarly, the 95th percentile is the value below which 95% of the data falls and is used as an upper bound while all other values are scaled linearly between those two bounds. This approach is less sensitive to outliers than methods that use the absolute minimum and maximum values. Figure 2.5 shows an example of a refreezing lake entry before and after normalization. It can be seen that this normalization approach increases the contrast of the images, making features within the image more distinct.

Figure 2.5: Example of refreezing lake entry before (up) and after (down) normalization. The blue rectangle indicates the entry before normalization.

2.6. Lake backscatter behavior

For some lakes, more than one orbit was available which provides an extra assurance if both orbits agree on the lakes' behavior as seen in Figures 2.6, 2.7. In both Figures, each plot corresponds to one example lake observed by two different orbits which means that for that lake there are two entries available which is not the case for all lakes. The mean value for each image within the entry is plotted with a solid line.

When looking at backscatter we need to take into account that it is influenced by several factors, such as surface roughness, density, dielectric properties, etc. In Figure 2.6, for the first 3-5 time steps/images (depending on the lake), the backscatter is relatively constant, followed by a sharp drop which indicates lake formation. This is due to the fact that water in C- band SAR imagery appears as low backscatter (**Brown and Johansson 2011**; **Miles et al. 2017**). In addition, when a lake forms (water fraction increases) the surface roughness decreases meaning that the surface becomes smoother. As the water fraction increases, the absorption increases as well leading to a smaller part of the radar signal traveling back, therefore backscatter decreases. Next, there are some fluctuations and a steady or sharp increase depending on the lake. Low backscatter due to the presence of a lake in one image changes to values similar to the surrounding ice once the lake has drained in the next images (**Miles et al. 2017**). This increase is due to the radar signal hitting the bottom of a drained lake and since little to no water is left, the biggest part of the signal travels back and does not get absorbed. Lastly, for lakes 1, 4, and 5 an intermediate stage is observed, where low backscatter goes to high backscatter values before changing to values that are similar to the surrounding ice.

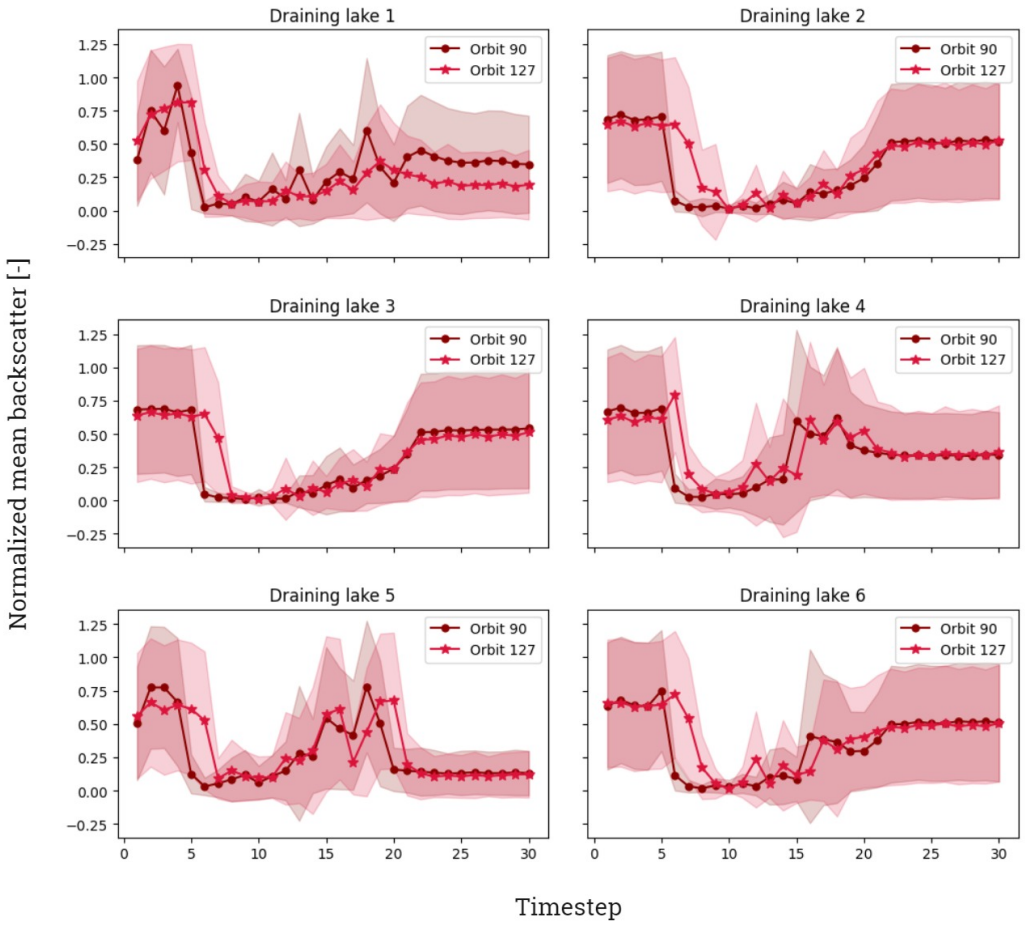


Figure 2.6: Normalized mean backscatter of six draining lakes as observed by two different orbits for the melt season (1st of May - 31st of October) of 2021 in Greenland, depicted with solid lines. The shaded regions indicate the 2σ confidence interval. The upper bound is +2σ, and the lower bound is -2σ.

In addition, when looking at the 2σ confidence interval, we can see that there are periods where the regions expand which means that the variability increases, specifically when the draining events take place (prominent in lakes 1, 4, 5, 6). After the draining takes place, the 2σ regions tend to become stable, indicating a reduction in variability. While both orbits generally agree in backscatter changes, the width of the 2σ regions can differ between orbits. For instance, in Draining lake 6, Orbit 127 shows a slightly wider 2σ interval compared to Orbit 90 during the initial time steps, indicating more variability in the measurements from Orbit 127 at that time. Overall, narrow intervals (e.g,. lake 1, 5) during most of the observed period suggest the presence of less noise. Whereas, wider intervals indicate more uncertainty and noise in the measurements. To conclude, summer lake drainage events have been observed to follow a pattern of low to high backscatter (**Miles et al. 2017**). For the time window chosen in this study, this is not exactly the case, since we observe the area before the lake is formed (very early in the melt season).

In Figure 2.7, for the first 5-10 time steps/images (depending on the lake) backscatter is relatively stable, followed by a sharp drop which indicates lake formation. As the water fraction increases, the absorption increases as well leading to a smaller part of the radar signal traveling back, therefore backscatter decreases. Next, there are some fluctuations and a steady increase. This happens as the lake's surface starts to freeze and scattering due to bubbles trapped in the ice increases. As long as the ice is not thick enough, C-band waves continue to reach the underlying water therefore backscatter steadily increases.

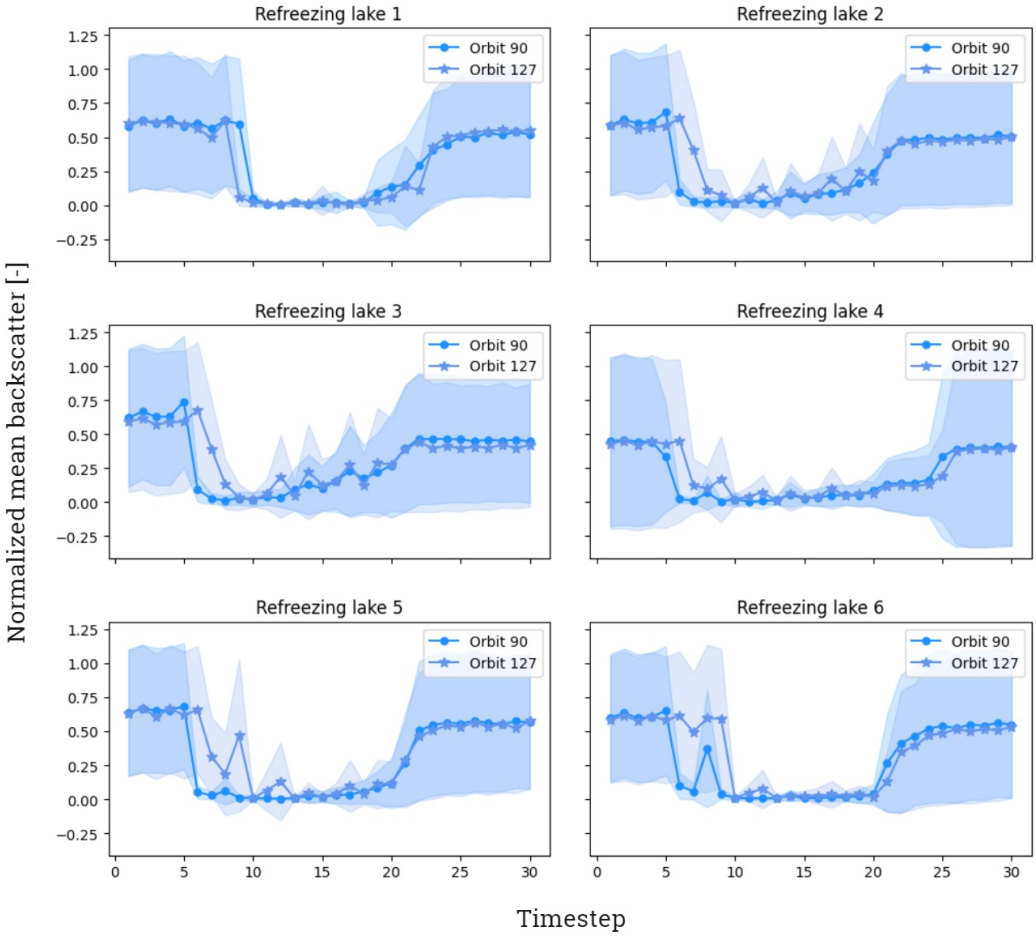


Figure 2.7: Normalized mean backscatter of six refreezing lakes as observed by two different orbits for the melt season (1st of May - 31st of October) of 2021 in Greenland depicted with solid lines. The shaded regions indicate the 2σ confidence interval. The upper bound is +2σ, and the lower bound is -2σ.

When the ice becomes thick enough, the radar signal hits the frozen surface and since no water is left, the biggest part of the signal travels back and does not get absorbed which leads to the increase in backscatter. However, as the penetration depth is limited to a few meters, the lakes might not have frozen fully, but only froze to the penetration depth. When looking at the 2σ confidence interval, we can see that the regions vary among different lakes. For example, lake 1 shows a relatively narrow interval for both orbits, indicating lower variability while lake 3 has a wider 2σ region, suggesting higher variability and greater uncertainty. Same as for the draining lakes, during the first time steps, the interval is wider and becomes more narrow when the lakes form and then increases again when the lakes refreeze.

When comparing Figure 2.6 and 2.7 some differences can be detected. When a lake refreezes there are fewer fluctuations (fewer peaks) and an overall more flat profile after the lake is formed and until it refreezes, with an overall steady increase. On the other hand, for draining lakes, there are more fluctuations and an overall higher and sharper increase in backscatter, especially when the lake starts draining. For both types of lakes, the normalized mean backscatter starts at approximately 0.50 (first time step). However, after a lake has drained or refrozen the backscatter does not go back to its starting value (0.50) and ends up being around 0.40 or lower from some draining lakes. This could be due to some water remaining after draining, therefore part of the signal is still being absorbed. To conclude, on the one hand, the backscatter signal of draining and refreezing lakes is sufficiently different with draining having a higher and not steady increase in backscatter before becoming steady. On the other hand, some draining lakes (Draining Lake 2, Draining Lake 3) showcase a similar behavior to refreezing lakes and this is one of the reasons deep learning is employed, to learn complex patterns that might not be so visible.

2.7. Developing a Deep Learning Model

2.7.1. Semantic Segmentation

Semantic segmentation is a computer vision task that assigns a category label to each pixel of an image, unlike image classification which assigns one or more category labels to the whole image. Due to this difference, semantic segmentation is considered more challenging than image classification because it has to understand the image at a pixel level and bridge the gap between low-level and high-level features. Generally, by feeding enough images and their corresponding pixel-wise labeling maps as training data, a deep learning network can be trained to learn the mapping between a label and its diversified visual representations (**Hao, Zhou, and Guo 2020**). This technique was chosen because even though the main focus of this study was to predict whether a lake is draining or refreezing we also wanted to see if the model is able to detect the lake's shape and position. Therefore, we needed the model to perform a pixel-wise prediction whereas with other techniques such as image classification the model only outputs one image-wise label (draining or refreezing). Additionally, an image-wise classification step was added after the model prediction which is explained in detail in Section 2.8.1.

2.7.2. Training, Validation, and Testing sets

Training, validation, and testing data sets were created using a stratified split method, which means that each of the three sets has approximately the same number of draining and refreezing entries (balanced split). More specifically, the training set included 64 draining and 65 refreezing entries. Lastly, the validation set included 18 entries of each while the testing set included nine draining and ten refreezing entries. The workflow is explained in detail below and can be seen in Figure 2.7.2.

- Training set: was used to train and make the model learn the hidden patterns within the data. In each epoch, it is fed to the neural network. It includes the biggest part of the dataset and in this thesis, 70% of the total dataset was assigned to training.
- Validation set: separate from the training set and was used to validate the model performance during training. It provides information that helps with hyperparameter tuning and is an indicator of whether training is moving in the right direction or not. The main idea behind splitting the dataset into training, and validation is to prevent the model from overfitting which means that the model becomes very good at learning the patterns in the training set but can not generalize and make accurate predictions on data that it has not been trained on. 20 % of the data was allocated to

validation.

- Test set: a separate set of data that was used to test the model after training was completed to assess model performance and confirm the results. 10 % of the data was allocated to testing.

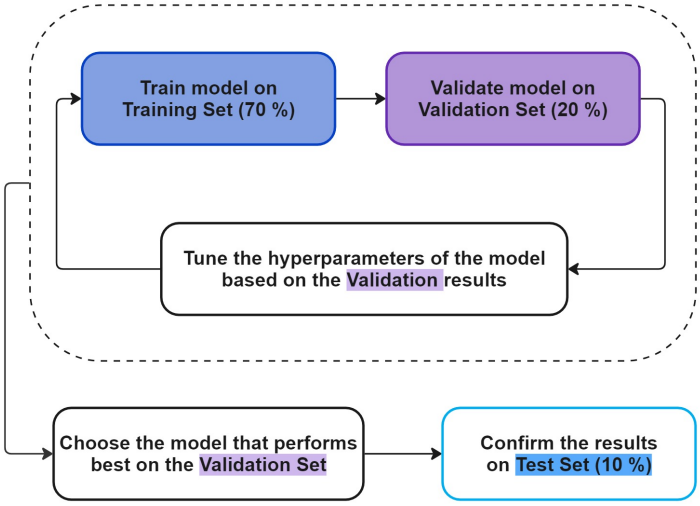


Figure 2.8: Workflow for training, validating, and testing the U-Net.

2.7.3. Sensitivity Analysis

In order to understand and evaluate why and when the model works, ten different instances of the testing dataset were created by exporting data again or by processing the existing testing dataset as listed below. An example entry of the original testing dataset can be seen in Figure 2.9.a in order to demonstrate how the dataset changes in each perturbation. The perturbations are grouped into two categories: temporal and spatial. Temporal indicates that we are investigating how removing the evolution of the lakes over time or how a specific time step affects model performance. On the other hand, spatial indicates that we investigate how the spatial patterns of the lakes affect model performance.

Temporal Perturbations:

- **Shuffled image sequence testing dataset:** 'np.random.shuffle' was used to randomly shuffle the first 30 bands of each entry but not the label band of the original testing dataset as can be seen in Figure 2.9.b.
- **Repeated time step testing dataset:** takes an entry of the original test dataset and generates new variations where each new entry consists of one of the original image bands repeated 30 times followed by the ground truth band. This results in multiple new entries, each containing the same image repeated 30 times and ending with the ground truth as can be seen in Figure 2.9.c.

Spatial Perturbations:

- **Decreased intensity testing dataset:** A value of 0.2 was subtracted from all the images of all entries but not from the labels of the original testing dataset. The result was then passed through 'np.clip'. An example can be seen in Figure 2.10.d.
- **Increased intensity testing dataset:** All the images of all entries but not the labels of the original testing dataset were multiplied by a factor of 1.5 which increased the intensity by 50%. The result is passed through 'np.clip' which ensures pixel values remain within the normalized range [0, 1]. An example entry can be seen in Figure 2.10.e.
- **Zoomed in testing dataset:** Data was exported again for the 19 entries of the testing dataset by using bounding boxes of 1 km x 1 km making the final images and labels zoomed in as seen in Figure 2.10.f.
- **Zoomed out testing dataset:** Data was exported again for the 19 entries of the testing dataset by using bounding boxes of 2 km x 2 km making the final images and labels zoomed out as seen

in Figure 2.10.g.

- **Shifted to the left testing dataset:** Two different approaches were implemented to shift the images and labels 10 pixels to the left as seen in Figure 2.10.h and i. The first approach (h) shifts the image 10 pixels to the left and it wraps the shifted pixels around the right side of the image. This approach has the limitation of lake pixels appearing on the right side of the image. The second approach (i) shifts the image 10 pixels to the left and fills the newly created space on the right with zeros.
- **Flipped testing dataset:** All images and labels of the original testing dataset were flipped upside down using 'np.flipud' which is a function that reverses the order of elements along the vertical axis. An example of a resulting entry can be seen in Figure 2.10.j.
- **Randomly shuffled pixels testing dataset:** The pixels of each band/image along with the ground truth of each entry were shuffled randomly as seen in Figure 2.10.k.

