Toward Enhanced Sea Ice Parameter Estimation: Fusing Ice Surface Temperature with the AI4Arctic Dataset using Convolutional Neural Networks

by

Lily de Loë

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2025

© Lily de Loë 2025

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Lily de Loë is the sole author of Chapters 5 and 6, which were not written for publication. All chapters were written under the supervision of Dr. Andrea Scott and Dr. David Clausi.

Components of Chapters 1-4 were submitted to OCEANS 2024 - Halifax as part of [1]. Lily de Loë wrote this paper under the supervision of Dr. Andrea Scott and Dr. Linlin Xu. This work and the thesis share common motivations, dataset scenes, architectures, and methods for evaluation. However, the scope of this thesis extends beyond the preliminary study, which is why only parts of these chapters are shared with the published work.

Chapters	**Citation**	**Author's Contributions**
1-4	Lily de Loë, Linlin Xu, and K. Andrea Scott, "Fusing Ice Surface Temperature with the AI4Arctic Dataset for Enhanced Sea Ice Concentration Estimation: A Preliminary Assessment," in *OCEANS 2024 - Halifax*, pp. 1-9, 2024. [1]	Primary author. Developed the methodology, implementation, and analysis.

Abstract

Arctic sea ice mapping is essential for supporting several key applications. These include facilitating safe marine navigation, providing accurate data for climate monitoring, and assisting efforts by remote northern communities to adapt to variable ice conditions. Automated mapping approaches can leverage an abundance of freely accessible satellite data, with the potential to supplement navigational ice charts, improve operational forecasting, and produce high-resolution estimates of sea ice parameters. However, current approaches rely on synthetic aperture radar (SAR) and passive microwave (PM) data, which can struggle to distinguish ice features due to ambiguous textures, atmospheric effects, and sensor limitations.

This thesis explores the potential for thermal-infrared data to improve estimates of sea ice concentration, stage of development, and floe size produced by multi-task deep learning architectures. Work builds on the recent AI4Arctic dataset, which combines Sentinel-1 SAR, AMSR2 brightness temperature, ERA-5 reanalysis data, and ice charts to enhance deep learning-based mapping approaches. VIIRS ice surface temperature (IST) is investigated for its potential to improve predictions in regions where SAR and PM measurements are challenging to interpret.

A VIIRS-AI4Arctic dataset is developed, which consists of 84 scenes, and demonstrates overlap between VIIRS, Sentinel-1, and AMSR2 products. Three variations on the U-Net architecture are introduced, which incorporate IST features at the input- and feature-levels. These models are evaluated against the winning AI4EO AutoICE Challenge architecture, which acts as an AI4Arctic baseline. A SIC accuracy metric is introduced to provide an additional assessment of model performance. Results demonstrate that models incorporating IST consistently reduce classification errors across all three tasks, particularly when identifying open water under conditions with low-incidence angle (SAR), high atmospheric moisture (PM), and wind roughening (SAR and PM). A single, shared decoder improves contextual awareness, although multi-decoder architectures effectively reconstruct task-specific features. The DEU-Net-V architecture, which learns IST features separately from AI4Arctic channels, is most effective at mitigating ambiguity introduced by SAR and PM data. Finally, estimation of aleatoric uncertainty yields heightened variance in marginal ice zones, highlighting potential discrepancies between ice chart labels and pixel-level conditions, and demonstrating the value of quantifying uncertainty from observation noise.

IST ultimately enhances sea ice classification, but is limited by cloud contamination and the resolution of current products. These findings support the continued development of deep learning approaches incorporating IST, and highlight the potential for next-generation thermal-infrared instruments to further improve automated sea ice mapping.

Acknowledgements

I would like to express my gratitude to those who supported me throughout this journey. My undergraduate degree shaped me as an engineer, but my master's gave me the space and time to deepen my love for research.

First, I extend my sincere thanks to my advisors, whose encouragement and guidance have been invaluable. To Dr. Andrea Scott, thank you for matching my desire for rigorous methodology with consistent support and feedback. To Dr. David Clausi, thank you for fostering an environment within VIP that enriches our graduate experiences through seminars, partnerships, and the pursuit of opportunities outside of our theses. Additionally, thank you to my committee members, Dr. Linlin Xu and Dr. Claude Duguay, whose contributions enhanced the quality of this work.

I feel immensely fortunate to have met my colleagues in the VIP Lab and SYDE, among whom I found a strong community. Your willingness to collaborate, share success, and support each other speaks to your calibre as researchers and friends. To the remote sensing group in particular: your insights throughout this process have made all the difference.

I would also like to acknowledge the significant role that financial support plays in developing the next generation of Canadian scientists, engineers, and researchers. Generous scholarships from the Vector Institute, Natural Sciences and Engineering Research Council of Canada, and the University of Waterloo enabled me to pursue this degree. Additionally, this research was funded by Environment and Climate Change Canada, with computing resources provided by the Digital Research Alliance of Canada.

On a personal note, I benefited tremendously from the support of my friends and family during this time. Thank you to my parents, who have been a constant source of encouragement, strength, and wisdom throughout my academic journey. I am incredibly proud that we can share the legacy of earning a master's degree from the University of Waterloo. To my partner, Owen: thank you for your unwavering patience, endless support, and for shaping your future around my goals. It's your turn to pick the next adventure, and I'm excited to see where it takes us.

I remain grateful, as always, for my friends. Thank you for your empathy, interest, and willingness to learn about niche research topics. Lastly, thank you to my primary thesis reader, Java Bean, for only sometimes sitting on my keyboard.

Dedication

This thesis has several dedications:

To my parents, who fostered my intellectual curiosity.

To Owen, who encouraged me to explore my technical capabilities.

And to my grandparents, not all of whom are here, but all of whom would be immensely proud of this accomplishment.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

Dedication vi

List of Figures xi

List of Tables xiv

1 Introduction 1

1.1 Motivations 1

1.2 Limitations in Remote Sensing Data 2

1.3 Thesis Organization 3

1.4 Related Work 4

2 Background 5

2.1 Sea Ice Parameters 5

2.1.1 Ice Charts 6

2.1.2 AI4Arctic Ice Chart Labels . . . 6
2.2 Deep Learning . . . 11
2.2.1 Neural Networks and Deep Learning . . . 11
2.2.2 Convolutional Neural Networks . . . 12
2.2.3 U-Net Architectures . . . 13
2.2.4 U-Net Variants . . . 14
2.2.5 Fusion Approaches . . . 15
2.3 Uncertainty . . . 16
2.3.1 Gaussian Negative Log Likelihood Loss . . . 17

3 Dataset **18**
3.1 Study Region and Significance . . . 18
3.2 AI4Arctic Dataset . . . 20
3.3 VIIRS Dataset . . . 21
3.3.1 Ice Surface Temperature . . . 21
3.3.2 Scene Selection Criteria . . . 23
3.3.3 Data Co-registration . . . 24
3.3.4 Overview of the Co-registered Dataset . . . 25

4 Methodology **27**
4.1 Dataset Organization . . . 27
4.2 Experimental Setup . . . 32
4.2.1 Data Pre- and Post-processing . . . 32
4.2.2 Hyperparameter Tuning . . . 32
4.2.3 Evaluation . . . 33
4.3 Summary of Architectures . . . 34
4.3.1 Shared Specifications . . . 34
4.3.2 Input-level Fusion of IST . . . 36

4.3.3 Feature-level Fusion of IST . . . 36
4.4 Model Evaluation . . . 38
4.4.1 AI4Arctic Metrics . . . 38
4.4.2 SIC Accuracy . . . 43

5 Results and Discussion **46**
5.1 Performance on the VIIRS-AI4Arctic Dataset . . . 46
5.2 Comparison of Architectures . . . 49
5.2.1 Improvements from Input-level Fusion of IST . . . 49
5.2.2 Predictive Differences in Dual-Encoder U-Net Variants . . . 53
5.2.3 Impacts of Task-specific Decoders . . . 54
5.3 Improvements from DEU-Net-V . . . 58
5.3.1 Improvements over Input-level Fusion . . . 58
5.3.2 Impacts of GNLL Loss . . . 61
5.4 Impacts on Sea Ice Parameters . . . 62
5.4.1 Sea Ice Concentration . . . 62
5.4.2 Stage of Development . . . 65
5.4.3 Floe Size . . . 67
5.5 Impacts of Novel Evaluation Methods . . . 70
5.5.1 Sea Ice Concentration Accuracy . . . 70
5.5.2 Uncertainty Quantification . . . 71

6 Conclusion **73**
6.1 Summary . . . 73
6.2 Future Work . . . 75
6.2.1 Expanding on Current Architectures . . . 75
6.2.2 Further Assessing Uncertainty Quantification . . . 75
6.2.3 Leveraging Next-Generation Thermal-Infrared Instruments . . . 76

Letters of Copyright Permission **77**

References **79**

Appendix **89**

.1 Results and Analysis Additional Figures 89

List of Figures

2.1 Example of a Canadian Ice Service (CIS) ice chart. 7

2.2 Illustration of a basic U-Net architecture. 14

3.1 Illustration of the Arctic regions covered by the AI4Arctic dataset. 19

3.2 Example of the original VIIRS scene data (left) and the output of the co-registration process (right) from the west coast of Greenland. 25

4.1 Dataset scene distribution by month (a-g) and region (h). 29

4.2 Comparison of sea ice concentration class distribution for the train-validation and test datasets. 30

4.3 Comparison of stage of development class distribution for the train-validation and test datasets. 31

4.4 Comparison of floe size class distribution for the train-validation and test datasets. 31

4.5 U-Net architecture diagram illustrating the input-level fusion of VIIRS IST with the AI4Arctic dataset, corresponding to AutoICE and AutoICE-V. . . 37

4.6 Diagram of the dual-encoder (DE) U-Net architecture. 40

4.7 Diagram of the tri-decoder (TD) U-Net architecture. 43

5.1 Example ice charts for a March 17, 2019 scene off the west coast of Greenland, demonstrating the difference in masked data between SIC (left), SOD (middle), and FLOE (right). 47

5.2 Summary of AutoICE and AutoICE-V model predictions for 2021-04-19 on the west coast of Greenland. 51

5.3 HH SAR (left), 36.5 GHZ H BT (middle), and IST (right) input channels for 2021-04-19. 51
5.4 Summary of AutoICE and AutoICE-V model predictions for 2018-12-31 in the Hudson straight. 52
5.5 HH SAR (left), 36.5 GHZ H BT (middle), and IST (right) input channels for 2018-12-31. 52
5.6 Summary of DEU-Net-SV and DEU-Net-V model predictions for 2020-01-24 off the coast of Labrador. 54
5.7 HH SAR (left), 36.5 GHZ H BT (middle), and IST (right) input channels for 2020-01-24. 54
5.8 Summary of DEU-Net-V and TDU-Net-V model predictions for 2021-03-01 off the west coast of Greenland. 56
5.9 HH SAR (left), 36.5 GHZ H BT (middle), and IST (right) input channels for 2021-03-01. 56
5.10 Summary of DEU-Net-V and TDU-Net-V model predictions for 2020-01-24 off the coast of Labrador. 57
5.11 SOD model predictions for 2019-01-19 off the coast of Labrador, comparing AutoICE and DEU-Net-V. 59
5.12 Summary of AutoICE, AutoICE-V and DEU-Net-V model predictions for 2020-11-12 on the east coast of Greenland. 60
5.13 HH SAR (left), 36.5 GHZ H BT (middle), and IST (right) input channels for 2020-11-12. 60
5.14 SIC and SIC accuracy model predictions for 2020-01-24 off the coast of Labrador. 61
5.15 Variance map produced by DEU-Net-V-GNLL for 2020-01-24. 62
5.16 Confusion matrices for sea ice concentration: AutoICE, DEU-Net-V, and DEU-Net-V-GNLL. 63
5.17 Confusion matrices for sea ice concentration accuracy: AutoICE, DEU-Net-V, and DEU-Net-V-GNLL. 64
5.18 Confusion matrices for stage of development: AutoICE, DEU-Net-V, and DEU-Net-V-GNLL. 66

5.19 Confusion matrices for floe size: AutoICE, DEU-Net-V, and DEU-Net-V-GNLL. 68
5.20 Comparison of the two scenes containing berg data: 2021-04-19 (left) and 2019-01-29 (right) on the west coast of Greenland. 69
5.21 Variance maps produce by DEU-Net-V-GNLL for 2021-03-01 (left), 2021-04-19 (middle), and 2018-12-31 (right). 72

1 Comparison of the sea ice concentration class distributions for folds 3 and 6. 89
2 Comparison of the stage of development class distributions for folds 3 and 6. 90
3 Comparison of floe size class distributions for folds 3 and 6. 90
4 SIC prediction from DEU-Net-V for 2018-12-31 in the Hudson Strait. . . . 91
5 Comparison of SIC, SOD, and FLOE ice charts for 2019-01-19 off the coast of Labrador. 91

List of Tables

2.1 SIC classes used in the AI4Arctic dataset [2]. 8

2.2 SOD classes used in the AI4Arctic dataset [2]. 9

2.3 FLOE classes used in the AI4Arctic dataset [2]. 10

3.1 Summary of coefficients used in the IST algorithm, by temperature range, for the Northern hemisphere. 22

3.2 Summary of criteria used in evaluating VIIRS data. 23

3.3 Descriptions of the data types and features used by the baseline architecture from Chen et al. [3], in addition to VIIRS IST data. 26

4.1 Summary of acquisition months for the 84-scene dataset. 30

4.2 Unmasked data, reported by percentage (unmasked pixels over total pixels) and number of pixels, for SIC, SOD, and FLOE. 33

4.3 Summary of models, by name and description. 35

4.4 Model specifications corresponding to the AutoICE and AutoICE-V architectures. 38

4.5 Input and output feature map dimensions, by layer, for the AutoICE and AutoICE-V architectures. 39

4.6 Model specifications corresponding to the following architectures: DEU-Net-SV*, DEU-Net-V*, DEU-Net-V-GNLL**, and TDU-Net-V*. 41

4.7 Input and output feature map dimensions by layer for the DEU-Net architecture. 42

4.8 Summary of thresholds associated with the three ice formation classes. . . 44

5.1 Validation and test metrics averaged over the six cross-validation folds, reported as the average and standard deviation. 48
5.2 Weighted validation metrics, averaged over the six cross-validation folds. . 49
5.3 Summary of class-wise F1 scores for SIC, averaged over cross validation results for the test set. 63
5.4 Summary of class-wise F1 scores for SIC accuracy, averaged over cross validation results for the test set. 64
5.5 Summary of class-wise F1 scores for SOD, averaged over cross validation results for the test set. 66
5.6 Summary of class-wise F1 scores for FLOE, averaged over cross validation results for the test set. 68

Chapter 1

Introduction

1.1 Motivations

There is a wealth of free, publicly available remote sensing data, which enables regular monitoring of land, ocean, and atmospheric conditions. Satellite missions led by the European Space Agency (ESA) [4, 5, 6], National Aeronautics and Space Administration (NASA) [7, 8, 9, 10], Canadian Space Agency (CSA) [11, 12], and Japanese Aerospace Exploration Agency (JAXA) [13], have been instrumental in collecting and disseminating data to improve Arctic sea ice mapping. These efforts support several key applications:

Marine navigation: Diminishing ice extent contributes to increasing variability in Arctic conditions [14, 15], which necessitates accurate mapping of sea ice cover to support safe marine navigation. Further, loss of sea ice cover is expected to expand the season where ships can navigate the Arctic [16]. These trends would increase demand for navigational aids, such as ice charts, which are time- and resource-intensive products. Thus, alternative mapping approaches are required to supplement ice charts, as well as to support strategic planning activities with shorter lead times [16].

Environmental monitoring: Seasonal ice loss is exacerbated by the ice-albedo feedback effect, wherein ice melt exposes the darker ocean surface, decreasing surface reflectivity. In turn, heat absorption increases, contributing to further melting [17]. Among other factors, this feedback loop contributes to Arctic amplification, where this region experiences temperature changes at rates that exceed the global average [18]. Thus, our ability to accurately monitor changes in Arctic sea ice is important for understanding and forecasting climate change [19].

Climate adaptation: Increasing variability in Arctic sea ice presents challenges for remote northern communities. Changing ice dynamics make conditions unpredictable, which increases the risks of encountering hazards during on-ice travel [20]. Further, sea ice acts as habitat for endemic species [21]; changes to regional ecosystems, driven by reduced ice extent, contribute to food insecurity and loss of traditional practices [22, 23]. In response to these challenges, advancements in Arctic mapping can be used to support Indigenous-led community initiatives by augmenting traditional knowledge [20].