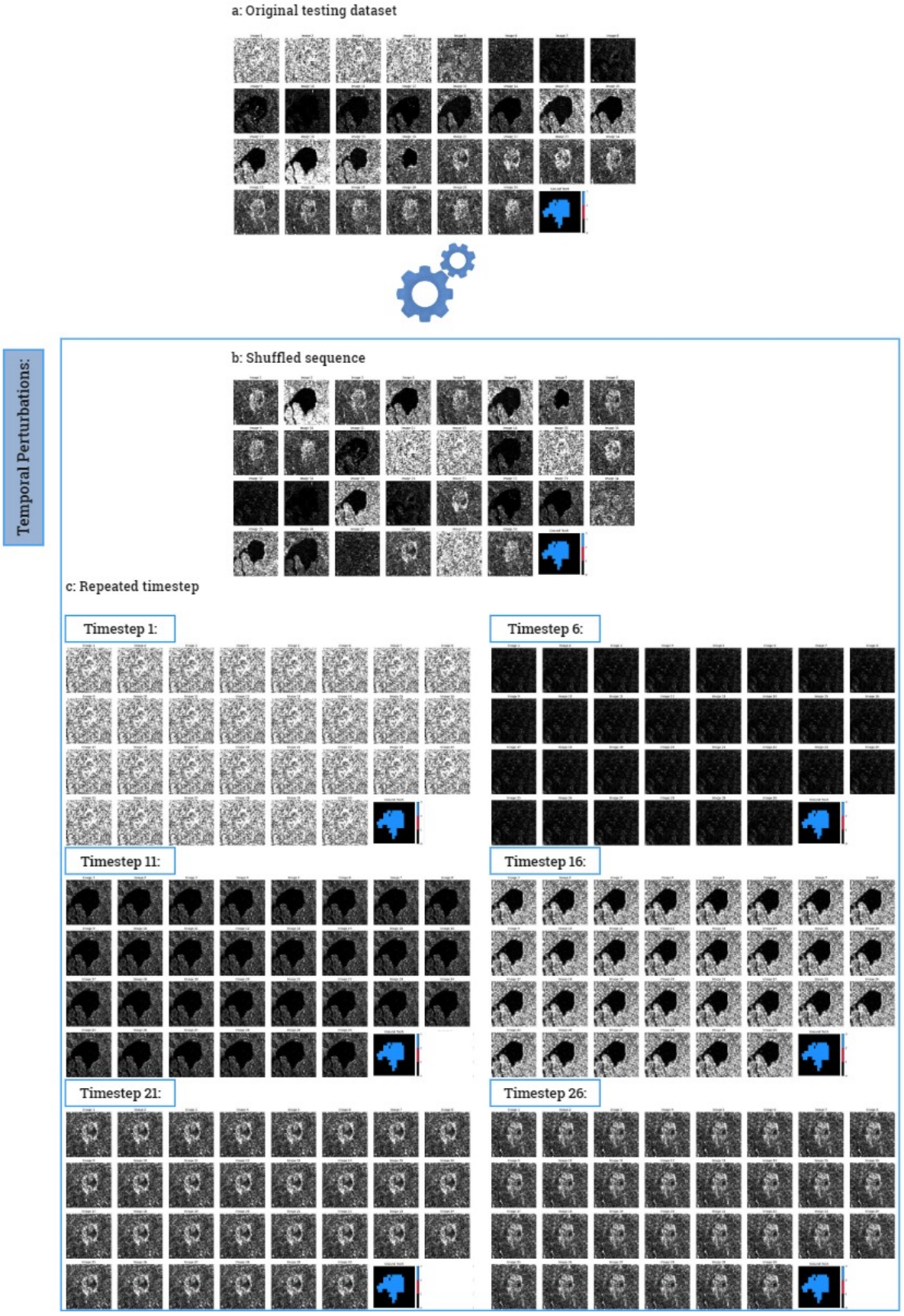


Figure 2.9: Representation of an example entry refreezing lake from the original test dataset (a) along with the two different temporal perturbations (c, b) created from this dataset. The Repeated time step dataset (c) example is plotted only for six time steps however there are 30 different entries created for each original entry.

Figure 2.10: Representation of an example entry refreezing lake from the original test dataset (a) along with the eight different spatial perturbations created from this dataset.

2.7.4. U-Net architecture

An Attention U-Net was utilized for this thesis as implemented by (**de Roda Husman et al. 2024**), which is an extension of a traditional U-Net (**Ronneberger, Fischer, and Brox 2015**) that uses attention gates. Attention gates help highlight relevant features while suppressing irrelevant regions within an input image. The U-Net consists of two parts: the contracting path for encoding (encoder) and the expansive path for decoding (decoder). Within the contracting path, the input is downsampled by convolutional layers and max pooling operations resulting in the extraction of high-level features which are important for understanding the overall scene and accurately classifying each pixel. More specifically, high-level features in this task capture information about the shape and size of the lakes and ignore smaller-scale details (**de Roda Husman et al. 2024**). Within the expansive path, downsampled features coming from the encoder are reconstructed by transposed convolutions to increase the image size. Those are then combined with corresponding features from the contracting path using skip connections which connect

corresponding encoder and decoder layers. This process preserves low-level features, such as the edges and boundaries of the lakes. By combining high and low-level features through the U-Net, the model is able to segment and differentiate between draining and refreezing lakes.

Common algorithms were used for activation, loss, and optimization. More specifically, the Rectified Linear Unit (ReLU) activation function was used within the hidden layers which allows only for positive values to propagate through the network. Furthermore, the softmax activation function was employed in the final layer to output a vector containing the probabilities of each possible outcome for the prediction step. In this case, for each pixel a vector was outputed that included three probabilities since three classes were present in the dataset (0: no lake pixel, 1: draining lake pixel, 2: refreezing lake pixel). Next, 'np.argmax' was applied to the output (prediction) in order to return the maximum probability for each pixel. Sparse categorical cross-entropy was used in terms of the loss function, which is suitable for multi-class classification tasks where the labels are represented as integers (int32 in our case). In terms of optimization, the Adam optimization algorithm was utilized. The implementation of the code for training the U-Net segmentation model was done using TensorFlow on Google Colab, which is a cloud-based platform.

2.8. Evaluating the Deep Learning Model

2.8.1. Evaluation Metrics

As mentioned in Section 2.7.2, in order to assess the performance of the U-Net the test set was utilized. Various performance metrics were calculated such as accuracy, precision, recall, F1 score as well as the confusion matrix and the Structural Similarity Index (SSIM).

Performance Metrics

The metrics were calculated in two different ways: pixel-wise and image-wise. For the calculation of pixel-wise metrics, 'sklearn.metrics' was utilized which assesses prediction error. Metrics are calculated for each class (no lake, draining lake, refreezing lake) and then the average (weighted by the number of true instances for each class) is calculated, which accounts for class imbalance.

The confusion matrix for our multi-class classification problem with three classes (0: no lake, 1: draining lake, 2: refreezing lake) can be seen in Figure 2.11.

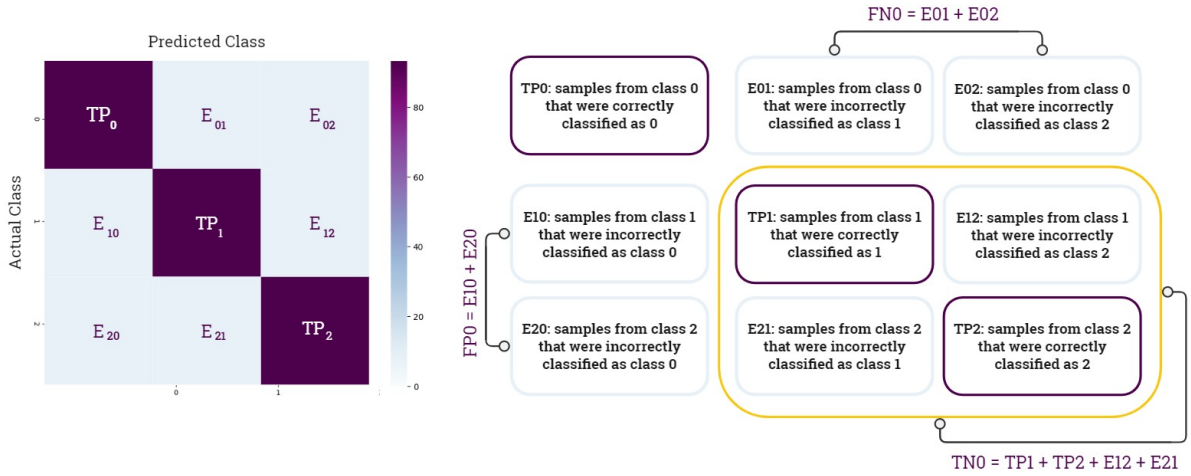


Figure 2.11: Example confusion matrix for a multiclass classification task (three classes) along with the explanation of TP, FP, FN, TN.

As shown, TP_0 is the number of true positive samples in class 0, i.e., the number of samples that are correctly classified from class 0, and $E01$ is the samples from class 0 that were incorrectly classified as class 1, i.e., misclassified samples. Thus, the false negatives in the 0 class ($FN0$) are the sum of $E01$ and $E02$ ($FN0 = E01 + E02$) which indicates the sum of all class 0 samples that were incorrectly classified as class 1 or 2.

Furthermore, the FP0 is the sum of E10 and E20 (FP0 = E10 + E20) which indicates the sum of all class 1 and class 2 samples that were incorrectly classified as class 0. Lastly, the TN0 is the sum of TP1, TP2, E12, and E21 (TN0 = TP1 + TP2 + E12 + E21) which indicates the samples that are correctly identified as not belonging to class 0. Simply, the FN of any class can be calculated by adding the errors in that row except for the TP value. In a similar manner the FP of any class can be calculated by adding the errors in that column except for the TP value. Lastly, to calculate the TN of any class simply sum all the values of all columns and rows except for the values of the class that we are calculating the value for (**Tharwat 2020**).

For the image-wise metrics, a custom function was created that classifies the model predictions based on the majority of the pixels present (draining or refreezing). First, if the prediction only included 1 (0 is always present - background) it was classified as draining, whereas if it included only 2 (and 0) it was classified as refreezing. In the case that the prediction included both 1 and 2 (0 is always present - background), the sum of each class was calculated and the class that is dominant was chosen. To calculate the evaluation metrics image-wise we needed to know the four quantities that are shown in Figure 2.12. The predictions were assigned to TP, FP, TN, and FN according to certain criteria. For the image-wise metrics calculation, the problem becomes binary since we care about the presence of 1 and 2. When the ground truth label (Draining) matched the model prediction label it was assigned to TP. The same applied when the ground truth label and the prediction label were ”Refreezing”, in which case it was assigned to TN. In the case that the model incorrectly predicted a ”Draining” lake as ”Refreezing it was assigned to FP and when a ”Refreezing” label was predicted as ”Draining” it was assigned to FN. The five metrics that were calculated are listed below:

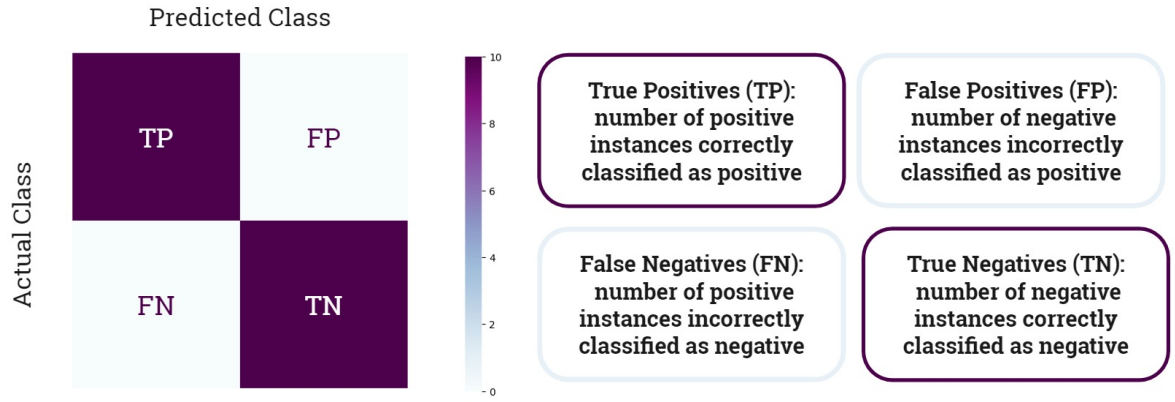


Figure 2.12: Example confusion matrix for a binary classification task along with the explanation of TP, FP, FN, TN.

- **Accuracy** is the ratio of the total correct predictions to all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Recall** is also known as sensitivity and it measures the proportion of true positive predictions among the total actual positive instances.

$$Recall = \frac{TP}{TP + FN}$$

- **Precision** assesses the proportion of true positive predictions over the total predicted positive instances.

$$Precision = \frac{TP}{TP + FP}$$

- **F1 score** is the harmonic mean of precision and recall and provides a balance between the two.

$$F1Score = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

- **Structural Similarity Index Measure (SSIM):**

  After the prediction, the SSIM was also calculated for each entry, and also the average SSIM for the test dataset and its eight different perturbations using 'tf.image.ssim'. The SSIM quantifies the difference in structure, luminance, and contrast between two images. In our case, this is focused on structure since the labels (ground truth) as well as the predictions only include values of 0, 1, and 2. 'tf.image.ssim' returns a score from -1 to 1 where 1 indicates perfect similarity (**Zhao et al. 2019**).

  The SSIM between two images x and y is defined using the following three components:

  1. Luminance Comparison:
  $$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

  2. Contrast Comparison:
  $$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

  3. Structure Comparison:
  $$s(x,y) = \frac{\sigma_{xy} + C_2/2}{\sigma_x\sigma_y + C_2/2}$$

  The overall SSIM index is then given by:

  $$SSIM(x,y) = l(x,y) \cdot c(x,y) \cdot s(x,y)$$

  Expressed in a single formula, the SSIM index is:

  $$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

  where:

  μ_x, μ_y is the mean of image x, y, respectively,
  σ_x^2, σ_y^2 is the variance of image x, y, respectively,
  σ_{xy} is the covariance of images x and y,
  $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$ are two constants,
  L is the dynamic range of the pixel values,
  K_1 and K_2 are small constants (e.g., $K_1 = 0.01$ and $K_2 = 0.03$).

  The function used computes this index on small patches of the images and then averages the results to produce an overall SSIM.

Each of these metrics offers unique insights into the model performance, helping to understand its strengths and weaknesses in different aspects of the prediction.

2.8.2. Hyperparameter Tuning

Hyperparameter tuning was performed on the validation dataset to optimize the performance of the U-Net. Different experiments were conducted to try all the possible combinations of hyperparameters which included batch size (2, 8, 16, 32, 64), learning rate (0.00001, 0.0001, 0.001, 0.01), number of epochs (10 - 100), and a constant dropout rate of 0.5 was applied to the 4th and 5th convolution layers (**de Roda Husman et al. 2024**). An example of model runs that had the same learning rate and number of epochs but different batch sizes can be seen in Figure 2.13. It can be seen that when increasing batch size the model performs worse, with higher validation loss.

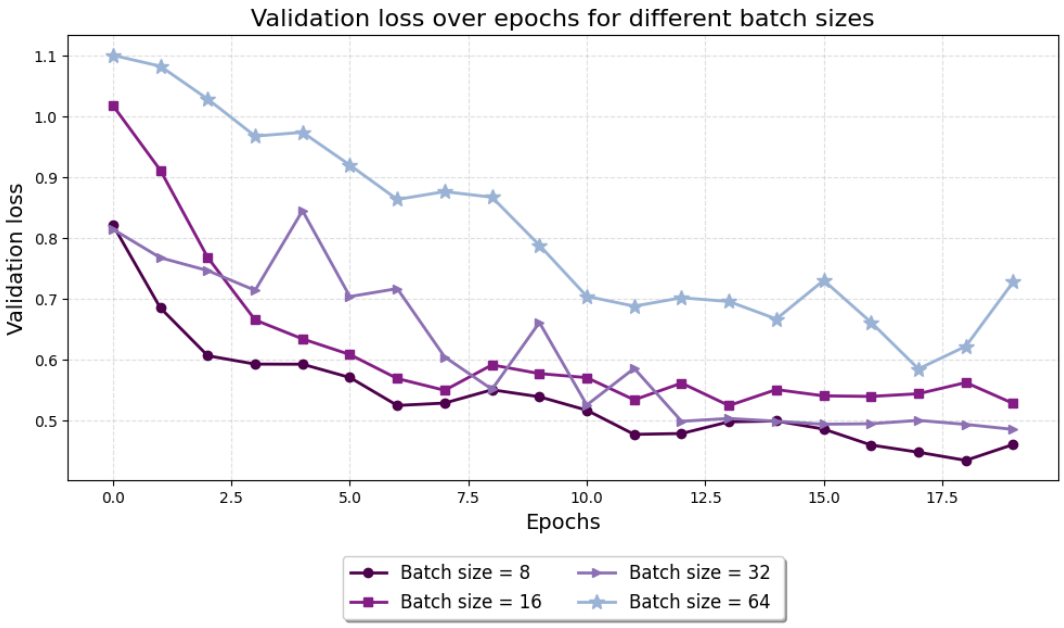


Figure 2.13: Validation loss for four different model runs (model trained for 20 epochs with learning rate 0.0001). The model runs had different batch sizes, starting from 8 to 64. The optimal validation loss (lowest) is achieved when having a batch size = 8. The highest validation loss is achieved for batch size = 64.

The model was trained for a maximum of 100 epochs with each epoch taking around 1 minute. Early stopping was implemented to cease training if validation loss did not substantially improve for six consecutive epochs. For each hyperparameter combination, the model was trained on the training dataset and its performance was evaluated on the validation dataset to determine the optimal model configuration. The model that achieved the highest accuracy and the lowest training and validation loss was trained for 60 epochs with a learning rate of 0.0001 and batch size 8 and can be seen in Figure 2.14

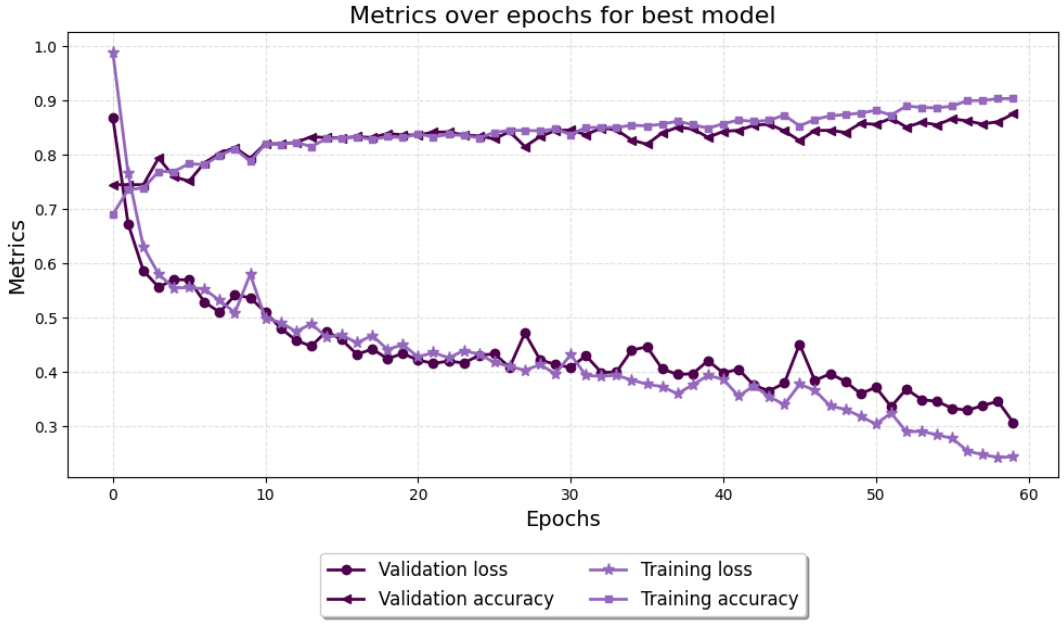


Figure 2.14: Metrics over training epochs for the best model.