Advancements in automated sea ice monitoring, driven by machine learning, are increasingly important for supporting marine navigation, climate modelling, and remote northern communities. The recent AI4Arctic dataset combines Sentinel-1 synthetic aperture radar (SAR) imagery, AMSR2 brightness temperature (BT) measurements, ERA-5 reanalysis data, and ice charts to improve deep learning-based mapping approaches [24]. This dataset marks a crucial step forward in the field's ability to compare deep learning approaches, and incorporate measurements from different sensors to improve sea ice parameter estimation. Nevertheless, AI4Arctic does not include imagery from the optical, near, or mid-infrared -ranges of the electromagnetic spectrum. In particular, thermal infrared data may provide insights into ice age and thickness [25, 26]. Further, current products (e.g., MODIS [8] and VIIRS [9]) offer higher resolution measurements than passive microwave (PM) imagery. Thus, it is critical to explore the use of these products, to determine if thermal infrared data, in combination with SAR and PM measurements, can improve the prediction of sea ice parameters.

1.2 Limitations in Remote Sensing Data

Algorithm-based sea ice concentration products exist, which support forecasting efforts and supplement ice charts. Passive microwave-derived sea ice maps (e.g., AMSR2 ASI sea ice concentration data [27]) are important tools in operational planning and climate research. More recently, daily mean SIC maps have emerged, which combine predictions from thermal-infrared and passive microwave [28], and optical and SAR data [29]. Additionally, products that use solely SAR, such as the approach used by Environment and Climate Change Canada [30], offer finer spatial resolution estimates. Recent advancements in deep learning have demonstrated promising results in estimating the three sea ice parameters relevant to ice charts: sea ice concentration (SIC), stage of development (SOD), and floe size (FLOE) [31]. These studies leverage convolutional neural networks [32, 33], U-Net architectures, [34, 35, 36, 3], and U-Net variants [37].

The works discussed above predominantly use SAR and passive microwave data; how-

ever, both present limitations. SAR effectively balances a finer spatial resolution that is sensitive to changes in ice conditions, with swath size and data storage considerations. Further, its measurements are not reliant on reflected solar radiation. SAR imagery measures the backscatter coefficient of the emitted radar pulse, which is dependent on surface roughness, incidence angle, and the physical and dielectric properties of the target medium [38]. Consequently, SAR measurements of wind-roughened water, smooth ice, and surface melt may produce ambiguous textures and backscatter responses, which can be misinterpreted [34]. The dielectric properties of snow on the ice surface can also influence backscatter mechanisms, with brine-wetted snow pack contributing to ambiguity [39]. Additionally, backscatter signatures for ice and water vary based on incidence angle [40]; this variation results in the appearance of open water and different ice types changing by scene.

Passive microwave sensors record brightness temperature, which is commonly used for estimating SIC due to its distinct contrast in ice-water measurements [32, 41, 42, 43]. Further, surface microwave radiation is capable of penetrating clouds at lower frequencies, and instruments operate independently of incident solar radiation [44]. However, the coarse resolution of these sensors (on the order of kilometres) makes them ineffective in regions that require high spatial resolution, such as near land or narrow waterways [45]. Similarly to SAR, open water measurements can be impacted by wind-roughening [44]. Studies have found that SIC estimates using passive microwave data are biased for thin ice due to low brightness temperature emissivity [46, 47]. Additionally, at certain frequencies, these sensors are sensitive to emittance from atmospheric moisture. Thus, machine learning models may misinterpret readings from specific channels, and must rely on learning features from other data sources to correct for inaccurate surface measurements [34, 44].

1.3 Thesis Organization

This thesis investigates if incorporating VIIRS ice surface temperature (IST) data can enhance predictions of SIC, SOD, and FLOE in regions that pose challenges for SAR and PM data. Although thermal-infrared data relies on cloud-free conditions to obtain accurate measurements [48], the potential benefits of the product's resolution and ability to distinguish ice thickness presents an opportunity to improve predictive accuracy where data is present. This capacity for IST data to improve sea ice parameter predictions is assessed against the performance baseline for the AI4Arctic dataset, and investigated using a variety of deep learning architectures and data fusion approaches.

The remainder of this thesis is structured as follows:

- Chapter 2 provides background on ice charts, sea ice parameters, and deep learning approaches relevant for understanding the study.
- Chapter 3 introduces the study region, remote sensing data types, and procedure for creating a co-registered dataset for VIIRS IST and AI4Arctic data.
- Chapter 4 describes the experimental setup, deep learning architectures, and process for evaluating model predictions. Additional evaluation methods are introduced, on top of those used by the performance baseline.
- The results of these experiments are discussed in Chapter 5. Models are assessed using a combination of evaluation metrics and visual analysis of the predicted sea ice maps. Predictions are evaluated against similar architectures, followed by comparisons to the best performing model. Finally, strengths and weaknesses are discussed for model predictions of the three sea ice parameters and new evaluation methods.
- Chapter 6 concludes the thesis. A discussion of insights and limitations from the study is used to guide future work.

1.4 Related Work

This study assesses the impacts of IST data against the multi-task U-Net architecture introduced by Chen et al. [3], which estimates SIC, SOD, and FLOE. This architecture won the Artificial Intelligence for Earth Observation (AI4EO) AutoICE Challenge in 2023, which evaluated the accuracy of multi-task deep learning models, with the goal to advance automated sea ice mapping. The model achieved the highest overall score (86.39%), SIC R^2 score (92.02%), and SOD F1 score (88.61%) [24]. These results were achieved training on the full, 533 scene AI4Arctic dataset, which includes 20 held-out test scenes. This model's performance on the AI4Arctic dataset provides a baseline comparison for sea ice maps incorporating VIIRS IST data.

The work presented in this thesis builds on a preliminary study with a 30-scene dataset, which verified the potential for VIIRS ice surface temperature data to improve sea ice concentration estimation. The method for dataset co-registration is consistent across studies, and the original 30 scenes are included in the VIIRS-AI4Arctic dataset discussed below. Additionally, both works compare to the Chen et al. [3] architecture, as well as this architecture incorporating IST data. Results from the preliminary study are not discussed in the thesis body, but can be reviewed in de Loë et al. [1].

Chapter 2

Background

2.1 Sea Ice Parameters

Sea ice refers to frozen sea water, which forms a barrier between the ocean and atmosphere [14]. Sea ice can either form with a direct connection to a shoreline (land-fast ice) or be free floating (ice pack). This formation–referred to as freeze-up–begins in early fall; conversely, the break-up of ice begins in the spring, with melt accelerating throughout the summer [21, 14]. While ice freeze-up can be tracked through the formation of young ice and subsequent progression through the stages of development (Table 2.2), melt is characterized by degradation in the ice pack, ranging from surface puddles to flooded ice [31]. In the context of this study, sea ice is characterized in terms of three parameters: sea ice concentration, stage of development, and floe size.

Sea ice concentration: SIC, which measures the percentage of total ice cover in a region, is an important indicator in climate monitoring. In ice charts, the total concentration is the sum of partial concentrations contained in a polygonal region, reported in tenths [31]. This differs from SIC estimates derived from satellite products (e.g., passive microwave) using algorithmic or data-driven methods, which report gridded concentrations at the resolution of the instrument. Accurate estimates of SIC can be used to estimate sea ice extent, which is a key climate change indicator for the Arctic region [49]. Further, high resolution maps of SIC are effective at illustrating the marginal ice zone (MIZ): the transitionary region between full ice-cover and open water [50].

Stage of development: SOD relates to the age and thickness of ice, making it an important parameter in marine navigation and the design of safe shipping routes [24].

Long-term measurements of ice age and thickness are also valuable for climate monitoring and modelling [19].

Floe size: FLOE characterizes the degree of break-up of sea ice, which acts as a mediator in the exchange of momentum between ocean and atmosphere [51]. The impacts of floe size are particularly noticeable in the MIZ, where small floes accelerate the lateral melt rate [17].

2.1.1 Ice Charts

Information on sea ice conditions is disseminated using ice charts, such as the example in Figure 2.1. The Canadian Ice Service (CIS) produces daily and regional ice charts. Daily charts provide an estimate of ice conditions, and are intended for short-term (24 hour) strategic planning of marine activities. These charts leverage software-based mapping and image analysis tools, and use a combination of remotely sensed data and ground truthing (e.g., from ships, helicopters, or aircraft). When available, image analysis charts, which interpret conditions from remote sensing data (predominately radar), are also used. Regional charts are prepared on a weekly to monthly basis, and provide an analysis of ice conditions for a specific date. These charts are used for advance planning initiatives (on the scale of weeks), incorporating all available data within approximately three days of the chart date. In addition to the resources described above, ice analysts rely on meteorological data and historical patterns to inform their analysis. [31]

Ice charts report conditions in polygonal regions, which are defined in terms of total concentration, partial concentration, stage of development, and floe size. This information is reported using egg codes (magnified in Figure 2.1). Total concentration for the polygon is reported in the top row of the oval, while partial concentrations in the next row are reported from the thickest (leftmost column) to third thickest ice conditions (rightmost column). During freeze-up and break-up, the concentration associated with the fourth thickest ice can be included outside of the oval on the right side, under additional groups (Figure 2.1). Sets of stage of development and floe size types are reported in the chart, each associated with a partial concentration. Additional SOD symbols, each with an associated floe size, can be reported in exceptional cases. [31]

2.1.2 AI4Arctic Ice Chart Labels

The AI4Arctic dataset provides pixel-level ice chart labels for total SIC, dominant SOD, and dominant FLOE. Dominant SOD and FLOE classes are assigned if there is a dominant

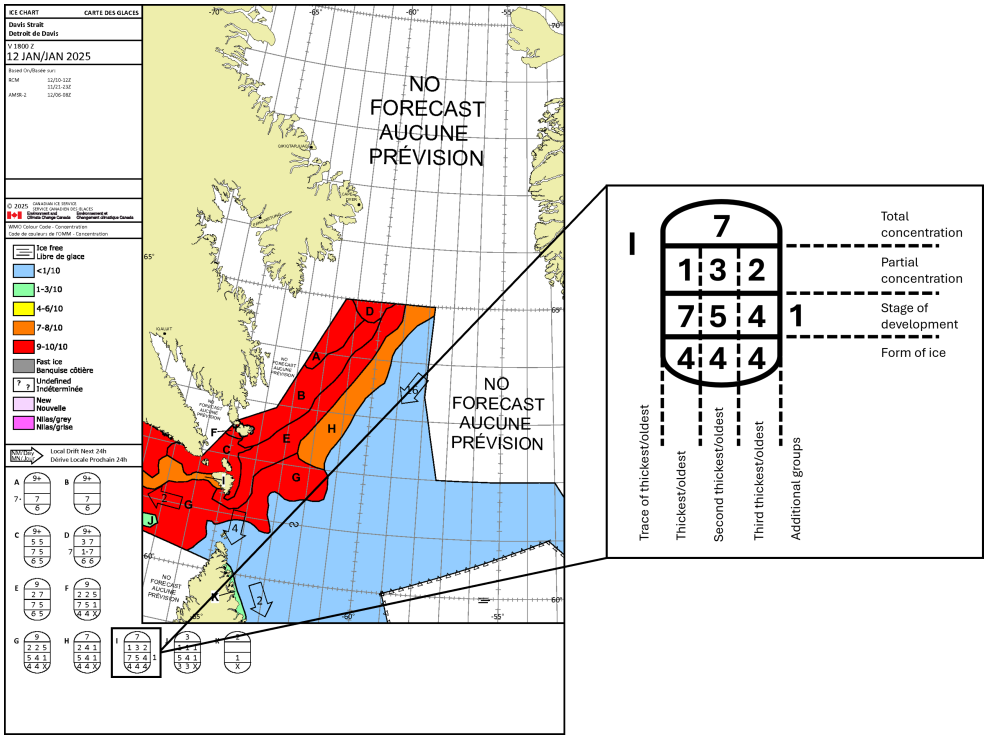


Figure 2.1: Example of a Canadian Ice Service (CIS) ice chart, Daily Ice Chart colour WMO CT - Davis Strait - WIS33CT - 2025/01/12, overlaid with a magnified egg code. Adapted from [52]. Available: https://ice-glaces.ec.gc.ca/prods/WIS33CT/20250112180000_WIS33CT_0013407737.pdf

partial concentration in the polygon, representing ≥65% of the region. Otherwise, SOD and FLOE types are not assigned. Ice charts are co-registered with Sentinel-1 SAR imagery, and an egg code's information is applied to every pixel contained in that polygon. This data is an invaluable resource for training deep learning models to estimate these three parameters, incorporating insights that may not be evident from remotely sensed data. However, conditions at a single pixel in a SAR image may deviate from the ice chart labels due to their total/dominant class assignments. Further, these labels may exacerbate existing uncertainty in ice concentration estimates, wherein ice analysts overestimate low concentrations, and report greater variability for mid-range concentrations [53]. Ultimately, while these limitations may introduce discrepancies between labels and pixel-level remote sensing measurements, data pre-processing can be leveraged to mitigate these impacts (Section 4.2.1).

In AI4Arctic, SIC is reported as 12 discrete classes, ranging from open water (0%) to full ice-cover (100%) in tenths, and including fast ice. SOD is reported as either open-water or one of five classes: new ice, young ice, thin first-year ice, thick first-year ice, and old ice. Similarly, FLOE is described as either open-water or one of six floe sizes: cake ice, small floe, medium floe, big floe, vast floe, and bergs [2]. These classes are summarized in Tables 2.1, 2.2, and 2.3, with accompanying descriptions.

Table 2.1: SIC classes used in the AI4Arctic dataset [2]. Note: sub-classes are assigned the label of their broader sea ice concentration.

SIC	**Sub-class**	**Description**
Ice Free		0% SIC
	Less than 1/10	Negligible ice presence
	Bergy water	Open sea; used in DMI ice charts because icebergs can appear in anywhere Greenland waters
1/10		10% SIC
2/10		20% SIC
3/10		30% SIC
4/10		40% SIC
5/10		50% SIC
6/10		60% SIC
7/10		70% SIC
8/10		80% SIC
9/10		90% SIC
Full Ice Cover		100% SIC
	Fast Ice	10/10; ice that has formed and fastened to the coast
	Fully Compacted Ice	9+/10 ;only used in DMI ice charts

Table 2.2: SOD classes used in the AI4Arctic dataset [2]. Note: sub-classes are assigned the label of their broader stage of development, and glacier ice is excluded from the labels.

SOD	**Sub-class**	**Description**
Open-water		Ice free/no stage of development
New Ice		Recently formed; thickness < 10 cm
	Nilas Ice	Thickness < 10 cm
Young Ice		Thickness of 10-30 cm
	Grey Ice	Thickness of 10-15 cm
	Grey-White Ice	Thickness of 15-30 cm
Thin First Year Ice		Persistence of no more than one winter; thickness of 30-70 cm
	Stage 1 First Year Ice	Thickness of 30-50 cm
	Stage 2 First Year Ice	Thickness of 50-70 cm
	Medium First Year Ice	Thickness of 70-120 cm
Thick First Year Ice		Persistence of no more than one winter; thickness > 120 cm
Old Ice		Persistence beyond one summer's melt

Table 2.3: FLOE classes used in the AI4Arctic dataset [2]. Note: sub-classes are assigned the label of their broader floe size. Floes < 2 m (pancake ice, small ice cakes/shuga, and brash ice) and fast ice are both excluded from the labels.

Floe Size	**Sub-class**	**Description**
Open-water		Ice Free
Cake Ice/Ice Cake		Size $<$ 20 m across
Small Floe		Size 20-100 m across
Medium Floe		Size 100-500 m across
Big Floe		Size 500 m-2 km across
Vast Floe		Size 2-10 km across
	Giant Floe	Size $>$ 10 km across
Icebergs		Glacier break-off; irregular in shape and floating $\geq$ 5 m above sea level
	Growlers, Floebergs, or Floebits	Smaller ice fragments/iceberg variants

2.2 Deep Learning

This thesis investigates methods of fusing IST data within the U-Net architecture. The prediction of SIC, SOD, and FLOE using these architectures is classified as a multi-task, supervised learning problem:

- **Multi-task learning:** models are trained to perform multiple, concurrent tasks (e.g., prediction of SIC, SOD, and FLOE).
- **Supervised learning:** models are trained on a set of "ground truth" labels (ice charts) to complete a task (generate sea ice maps).

The following subsections provide comprehensive background on machine learning and deep learning concepts related to this thesis.

2.2.1 Neural Networks and Deep Learning

Convolutional neural networks (CNNs) remain a leading deep-learning approach for handling image problems. Fundamentally, this architecture extends artificial neural networks (ANNs): sequential layers of connected units, each functioning as an artificial "neuron", with learnable weights. A basic network consists of input and output layers, in addition to at least one hidden (or intermediate) layer. The output of a neuron, o_i, at layer, l, is described by Equation 2.1:

$$o_i^l = f(\sum_{j=1}^{N} w_{ij} o_j^{l-1} + b_i) \tag{2.1}$$

This relationship is dependent on the output of each connected neuron, o_j, in the previous layer, $l-1$, in addition to its associated connection weight w_{ij}, and a bias, b_i, which offsets the result. This result is passed through an activation function, f, which introduces non-linearity, allowing the model to learn more complex relationships. The field of deep learning refers to the subset of machine learning that leverages these multi-layered neural networks. [54]

Deep learning consists of four key components: forward propagation, loss functions, back propagation, and optimizers. Forward propagation passes input data through the network layers using the relationship in Equation 2.1. When training supervised models,

the output prediction is compared to the correct result (referred to as the labels or ground truth) using the loss function. The backpropagation algorithm then calculates how much each model parameter (weight and bias) contributed to the error. It completes this task by computing the gradient of the loss function with respect to each parameter, propagating the error backwards through the network. The optimizer then adjusts the model parameters to minimize the loss–effectively learning from the model's last mistake. During inference (evaluation), forward propagation is used on the trained model to achieve a prediction, but backpropagation is not required. [54]

2.2.2 Convolutional Neural Networks

CNNs address the efficiency limitations of using fully connected ANNs for image data. Specifically, CNNs utilize the two-dimensional convolution: a linear operation defined in Equation 2.2.

$$(f * g)[i, j] = \sum_{m=-M}^{M} \sum_{n=-N}^{N} f[i - m, j - n]g[m, n] \tag{2.2}$$

Here, f is a two-dimensional image represented as a matrix of dimension MxN, and g represents the convolutional kernel. The kernel–otherwise known as a filter–is a matrix (commonly size 3x3) of translation-invariant weights, which is centred at index $[i, j]$.[1] The convolution operation produces a new matrix, referred to as a feature map, by sliding the kernel across the image; the output of the convolution at each index $[m, n]$ is the sum of element-wise products of the input, f, with the weights, g. If the border of the image is padded (e.g., with zeros), then the output of this operation maintains the same size as the input matrix. [55]

Comparing to an ANN, a CNN's neurons model local interactions between neighbouring pixels using the sliding kernel. This local connectivity, combined with the kernel's shared weights, improves the architecture's efficiency by reducing the number of model parameters [56]. For example, consider a standard convolution on a 256x256 pixel image using a single 3x3 kernel: Here, we have 256^2 input features and 3^2 output features. While a convolutional layer uses only nine parameters, a fully connected layer would require 256^2*9 parameters because each output feature would depend on every input feature.

[1] Translational invariance is the property where an operator produces the same output regardless of if the subject is shifted or translated.

A standard CNN architecture is composed of convolutional layers, each with a number of filters, which are used to learn features. Early layers are used to extract low-level features (e.g., edges), while deeper layers identify complex patterns. CNNs commonly use pooling layers, which reduce the spatial dimensions (and thus the parameters) of input data via an aggregation operation.[2] With subsequent layers, the region–or receptive field–that influences a feature increases, leading the neuron to capture broader contextual information [55]. Building on ANNs, this architecture leverages non-linear activation functions, typically followed by batch normalization to regularize the model and improve training stability. Fully connected layers are also commonly used as the final layer of the network to perform a classification based on the features extracted by earlier layers.

2.2.3 U-Net Architectures

Early deep learning approaches to sea ice mapping leveraged CNN architectures. Contributions include Wang et al. [57] (estimation of SIC from SAR imagery), Boulze et al. [33] (estimation of SOD from SAR imagery), and Malmgren-Hansen et al. [32] (estimation of SIC from SAR and passive microwave data). However, sea ice mapping can be classified as a segmentation task, and the field has progressively shifted to adopt the U-Net architecture.[3]

U-Nets are an extension of fully convolutional networks, which use locally connected layers. Originally designed for segmentation in medical imaging [58], the architecture (Figure 2.2) is characterized by its U-shaped structure, which is constructed from its encoding and decoding paths. The encoder path uses contracting layers to extract important features, while the decoder path up-samples the latent representation to its original dimensions, producing a segmentation mask. Skip connections between encoder and decoder blocks concatenate feature maps from the encoder path. This process prevents information loss during down-sampling by combining low-level features from the encoder with high-level semantic information. [58]

U-Net architectures have been used to estimate sea ice cover from SAR, improving the accuracy of estimates in the MIZ [34, 35]. More recently, the AI4SeaIce article series [36, 59, 60, 61] established the basis for the use of U-Nets in the AI4EO AutoICE Challenge [24]. As a result of this competition, which drew 129 entries, U-Net architectures such as

[2] Averaging, returning the maximum value, etc.

[3] Segmentation is the grouping of pixels into discrete regions for object recognition and other related tasks.

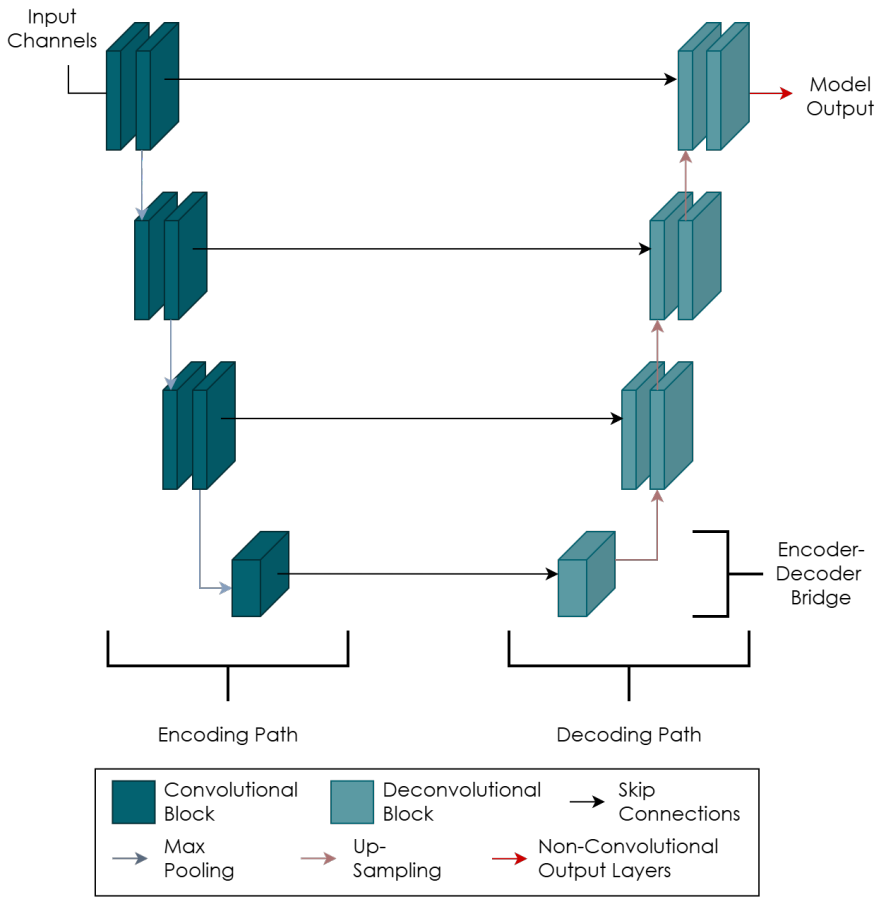


Figure 2.2: Illustration of a basic U-Net architecture.

Chen et al. [3] have emerged as a leading method for automating sea ice mapping based on multimodal remote sensing data.

2.2.4 U-Net Variants

Variations of the U-Net's encoding and decoding paths are also being explored as a method for improving the architecture. This includes the use of multiple encoders for fusing multimodal data, and the use of multiple decoders for multiple tasks. In the field of sea ice mapping, Rogers et al. [37] have demonstrated the successful application of parallel encoders for fusing and concatenating remote sensing data at different spatial resolutions for

ice-water classification. Their ViSual_IceD architecture outperforms the input-level fusion of SAR and multispectral imagery (MSI) on binary accuracy, precision, and F1 score.[4] Further, the field of medical image segmentation continues to produce advancements in the U-Net architecture. For example, Mahbod et al. [62] introduce a dual-decoder U-Net for nuclei instance segmentation. The decoders simultaneously predict foreground and distance maps for nuclei in hematoxylin- and eosin-stained histological images. These outputs are then merged to improve segmentation performance for nuclei classification. Because ice chart polygons report shared regions of consistent SIC, SOD, and FLOE, further work is necessary to explore if segmentation can be improved by adopting a multi-decoder approach.

2.2.5 Fusion Approaches

In the context of deep learning, data fusion refers to a model using data from multiple input sources to make predictions. The benefit of data fusion for applications in remote sensing is that models can obtain predictions based on multiple different sensor measurements, overcoming the limitations of a single product. Fusion methods can be classified as one of three approaches: input-level, feature-level, and decision-level.

Chen et al. [3] employ input-level data fusion, where the model combines raw input channels prior to feature selection. Using this approach, the model learns features based on the combined data, which may capture complementary information that cannot be extracted from a single input [63].

Conversely, feature-level fusion combines features learned from separate inputs, in different parts of the network. This approach is beneficial for fusing inputs from different data sources, allowing the model to better identify salient features. Malmgren-Hansen et al. [32] and Rogers et al. [37] both employ feature-level fusion: each model accepts input channels with different resolutions, which are combined to make pixel-wise predictions. The success of these models in fusing high resolution SAR imagery with coarser passive microwave measurements [32], and multi-spectral imagery [37] demonstrates the benefits of this approach for satellite products.

Finally, decision-level fusion combines results from multiple classifiers, often with a weighting scheme, to improve the final decision for a single task [63]. For example, Wang et al. [64] use a Bayesian framework to combine ice-water information from SAR data with

[4]Input-level fusion of SAR and MSI using the traditional U-Net architecture with one encoding and one decoding branch.

the AMSR-E daily ice concentration product, improving the detail and accuracy of SIC estimates. This approach is also common in algorithm-based satellite sensor fusions, which do not use machine learning. Both Ludwig et al. [28] and Konig et al. [29] (Section 1.2) merge SIC estimates obtained from multiple remote sensing products.

2.3 Uncertainty

Observation noise between remotely sensed measurements, ice chart annotations, and real-world conditions introduces uncertainty in predictions from machine learning models. This statistical uncertainty is inherent to the data, and is referred to as aleatoric uncertainty [65]. Noise associated with aleatoric uncertainty can either be constant (homoscedastic) or data-dependent and spatially-varying (heteroscedastic). Conversely, epistemic uncertainty refers to uncertainty in the model's weights, and can be reduced by improving the dataset or model. Estimating aleatoric uncertainty allows for a more discerning approach to utilizing model predictions in safety-critical and scientific applications (e.g., navigation and climate modelling) because it acknowledges the noise inherent to remote sensing data. Consequently, uncertainty is increasingly considered when applying deep learning approaches to Arctic mapping [66, 44, 65].

Quantification of heteroscedastic aleatoric uncertainty requires estimating the variance associated with a model prediction. One method for doing so is to use either log-likelihood or negative log-likelihood estimation. Labels are treated as samples in a Gaussian distribution, with an associated mean and variance that the model predicts. A high log-likelihood (or conversely, a low negative log-likelihood) represents a better fit for the training data, and is associated with lower uncertainty. For image segmentation, the negative log-likelihood is calculated as the average result of each pixel. Whether the task is a regression or a classification problem (e.g., SIC versus SOD and FLOE) changes the loss function required to predict the mean of the distribution. For example, Gaussian negative log likelihood (GNLL) loss, used for regression tasks, is described below. This loss function has previously been applied to modelling SIC uncertainty, with the highest uncertainty being reported along the MIZ and ice edge [44, 66].

An alternative method for estimating variance with classification problems is to use the probabilities from the task's softmax output layer. For example, Chen et al. [66] leverage the method proposed by Kwon et al. [67] for decomposing the variance associated with the predictive distribution into aleatoric and epistemic uncertainty. The predictive uncertainty of their ice-water classification is then obtained from this variance.

2.3.1 Gaussian Negative Log Likelihood Loss

In the context of log likelihood estimation, mean squared error (MSE) loss (Equation 2.3) can be used to predict the mean of a Gaussian distribution under the homoscedastic assumption:

$$L_{MSE} = 1/N \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2.3}$$

where N refers to the number of pixels in the scene, y_i is the true label at pixel i, and $\hat{y}_i$ is the model prediction [24]. Conversely, heteroscedastic variance can be estimated using the GNLL loss function (Equation 2.4), defined by Kendall et al. [68] as:

$$L_{GNLL} = 1/N \sum_{i=1}^{N} 1/2(log(\hat{\sigma}_i^2) + (y_i - \hat{y})^2/\hat{\sigma}_i^2) \tag{2.4}$$

where N refers to the number of pixels in the scene, y_i is the true label at pixel i, $\hat{y}_i$ is the predicted mean of the distribution, and $\hat{\sigma}_i^2$ is the predicted variance. Use of this loss function requires adapting the model output. For MSE loss, the network's output neuron returns the mean of a Gaussian distribution with constant variance. For GNLL loss, the network now outputs two neurons: the predicted mean and variance.

In practice, Kendall et al. recommend that the network is trained to predict log variance, given by Equation 2.5:

$$s_i = log(\hat{\sigma}_i^2) \tag{2.5}$$

Training on the log variance provides greater numerical stability by avoiding a division by zero in the loss term. An exponential mapping of $exp(-s_i)$ is then applied prior to calculating the loss. This mapping constrains variance values predicted by the network to the positive domain, which is necessary for valid solutions [68].

Chapter 3

Dataset

This chapter provides an overview of the datasets used in this study. Specifically, regional coverage of the Arctic, remote sensing products, and dataset selection are discussed. Because this study builds on the existing AI4Arctic dataset, the study region is discussed in terms of the full dataset. However, only a subset of these scenes are co-registered with ice surface temperature data from the VIIRS instrument. Thus, the chapter closes with an overview of the VIIRS-AI4Arctic dataset.

3.1 Study Region and Significance

The AI4Arctic dataset covers the charting regions of the Canadian and Greenland Ice Services (CIS and GIS). Coverage can be broken into six Arctic regions: the Central Arctic, Canadian Arctic Archipelago, Baffin Bay, Hudson Bay, Beaufort Sea, and Greenland Sea, which are seen illustrated in Figure 3.1 [2]. Differing ice conditions, ecosystems, and climatological trends necessitate sea ice monitoring in these regions. Further, regional differences in ice types and trends (e.g., freeze-up, break-up, and melt) within the dataset allow for a comprehensive assessment of mapping strategies.

Recent studies in the Hudson Bay demonstrate changing freeze-up and break-up trends, while regional factors (complex coastal orientations and bathymetry) are seen to influence the persistence of fast ice on the bay's eastern shore [69]. Climatological trends influence the adjacent Foxe Basin, where diminishing land-fast ice poses a major threat to the food security and cultural practices of Indigenous communities [23]. Similarly, the Baffin Bay region is experiencing delayed and reduced seasonal ice formation, driven by warming



Figure 3.1: Illustration of the Arctic regions covered by the AI4Arctic dataset.

surface currents originating from the Atlantic ocean [70]. Currents in this region then flow south along the west coast of Baffin Island towards Newfoundland, which is significant due to its seasonal ice formation and melt [71].

The high Arctic, including the Canadian Arctic Archipelago and Northern Greenland, features thicker, multi-year and land-fast ice. Diminishing sea ice cover is a threat to ecosystems and wildlife, and these areas may become one of the last remaining refuges for species [71]. Significant reductions in sea ice extent and multi-year ice are also seen in the mid Arctic, including the Beaufort Sea [71]. The region's ecological importance and role in regulating climate variability necessitates the study of these concerning long-term trends [72]. Above Greenland, the dataset extends into the Central Arctic, which experiences

less seasonal variability and is characterized by its more consistent ice cover [71]. Finally, Greenland encompasses a diverse range of sea ice types, including multi-year (Central East, Central West, Northeast, North), seasonal (Southeast, Central West, Northeast), and land-fast ice (Northwest, Qaanaaq). Greenland is influenced by sea ice dynamics and is negatively impacted by the loss of multi-year ice, which has consequences for weather patterns, ecosystems, and communities [73].

The unique characteristics of these Arctic regions, including the presence of varying ice types (e.g., thin and land-fast ice), their distinct regional significance, and the potential for comparative analysis with other studies [3, 36, 59, 60, 61], supports the use of the AI4Arctic dataset. The Central Arctic, Canadian Archipelago, Baffin Bay, Hudson Bay, Beaufort Sea, and Greenland Sea offer distinct conditions under which to evaluate automated sea ice mapping approaches. However, use of AI4Arctic scenes is restricted to those with spatially and temporally consistent, high quality measurements from the VIIRS instrument. Thus, Section 3.3.4 provides an updated summary of the regions covered in this study.

3.2 AI4Arctic Dataset

The AI4Arctic dataset includes SAR imagery and auxiliary data, brightness temperature (BT) measurements, reanalysis data, and ice charts from the Canadian and Greenland Ice Services, which are used as labels [2]. The dataset is comprised of 533 co-registered scenes from January 2018 to December 2021; the frequency of acquisitions ranges from multiple scenes a day to a couple scenes a month, and does not evenly survey the six Arctic regions.

The AI4Arctic ready-to-train dataset, used in this study, provides data that is normalized using the mean and standard deviation. AMSR2 brightness temperature measurements and ERA5 reanalysis data are gridded to a 2 km resolution. The ice charts, distance maps, incidence angle measurements, and dual polarized Sentinel-1 SAR images are gridded to 80 m pixel spacing, which is downsampled from the original SAR pixel spacing of 40 m. Gridded latitude and longitude coordinates, as well as the acquisition date and time, are provided for reference and use as spatial-temporal encodings. These products are discussed in greater detail below; a subset of these data are used in the study (Table 3.3).