2.9. Applying the Deep Learning Model to Antarctica Data

After the model was successfully trained, evaluated, and tested on the Greenland dataset and the sensitivity analysis was conducted, the model was applied to the Antarctic dataset. In order to assign the predictions to a specific class, the same approach was used as in Section 2.8.1 which assigns a ”Draining” or ”Refreezing” label to the majority class present. The prediction of the model was then used to create an Antarctic-wide map depicting where draining and refreezing events occurred.

3 Results

In the following chapter we assess the performance of the model on the testing dataset, check the sensitivity of the model on perturbing the testing dataset and lastly apply the trained model on the Antarctic Ice Sheet.

3.1. Performance of the model

The performance of the model on the testing dataset (a) was evaluated both pixel-wise and image-wise as mentioned in Section 2.8.1. For the calculation of the pixel-wise model metrics 'sklearn.metrics' was utilized which assesses prediction error and can be seen on the right side of 3.1. The model achieved an accuracy of 90.6 %, precision of 90.9 %, recall of 90.6 %, F1-score of 90.0 % and an average SSIM of 0.56.

Figure 3.1: Image-wise (left) and pixel-wise (right) model evaluation metrics for the testing dataset along with the corresponding confusion matrix.

In the confusion matrix it can be seen that for class 0 (no lake pixels) and class 1 (draining lake pixels), a high percentage of the pixels were classified correctly (97.5 % and 84.4 %, respectively). However, the percentage of correctly classified pixels for class 2 (refreezing lake pixels) is lower (55.4 %). This might be due to the class imbalance present in the training dataset (0: 388735, 1: 78419, 2: 61230). Therefore, the model might be biased towards the majority class (class 0), because it is seen more frequently during training. This can lead to poor performance in the minority classes (classes 1, 2).

The image-wise model metrics were calculated by a custom function that assigns the predicted images to TP, FP, FN, and TN and can be seen in Figure 3.1 on the left. The model correctly classified all refreezing (ten) and draining (nine) lakes as seen in the diagonal of the confusion matrix. This results in an accuracy, precision, recall, and F1-score of 100% with nine TP, zero FP, zero FN, and ten TN.

Given that the scope of this thesis is to create a model that can distinguish between draining (class 1) and refreezing (class 2) lakes, class 0 (no lake pixels) is not as important (image-wise) since when looking at the image-wise metrics the model manages to correctly classify all images.

In addition, the average SSIM was calculated which takes values from [-1, 1] for non-negative pixels. An SSIM value between 0 (no similarity) and 1 (perfect similarity) indicates high similarity between the images being compared meaning the ground truth and the prediction. Lastly, negative values are rarely encountered and indicate dissimilarity. The SSIM between the ground truth and the prediction for the test dataset is 0.56 indicating an overall high similarity.

3.2. Sensitivity of the model

As mentioned in Section 2.7.3, in order to understand and evaluate why and when the model works, ten different perturbations of the testing dataset were created and were split into temporal and spatial. For all those perturbed datasets, a prediction was made using the already trained model.

3.2.1. Classification Metrics

The metrics were calculated in the same way as for the test dataset in order to evaluate the performance of the model. The image-wise metrics along with the corresponding confusion matrices for the test dataset and its ten perturbations are shown in 3.2. The confusion matrices for some datasets ([a, f, g] and [d, e, h, i, j]) look the same because the model predicted the same number of TP, FP, FN, and TN therefore, they are plotted once. Furthermore, the pixel-wise metrics along with the corresponding confusion matrices for the perturbed datasets are shown in Figure 3.3 and 3.4.

Image-wise:

As can be seen in Figure 3.2, the best performance when looking at the image-wise confusion matrices, corresponds to the original test dataset (a), 'Zoomed in' (f), and 'Zoomed out' (g) datasets for which the model correctly classified all refreezing (ten) and draining (nine) lakes as seen in the diagonal of the first confusion matrix. This results in an accuracy, precision, recall, and F1-score of 100% with nine TP, zero FP, zero FN, and ten TN.

The worst image-wise performance corresponds to the two temporal perturbations. More specifically, the 'Shuffled sequence' (b) dataset and the 'Repeated time step' (c) dataset led to model performance of less than 67 % across all metrics.

For the 'Shuffled sequence' (b) dataset the model only managed to correctly predict five (TN) out of the ten refreezing lakes, and six (TP) out of the nine draining lakes. The remaining five (FN) lakes were classified incorrectly as draining, and the three (FP) were classified incorrectly as refreezing. For the 'Repeated time step' (c) dataset, there were in total 570 (19 original entries times 30 images per entry) entries/ predictions. In the confusion matrix, the total number of entries corresponds to 521, meaning that 49 entries are not classified as TP, FP, TN, or FN because the model did not predict draining or refreezing.

For the 'Randomly shuffled pixels' (k) dataset, the model correctly classified all nine draining lakes (TP) but only four out of the ten refreezing lakes. Lastly, one refreezing lake was not classified as either leading to a total number of 18 entries in the confusion matrix. Generally, the presence of distinct spatial patterns helps the model learn meaningful features, and as can be seen when those spatial

patterns are removed the model performs poorly. For the remaining datasets ('Flipped' (j), 'Decreased intensity' (d), 'Increased intensity' (e), 'Shifted left variation 1' (h), 'Shifted left variation 2 ' (i)), the model performed very well by correctly predicting the nine draining (TP) and refreezing (TN) lakes and only misclassified one (FN) refreezing lake as draining. This results in an accuracy, F1-score of 94.7%, precision of 100%, and recall of 90%.

To conclude, when examining the model's sensitivity to image-wise predictions, for eight out of the eleven datasets the model performs extremely well, whereas for the 'Shuffled Sequence' (b), 'Repeated time step' (c), and the 'Randomly shuffled pixels' (k) datasets it performs poorly. This result supports the behavior shown in the examples in Section 3.2.2, where we saw that the prediction of the shape and class for this dataset was not adequate, even though the average SSMI is not the lowest.

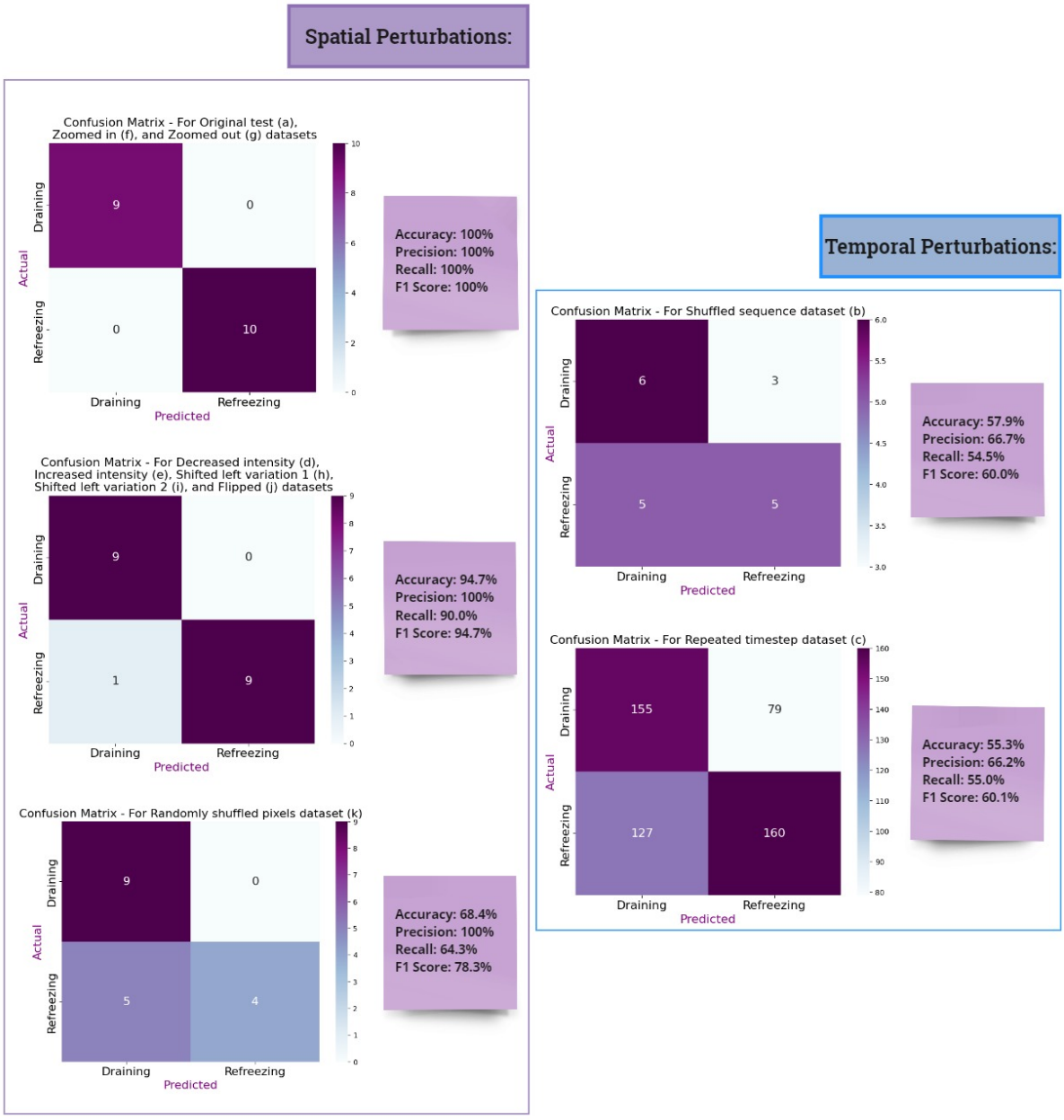


Figure 3.2: Image-wise model evaluation metrics for all nine datasets along with their corresponding confusion matrix.

Pixel-wise:
As can be seen in Figures 3.3, 3.4, the best performance corresponds to the 'Zoomed out' (g) dataset. It has the overall highest accuracy (92.9%), precision (93.6%), recall (92.9%), and F1-score (92.9%). Furthermore, in the confusion matrix, it can be seen that for class 0 (no lake pixel) and 1 (draining lake pixel), approximately 95% of the pixels were classified correctly, while for class 2 (refreezing lake pixel) only 62.8% which is due to the class imbalance present in the dataset as mentioned in 3.1. Even though there are datasets that might have a higher percentage of correctly classified pixels for one specific class, the 'Zoomed out' (g) dataset is the one that has the highest percentages across all classes. In addition, when looking at class 2 (refreezing lake pixel), it is evident that across all datasets, it is the one class that the model finds difficult to classify correctly. The prediction for the 'Decreased intensity' (d) dataset has the highest percentage of correctly classified pixels for class 2 with 68.4%, as seen in Figure 3.4.

The prediction for the 'Shuffled sequence' (b) and the 'Zoomed in' (f) datasets shows the overall worst performance when looking at the problem pixel-wise with an accuracy of 77.3%, precision of 72.4%, recall of 77.3%, and F1-score of 73.2%. The pixel-wise evaluation of the model aligns with the image-wise evaluation of the model. In both cases, the 'Shuffled sequence' (b) dataset performs the worst. In the pixel-wise approach, only 25.7% of the pixels that belong to class 1 and 13.7% of the pixels that belong to class 2 were correctly classified. Given that the scope of this thesis is to create a model that can distinguish between draining (class 1) and refreezing (class 2) lakes, class 0 (no lake pixels) is not as important (96.8% of the pixels that belong to class 0 were classified correctly). For the remaining datasets, the model performance lies somewhere between the 'Zoomed out' (g) dataset (best overall performance) and the 'Shuffled sequence' (b) dataset (worst overall performance).

Lastly, Figures 3.3, 3.4 also show the average SSIM for all test dataset perturbations. When looking at the average SSIM values it can be concluded that the perturbed dataset that makes the model perform the worst when looking at the SSIM is the 'Zoomed in' (f) dataset with a value of 0.35. This might be happening due to the fact that when zooming in, images include more detail in the shape and outline of the lakes as well as lower resolution. The amount of detail might complicate the process. On the other hand, the model seems to perform the best for the 'Zoomed out' (g) dataset which is the exact opposite of the 'Zoomed in' (f) dataset. When the images are zoomed out, there is less detail captured when it comes to the shape and outline of the lake, therefore it is easier for the model. The difference in average SSIM for the original dataset and the 'Zoomed out' (g) which is the dataset with the highest average SSIM is 0.08 which indicates that creating slightly bigger bounding boxes helps produce slightly better results and model performance. For the remaining datasets, the average SSMI value is higher than 0.40 and lower than 0.53 indicating an overall good similarity between prediction and ground truth.

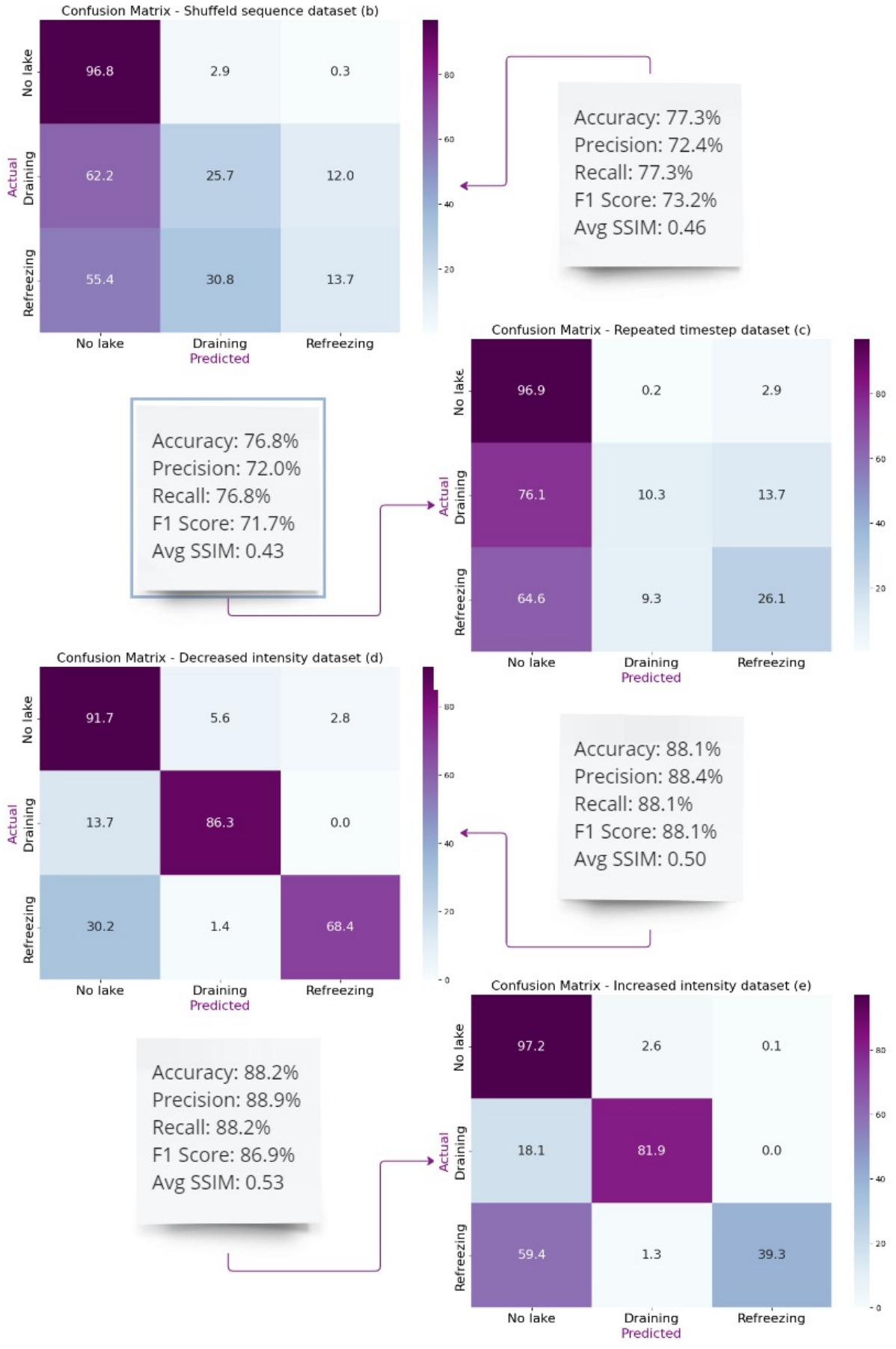


Figure 3.3: Pixel-wise model evaluation metrics for datasets b, c, d, and e along with their corresponding confusion matrix. The grey box around the metrics for the 'Repeated time step' (c) dataset indicates that the model performed the worst for this dataset.

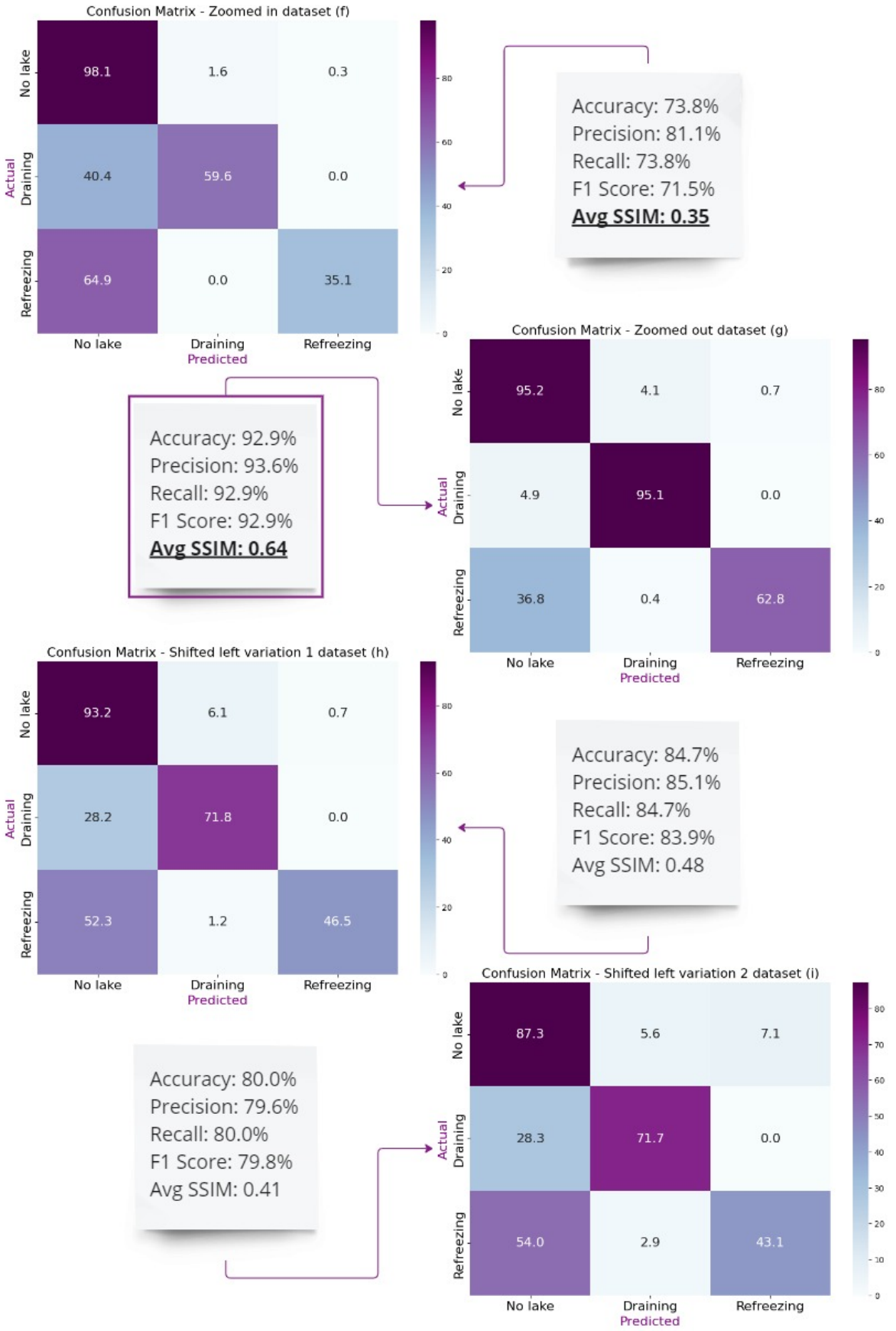


Figure 3.4: Pixel-wise model evaluation metrics for datasets f, g, h, and i along with their corresponding confusion matrix. The purple box around the metrics for the 'Zoomed out' (g) dataset indicates that the model performed the best for this dataset. The highest (Zoomed out dataset) and lowest (Zoomed in dataset) average SSIM is underlined.