Sentinel-1 Synthetic Aperture Radar

The AI4Arctic dataset uses Sentinel-1 HH and HV, Extra Wide Swath images. These images report the backscatter coefficient in dB, and are noise corrected using the NERSC

algorithm described in Korosov et al. [74]. Sentinel-1 offers C-band SAR data, which is effective in areas of thin ice and open water [35, 32, 75, 36]. Additionally, incidence angle and distance-to-nearest-coast measurements are provided as supporting data. Incidence angle is reported due to its impact on radar backscatter, while distance-to-land maps are useful in combination with coarse AMSR2 data, which becomes inaccurate near land [24]. Each scene's corresponding ice charts are co-registered with the Sentinel-1 image.

AMSR2 Passive Microwave

The Advanced Microwave Scanning Radiometer 2 (AMSR2) is an instrument onboard the JAXA GCOM-W satellite. AMSR2 provides H and V polarized brightness temperature measurements, in Kelvin, at 6.9, 7.3, 10.65, 18.7, 23.8, 36.5, and 89 GHz. Spatial resolution varies by channel; 6.9 GHz offers the coarsest ground resolution (35 x 62 km), whereas 89 GHz offers the finest resolution (3 x 5 km) [76]. The 18.7 and 36.5 GHz channels, which are used in this study, are sensitive to sea ice concentration. While the finer 89 GHZ channel is also sensitive to sea ice, higher frequency channels are more impacted by emissions from atmospheric phenomena (e.g., water vapour) [77]. These ambiguities are reduced for lower frequency channels in the microwave spectrum; however, overall energy emission registered by the sensor is also lower, which is part of why a larger footprint is required [35].

ERA5 Reanalysis Data

The AI4Arctic dataset also includes five numerical weather prediction (NWP) parameters from the ERA5 reanalysis dataset: 10 m wind speed, 2 m air temperature, total column water vapour, total column cloud liquid water, and skin temperature [24]. ERA5 is based on four-dimensional variational data assimilation, utilizing Cycle 41r2 of the Integrated Forecasting System. A subset of its reported climate variables can influence ice formation and break-up, including wind speed and temperature [78].

3.3 VIIRS Dataset

3.3.1 Ice Surface Temperature

The Visible Infrared Imager Radiometer Suite (VIIRS) flies on the Suomi National Polar-Orbiting Partnership (SNPP). IST data is available in the form of 6-minute, L2 Swath

scenes with a pixel spacing of 750 m. The VIIRS product follows a similar approach to MODIS to calculate IST: The algorithm first converts VIIRS calibrated radiance data from band M15 (10.763 µm) and M16 (12.013 µm) into brightness temperatures using a method by Key et al. [79], which derives brightness temperature from the Planck function, incorporating an emissivity term:

$$T = c_2\lambda / ln(1 + (\epsilon c_1 \lambda^3)/L) \tag{3.1}$$

Here, brightness temperature (T), is a function of the calibrated radiance from the sensor, L, the central wavelength from each band, λ, and the emissivity for each band, ϵ. The constants c_1 and c_2 are given as $1.1910659 * 10^{-5} mW/m^2 * sr * cm^{-4}$ and 1.438833 $cm * K$, respectively [48].

Next, IST is calculated using a split-window method adapted from Yu et al. [80]. This process is summarized in Equation 3.2 [9]:

$$IST = a + b * T_{11} + c * (T_{11} - T_{12}) + d * [(T_{11} - T_{12}) * (sec(q) - 1)] \tag{3.2}$$

Where T_{11} and T_{12} are the brightness temperatures, in Kelvin, for bands M15 and M16, q is the scan angle from nadir, and the coefficients a, b, c, and d compensate for atmospheric effects (primarily water vapour). Separate sets of coefficients are used for three different T_{11} temperature ranges, which are summarized in Table 3.1 [48]. These coefficients are derived from modelled radiances in the split-window channels, incorporating Arctic temperature and humidity profiles, along with angular emissivity models for snow [81]. Data quality flags such as those for cloud and day/night (applied using a solar zenith angle $< 85°$ for day) are generated after IST is calculated [48].

Table 3.1: Summary of coefficients used in the IST algorithm, by temperature range, for the Northern hemisphere. Note: separate coefficients are used in the southern hemisphere [82].

Temperature Range	**a**	**b**	**c**	**d**
$<$ 240 K	-7.335613	1.030383	1.264255	-0.438851
240 – 260 K	-8.606919	1.03532	0.641668	1.838797
$>$ 260 K	-6.629177	1.027197	1.082237	2.159417

Calibrated radiance data is accessible through the VIIRS L1B calibrated radiances product, which is comprised of eleven moderate-resolution reflective solar bands (M1-M11),

and five thermal emissive bands (M12-M16) [83]. IST data is expressed in Kelvin; the product also contains flags for features such as land, ocean, and cloud. The VIIRS instrument measures IST up to 310 K, but targeted uncertainty for this product is ±1 K over the range of 213-275 K [9]. Prior to use in this study, flag data is masked by the model and the remaining IST data is normalized using the mean and standard deviation.

A major limitation of passive thermal-infrared products is that atmospheric phenomena result in ambiguous measurements of surface EM emissions in this spectrum. Consequently, IST measurements in cloud-contaminated pixels are inaccurate, and must be discarded. Further, conservative cloud masking must be applied at night, where it is more difficult to accurately identify thin clouds [48]. The quality of IST measurements has been validated [84]. However, inaccuracies in the product's cloud mask may result in erroneous estimates in sea ice state. The MODIS-VIIRS Cloud Mask (MVCM) uses the 14 spectral bands shared by the two instruments; thus, VIIRS's cloud mask achieves comparable masking accuracy to its predecessor [85].

3.3.2 Scene Selection Criteria

Scene selection criteria include the time between sensor acquisitions, overlap in Sentinel-1 and VIIRS coverage regions, and the percentage of a VIIRS scene that 1.) is unobstructed by cloud cover and 2.) contains IST measurements that register as ice. For each Sentinel-1 scene, the acquisition date and latitude-longitude bounds are extracted; retrieval of the corresponding VIIRS scenes from the National Snow and Ice Data Center (NSIDC) is automated. Scene selection criteria are summarized in Table 3.2.

Table 3.2: Summary of criteria used in evaluating VIIRS data. Visual analysis is used following the cloud and IST thresholds to assign data quality grades to viable scenes.

Criterion	Condition
Acquisition Date	Same day
Period between Sentinel-1 and VIIRS Acquisitions	$\leq$ 8 hours
Overlap between Sentinel-1 and VIIRS Scenes	$\geq$ 75% of the Sentinel-1 scene
Percentage of Cloud Cover in VIIRS Scene	$\leq$ 60%
Range of IST Values in VIIRS Scene	$\geq$ 50% below 271 K; $\geq$ 15% below 260 K

Same day acquisition matches the retrieval criteria for the AI4Arctic dataset [2]. The maximum acquisition window is set in line with thresholds from similar studies. Malmgren-Hansen et al. [32] use acquisition thresholds of 7 hours preceding and 2 hours following the Sentinel-1 image, and all scenes in their dataset are found within a 6 hour and 40 minute window. Rogers et al. [37] use two acquisition thresholds: up to 8 hours in regions with low drift velocities, and up to 2 hours in regions with high drift velocities. Based on these findings, and considering drift within a 750 m pixel-spaced scene, a threshold of 8 hours is used for this study. Ultimately, the final dataset has a composition of 62% scenes acquired within 5 hours, 32% scenes acquired within 7 hours, and 6% scenes acquired within 8 hours.

A minimum area of overlap between the Sentinel-1 and VIIRS scenes ensures that sufficient VIIRS IST data is present in the image, as regions of fill are masked and do not contribute to model predictions. Finally, rough thresholds for the percentage of cloud and ice are defined to filter out inadequate scenes prior to visual analysis. Percentage of cloud cover is calculated using the cloud mask provided by the VIIRS IST product. Two IST thresholds are defined to identify pixels containing ice: the first because temperatures near to or exceeding 271 K may yield limited useful features [9], and the second to confirm the presence of thicker ice. Temperature measurements exceeding 271.4 K are used to identify water [48]. Visual analysis is still required for both criteria because the percentage of cloud or IST data present in the region does not consider the distribution of this data (e.g., concentrated regions versus small cloud formations evenly distributed throughout). Accounting for this fact, scenes are assigned a data quality class based on cloud cover; this process is described in Section 4.1, and ensures that poorer quality scenes with limited IST data are not allocated to the test set.

3.3.3 Data Co-registration

The overlapping region of each VIIRS scene is co-registered per pixel with its associated Sentinel-1 SAR image using a nearest neighbour approach. Viable scenes are selected using the criteria defined above. VIIRS data at coordinates outside of the Sentinel-1 scene are discarded. This process is achieved by defining a polygonal region of intersection between the Sentinel-1 and VIIRS scenes. VIIRS latitude-longitude coordinates and data outside of this polygon are set to NaN, which omits these points from the interpolation. Finally, the AI4Arctic dataset provides latitude-longitude grids that contain approximately 21x21 coordinates. These grids are up-sampled to match the resolution of the SAR image, and the nearest neighbour algorithm is applied to every coordinate in the Sentinel-1 grid to obtain the co-registered VIIRS scene. This approach allows for custom nearest neighbour criteria, differentiating between IST data (given in Kelvin), flags (given as integers outside

of the valid temperature range), and coordinates that exceed the bounds of the Sentinel-1 scene (set to NaN).

Visual analysis of the interpolated scenes was used to determine the maximum nearest neighbour radius. If no radius was set, this would result in smearing, where the outer boundary of the VIIRS image is used for every point outside the polygon in the Sentinel-1 grid. This radius is necessary because VIIRS scenes are co-registered over the entire Sentinel-1 grid so that the AI4Arctic scenes do not need to be cropped to match the VIIRS coverage region. Additionally, co-registration with one, two, and four nearest neighbours was assessed. Ultimately, one nearest neighbour provides sufficient accuracy for co-registering an image with 750 m pixel spacing (VIIRS) with an image with 80 m pixel spacing (down-sampled Sentinel-1). An example of the interpolation process output is provided in Figure 3.2.

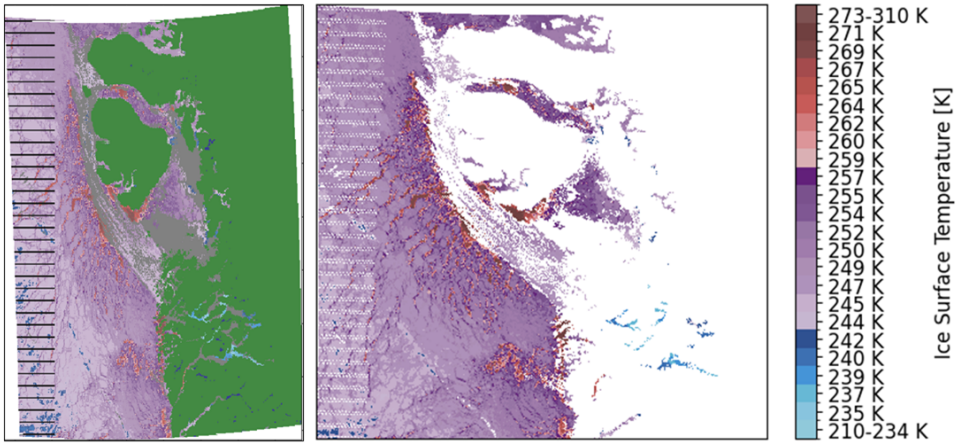


Figure 3.2: Example of the original VIIRS scene data (left) and the output of the co-registration process (right) from the west coast of Greenland. Land, which appears in green in the original scene, is masked based on the AI4Arctic data.

3.3.4 Overview of the Co-registered Dataset

Of the original 533 scenes, 84 met the scene selection criteria outlined in Section 3.3.2. The VIIRS-AI4Arctic co-registered dataset can be seen summarized in Figure 4.1, in the following chapter. Comparing to the full AI4Arctic dataset (Section 3.1), the VIIRS-AI4Arctic dataset has coverage in the regions of Hudson's Bay, Baffin Bay, the Canadian Archipelago,

and the Greenland Sea. The dataset excludes the Central Arctic, Beaufort Sea, and Newfoundland based on acquisition month and cloud cover. Specifically, June through October were omitted due to extensive cloud cover. Many scenes between November and May also do not meet the selection criteria, which contributes to the absence of certain Arctic regions in the dataset. Of the original input channels described in Section 3.2, this study uses the same subset as Chen et al. [3]. These input channels are summarized below in Table 3.3. Additional description of the dataset composition is provided in Section 4.1.

Table 3.3: Descriptions of the data types and features used by the baseline architecture from Chen et al. [3], in addition to VIIRS IST data.

Data Type	**Feature**	**Number of Channels**
Sentinel-1	Dual-polarized (HH and HV) SAR images	2
	Gridded incidence angle measurements from the instrument to the ground, corresponding to each SAR image	1
	Distance-to-land map indicating the nearest coast, per pixel	1
AMSR2	18.7 and 36.5 GHz H and V brightness temperature measurements	4
ERA5	East and northward 10 m wind speed components	2
	2 m air temperature	1
	Total column water vapour	1
	Total column cloud liquid water	1
Auxiliary	Gridded Latitude	1
	Gridded Longitude	1
	Acquisition Month	1
VIIRS	Ice surface temperature measurements	1

Chapter 4

Methodology

This chapter summarizes the methodology applied to the following experiments:

- Input-level fusion of IST with the AI4Arctic dataset using the U-Net architecture
- Feature-level fusion of IST with the AI4Arctic dataset using variants of the U-Net architecture:
  - A dual-encoder U-Net with varied input combinations
  - A dual-encoder U-Net incorporating estimation of aleatoric uncertainty
  - A dual-encoder U-Net with separate decoders for SIC, SOD, and FLOE

4.1 Dataset Organization

This study separates the 84-scene VIIRS-AI4Arctic dataset into a 72-scene train-validation set, and 12-scene test set, which is used solely for evaluation. Due to the limited size of the dataset, the train-validation set is also used for training, validation, and testing during hyperparameter tuning. Several considerations, summarized below, inform the selection of these two sets. The following discussion emphasizes the selection of a representative test set, because the study aims to assess the impacts of IST data.

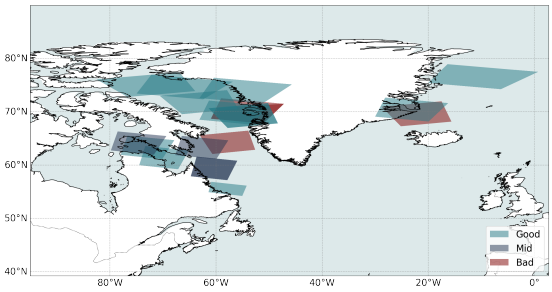
Data quality: VIIRS scenes are assigned a good, mid, or poor classification, referring to the presence and distribution of clouds within a scene. This classification is made based on visual inspection of the data. "Bad" scenes, which still meet the scene selection

criteria (Table 3.2), are included in the train-validation dataset because patches of the scene are used in training (Section 4.2.1), and patch selection criteria prevents the use of a patch containing only clouds.[1] "Mid" and "Good" scenes are used across training, validation, and evaluation. Use of "Good" scenes is prioritized for the test datasets (both for hyperparameter tuning and evaluation) to maximize the percentage of VIIRS IST data used in predicting the three sea ice parameters. Because IST measurements are not retrieved for cloud-contaminated pixels, use of cloudy scenes would lead to analysis of predictions that do not depend on IST measurements. Consequently, any assertions regarding the impacts of IST data would be inconclusive or unreliable if predominantly cloudy scenes are used.

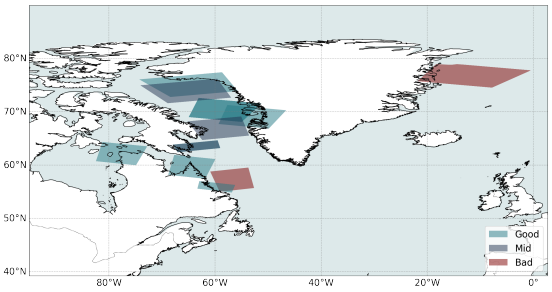
Spatial-temporal distribution: The distribution of scenes by year is summarized in Table 4.1, while the distribution by month and region is illustrated in Figure 4.1. The train-validation and test sets were selected to adequately sample across space and time and therefore allow spatial-temporal encodings to increase the accuracy of SOD predictions, in line with Chen et al [3]. In addition to dataset selection, this finding informed the exclusion of June through October from the dataset, as these months yielded few non-cloudy scenes. In order to comprehensively assess models across dataset conditions, the test set has a comparable distribution of years and regions to the entire dataset, and evenly samples acquisition months.