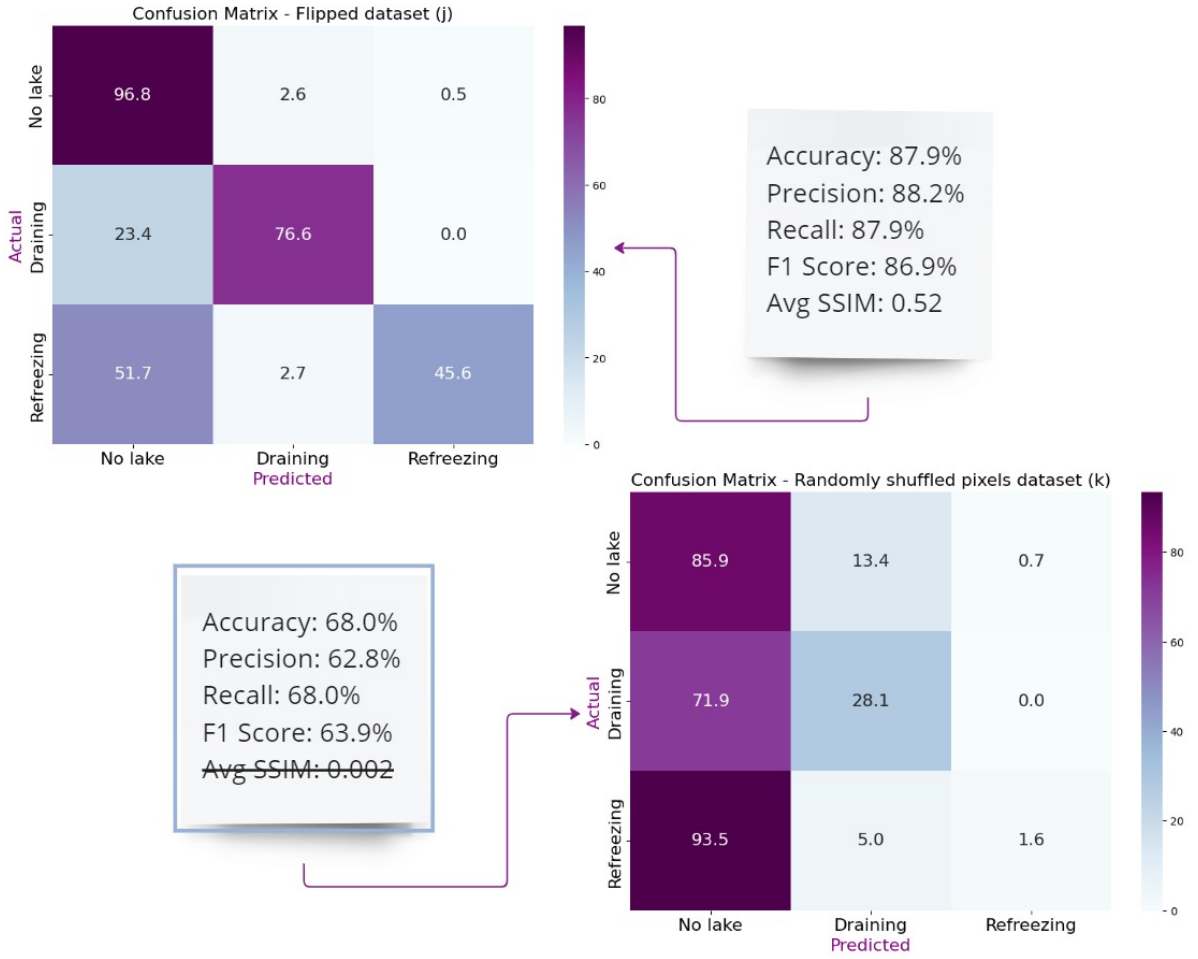


Figure 3.5: Pixel-wise model evaluation metrics for datasets j and k along with their corresponding confusion matrix. The grey box around the metrics for the 'Randomly shuffled pixels' (k) dataset indicates that the model performed the worst for this dataset. The SSIM for this dataset is crossed out because there is no distinct spatial structure present in the data after shuffling, leading to an SSIM of 0.002 that corresponds to no similarity. This dataset is not taken into account when looking at the pixel-wise classification performance of the model since there is no longer a lake shape present in the data.

3.2.2. Example Use Cases

For Greenland, three example lakes were chosen in order to illustrate how the model performs for each of the ten datasets. The choice of lakes was made based on the difference in shape and behavior when it came to the predictions. The figure structure of the three example use cases follows the same logic. The ground truth and the prediction for the original test dataset (a) and the nine perturbed datasets (b, d, e, f, g, h, I, j, k) are plotted in pairs along with their corresponding SSIM value which indicates the similarity between the ground truth and the model prediction. However, the 'Repeated time step' (c) dataset, is only included in the second example use case and can be seen in Figure 3.8.

The first example can be seen in Figure 3.6 and indicates a refreezing lake depicted in blue. The ground truth does not have a typical circular shape and looks like a centered irregular polygon with a non-uniform outline. This lake has been classified correctly as refreezing for all the dataset predictions except for the 'Randomly shuffled pixels' (k) dataset, for which the model predicted neither draining nor refreezing. It has a corresponding SSIM value of 0 which indicates no similarity. In the case of the 'Shifted left variation 1' (h) and 'Shifted left variation 2' (i) datasets, we can see that draining pixels are present in the prediction but in order for an image to be classified as draining or refreezing the majority of the pixels is taken into account. In this case, the majority of the pixels were refreezing therefore the lake is classified as refreezing.

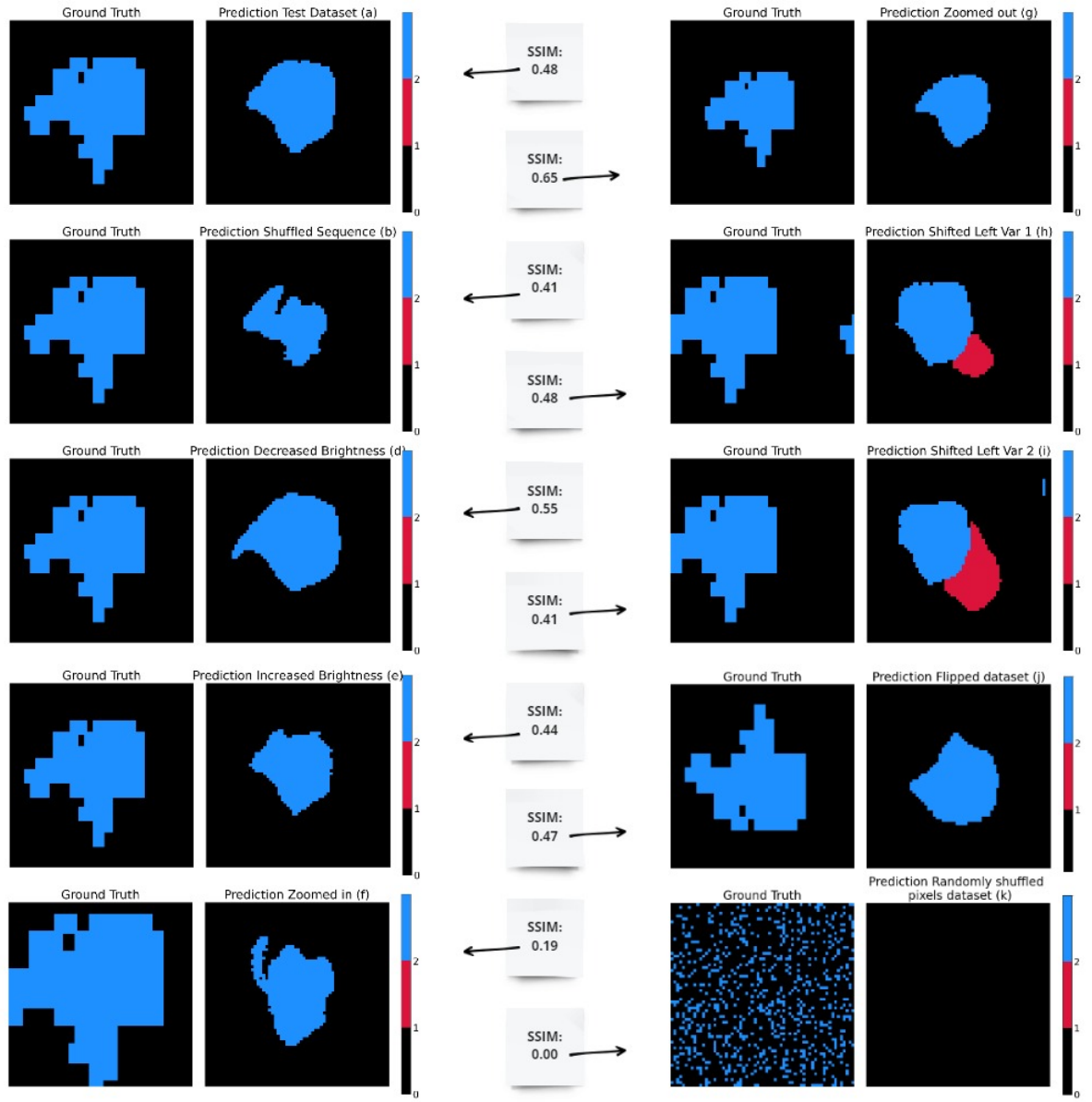


Figure 3.6: Comparison of prediction and ground truth for the original test dataset (a) and its nine perturbed datasets for a refreezing lake example.

It seems that for this specific lake, the model does not accurately predict the shape of the lake. In the case of the 'Shuffled sequence' (b) dataset the images for the lake entry were given to the model randomly shuffled and not in chronological order and it seems to affect the prediction way more than the other perturbed datasets. This is an indication that the model learns based on the spatio-temporal patterns present in the data. More specifically, for each entry in the dataset, the first images usually do not depict a lake, then a lake if formed, and after some time the lake either refreezes or drains as seen in Figure 2.9.a. In the case of the 'Zoomed in' (f) dataset, the images and ground truth are zoomed in maybe making it more difficult for the model to detect the lake boundaries since it is no longer centered. Lastly, when looking at the prediction for the 'Randomly shuffled pixels' (k) dataset, it can be seen that the model predicts neither draining nor refreezing. This might be an indication that the model heavily relies on the spatial arrangement of pixels to make predictions and is unable to find any recognizable patterns and defaults to predicting no lake pixels. Furthermore, the model has likely overfitted to the specific spatial patterns present in the training data which results in it being unable to generalize to new, unseen spatial arrangements, such as those that are created when the pixels are randomly shuffled.

Overall, the predictions seem to have a similar shape with the ground truths for the original test (a), the 'Increased intensity (e), the 'Decreased intensity' (d), the 'Flipped' (j), and the ' Zoomed out' (g) datasets which also correspond to higher SSIMs (above 0.44). Another prediction that has a high SSIM can be seen for the 'Shifted left variation 1' (h) dataset, with an SSIM value of 0.48 even though when looking at the graphs the shape does not resemble the ground truth as much as others. One of the limitations of the SSIM is that it takes into account specific parameters as mentioned in Section 2.8.1 but does not explicitly take into account the absolute position of objects within the image, which justifies the high score. Therefore, visual inspection still plays a key role when trying to conclude how the model performs. It is evident that when looking at the SSIM calculation and the visual representation of the prediction we do not conclude the same thing. The highest SSIM score is calculated for the 'Zoomed out' (g) dataset with a value of 0.65. On the other hand, the lowest SSIM scores are calculated for the 'Shuffled sequence' (b), 'Shifted left Variation 2' (i), and 'Zoomed in' (f) datasets and are 0.41, 0.41, and 0.19 respectively.

The second example can be seen in Figure 3.7 and indicates a draining lake depicted in red. In this example, the ground truth has a more elongated polygon shape that has a vertical orientation and is not exactly centered. Again it does not have a uniform outline.

This lake has been classified correctly as draining for all datasets except for the 'Shuffled sequence' (b) dataset. In the case of the 'Shift left variation 2' (i) dataset, refreezing pixels are present in the prediction but draining pixels are dominant therefore it is classified as draining. Overall, the predictions seem to have a very similar shape with the ground truth for datasets (a), (b), (d), (e), (f), (g), (h), (i), and (j) which also correspond to higher SSIM values (above 0.46).

The prediction for the 'Shuffled sequence' (b) dataset except for the fact that it is misclassified, is also slightly more centered and not as elongated, which is an indication that the model does not perform well when the order of the images is shuffled randomly because the spatio-temporal evolution of the lake is lost. The lowest SSIM calculation corresponds to the 'Zoomed in' (f) dataset with a value of 0.46. For the 'Zoomed in' (f) dataset, the model detects that the whole lake is within the predicted region even though the way the ground truth and the images are zoomed in, part of the lake is no longer visible (in the 1 by 1 km region depicted). For the 'Shifted left variation 2' (i) dataset, the model predicts refreezing pixels on the right which is where the ten-pixel shift was applied, and the ten pixels were replaced by constant (value of 0 - black) pixels. This is a non-ideal way to create a shift of the pixels because the model views it as something more than background space. This is a limitation of this approach and if more time was available this would have been implemented with k-Nearest Neighbors (k-NN) to interpolate the missing pixel values on the right side by considering the neighboring pixels. This would involve filling in the pixels based on the values of their nearest neighbors in the image which would help to maintain the visual coherence of the image. Last but not least, for the 'Randomly shuffled pixels' (k) dataset, the model predicts a centered circular structure that does not correspond to the ground truth. Again, this suggests that the model is highly sensitive to the spatial patterns present in the data. Furthermore, this might indicate overfitting as the model seems to rely heavily on the specific spatial patterns learned during training and can not generalize to altered spatial arrangements.

Lastly, the predictions for the 'Repeated time step' (c) dataset can be seen in Figure 3.8. On top can be seen the original test dataset entry along with the ground truth, which is identical for all 30 entries. It can be seen that the model manages to predict a similar shape to the ground truth only for entries 19-23. However, when looking at the original test dataset entry, we can see that this shape should have been predicted for time steps 11-15 instead. It can be concluded that the model learns from the spatio-temporal patterns present in the data and not from a specific time step in each entry. If we compare with the predictions for the original test set and most of the perturbed datasets, for this specific lake it can be seen that they correctly classify the lake as draining. Whereas, when looking at the predictions for the 'Repeated entry' (c) dataset, it can be seen that the model classifies as draining 19 out of the 30 entries created from the original entry.

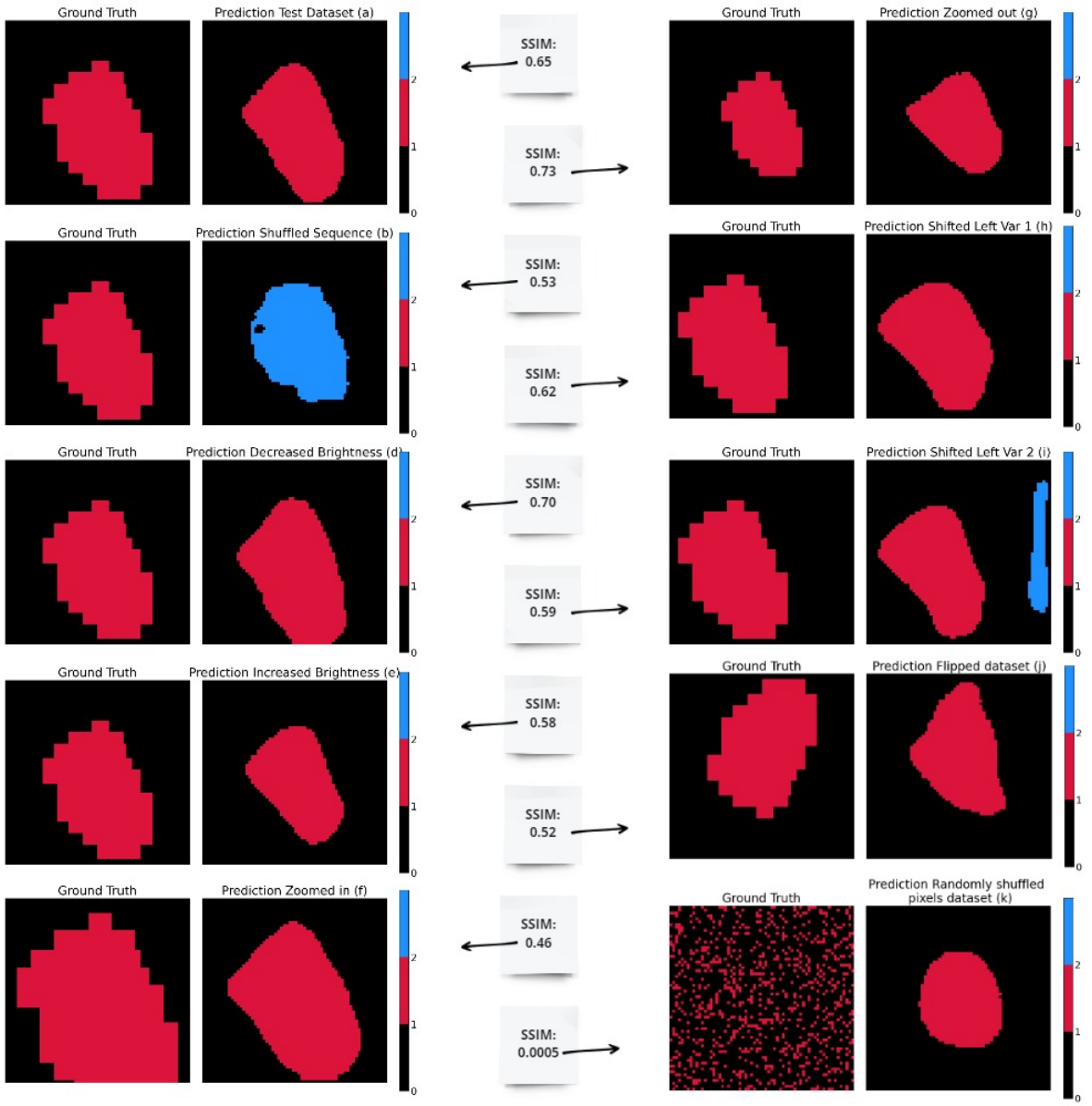


Figure 3.7: Comparison of prediction and ground truth for all nine datasets for a draining lake example.