Pixel class distribution: Figures 4.2- 4.4 illustrate the distribution of pixels, by class, across the train-validation and test datasets for each of SIC, SOD, and FLOE. While the datasets vary in class composition, the test set was selected such that there would not be any large class imbalances compared to the data the models are trained and validated on. Notably, variation in class distribution does impact model predictions (see Section 5.1). Further, the test set was selected such that the scenes depicted varied ice conditions (full ice cover, mixed ice conditions, marginal ice zones, landfast ice, etc.).

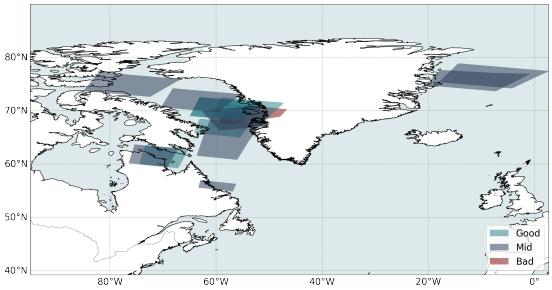
[1] Image patches can be used in training to improve efficiency and increase the amount of training data. In evaluation, the entire scene is provided as the input to the trained model, which must predict the associated ice chart labels.



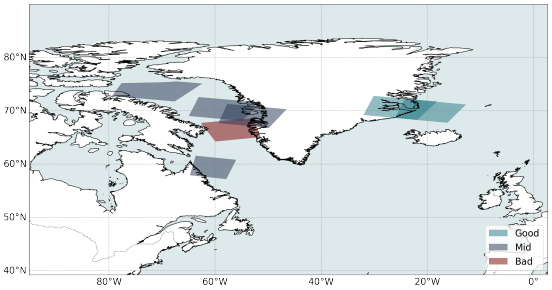
(a) Scene distribution for January.



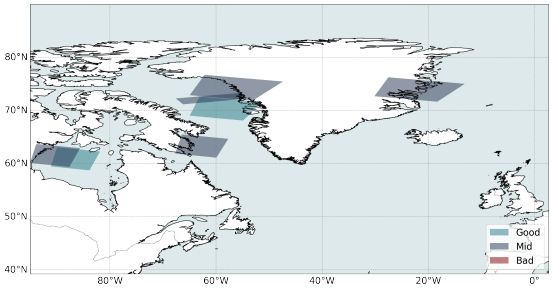
(b) Scene distribution for February.



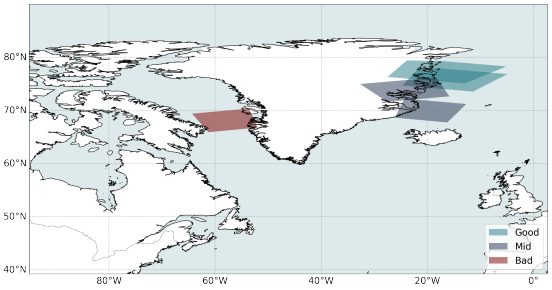
(c) Scene distribution for March.



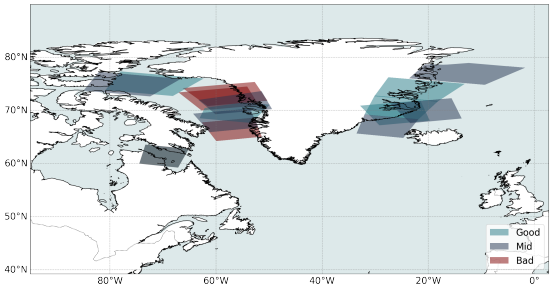
(d) Scene distribution for April.



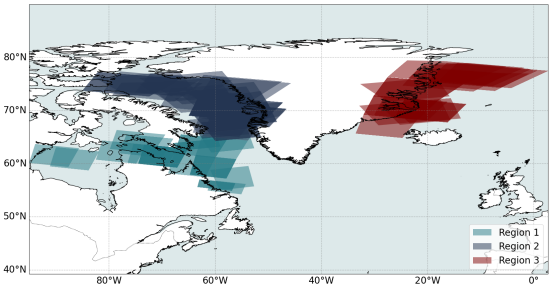
(e) Scene distribution for May.



(f) Scene distribution for November.



(g) Scene distribution for December.



(h) Scene distribution by region.

Figure 4.1: Dataset scene distribution by month (a-g) and region (h). In order of figures, monthly scene percentages are 24, 15, 18, 10, 8, 6, and 19%, with scenes plotted by their data quality classification. The total scene distribution (h) is 22% (Region 1: Hudson Bay & lower Baffin Bay), 49% (Region 2: upper Baffin Bay & Canadian Archipelago), and 29% (Region 3: Greenland Sea).

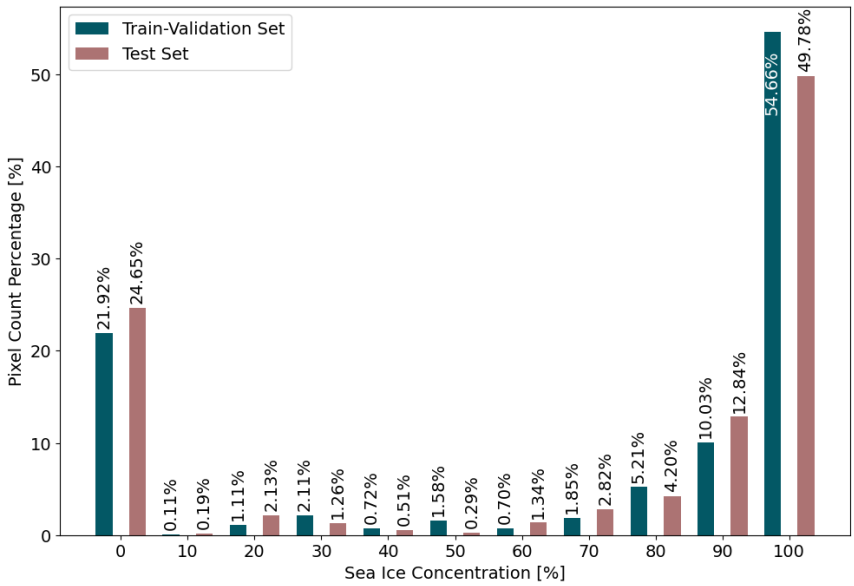


Figure 4.2: Comparison of sea ice concentration class distribution for the train-validation and test datasets.

Table 4.1: Summary of acquisition months for the 84-scene dataset. Scenes from June to October are omitted due to extensive cloud cover and limited availability, which would increase temporal sparsity.

Month	**2018 Scenes**	**2019 Scenes**	**2020 Scenes**	**2021 Scenes**
January	2	6	5	7
February	1	2	4	6
March	3	4	2	6
April	1	3	1	3
May	1	1	1	4
November	1	1	1	2
December	6	3	4	3

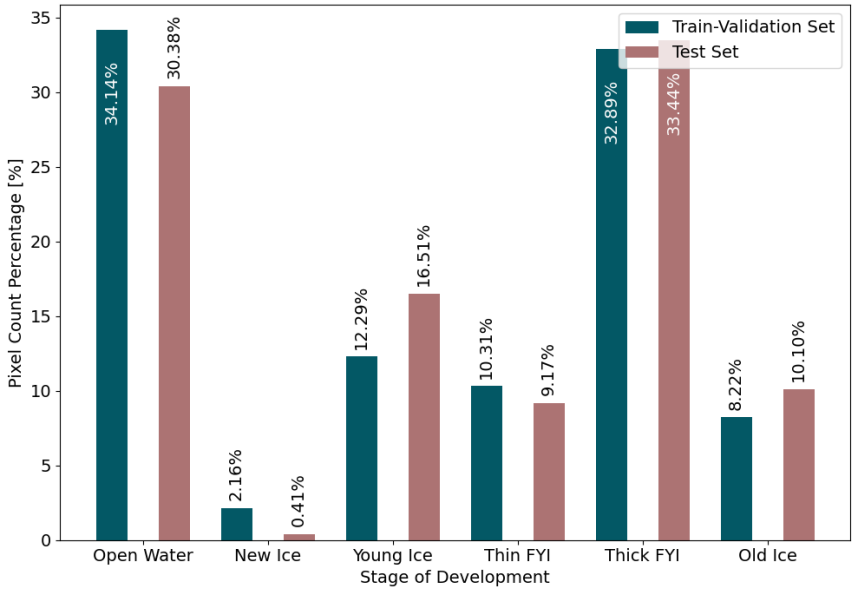


Figure 4.3: Comparison of stage of development class distribution for the train-validation and test datasets.

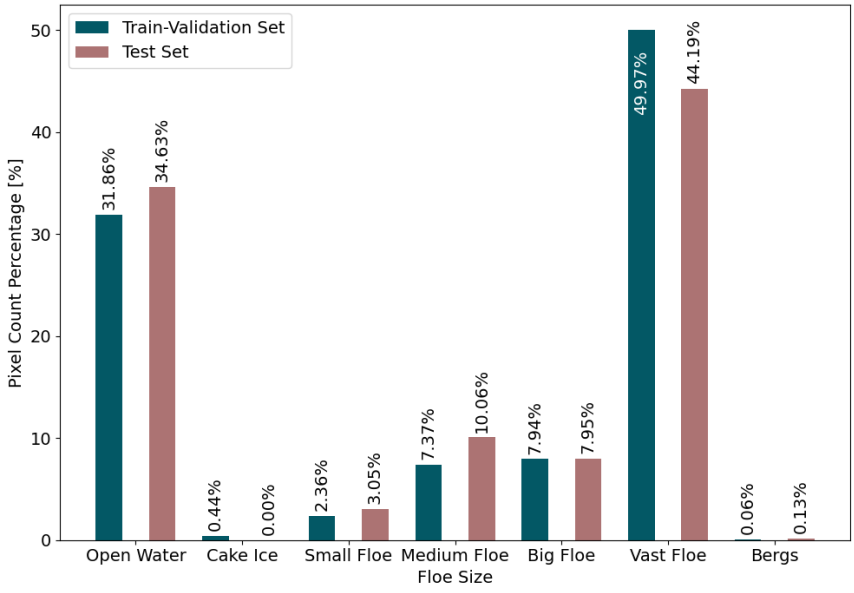


Figure 4.4: Comparison of floe size class distribution for the train-validation and test datasets.

4.2 Experimental Setup

All experiments are conducted using an NVIDIA A100-SXM4-40GB GPU, 128 GB of memory, and Pytorch version 2.0.1 with CUDA 11.8. These resources are accessed using the Narval cluster, supported by the Digital Research Alliance of Canada.

4.2.1 Data Pre- and Post-processing

The AI4Arctic data and corresponding, co-registered VIIRS IST scenes are stored in separate netCDF files. Prior to entering the network, SAR input channels (Table 3.3), IST measurements, and ice chart labels are downsampled by a factor of 10, reducing these scenes to approximately 500x500 pixels, with 800 m pixel spacing. This downscaling improves the model's ability to predict the ice chart labels, which are inconsistent with pixel-level SAR measurements (Section 2.1.2). Gridded input channels (those remaining in Table 3.3) are then upsampled to match this new pixel spacing. Uniform height and width dimensions for each input channel are enforced by extracting randomly cropped, 256x256 patches from scenes in the training set. When carrying out inference, validation and test scenes are still downsampled. However, entire scenes are accepted as inputs to the trained model, instead of patches. As a result of this pre-processing, predicted sea ice maps are also downscaled. At the output of the network, predictions are interpolated such that maps are evaluated at the original 80 m pixel spacing. Ultimately, this approach, seen in Figures 4.5- 4.7, follows previous work using the full AI4Arctic dataset and labels, which achieved enhanced accuracy and improved the capture of contextual information [3].

4.2.2 Hyperparameter Tuning

Tuning involved a systematic grid-search approach to evaluate key hyperparameters (learning rate, optimizer, batch size, number of convolutional filters, etc.). Referring to Table 4.2, the 84-scene dataset is split into a 12-scene test set that is held-out from tuning and a 72-scene dataset of six folds, each comprised of 12 scenes. To tune the models, the six folds are used for holdout k-fold cross validation, which ensures a robust evaluation of each model configuration. Fold 6 is used as the test set, while folds 1-5 are used for training and validation. For consistency, models were compared over the optimal training length for the Chen et al. [3] AutoICE architecture. Different network depths were also assessed, but no meaningful improvements from the AutoICE configuration were found. Optimized model specifications are summarized in Section 4.3.

4.2.3 Evaluation

Evaluation metrics are reported in Section 5.1 as the average result of holdout k-fold cross validation. During evaluation, the 72-scene dataset becomes the train-validation set. Each run, one fold is assigned to validation, while the remaining scenes are used for model training. The test set is fixed in order to evaluate model performance when trained on different class distributions. This process is repeated six times for each architecture, which allows each fold (Table 4.2) to be assigned to the validation set. Test and validation scores are then averaged across the six folds, which are consistent across the models.

Table 4.2: Unmasked data, reported by percentage (unmasked pixels over total pixels) and number of pixels, for SIC, SOD, and FLOE. Statistics are reported by 12-scene fold, where folds 1-6 are used for training and validation, and the test set is fixed.

	Unmasked SIC		**Unmasked SOD**		**Unmasked FLOE**	
SET	**% Pixels**	**# Pixels**	**% Pixels**	**# Pixels**	**% Pixels**	**# Pixels**
Fold 1	40.74	120328141	15.82	46725225	22.77	67236761
Fold 2	46.48	140277988	25.67	77473992	30.36	91638717
Fold 3	58.98	168638610	40.17	114860267	42.08	120327299
Fold 4	53.74	171800825	36.92	118029866	41.46	132552269
Fold 5	58.45	173827836	40.12	118992931	38.93	114984055
Fold 6	47.00	149845178	28.42	90622618	30.57	97448349
Test Set	57.75	165651870	46.87	136510447	41.12	420097905

This study reports validation scores as both the average and weighted average for each metric to account for differences in the percentage of unmasked pixels in each fold. Specifically, AI4Arctic ice chart labels mask certain classes (e.g., pancake ice - Table 2.3), pixels that are not assigned a dominant ice type, and land. As a result, the amount of masked data in a scene is inconsistent; Table 4.2 demonstrates that a fold may contain as much as 58.98% unmasked data, or as little as 15.82%. Consequently, reported metrics for folds with a low percentage of unmasked data are disproportionately impacted by misclassified pixels. Use of a weighted average (Equation 4.1) reduces the impact of these misclassified pixels on the overall cross-validation scores:

$$W = \frac{\sum_{i=1}^{N} w_i S_i}{\sum_{i=1}^{N} w_i} \tag{4.1}$$

Here, the weighted average, W, over N folds is calculated as the sum of each fold's weight, w_i, multiplied by its score, S_i, over the sum of the weights. The weight of the i^{th} fold is the ratio of unmasked pixels in the fold to unmasked pixels in the test set. Models are evaluated using five scores, which are introduced in Section 4.4: the SIC R^2 coefficient, SIC accuracy score, SOD F1 score, FLOE F1 score, and a weighted combined score. If the metric assesses a single sea ice parameter (e.g., SIC), weights are calculated based on the number of unmasked pixels for that parameter. Conversely, weights for the combined score are calculated from the total number of unmasked pixels across SIC, SOD, and FLOE. Statistics used in computing the weights are summarized in Table 4.2.

4.3 Summary of Architectures

This study assesses the impacts of IST data on sea ice parameter predictions using both input- and feature-level data fusion. Further, it assesses the use of multiple encoders, decoders, and loss functions. While the following sections summarize model specifications in detail, Table 4.3 introduces and differentiates between the six models.

4.3.1 Shared Specifications

Models build on the base AutoICE U-Net architecture (Chen et al. [3]), which features four sets of encoder and decoder blocks, connected by a bridge layer. Each block is comprised of two convolutional layers, each of which is followed by batch normalization and the ReLU activation function. As discussed in Section 4.2.2, changing this structure yielded no performance improvements. Thus, all architectures consist of these four-block sets, although the number of encoders or decoders changes. The number of filters associated with each block are specified, by model, in Tables 4.5 and 4.7. For SOD and FLOE, outputs from the final decoder block are fed into separate 1x1 convolution layers followed by the softmax activation function, with output channels corresponding to the number of discrete classes. Concurrently, the decoder output for SIC is fed into a regression layer. Prior to evaluation, SIC predictions are rounded to the nearest tenth, matching the ice chart classes.

Table 4.3: Summary of models, by name and description.

Fusion Type	U-Net Variant	Model	Description
Input-level fusion	Base U-Net	AutoICE	The Chen et al. [3] architecture, which is used as a performance baseline.
		AutoICE-V	AutoICE, incorporating VIIRS IST data.
Feature-level fusion	Dual-encoder (DE) U-Net	DEU-Net-SV	DEU-Net where encoder 1 contains SAR and VIIRS data.
		DEU-Net-V	DEU-Net where encoder 1 contains only VIIRS data.
		DEU-Net-V-GNLL	DEU-Net-V using GNLL loss in place of MSE, and estimating SIC and variance.
	Tri-decoder (TD) U-Net	TDU-Net-V	DEU-Net-V encoders, separate decoders for SIC, SOD, and FLOE.

This study uses three different loss functions: mean squared error (MSE), cross entropy (CE), and Gaussian negative log likelihood (GNLL). Chen et al. [3] use mean squared error (MSE) and cross entropy (CE) loss in their multi-task architecture. The authors observed that treating SIC classification as a regression task with MSE loss (Equation 2.3), and SOD and FLOE as classification tasks with CE loss (Equation 4.2) achieves the highest test accuracy. These findings are supported by the evaluation metrics defined by the AutoICE Challenge (Section 4.4.1), which capture the regression aspect of SIC, and the classification aspects of SOD and FLOE. While MSE and GNLL loss are described in Equations 2.3 and 2.4, above, CE loss is defined as follows:

$$L_{CE} = -\sum_{i=1}^{N} y_i log(\hat{y}_i) \tag{4.2}$$

where N refers to the number of pixels in the scene, y_i is the true label at pixel i, and $\hat{y}_i$ is the model prediction. For CE loss, $\hat{y}_i$ is the predicted probability of the true SOD or FLOE class. Returning to MSE loss, $\hat{y}_i$ is the predicted SIC. In place of MSE, GNLL loss is used in the best performing architecture to provide an estimate of aleatoric uncertainty.

Finally, use of down-scaling and patches in training (Section 4.2.1) improves model predictions due to the resultant, larger receptive field. This approach broadly captures a scene's contextual information, improving predictions that are scored against coarse, polygon-based ice chart labels [3]. Models are trained for 300 epochs, with each epoch consisting of 500 iterations: the number of mini-batches processed during a pass through the training set. This training approach is consistent with Chen et al. [3], in order to compare their baseline model to predictions made using IST data. Each iteration, a batch of 256x256 patches–selected based on the AutoICE model's receptive field–is extracted from the train set, and the three task-specific losses are combined to form a total loss, which is minimized. A summary of each model, including optimal hyperparameters and architecture diagrams, is presented in the following sections.

4.3.2 Input-level Fusion of IST

Input-level fusion (Figure 4.5) appends VIIRS IST as an additional channel in the Chen et al. [3] AutoICE architecture. The model's optimized hyperparameters are summarized in Table 4.4, while its input and output feature map dimensions are summarized in Table 4.5. AutoICE is re-trained on this study's 84-scene dataset to provide fair comparison to the baseline. Input-level fusion of IST in this architecture, referred to as AutoICE-V, then allows for a direct comparison of the impacts of IST data on model predictions.

4.3.3 Feature-level Fusion of IST

Feature-level fusion of VIIRS IST data is investigated using several U-Net variants. These models share optimized hyperparameters (Table 4.6). Input and output feature map dimensions are summarized in Table 4.7. Each model is described, by name, below:

DEU-Net-SV: This model (Figure 4.6) features two encoding paths, grouping the co-registered VIIRS data with Sentinel-1 SAR imagery in Encoder 1, and the gridded AMSR2, ERA5, and auxiliary data in Encoder 2. This approach follows the findings from studies discussed in Section 2.2.5, which recommend separating input data by resolution.

DEU-Net-V: This model (Figure 4.6) similarly features two encoding paths, but groups VIIRS IST data in Encoder 1, and all AI4Arctic input channels in Encoder 2. Similarly to AutoICE-V, this approach compares the impacts of IST data within the same architecture. Returning to Section 2.2.5, because this model treats IST as a separate input, the impacts of salient features learned solely from IST data can be assessed.

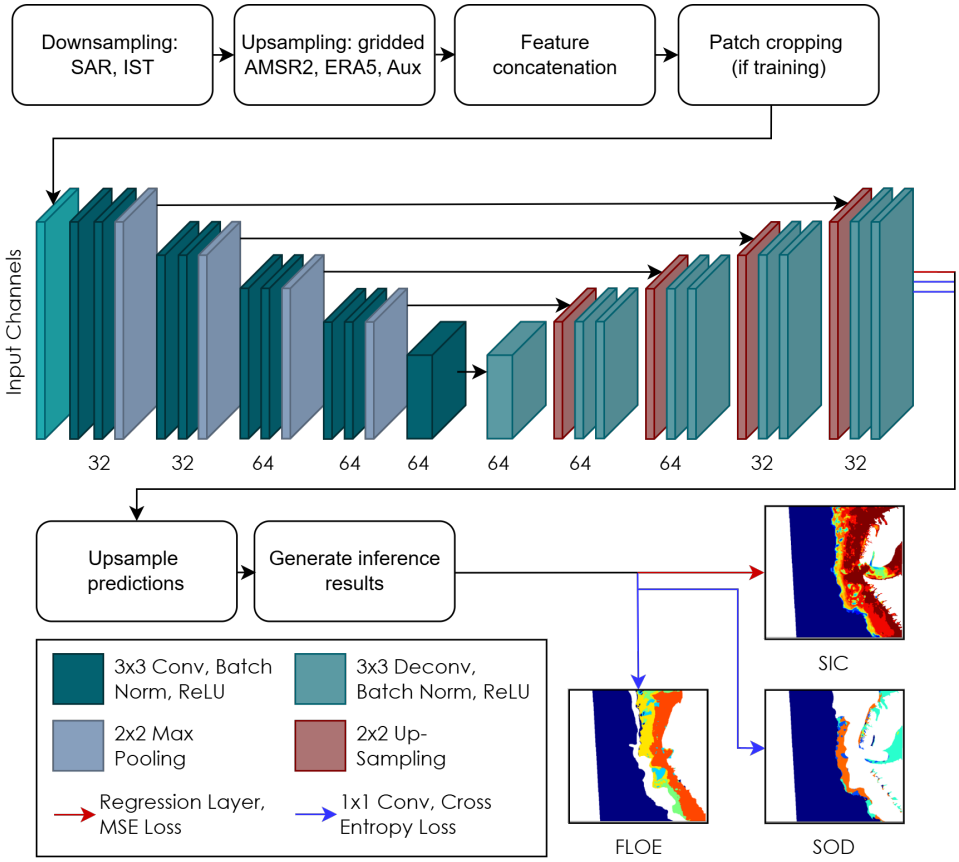


Figure 4.5: U-Net architecture diagram illustrating the input-level fusion of VIIRS IST with the AI4Arctic dataset (AutoICE-V). Model and layer specifications are summarized in Tables 4.4 and 4.5. Note: this architecture is consistent with the Chen et al. [3] AutoICE model if the IST input channel is excluded.

The final two architectures build on the best-performing model out of AutoICE-V, DEU-Net-SV, and DEU-Net-V, following optimization and cross-validation. DEU-Net-V-GNLL is intended to be a direct comparison to DEU-Net-V; thus, the model did not undergo hyperparameter tuning. TDU-Net-V uses the same encoder structure as DEU-Net-V, but hyperparameter tuning was used to see if the model could be further optimized. Ultimately, no performance improvements were found.

DEU-Net-V-GNLL: This model (Figure 4.6, inclusive of the boxed region) assesses the impacts of applying the heteroscedastic assumption to DEU-Net-V. DEU-Net-V-GNLL differs from and DEU-Net-V by incorporating estimation of aleatoric uncertainty. The SIC

Table 4.4: Model specifications corresponding to the AutoICE and AutoICE-V architectures. These specifications are consistent with those published by Chen et al. [3].

Model Specification	Details
Optimizer	Stochastic Gradient Descent with Momentum
Learning Rate	0.001
Weight Decay Coefficient	0.01
Batch Size	16
Learning Rate Scheduler	Cosine Annealing (CA)
CA Epochs till Restart	20
Down-scaling Factor	10
Data Augmentation Operations	Rotation, flip, random scale, cutmix
Patch Size	256
Loss Functions	SIC: MSE, SOD and FLOE: CE
Loss Function Weights	SIC: 1, SOD and FLOE: 3

task predicts both SIC and SIC variance, and uses the GNLL loss function in place of MSE. The two models are compared in Sections 5.1 and 5.3.2 using a combination of evaluation metrics, predicted sea ice maps, and the model's predicted variance.

TDU-Net-V: Finally, this model (Figure 4.7) addresses the question posed in Section 2.2.4: can the use of separate decoders improve segmentation of SIC, SOD, and FLOE, when the parameters share common ice chart polygons? Building on the best single-decoder architecture (DEU-Net-V), this model features three decoder branches for the SIC, SOD, and FLOE tasks.

4.4 Model Evaluation

4.4.1 AI4Arctic Metrics

The AutoICE challenge evaluated model accuracy using the R^2 coefficient (Equation 4.3) for SIC, and the F1 score (Equation 4.4) for SOD and FLOE. A combined score (Equation 4.5) was developed by ice chart experts, which assigns greater importance to SIC and SOD. R^2 is used as the measure of accuracy for SIC because it accounts for relative proximity between target and predicted classes. F1 is a measure of predictive performance,

Table 4.5: Input and output feature map dimensions, by layer, for the AutoICE and AutoICE-V architectures.

	Layer	**Input feature maps**	**Output feature maps**
Encoder	Convolution block 1	17×256×256	32×256×256
	Max-pooling 1	32×256×256	32×128×128
	Convolution block 2	32×128×128	32×128×128
	Max-pooling 2	32×128×128	32×64×64
	Convolution block 3	32×64×64	64×64×64
	Max-pooling 3	64×64×64	64×32×32
	Convolution block 4	64×32×32	64×32×32
	Max-pooling 4	64×32×32	64×16×16
	Bridge	64x16x16	64x16x16
Decoder	Up-sample 1	64×16×16	64×32×32
	Convolution block 5	128×32×32	64×32×32
	Up-sample 2	64×32×32	64×64×64
	Convolution block 6	128×64×64	64×64×64
	Up-sample 3	64×64×64	64×128×128
	Convolution block 7	96×128×128	32×128×128
	Up-sample 4	32×128×128	32×256×256
	Convolution block 8	64×256×256	32×256×256
Outputs	Regression layer (SIC)	32×256×256	12×256×256
	1 × 1 convolution+ Softmax (SOD)	32×256×256	7×256×256
	1 × 1 convolution+ Softmax (FLOE)	32×256×256	8×256×256

which accounts for imbalances between the number of pixels belonging to each class [24].

The R^2 coefficient is reported as a percentage, and is defined as follows:

$$R^2 = 1 - (\sum_{i=1}^{N}((y_i^t - y_i^p)^2)/(\sum_{i=1}^{N}((y_i^t - \hat{y}^t)^2)) \tag{4.3}$$

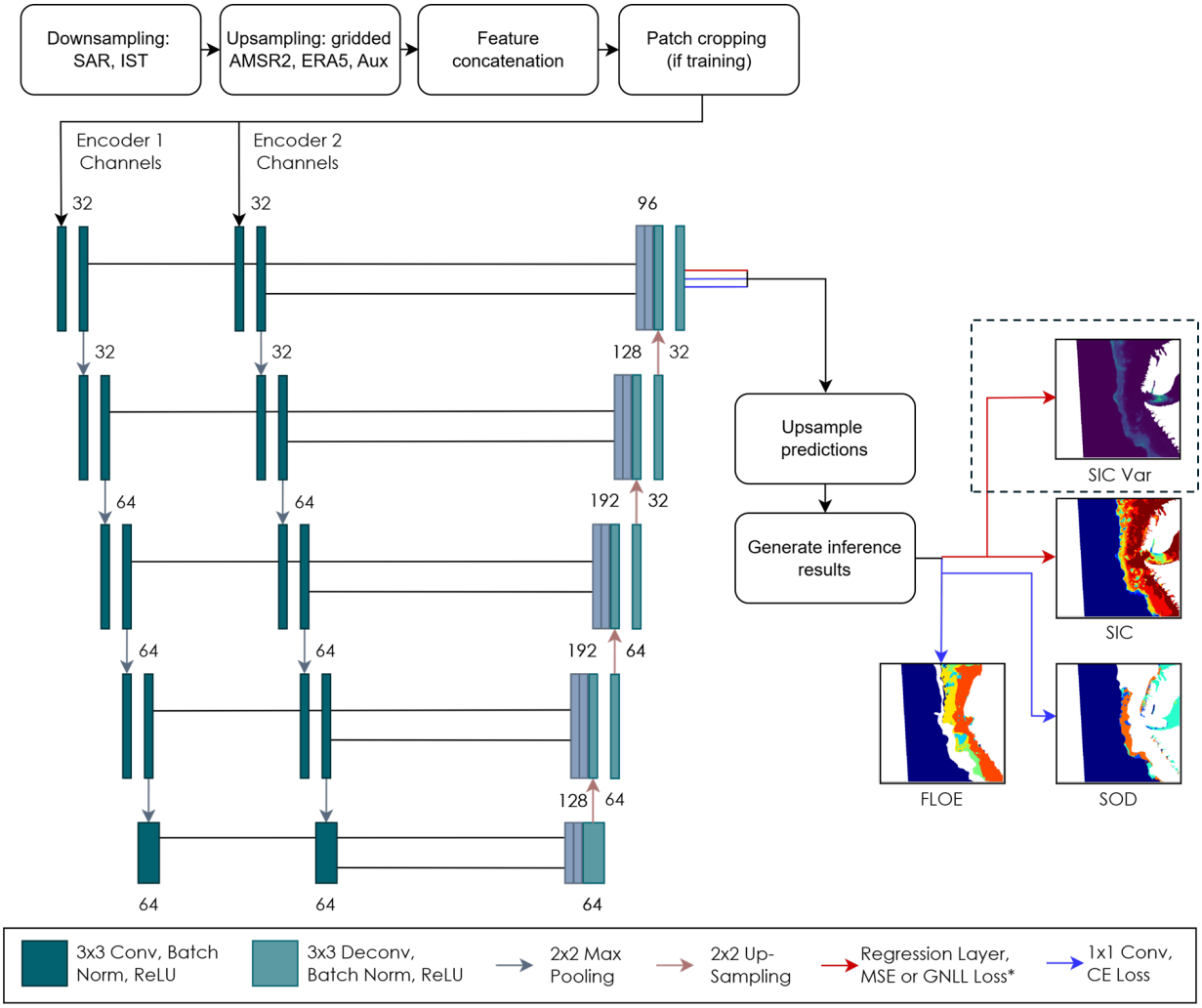


Figure 4.6: Diagram of the dual-encoder (DE) U-Net architecture. Model and layer specifications are summarized in Tables 4.6 and 4.7. Input channels and loss functions differ by model. **DEU-Net-SV:** Encoder 1 accepts VIIRS IST and Sentinel-1 SAR, while Encoder 2 accepts AMSR2 brightness temperature, ERA5 reanalysis data, and auxiliary channels. MSE loss is used. **DEU-Net-V:** Encoder 1 accepts only VIIRS IST, while encoder 2 accepts the AI4Arctic input channels. MSE Loss is used. **DEU-Net-V-GNLL:** Encoder 1 and 2 input channels are consistent with DEU-Net-V. GNLL loss is used, and an additional variance output (boxed) is produced by the model.

Table 4.6: Model specifications corresponding to the following architectures: DEU-Net-SV*, DEU-Net-V*, DEU-Net-V-GNLL**, and TDU-Net-V*.

Model Specification	**Details**
Optimizer	AdamW
Learning Rate	0.0005
AdamW Betas	0.9, 0.99
Weight Decay Coefficient	0.01
Batch Size	4
Learning Rate Scheduler	Cosine Annealing (CA)
CA Starting Epochs till Restart	10
CA Restart Multiplier	2
Down-scaling Factor	10
Data Augmentation Operations	Rotation, flip, random scale, cutmix
Patch Size	256
Loss Functions	SIC: MSE* or GNLL**, SOD and FLOE: CE
Loss Function Weights	SIC: 1, SOD and FLOE: 3

Where, N denotes the number of pixels, y_i^t denotes the true value of the i^{th} pixel according to the ice chart, $\hat{y}^t$ denotes the mean value over all true pixels containing sea ice, and y_i^p denotes the pixel's predicted SIC class [24]. An R^2 score of one (100%) corresponds to a perfect fit of the data, whereas a score of zero reflects no linear relationship between the target and predicted values.

The F1 score, which is similarly expressed as a percentage, is defined as:

$$F1 = (TP)/(TP + 1/2(FP + FN)) \tag{4.4}$$

Where TP denotes the number of true positive pixels, FP denotes the number of false positive pixels, and FN denotes the number of false negative pixels [24]. Here, a high F1 score (around one or 100%) reflects an accurate prediction of the ice chart labels (TP), and minimal misclassification (FP and FN). For a multi-class classification problem, F1 scores are calculated on a per-class basis and averaged.