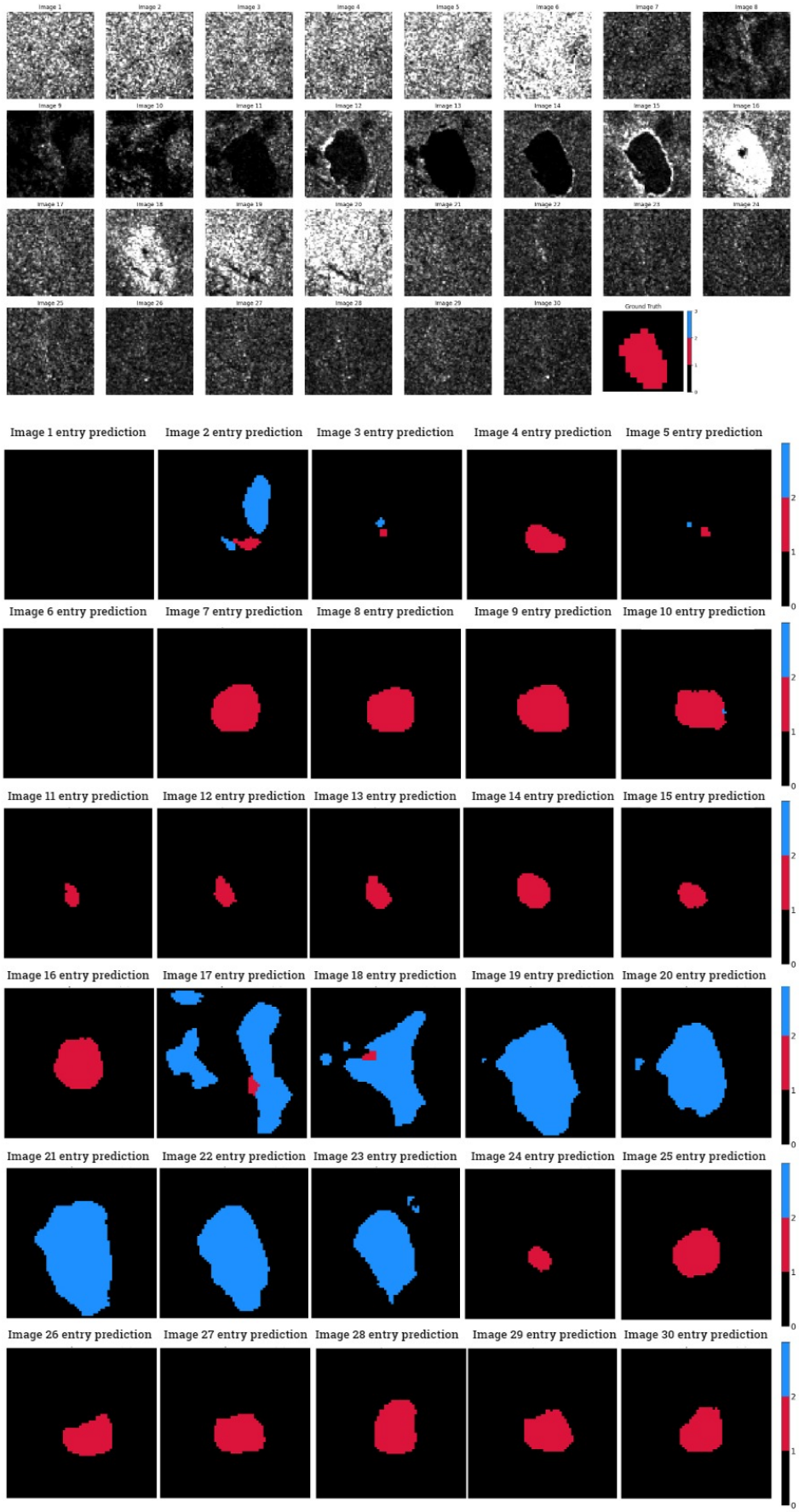


Figure 3.8: Original testing dataset entry comprised of 30 SAR images and the ground truth on top. Predictions for each of the 30 entries that were created from the one original entry for the 'Repeated time step' (c) dataset on the bottom.

The third example can be seen in Figure 3.9 and indicates a refreezing lake depicted in blue. The ground truth does not have a typical circular shape and looks like a centered irregular curved polygon with a non-uniform outline and a horizontal orientation.

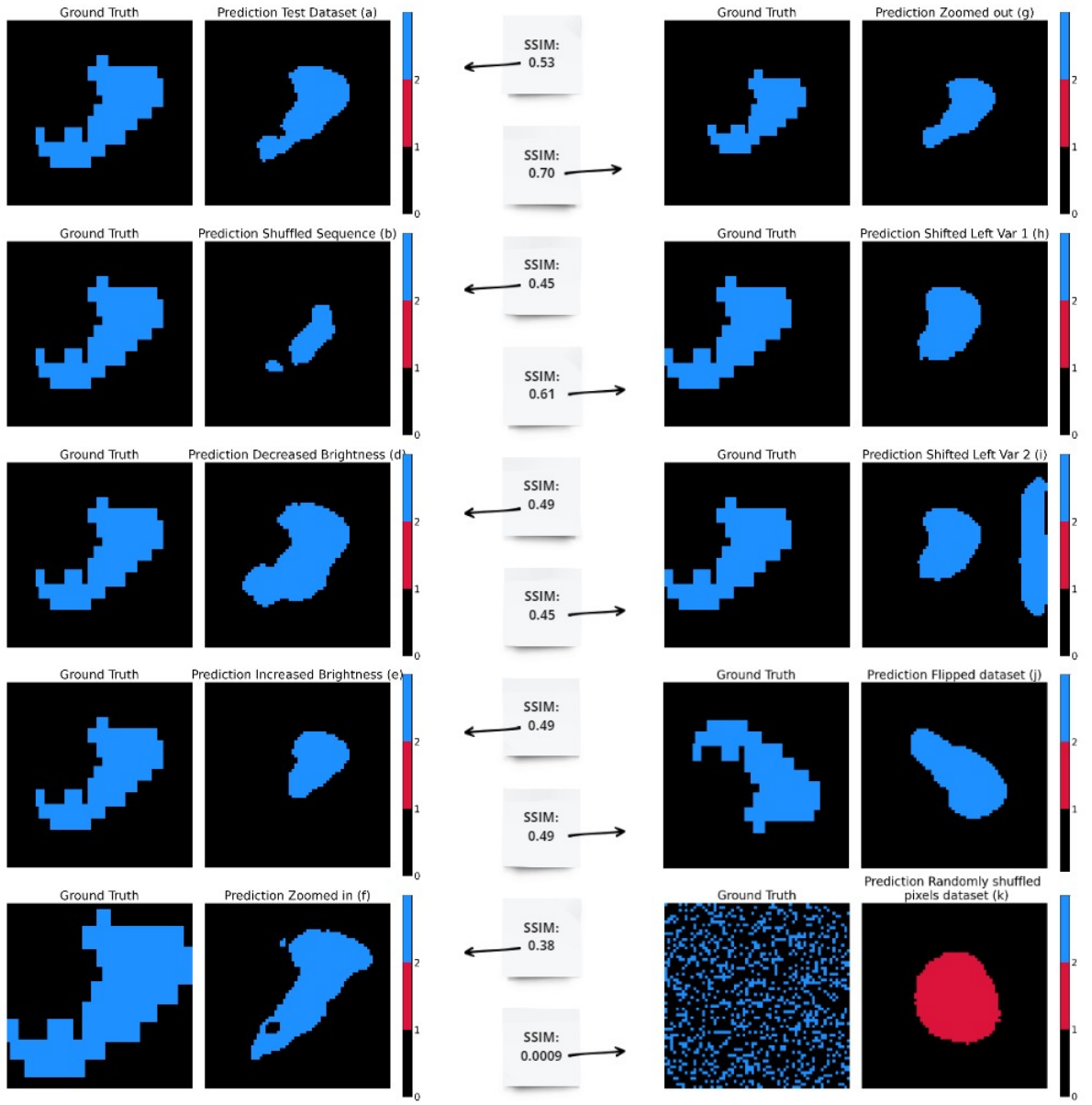


Figure 3.9: Comparison of prediction and ground truth for all nine datasets for a refreezing lake example.

This lake has been correctly classified as refreezing for all the perturbed datasets, and no draining pixels are present. Overall, the predictions seem to have a similar shape with the ground truth for all datasets except for the 'Zoomed in' (f) dataset which has the lowest SSIM value of 0.38. This might indicate that when a lake is viewed from a slightly zoomed-in perspective (more detail is captured in the image) the model is not able to capture all of it. Furthermore, the highest SSIM value corresponds to the 'Zoomed out' (g) dataset which is an indication that the model works best when the lake geometry is viewed from slightly further away when the shape has less detail and the whole lake is centered and included in the image. For the 'Shuffled sequence' (b) dataset, the prediction captures only part of the lake which is the case for all three examples, meaning that the model prediction highly depends on the spatial evolution of the lakes over time. Next, for the 'Increased intensity' (e), 'Shifted left variation 1' (h), and 'Shifted left variation 2' (i) dataset only the top part of the lake is predicted which happens for a

variety of reasons depending on the dataset. For the 'Increased intensity' (e) dataset, which is created by increasing the intensity of the images, the bottom part of the lake becomes brighter and is no longer dark (in SAR images, bodies of water appear dark - low backscatter), leading to the model viewing it as "no lake pixels". For the 'Shifted left variation 1 and 2' (h, i) datasets, this happens because when the image is shifted to the left, the model is unable to detect the part of the lake that is close to the image border. This might be happening due to the fact that most lakes within the training dataset are centered and when the model encounters lakes that are near the image borders it can not generalize. Last but not least, for the 'Randomly shuffled pixels' (k) dataset the model predicts a very similar circular, centered shape as in the previous example but this time it does not correctly predict refreezing. This further indicates that the model is overfitted on the training data where most lakes are more circular and centered. Lastly, it can be seen from all three examples that the average SSIM shows a similar behavior as the individual SSIMs, at least when it comes to the highest and the lowest. This indicates that overall the model predictions are showing a similar behavior within each dataset.

3.3. Applying model on Antarctic lakes

As mentioned in Section 2.9, the trained model (on Greenland data) was then applied to the Antarctica dataset to perform the prediction step. Using the prediction, an Antarctic-wide map showing where the model predicting drainage and refreezing was created as seen in Figure 3.10.

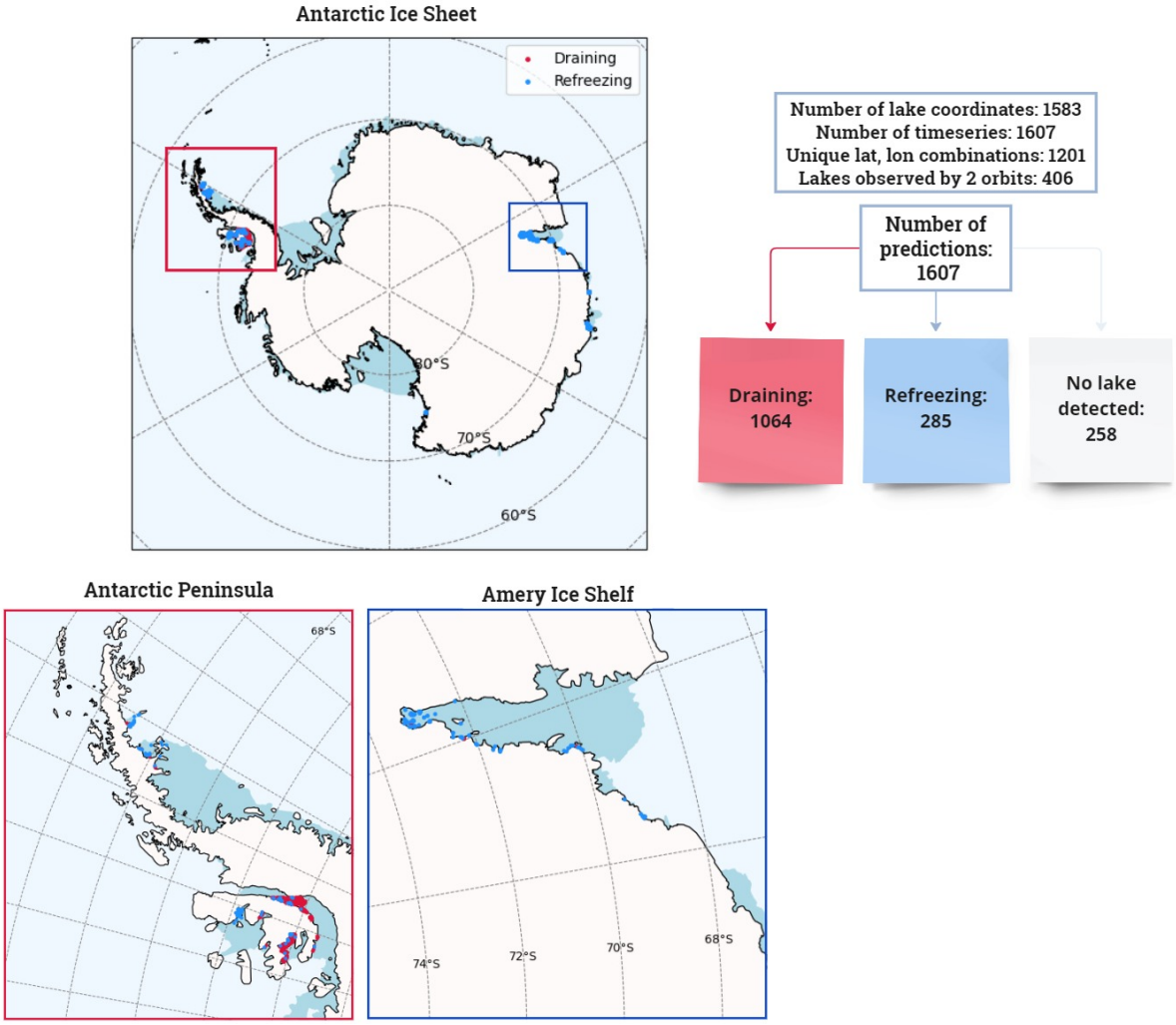


Figure 3.10: The model predictions for the AIS plotted. Red circles indicate draining events, and blue indicates refreezing events. The red box indicates the area of the AP zoomed in, while the blue box indicates the area of the Amery Ice Shelf zoomed in.

The initial number of lake centroids was 1583, which were used to extract Sentinel-1 data. The total number of Sentinel-1 time series (entries) was 1607 with 1201 unique longitude, and latitude combinations, meaning that data were not available for all lakes while 406 lakes were observed by two orbits. The model performed a prediction for each time series and from those, 66.3 % (1064) was draining, 17.7 % (285) was refreezing and no lake was detected for 16.0 % (258).

When looking at the map, we can see that the draining lakes are clustered together mainly in the Antarctic Peninsula. There are some draining lakes detected on the Amery Ice Shelf as well but the majority of lakes in that location are predicted as refreezing. Refreezing is also predicted on the Shackleton and West Ice Shelves.

Figure 3.11 shows four different entries from the Antarctica dataset along with the prediction for each entry. It is clear that the model performs very poorly when making a prediction, meaning that it can not even capture the shape or position of most lakes. This could be due to the fact that Antarctic SLs are very different than Arctic SLs, both in shape and size. In Antarctica, most lakes look like channels of meltwater or are very close together forming complex connected lake systems. Most of the time the model predicts a circular centered draining lake which makes sense since most lakes that it has been trained with had such a shape and positioning. On the other hand, in Figure 3.12, some lakes that look more than Greenland lakes can be seen and it is clear that the model performs slightly better for those lakes.

To conclude, for the majority of lakes the model predicts draining and an overall centered circular shape. For the majority of lakes that look like braided river channels or complex lake systems, the model is not really able to predict an accurate shape or position. Lastly, for lakes that slightly resemble Greenland lakes, the model performs slightly better but still not good enough.

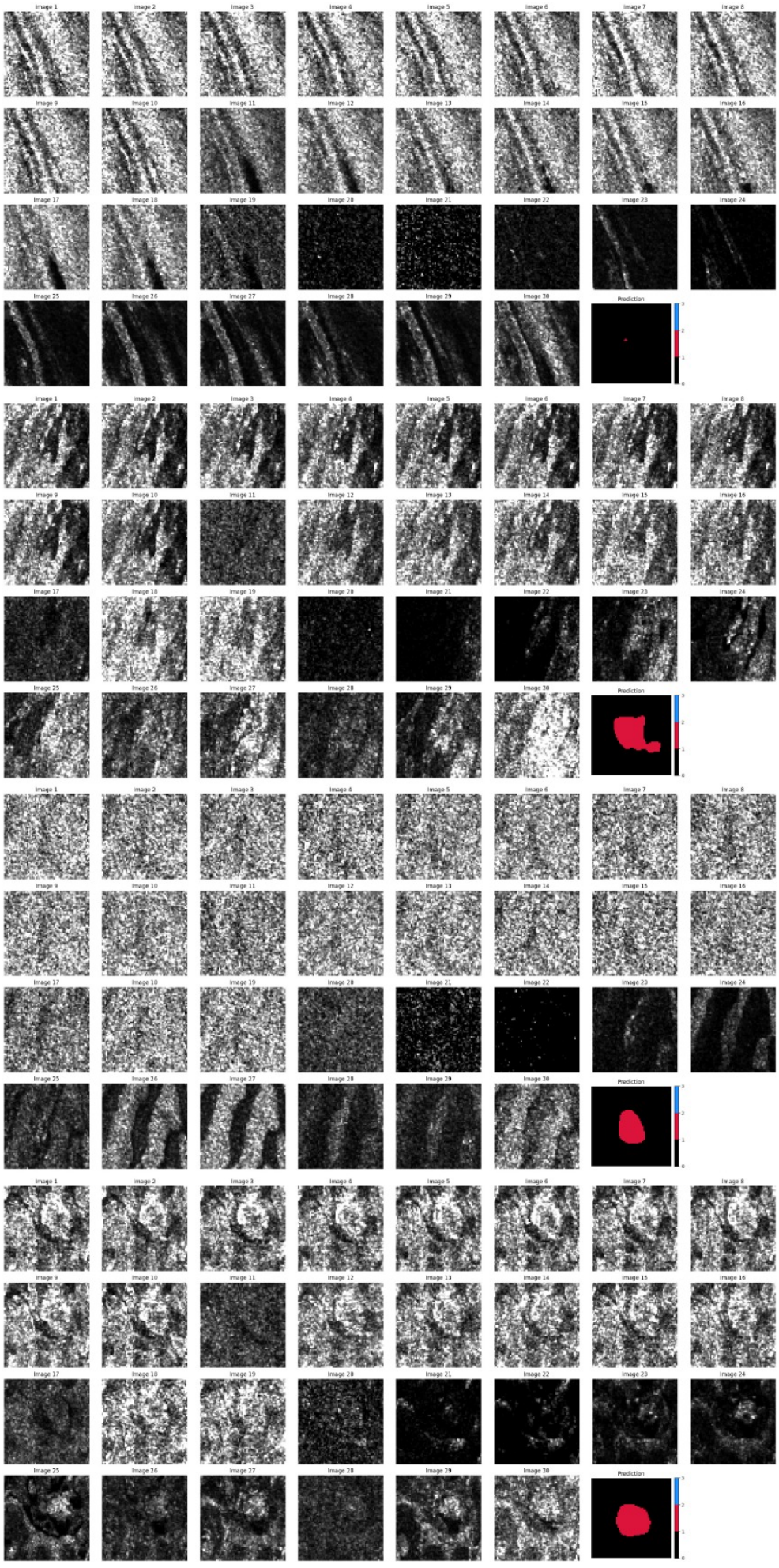


Figure 3.11: Four Sentinel-1 time series of Antarctic SLs (look like river or melt channels) along with the model prediction. Red in the prediction indicates draining lake pixels, blue indicates refreezing lake pixels, and black indicates no lake pixels.