Finally, a weighted sum of these metrics is used to generate the combined score [24]:

$$CombinedScore = (2/5) * R^2_{SIC} + (2/5) * F1_{SOD} + (1/5) * F1_{FLOE} \tag{4.5}$$

Table 4.7: Input and output feature map dimensions by layer for the DEU-Net architecture. Bridge L and Decoder K indicate the number of each element in the architecture. All DEU-Net models have L = K = 1, while TDU-Net-V has L = K = 3. Feature map dimensions remain consistent, but encoder input dimensions (M) vary by model: **DEU-Net-SV:** Encoder N = 1 has M = 5 input channels (Sentinel-1 SAR, VIIRS IST); Encoder N = 2 has M = 12 (AMSR2 brightness temperature, ERA5 reanalysis, auxiliary channels).**DEU-Net-V, DEU-Net-V-GNLL, TDU-Net-V:** Encoder N = 1 has M = 1 (VIIRS IST); Encoder N = 2 has M = 16 (all AI4Arctic channels).

	Name	**Input feature maps**	**Output feature maps**
Encoder N	Convolution block 1	M×256×256	32×256×256
	Max-pooling 1	32×256×256	32×128×128
	Convolution block 2	32×128×128	32×128×128
	Max-pooling 2	32×128×128	32×64×64
	Convolution block 3	32×64×64	64×64×64
	Max-pooling 3	64×64×64	64×32×32
	Convolution block 4	64×32×32	64×32×32
	Max-pooling 4	64×32×32	64×16×16
Bridge L	Bridge	128x16x16	64x16x16
Decoder K	Up-sample 1	64×16×16	64×32×32
	Convolution block 5	192×32×32	64×32×32
	Up-sample 2	64×32×32	64×64×64
	Convolution block 6	192×64×64	64×64×64
	Up-sample 3	64×64×64	64×128×128
	Convolution block 7	128×128×128	32×128×128
	Up-sample 4	32×128×128	32×256×256
	Convolution block 8	96×256×256	32×256×256
Outputs	Regression layer (SIC)	32×256×256	12×256×256
	1 × 1 convolution+ Softmax (SOD)	32×256×256	7×256×256
	1 × 1 convolution+ Softmax (FLOE)	32×256×256	8×256×256

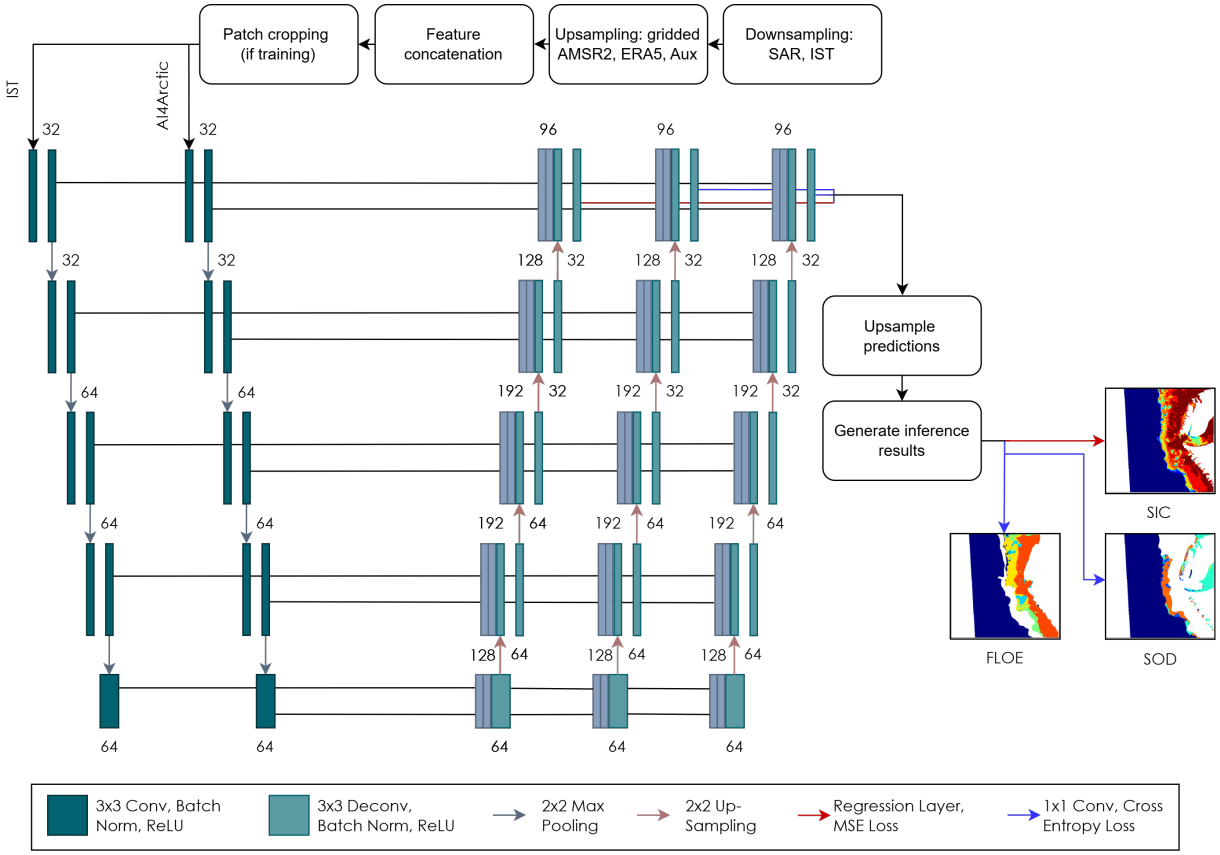


Figure 4.7: Diagram of the tri-decoder (TD) U-Net architecture. Encoder input channels are consistent with the DEU-Net-V architecture. Separate decoders are used for the prediction of SIC, SOD, and FLOE. Model and layer specifications are summarized in Tables 4.6 and 4.7.

4.4.2 SIC Accuracy

A major limitation of the R^2 metric is that incorrect predictions of underrepresented classes skew an individual scene's score. Equation 4.6 simplifies Equation 4.3 as the residual sum of squares (RSS) over the total sum of squares (TSS):

$$R^2 = 1 - RSS/TSS \tag{4.6}$$

For this relationship to be negative, the model's prediction (accounted for in the RSS) must be greater than the variability about the mean of the true pixel labels (accounted for

in the TSS). The models presented in this study are trained and tested on separate data, which differ in pixel class distributions at the dataset- and scene-levels. Thus, a model that is tested on a SIC class that was underrepresented in the train set may produce a prediction that has a higher RSS than TSS for the given SIC class. For a small dataset (e.g., calculation over one scene) this can produce a negative R^2, which does not allow for accurate comparisons of per-scene performance. This limitation presents an issue when analysing the model's performance across varied ice conditions.

In addition to R^2, this study introduces a basic accuracy metric for scene-level comparisons of SIC predictions, using thresholds for water, marginal ice, and consolidated ice (defined in Table 4.8). Sea ice concentrations between 15 and 80% are commonly accepted as thresholds for the marginal ice zone [50, 86, 51]. Because the ice chart labels report concentrations in tenths, SIC greater than or equal to 20% is used as the threshold for marginal ice. This metric (Equation 4.7) allows for comparison of predictions on a per-scene basis across these three categories. Instead of penalizing all misclassified pixels, this implementation of accuracy acknowledges that a prediction can be incorrect relative to the ice chart labels, but may still reflect the right category of ice formation (e.g., marginal ice). Thus, the model may correctly predict marginal ice at that location, but incorrectly predict the concentration due to class imbalances (e.g., low representation of 40% SIC within the training dataset). Additionally, this metric still penalizes predictions that fall outside of the ice chart label's category. Consequently, SIC accuracy can be loosely compared to top-k accuracy: instead of the metric assessing model accuracy for the correct label appearing within the top k predictions, it assesses model accuracy for predicting the correct broader category of ice formation, which is of greater interest to ice analysts.

Table 4.8: Summary of thresholds associated with the three ice formation classes.

Category	**Threshold**
Water	$0 \leq \text{SIC} < 20\%$
Marginal Ice	$20 \leq \text{SIC} \leq 80\%$
Consolidated Ice	$\text{SIC} > 80\%$

To compute SIC accuracy, the model's predictions ($\hat{y}$) and the ice chart labels (y) are first classified according to the three categories of ice formation using the thresholds in Table 4.8. Equation 4.7 then calculates the scene's accuracy by dividing the scene's correct pixels (the sum of $(y_i = \hat{y}_i)$ for each pixel i) by its total pixels, N.

$$Accuracy = 1/N * \sum_{i}^{N} 1(y_i = \hat{y}_i) \tag{4.7}$$

While R^2 is still used to compare model predictions for SIC, accuracy provides a clearer indication of the model's performance on individual scenes, and across the three categories. Its importance for analysis of model predictions is discussed in Section 5.

Chapter 5

Results and Discussion

This section is structured as follows: First, validation and test scores are reported across the five metrics and six models. These results are discussed in terms of weighted and unweighted averages in Section 5.1. Next, Section 5.2 compares the three base U-Nets (AutoICE, DEU-Net, and TDU-Net) by architecture, using test set predictions. Similarly, Section 5.3 highlights key impacts of IST data when used in combination with the DEU-Net-V architecture. Finally, Sections 5.4 and 5.5 summarize findings by sea ice parameter, and assess the impacts of SIC accuracy and GNLL loss.

5.1 Performance on the VIIRS-AI4Arctic Dataset

Table 5.1 reports the average and standard deviation of each metric, while Table 5.2 summarizes the weighted performance of each model in the validation set. Notably, higher standard deviation is reported across the validation set for all models, with the largest standard deviation across scores being reported for stage of development and floe size. Referring to Table 4.2, these parameters have the folds containing the least data, and the largest variation in percentage of unmasked data (15.82-46.87% for SOD and 22.77-42.08% for FLOE, relative to 40.74-58.98% for SIC). Further, as seen in Figure 5.1, the percentage of masked data across parameters differs at the scene level. Predominantly, this discrepancy results from AI4Arctic's definition of a dominant ice type, where there must be a dominant partial concentration (representing $\geq$65% of the polygon) for a SOD and FLOE to be assigned. Additionally, AI4Arctic combines and or excludes certain ice chart codes (e.g., combining young ice and grey ice) depending on representation within the dataset [2].

Thus, classification errors in folds containing minimal unmasked data result in low, outlying scores, which skew the reported average and standard deviation. Because pixel class distribution varies across the folds being used as the validation set (e.g., see Figures 1- 3 in the appendix), some variability in model accuracy is also attributed to this factor.

In comparison, test set scores report more consistent standard deviations across parameters, in line with comparable data representation of 57.75% unmasked SIC pixels, 46.87% unmasked SOD pixels, and 41.12% unmasked FLOE pixels. Some variability in model predictions for both validation and test can be attributed to different class distributions across train sets, in addition to training on random scene patches. Finally, referring to Table 5.2, use of the weighted average improves combined scores, suggesting that model performance on the validation set is more comparable to the test set when accounting for unmasked data. This improvement is largely attributed to SOD–the parameter with the least unmasked data. These findings support the use of a weighted average for comparing scores in the validation set. However, the variation in weighted and unweighted scores ultimately illustrates the importance of considering percentage of unmasked pixels in dataset design.

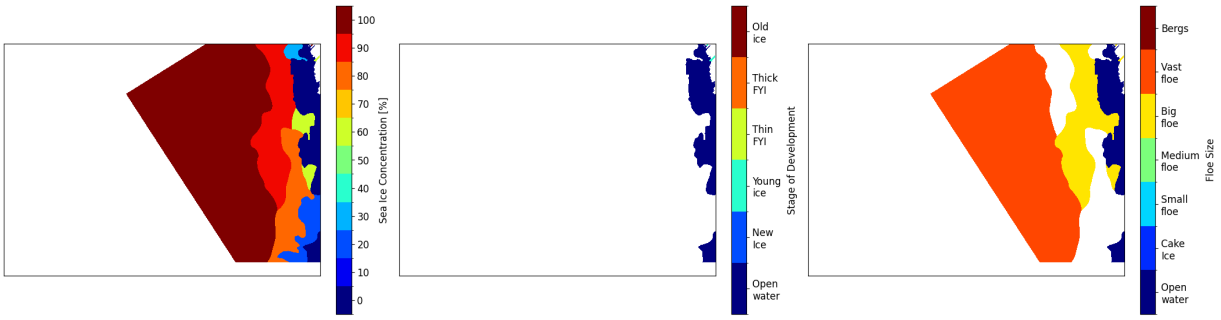


Figure 5.1: Example ice charts for a March 17, 2019 scene off the west coast of Greenland, demonstrating the difference in masked data between SIC (left), SOD (middle), and FLOE (right).

Table 5.1: Validation and test metrics averaged over the six cross-validation folds, reported as the average and standard deviation.

VALIDATION SET					
Model	**Combined Score**	**SIC R^2** [%]	**SIC Acc** [%]	**SOD F1** [%]	**FLOE F1** [%]
AutoICE	80.76±6.47	87.77±4.10	89.79±3.16	74.33±11.61	79.35±7.07
AutoICE-V	81.22±6.37	88.29±4.28	89.78±3.62	75.63±11.07	78.28±7.56
DEU-Net-SV	81.94±7.02	**88.97±3.68**	90.46±3.09	76.01±13.47	**79.95±7.71**
DEU-Net-V	80.90±7.03	88.72±4.03	90.55±3.07	76.49±9.60	78.66±7.01
DEU-Net-V-GNLL	**82.86±5.20**	88.72±4.37	**90.56±3.50**	78.99±7.29	78.88±8.21
TDU-Net-V	81.83±5.69	87.84±3.62	90.29±3.10	**79.79±6.05**	75.60±10.80
TEST SET					
Model	**Combined Score**	**SIC R^2** [%]	**SIC Acc** [%]	**SOD F1** [%]	**FLOE F1** [%]
AutoICE	82.58±1.44	90.33±1.40	84.44±3.15	75.28±3.00	81.69±1.64
AutoICE-V	83.79±1.03	91.37±0.95	85.99±1.00	76.79±2.64	82.39±1.53
DEU-Net-SV	83.40±1.78	91.95±0.93	86.59±2.20	74.82±2.70	83.48±2.73
DEU-Net-V	**84.42±0.72**	**92.53±0.76**	**88.67±0.56**	**76.92±1.15**	85.13±1.40
DEU-Net-V-GNLL	84.28±1.53	91.75±0.94	87.72±1.80	76.45±2.71	**85.23±1.83**
TDU-Net-V	83.38±1.86	91.99±0.80	87.19±2.22	74.59±3.56	83.73±2.34

Table 5.2: Weighted validation metrics, averaged over the six cross-validation folds.

Model	Combined Score	SIC R^2 [%]	SIC Acc [%]	SOD F1 [%]	FLOE F1 [%]
AutoICE	81.78	88.09	89.54	77.09	79.65
AutoICE-V	82.23	88.59	89.53	78.28	78.81
DEU-Net-SV	83.05	**89.25**	90.19	79.18	**80.14**
DEU-Net-V	81.97	89.03	**90.31**	78.67	79.02
DEU-Net-V-GNLL	**83.55**	89.00	90.29	**80.50**	79.40
TDU-Net-V	82.66	88.08	90.05	80.17	77.10

Comparing results, by metric, for the test set, the greatest overall performance improvement is found using the DEU-Net-V architecture, which improves predictive accuracy by 2.44% (SIC), 5.01% (SIC accuracy), 2.18% (SOD), 4.21% (FLOE), and 2.23% (combined score). The greatest improvement is seen in SIC accuracy, which does not impact the combined score. Additionally, DEU-Net-V-GNLL achieves comparable scores, and produces the greatest improvement to the FLOE F1 score at 4.33%. These improvements are discussed in the subsequent sections, in terms of specific test set examples.

5.2 Comparison of Architectures

The following subsections are grouped by architecture, providing a comparison of the impacts of IST using different fusion approaches.

5.2.1 Improvements from Input-level Fusion of IST

In order to directly assess the impacts of IST data on model predictions, AutoICE and AutoICE-V are compared across sea ice parameters.

2021-04-19

Figure 5.2 is a scene from the west coast of Greenland, which is characterized by a region of full ice cover, in addition to an inland region surrounding Disko Island that features

mixed marginal and consolidated ice conditions, and open water. Comparing predicted SIC maps, AutoICE-V more accurately models the consolidated ice region south of Disko Island, in addition to reducing misclassification of marginal ice north of Disko Island. For stage of development, AutoICE-V reduces misclassification of thin FYI in the same region. Neither model correctly predicts the inland regions as thin FYI; however, use of IST data produces a more accurate prediction of the thick FYI region. Finally, comparing floe size, the models achieve comparable accuracy on open water and vast flow. AutoICE accurately predicts the region south of Disko Island as big floe whereas AutoICE-V more accurately predicts the same ice condition in the region between Disko Island and the peninsula.