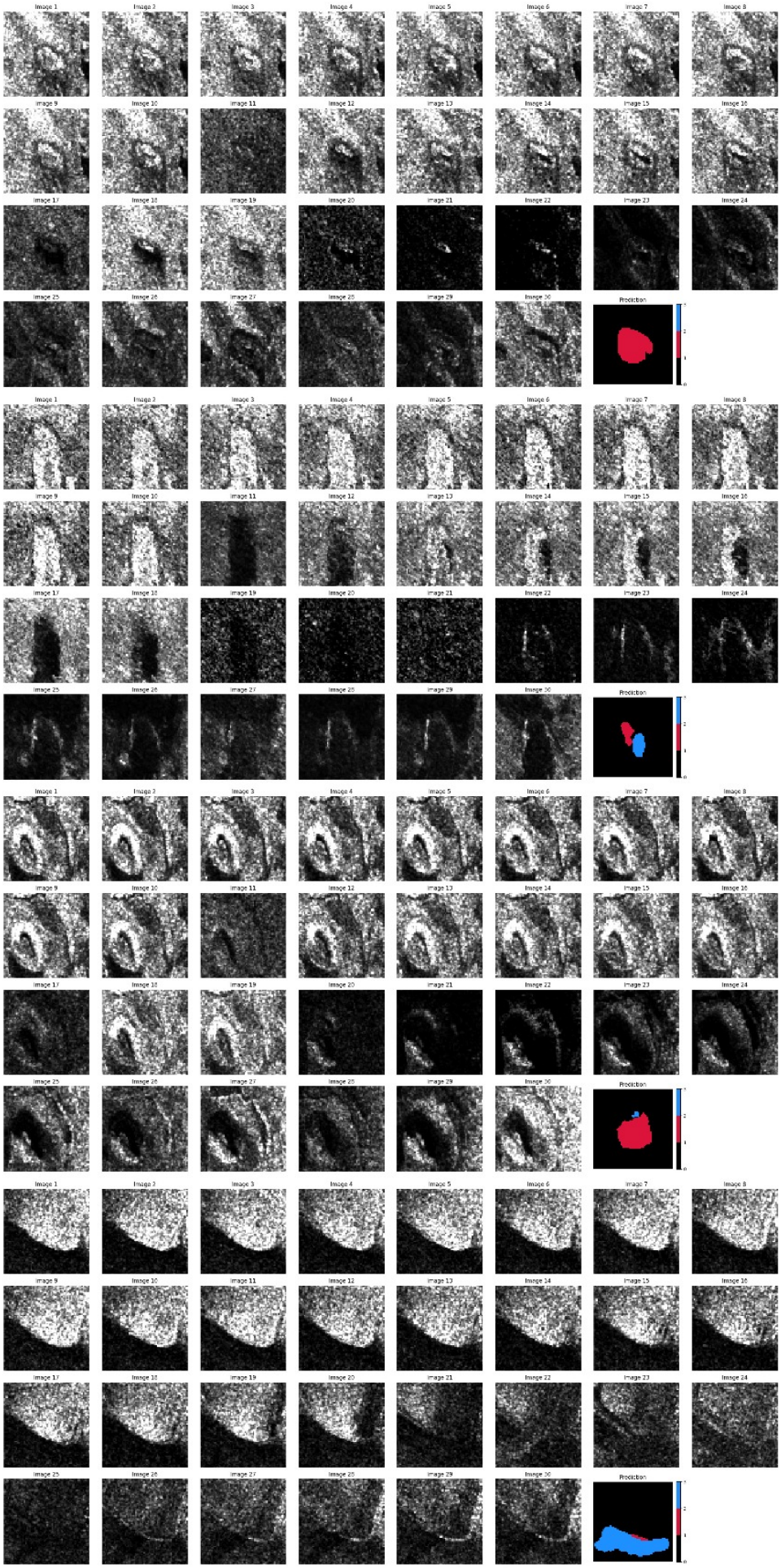


Figure 3.12: Four Sentinel-1 time series of Antarctic SLs that resemble Greenland SLs along with the model prediction. Red in the prediction indicates draining lake pixels, blue indicates refreezing lake pixels, and black indicates no lake pixels.

4

Discussion and Recommendations

4.1. Impact of Spatial and Temporal Factors on Model Performance

As mentioned in Section 2.7.3, a sensitivity analysis was conducted in order to determine how temporal and spatial factors affect model performance. When looking at the overall model performance, it can be concluded that the temporal aspect of the data is key in order for the model to learn. When the temporal aspect is removed by either randomly shuffling the images within the time series or repeating a specific time step of the time series, the model achieves less than 77 % across all metrics leading to poor model performance (pixel-wise). When randomly shuffling the images ('Shuffled sequence' (b) dataset), the spatial evolution of the lake over time is removed, and the poor model performance indicates that the model relies on it. This is further supported by the performance of the model when repeating a specific time step ('Repeated timestep' (c) dataset) which indicates that the model does not learn from a specific time step, rather than learning from the spatial evolution of the lakes that takes place over time. The image-wise model performance agrees, with overall metrics below 67%.

When looking at how the model performance (pixel-wise) is affected by changes in space it can be concluded that four out of eight perturbations created affect the model as much as the temporal perturbations. More specifically, the model achieves metrics below 85 % when the images are zoomed in, shifted to the left, and pixels are randomly shuffled. This could indicate that the model is overfitted on the training data and is not able to generalize to slightly different input data. Firstly, when zooming in, the model receives more detail ('Zoomed in' (f) dataset), and underestimates the size of the lake by predicting that the whole lake is within the predicted area. In most cases, after zooming in part of the lake is no longer within the region depicted. Secondly, when the lake is not centered ('Shifted left variation' (h, i) datasets) the model does not perform as well which indicates that it is overfitted on the training data, where most lakes are centered. Last but not least, when the spatial pattern of the lake is removed (Randomly shuffled pixels (k) dataset) as mentioned in Section 3, poor model performance is achieved because the model relies on distinct spatial patterns to extract features and when we randomly shuffle the pixels those spatial patterns are no longer present. Lastly, for the remaining spatial perturbations, the model achieves more than 85% across all metrics, indicating good model performance.

On the other hand, when looking at how the model performs image-wise for the spatial perturbations seven out of the eight perturbations lead to good model performance with with overall metrics above 90%. This is due to the fact that we only take into account if the model correctly predicts draining or refreezing, and not if the pixels are correctly classified. The dataset that leads to poor model performance is again the 'Randomly shuffled pixels' (k) dataset which achieves an accuracy of 68.4 %, precision of 100 %, recall of 64.3%, and, F1 score of 78.3 %. Precision is 100 % because in order to calculate it, only the predictions for draining are taken into account and the model correctly classifies all nine of them.

4.2. Comparison with Current Studies

One of the differences between our approach and the study by **Miles et al. (2017)** is that they use both HH and HV polarization. In their case, the HV polarization typically classified a much greater lake area than the HH polarization, while most pixels classified using the HH were also classified by the HV. This is due to the fact that when using the HV polarization of C-band SAR, water can be detected even if the lake's surface is frozen and snow-covered because it can penetrate several meters of ice

(**Eric Rignot, Echelmeyer, and Krabill 2001**). Since only HH polarization was used in this study, this might have contributed to underestimating the extent of the lakes. Consequently, the choice of HH polarization which was made due to the fact that the IW data product in the 10 m resolution is available in this polarization (over Antarctica) might pose a limitation of this method.

In addition, a study by **de Roda Husman (2024)** used a ConvLSTM model to predict draining and refreezing Antarctica-wide for the same lakes. The results indicated the opposite behavior. More specifically, most draining lakes were detected on the Amery ice shelf where we detected mainly refreezing. Furthermore, most refreezing lakes were detected on the Antarctic Peninsula where the majority of the draining lakes are concentrated in our prediction. Currently, it is not possible to determine which approach leads to the most accurate results since none of the two studies are validated. This could be done by comparing the results with optical data to confirm whether the lakes are draining or refreezing.

In further research, it would be interesting to examine when a draining or refreezing occurs during the melt season. This could prove helpful because a smaller time series length would be needed. Currently, a time series of length 30 is used but if we knew when the events are taking place this could be cut down to half. Furthermore, an investigation of relevant parameters such as ice thickness and surface temperature could result in a better understanding of where and why hydrofracture takes place.

4.3. Recommendations

Although the approach used in this thesis produced encouraging results, there are some limitations and ways to mitigate them. First of all the size of the dataset is limited by the use of reference data for one melt season (summer 2021) and since the accuracy and robustness of deep learning models heavily rely on the amount of data used, this approach probably does not lead to optimal performance. This can be mitigated by increasing the size of the dataset which can lead to a better overall model performance. Reference data from 2017 - 2020 can be used in the same way as for 2021, leading to five times the amount of data currently used.

In addition, the training dataset contains significantly more instances of one class (0: no lake pixels/388735) than the other two classes (1:draining lake pixels/ 78419, 2: refreezing lake pixels/61230) which means that the model might be biased toward the majority class and that is why it underperformed in correctly classifying a high percentage of especially class 2. More specifically, in Chapter 3, the performance of the model per class was presented, which showed that for class 2 (refreezing) the model classified correctly a maximum of 65 % of the pixels. If the class imbalance is addressed with resampling or class weights the model will probably perform better for the minority classes.

Also, the model overall performs well on Greenland data but when encountering data from different regions or different conditions (Antarctica) it does not manage to generalize and performs poorly. This could be solved by training the model on both ice sheets.

Furthermore, deep learning models, particularly CNNs are often considered ”black boxes” which means that it is really hard to understand why the model makes a certain prediction and further complicates the interpretation of the results. To mitigate this limitation, the eight perturbations of the test dataset were created which provided some insights into when the model is performing well and why.

Another change that would improve model performance would be to use slightly bigger bounding boxes (2 km by 2 km) in order to export the Sentinel-1 Greenland dataset. This became clear when looking at the results of the sensitivity analysis which showed that the model performed best across all metrics (accuracy, precision, recall, F1 score, SSIM) for the ’Zoomed out’ (g) dataset.

One of the most important limitations that affect the shape of the prediction is the fact that for each entry/time series, there is only one lake mask available for all time steps. This affects the prediction and does not lead to the most accurate pixel-wise prediction. This limitation can not be addressed because, for most cases, only one optical image might be available for one melt season.

Furthermore, as mentioned in Section 4.2, the study by **Miles et al. (2017)** uses both HV and HH polarizations which could also be applied in our study. We would need to use a coarser resolution (EW has a resolution of 40 m and HV, HH polarizations) in order to also use the HV polarization. However, this could pose a new limitation because the 40 m resolution might be too coarse to detect lakes in Antarctica, that are relatively small.

Another way to improve this method could be to explore advanced deep learning pre-trained models such as EfficientNet or ResNet which might lead to better performance and cut down training time for this task. Lastly, sophisticated data augmentation techniques such as random erasing, cutout, and rotation can be applied to the whole dataset, not just to the testing set as done in this thesis. This will make the model more robust to variations in the input data resulting in overall better generalization and model performance.

5
Conclusions

This study assessed whether a U-Net can be used in order to distinguish between draining and refreezing lakes in ice sheets using Sentinel-1 SAR data. The main conclusions are presented in this chapter. The research questions are answered and due to their nature recommendations are also included.

How can we develop a deep learning model to classify draining and refreezing lakes in Greenland using Sentinel-1 data?

To develop a deep learning model for classifying draining and refreezing lakes in Greenland using Sentinel-1 data, we need a high-resolution Sentinel-1 time series that focuses on areas with known lake behavior. Next, the images need to be normalized in order to have a similar scale across the whole dataset which is necessary because data comes from different orbits (backscatter intensity variability) which will also improve the convergence speed of the model. In addition, reference data (shapefile, Section 2.3) needs to be used in order to create a label band for each time series. Next, using a CNN architecture and more specifically a U-Net, the model is trained on the labeled dataset which is split into training (70%), validation (20%), and testing (10%). A form of data augmentation is then used in order to understand better when and why the model performs better. This is done by applying some transformations to the test set and creating eight perturbations of it. To optimize performance, the hyperparameters are tuned based on the validation dataset. Next, when the model is trained and achieves the optimal performance based on the validation dataset, the test dataset and its eight perturbations are utilized to perform the prediction. Lastly, the model's evaluation metrics (accuracy, precision, recall, f1 score, and SSIM) are calculated both image-wise and pixel-wise for all nine datasets.

How do spatio-temporal patterns and changes in the data affect model performance?

By conducting a sensitivity analysis we assessed how temporal and spatial factors influence model performance. It revealed that spatial lake evolution over time is crucial for the model's learning process, as removing the temporal aspect by shuffling images or repeating specific time steps significantly drops performance below 77% across all metrics, indicating reliance on those spatio-temporal patterns. Spatial factors also play a critical role, with four out of eight perturbations such as zooming in, shifting left, and randomly shuffling pixels causing metrics to drop below 85%. These perturbations showed that the model struggles with detailed or altered inputs and relies on specific spatial patterns, suggesting overfitting to the training data.

How can this method be improved?

To improve the model, the dataset can be expanded by including reference data from 2017-2020, increasing its size. Using larger bounding boxes (2 km by 2 km) for data export has been shown to enhance model performance. Creating a lake mask for each time step using optical data, rather than a single mask per melt season, can also improve accuracy. Furthermore, implementing advanced pretrained models like EfficientNet or ResNet may lead to better performance and reduced training time. Addressing class imbalance through resampling can improve model performance for minority classes, such as refreezing lake pixels. Lastly, applying sophisticated data augmentation techniques, such as random erasing, cutout, and rotation, to the entire dataset rather than just the testing set can increase model robustness and overall performance.

Can this model be generalized to other polar regions?

At its current state, this method showed that it can not be generalized to Antarctica. This is mainly due to the fact that Antarctica lakes look very different in terms of shape (complex river channels, complex interconnected lake systems) and positioning (not centered) as seen in Figure 3.11. Therefore, since the model has not seen similar data at all it struggles to generalize. Some lakes within the Antarctica dataset resemble Greenland lakes in terms of shape as seen in Figure 3.12, for which the model seems to pick up part of the shape.

What are the advantages of this approach?

One of the main advantages of this approach is the use of Sentinel-1 data which provides high temporal resolution which is crucial for monitoring dynamic processes like hydrofracturing. It also enables detailed analysis of the lakes, allowing for the identification of subtle features that might indicate draining or refreezing. Furthermore, Sentinel-1's SAR operates in all weather conditions, ensuring consistent data acquisition regarding cloud cover. Since this approach uses a deep learning model, it has the advantage of automatically learning and extracting relevant features from data which might be difficult to extract just by looking. In addition, when the model is trained it can be applied to new data and make a prediction.

What are the limitations of this approach?

The limitations of this approach include the restricted dataset size, as it only uses reference data from the summer of 2021, impacting the accuracy and robustness of the deep learning model. There is also a class imbalance, with significantly more instances of "no lake pixels" than "draining" and "refreezing lake pixels," which biases the model towards the majority class, resulting in poor classification of minority classes, especially class 2 (refreezing). Additionally, the model performs well on Greenland data but fails to generalize to other polar regions, such as Antarctica. The "black box" nature of deep learning models like CNNs makes it difficult to interpret predictions. Lastly, having only one lake mask per time series affects pixel-wise prediction accuracy.

References

Ali, Peshawa Jamal Muhammad et al. (2014). “Data normalization and standardization: a technical report”. In: *Mach Learn Tech Rep* 1.1, pp. 1–6.

Alley, K.E. et al. (2018). “Quantifying vulnerability of Antarctic ice shelves to hydrofracture using microwave scattering properties”. In: *Remote Sensing of Environment*, pp. 297–306. DOI: https://doi.org/10.1016/j.rse.2018.03.025. URL: https://www.sciencedirect.com/science/article/pii/S0034425718301263.

Benedek, C. L. and I. C. Willis (2021). “Winter drainage of surface lakes on the Greenland Ice Sheet from Sentinel-1 SAR imagery”. In: *The Cryosphere* 15.3, pp. 1587–1606. DOI: 10.5194/tc-15-1587-2021. URL: https://tc.copernicus.org/articles/15/1587/2021/.

Brown, Ian and A. Malin Johansson (Jan. 2011). “Observations of supra-glacial lakes in west Greenland using winter wide swath Synthetic Aperture Radar”. In: *Remote Sensing Letters* 3, pp. 531–539. DOI: 10.1080/01431161.2011.637527.

Church, John A. and Neil J. White (2011). “Sea-Level Rise from the Late 19th to the Early 21st Century”. In: *Surveys in Geophysics* 32.4-5, pp. 585–602. DOI: 10.1007/s10712-011-9119-1.

Davies, Bethan (June 2020). *Ice shelves*. URL: https://www.antarcticglaciers.org/glaciers-and-climate/changing-antarctica/shrinking-ice-shelves/ice-shelves/.

de Roda Husman, Sophie et al. (2024). “A high-resolution record of surface melt on Antarctic ice shelves using multi-source remote sensing data and deep learning”. In: *Remote Sensing of Environment* 301, p. 113950. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2023.113950. URL: https://www.sciencedirect.com/science/article/pii/S0034425723005023.

Dell, R. et al. (2020). “Lateral meltwater transfer across an Antarctic ice shelf”. In: *The Cryosphere* 14.7, pp. 2313–2330. DOI: 10.5194/tc-14-2313-2020. URL: https://tc.copernicus.org/articles/14/2313/2020/.

Fox-Kemper, B. et al. (2021). “Ocean, Cryosphere and Sea Level Change”. 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*. Ed. by V. Masson-Delmotte et al. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, pp. 1211–1362. DOI: 10.1017/9781009157896.011.

Fretwell, P. et al. (2013). “Bedmap2: improved ice bed, surface and thickness datasets for Antarctica”. In: *The Cryosphere*, pp. 375–393. DOI: 10.5194/tc-7-375-2013. URL: https://tc.copernicus.org/articles/7/375/2013/.

Hao, Shijie, Yuan Zhou, and Yanrong Guo (2020). “A Brief Survey on Semantic Segmentation with Deep Learning”. In: *Neurocomputing* 406, pp. 302–321. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2019.11.118. URL: https://www.sciencedirect.com/science/article/pii/S0925231220305476.

J. Sommer, S. de Roda Husman and M. Izeboud. (2023). “Regions of Ice Weakness and Hydrofracture on Shackleton Ice Shelf, Antarctica”. In.

Kuipers Munneke, Peter et al. (2014). “Firn air depletion as a precursor of Antarctic ice-shelf collapse”. In: *Journal of Glaciology* 220, pp. 205–214. DOI: 10.3189/2014JoG13J183.

Lee, H. and J. (Core Writing Team) Romero (2023). “Sections. In: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change”. In: *IPCC, Geneva, Switzerland*, pp. 35–115. DOI: 10.59327/IPCC/AR6-9789291691647.

Miles, Katie E. et al. (2017). “Toward Monitoring Surface and Subsurface Lakes on the Greenland Ice Sheet Using Sentinel-1 SAR and Landsat-8 OLI Imagery”. In: *Frontiers in Earth Science* 5. ISSN: 2296-6463. DOI: 10.3389/feart.2017.00058. URL: https://www.frontiersin.org/articles/10.3389/feart.2017.00058.

Moussavi, Mahsa et al. (2020). “Antarctic Supraglacial Lake Detection Using Landsat 8 and Sentinel-2 Imagery: Towards Continental Generation of Lake Volumes”. In: *Remote Sensing* 12.1. ISSN: 2072-4292. DOI: 10.3390/rs12010134. URL: https://www.mdpi.com/2072-4292/12/1/134.

NASA and JPL/Caltech (Aug. 2023). *Antarctic and Greenland Ice Loss*. Author: Felix W. Landerer, Visualizer: Marit Jentoft-Nilsen. Originally published on March 8, 2024. Last updated on March 13, 2024 at 11:21 AM EDT.

National Snow and Ice Data Center (2023). URL: https://nsidc.org/learn/parts-cryosphere/ice-shelves.

Potin, P. (2013). URL: https://sentinel.esa.int/documents/247904/349449/S1_SP-1322_1.pdf.

Rignot, E. et al. (2013). “Ice-Shelf Melting Around Antarctica”. In: *Science*, pp. 266–270. DOI: 10.1126/science.1235798. eprint: https://www.science.org/doi/pdf/10.1126/science.1235798. URL: https://www.science.org/doi/abs/10.1126/science.1235798.

Rignot, Eric, Keith Echelmeyer, and William Krabill (2001). “Penetration depth of interferometric synthetic-aperture radar signals in snow and ice”. In: *Geophysical Research Letters* 28.18, pp. 3501–3504. DOI: https://doi.org/10.1029/2000GL012484. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2000GL012484. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000GL012484.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab et al. Cham: Springer International Publishing, pp. 234–241.

Scambos, T. A. et al. (2004). “Glacier acceleration and thinning after ice shelf collapse in the Larsen B embayment, Antarctica”. In: *Geophysical Research Letters* 31.18. DOI: https://doi.org/10.1029/2004GL020670. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2004GL020670. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004GL020670.

Tharwat, Alaa (2020). “Classification assessment methods”. In: *Applied Computing and Informatics*. URL: https://api.semanticscholar.org/CorpusID:59212480.

Torres, Ramon et al. (2012). “GMES Sentinel-1 mission”. In: *Remote Sensing of Environment* 120. The Sentinel Missions - New Opportunities for Science, pp. 9–24. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2011.05.028. URL: https://www.sciencedirect.com/science/article/pii/S0034425712000600.

Trusel, Luke D., Zhuolai Pan, and Mahsa Moussavi (2022). “Repeated Tidally Induced Hydrofracture of a Supraglacial Lake at the Amery Ice Shelf Grounding Zone”. In: *Geophysical Research Letters* 49.7. e2021GL095661 2021GL095661, e2021GL095661. DOI: https://doi.org/10.1029/2021GL095661. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021GL095661. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021GL095661.

Yang, Kang and Laurence C. Smith (2013). “Supraglacial Streams on the Greenland Ice Sheet Delineated From Combined Spectral–Shape Information in High-Resolution Satellite Imagery”. In: *IEEE Geoscience and Remote Sensing Letters* 10.4, pp. 801–805. DOI: 10.1109/LGRS.2012.2224316.

Yang, Kang, Laurence C. Smith, et al. (2021). “Seasonal evolution of supraglacial lakes and rivers on the southwest Greenland Ice Sheet”. In: *Journal of Glaciology* 67.264, pp. 592–602. DOI: 10.1017/jog.2021.10.

Zhao, Shuai et al. (2019). “Correlation Maximized Structural Similarity Loss for Semantic Segmentation”. In: *CoRR* abs/1910.08711. arXiv: 1910.08711. URL: http://arxiv.org/abs/1910.08711.

Zheng, Lei (Apr. 2023). “Summer supraglacial lakes and winter buried lakes in the Greenland Ice Sheet”. In: DOI: 10.6084/m9.figshare.22017989.v1. URL: https://figshare.com/articles/dataset/Summer_supraglacial_lakes_and_winter_buried_lakes_in_the_Greenland_Ice_Sheet/22017989.

Zheng, Lei et al. (2023). “Multi-sensor imaging of winter buried lakes in the Greenland Ice Sheet”. In: *Remote Sensing of Environment* 295, p. 113688. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2023.113688. URL: https://www.sciencedirect.com/science/article/pii/S0034425723002390.

ChatGTP has been used as a tool in this master thesis to enhance code quality and refine text.

The scripts developed for this thesis are available on GitHub at: https://github.com/Fenia-Ps/UNet_Supraglacial_Lake_Classification

A

Extensive Results

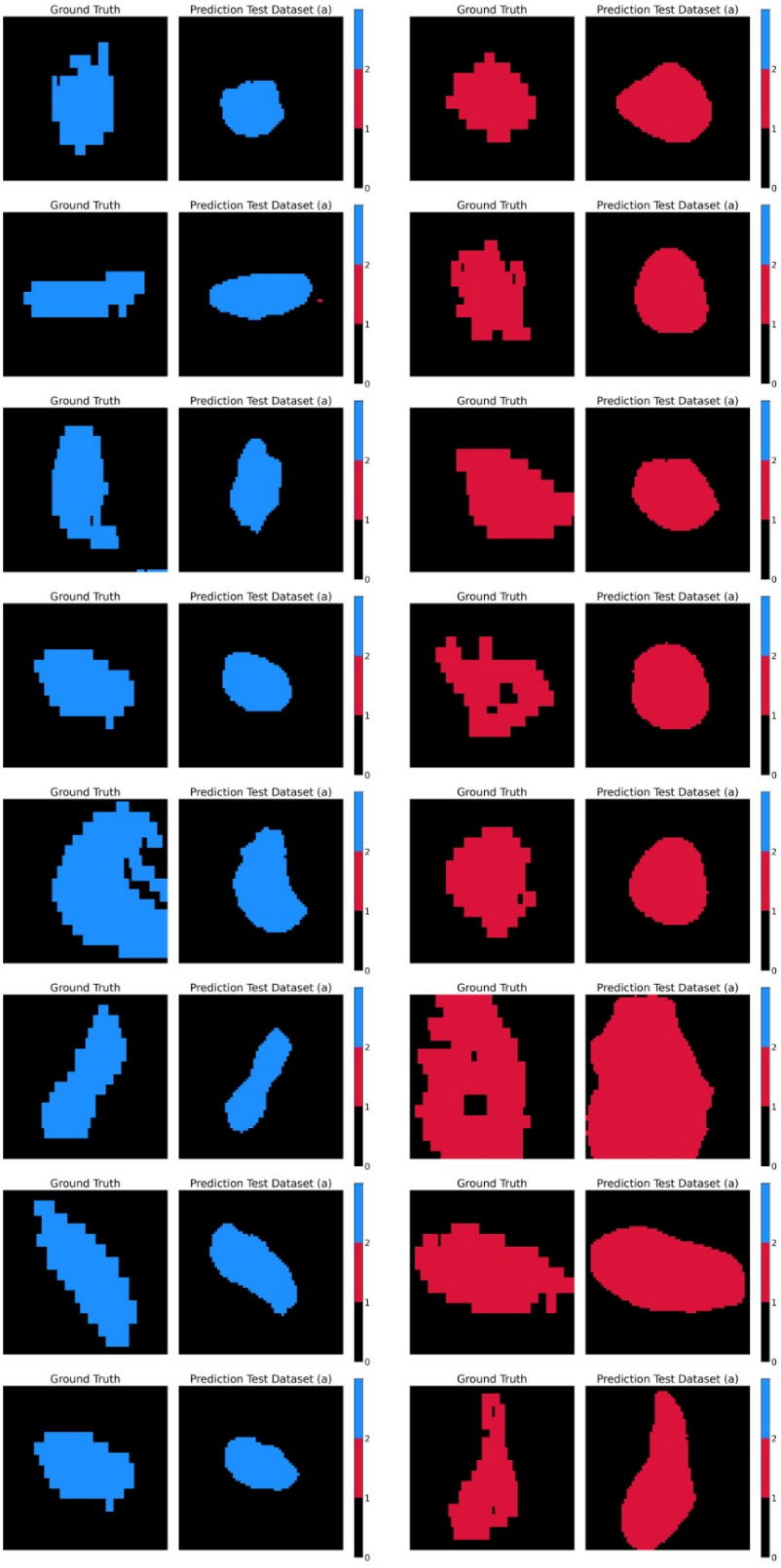


Figure A.1: Prediction for the Original test dataset (a).

Figure A.2: Prediction for 'Shuffled Sequence' (b) dataset.

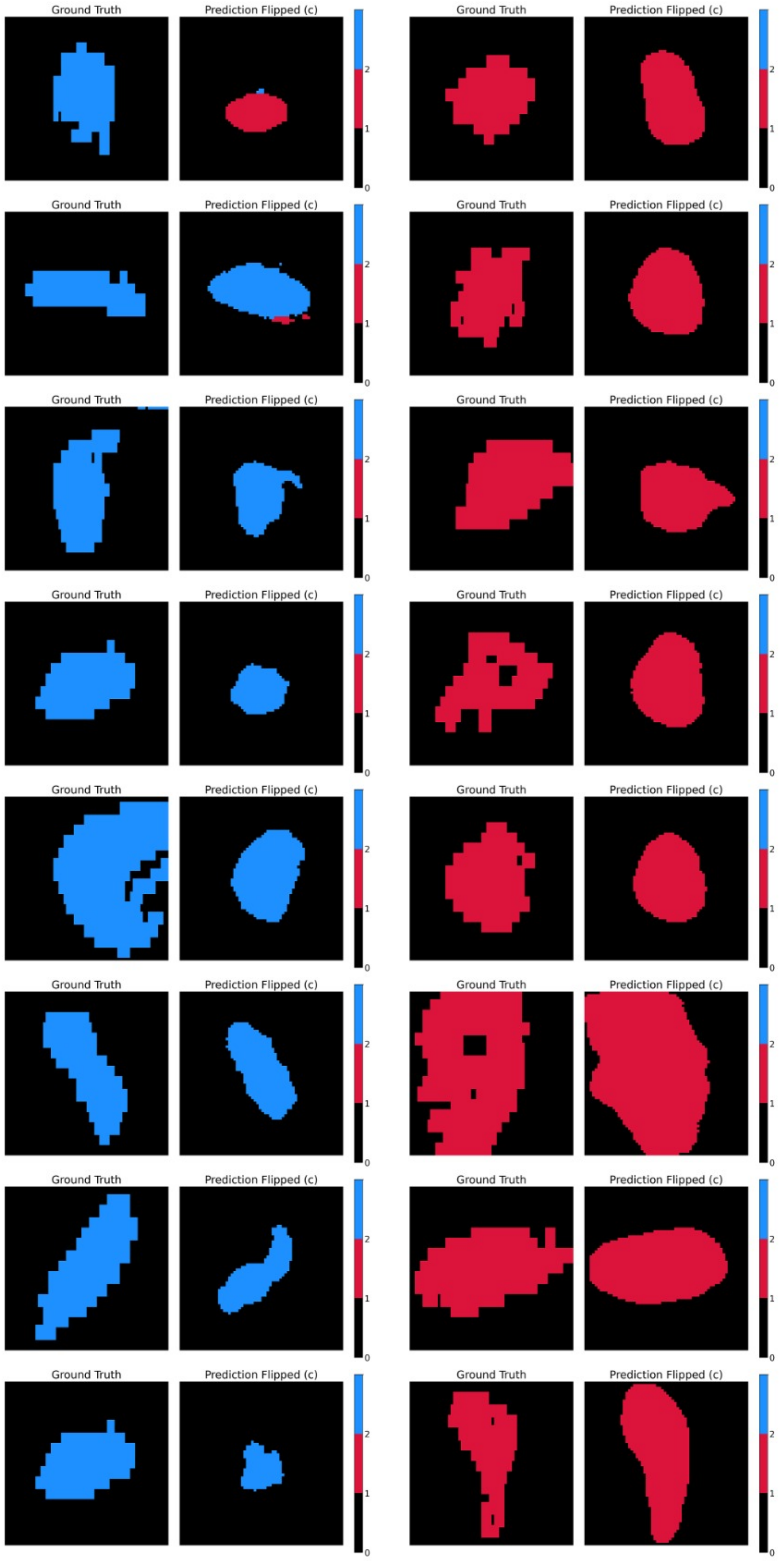


Figure A.3: Prediction for 'Flipped' (c) dataset.

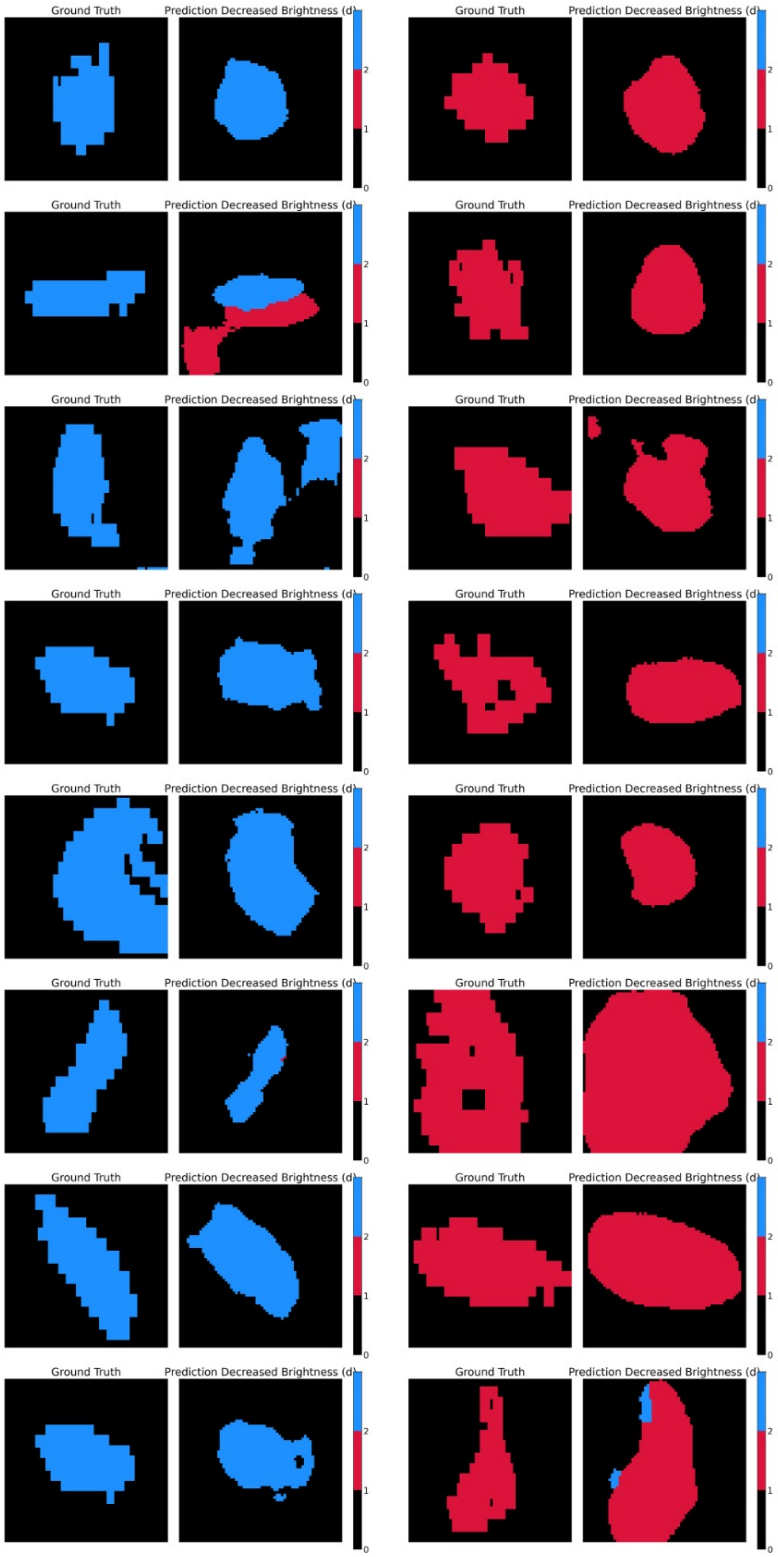


Figure A.4: Prediction for 'Decreased Brightness' (d) dataset.

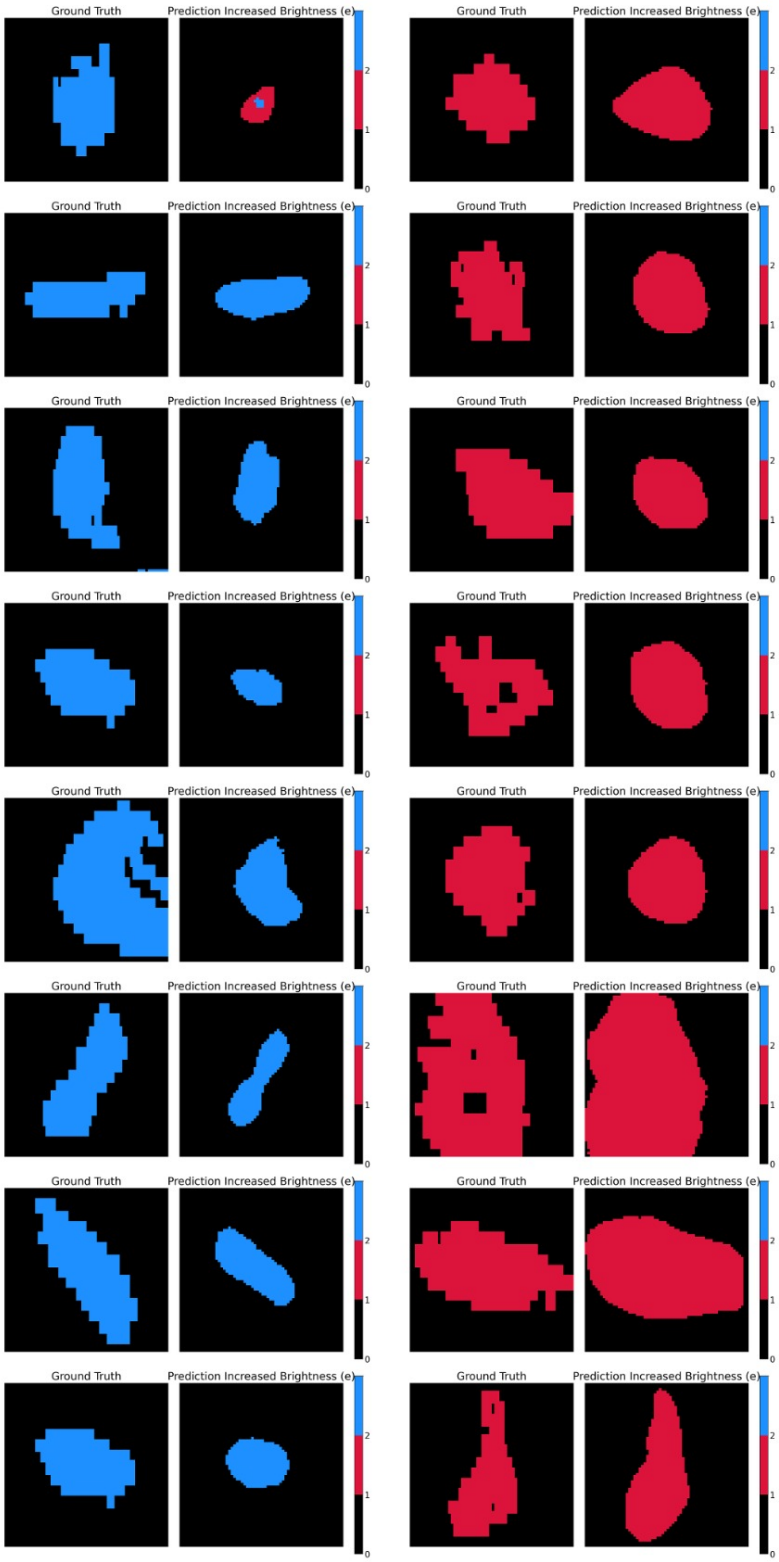


Figure A.5: Prediction for 'Increased Brightness' (e) dataset.

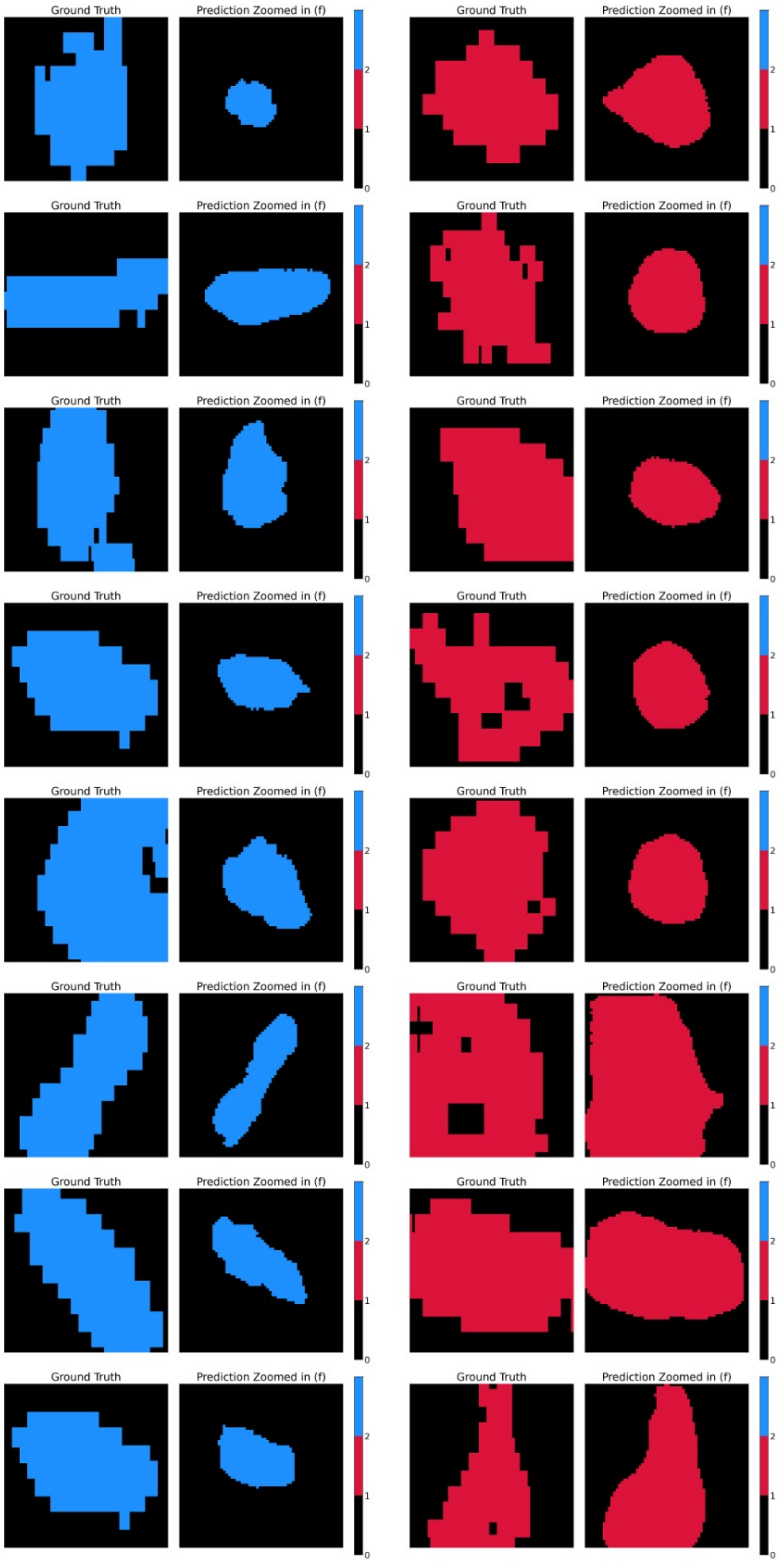


Figure A.6: Prediction for ’Zoomed in ’ (f) dataset.

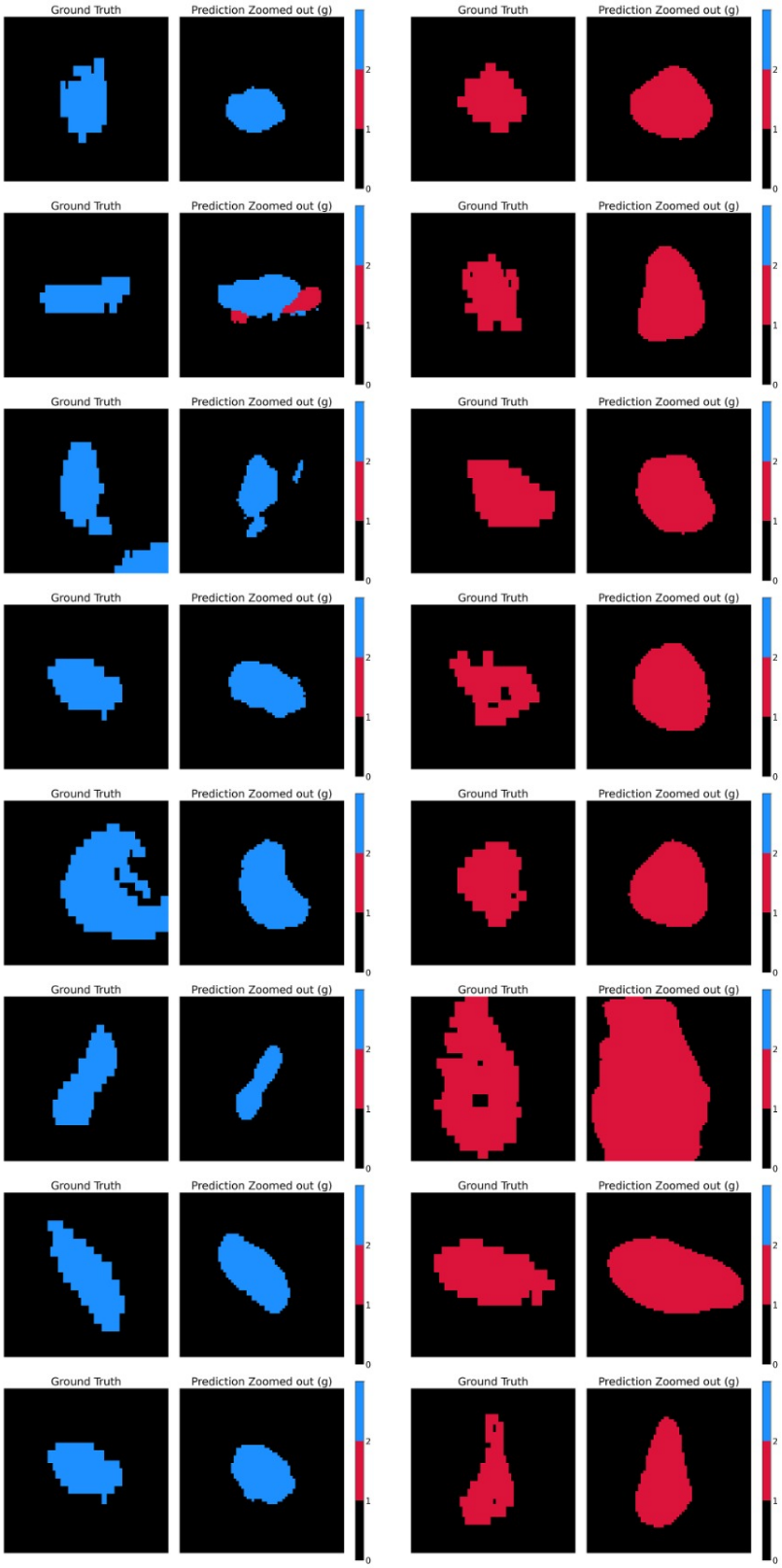


Figure A.7: Prediction for 'Zoomed out' (g) dataset.

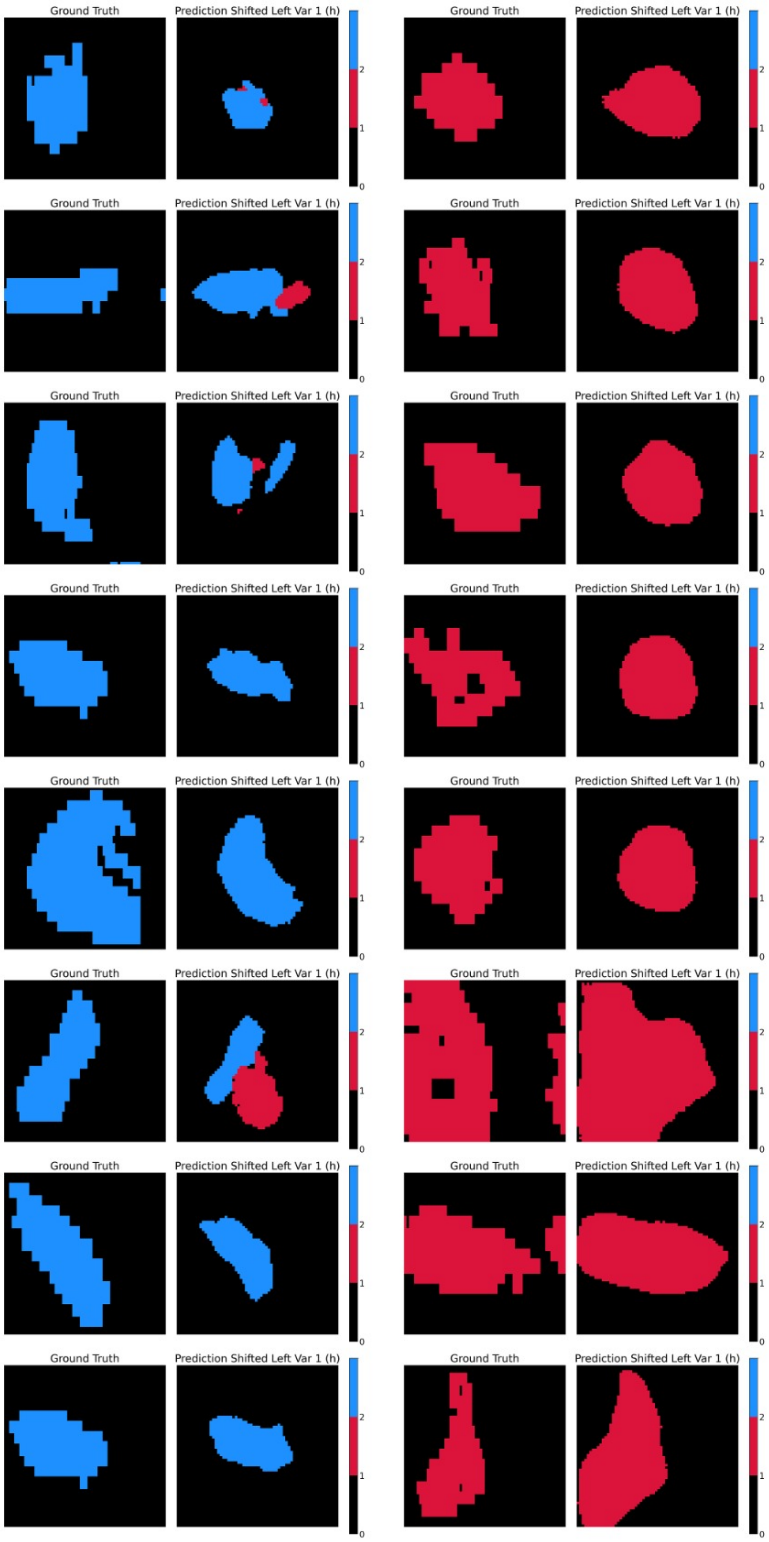


Figure A.8: Prediction for 'Shifted left variation 1' (h) dataset.

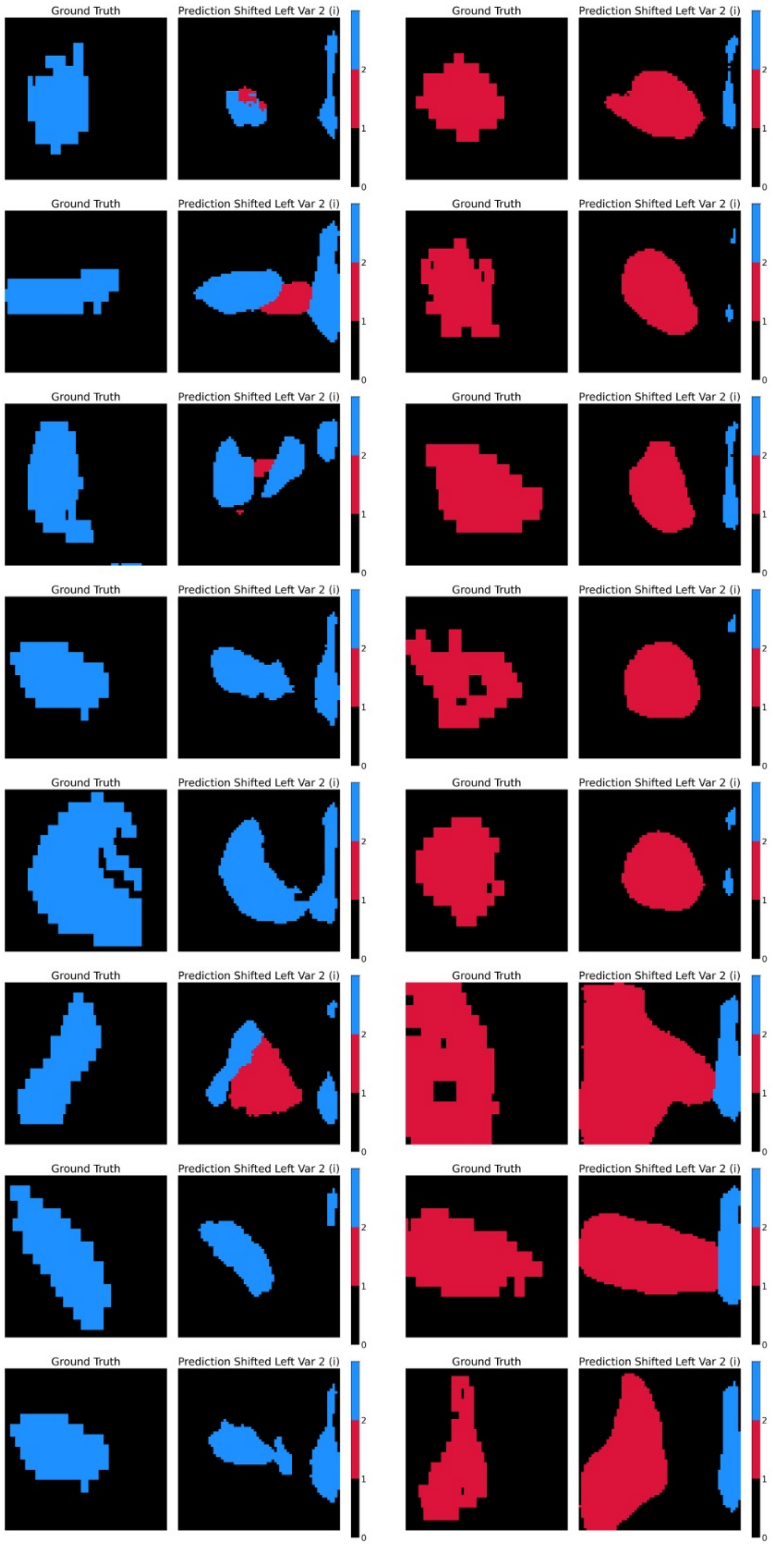


Figure A.9: Prediction for 'Shifted left variation 2' (i) dataset.