Referring to the input channels (Figure 5.3), the region of 90% concentration south of Disko Island contains ambiguous SAR textures. Regions of smooth ice appear similarly to the scene's smaller regions of open water, and the entire feature registers as darker than the large regions of open water and consolidated ice. Brightness temperature measurements distinguish this region from open water and consolidated ice; however, IST data in combination with BT reduces ambiguity. Stage of development data is not provided in this region, but the mid-range BT measurements alongside higher, ice-range IST estimates ($260 < \text{IST} < 267$ K) suggests the presence of consolidated FYI. Notably, both models appear to correlate SIC and FLOE predictions. For AutoICE-V, misclassification as vast floe is consistent with the classification of 90% SIC. Similarly for AutoICE, correct classification of big floe is more consistent with the model's misclassification of 90% SIC as 70-80%. Finally, comparing to the SAR imagery (Figure 5.3), the resolutions of the VIIRS and AMSR2 products are too coarse to improve predictions of ice conditions in the tributaries. Overall, fusion of IST at the input-level appears to improve predictions most significantly over regions with homogeneous conditions (e.g., in the case of SOD).

2018-12-31

Figure 5.4 is a scene from the Hudson Strait in Nunavut, Canada, which is characterized by predominantly young, consolidated ice. Higher IST values and lower BT measurements in the middle of the scene are consistent with the central region of 90% concentration reported by the ice chart; however, neither model correctly predicts this feature. Notably, use of IST data in DEU-Net-V does improve this prediction, demonstrating that salient features learned from solely IST data can improve model predictions (see Figure 4 in the appendix). AutoICE-V is seen to eliminate the incorrect region of thin FYI reported by AutoICE, which is likely learned from the higher BT measurements in this region. AutoICE-V also reduces misclassification as vast floe in the same region, which does not correspond to a lower concentration prediction from the AutoICE model. Similarly, the region of new ice

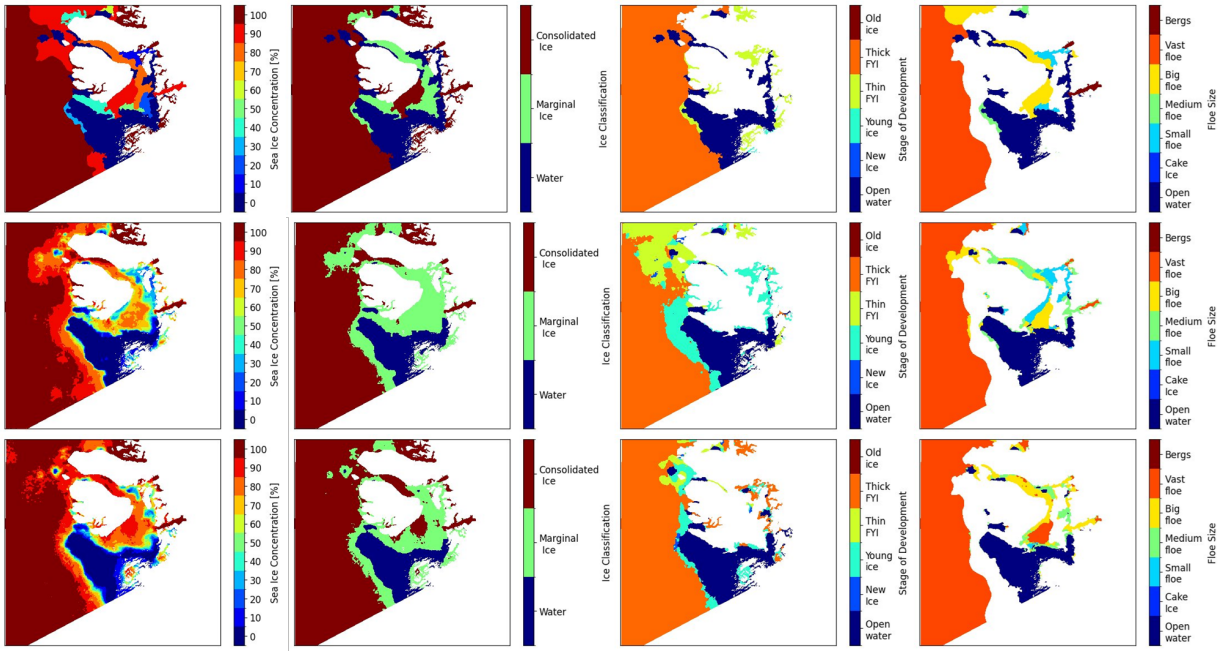


Figure 5.2: Summary of AutoICE (middle row) and AutoICE-V (bottom row) model predictions for 2021-04-19 on the west coast of Greenland. Predictions of SIC (left, column 1), SIC accuracy (column 2), SOD (column 3) and FLOE (column 4) are compared to ice chart labels (top row).

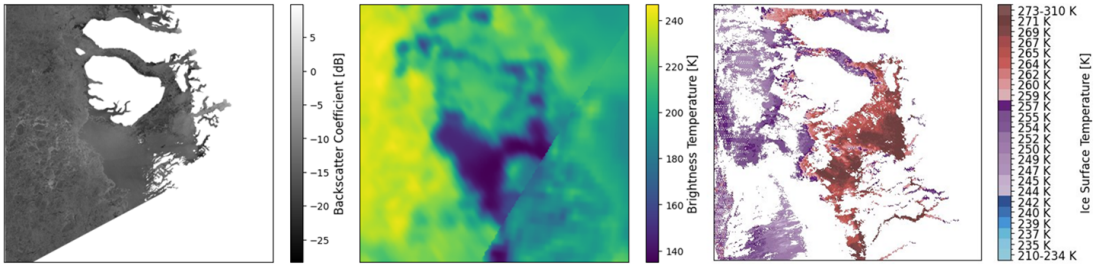


Figure 5.3: HH SAR (left), 36.5 GHZ H BT (middle), and IST (right) input channels for 2021-04-19.

reported by AutoICE is reduced with the inclusion of IST data; the remaining features are consistent with regions of higher (>260 K) IST measurements, which are seen in Figure 5.5. IST is largely masked in the inland regions, so the data does not impact predictions of these features.

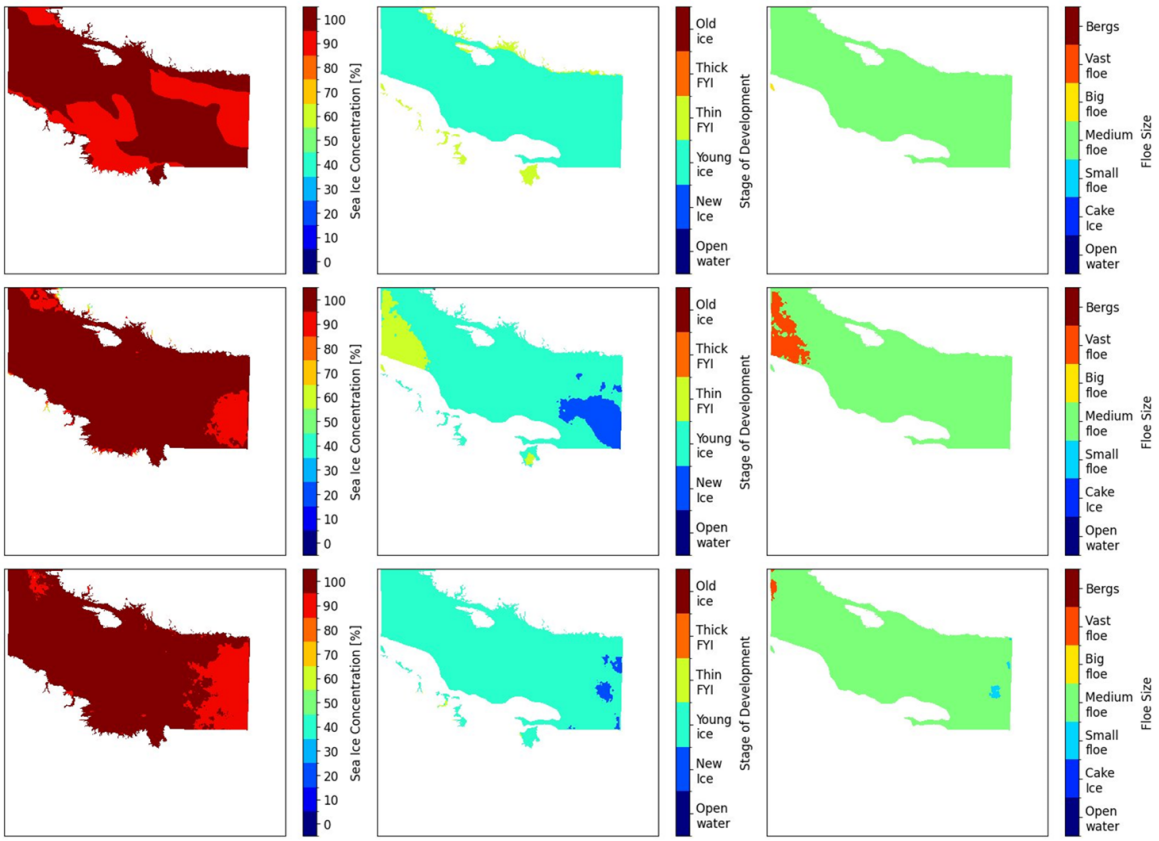


Figure 5.4: Summary of AutoICE (middle row) and AutoICE-V (bottom row) model predictions for 2018-12-31 in the Hudson straight. Predictions for SIC (left), SOD (middle) and FLOE (right) are compared to ice chart labels (top row).

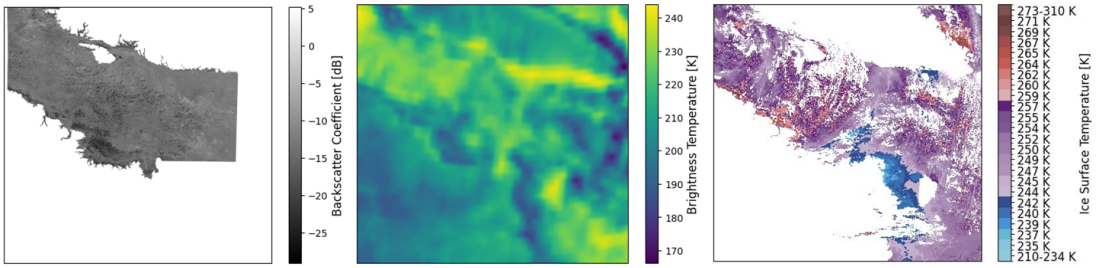


Figure 5.5: HH SAR (left), 36.5 GHZ H BT (middle), and IST (right) input channels for 2018-12-31.

5.2.2 Predictive Differences in Dual-Encoder U-Net Variants

Next, predictions from DEU-Net-SV and DEU-Net-V are compared to identify key differences in learning combined features from IST and SAR, and independent features from IST. Previous works fuse data from two instruments; thus, feature-level fusion methods learn features from each data type in different encoder branches. Because this study fuses data from four sources, two variations on this approach are trialled: grouping input channels by resolution (80-m pixel spacing and 2-km pixel spacing) and separating IST in its own encoder branch. Ultimately, three- or four-encoder architectures could also be explored. However, the scope of this study only considers grouping inputs by resolution (DEU-Net-SV), and determining if salient features learned from IST can reduce misclassification from SAR and passive microwave data (DEU-Net-V).

2020-01-24

Figure 5.6 is a scene off the coast of Labrador, Canada, which features mixed polygons of marginal and consolidated ice, in addition to an open water region. Referring to the input channels (Figure 5.7), the SAR imagery features open water with comparable-to-higher backscatter measurements than the ice (attributed to low incidence angle and or wind roughening), while the passive microwave measurements exhibit noise over water, likely due to atmospheric moisture or similar wind roughening. While IST data is largely masked in the open water region, the unmasked measurements exceeding 271 K do positively impact predictions, as seen in the more accurate DEU-Net-V SIC map. In contrast, because IST features are learned in combination with SAR for DEU-Net-SV, and IST data is largely masked, the model predominantly learns features from the ambiguous SAR data and is more likely to predict regions of low concentration ice. Regarding the misclassified region of 30% concentration, DEU-Net-SV incorrectly predicts young and FYI, as well as floes in the open water region. Even though DEU-Net-V predicts some low concentration marginal ice in this region, the model predicts predominantly correct open water for SOD and FLOE. Finally, while DEU-Net-V misclassifies the central region of big floe, predictions from other folds, which are trained on a different class distribution, do not predict this error. This variability is an example of sensitivity to the pixel class distribution, which contributes to the standard deviations reported in Table 5.1.

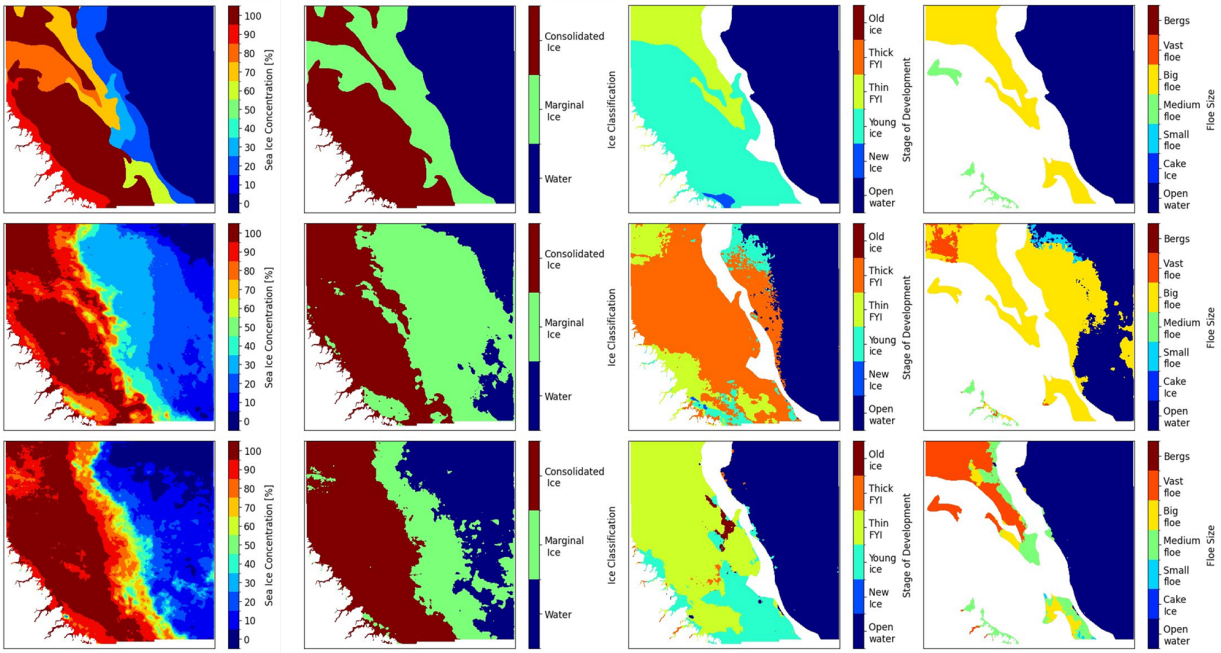


Figure 5.6: Summary of DEU-Net-SV (middle row) and DEU-Net-V (bottom row) model predictions for 2020-01-24 off the coast of Labrador. Predictions of SIC (left, column 1), SIC accuracy (column 2), SOD (column 3) and FLOE (column 4) are compared to ice chart labels (top row).

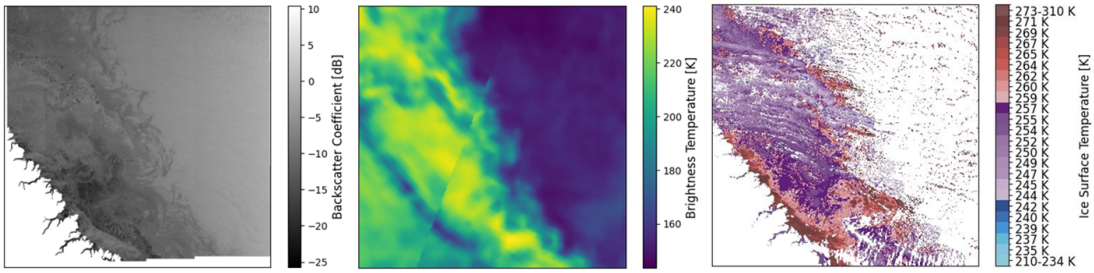


Figure 5.7: HH SAR (left), 36.5 GHZ H BT (middle), and IST (right) input channels for 2020-01-24.

5.2.3 Impacts of Task-specific Decoders

Because only one multi-decoder architecture is presented in this study, predictions from TDU-Net-V are compared to predictions from DEU-Net-V. Strengths and weaknesses of

using task-specific decoders are highlighted.

2021-03-01

Figure 5.8 is a scene off the west coast of Greenland characterized by a large region of consolidated ice, followed by a MIZ and a region of open water. For sea ice concentration, TDU-Net-V is more susceptible to noise in the BT measurements (Figure 5.9), resulting in incorrect predictions of 10% SIC. However, the use of separate decoders is seen to improve predictions of thin and thick FYI (SOD), as well as small floes (FLOE). Compared to DEU-Net-V, the single decoder architecture produces more accurate predictions over common classes and homogeneous regions (e.g., open water or 100% SIC/old ice/vast floe). Conversely, regions that differ between parameters (e.g., thin and thick FYI in a homogeneous region of 100% concentration) are more accurately estimated by using task-specific decoders. This trend may result for two reasons: 1) All models struggle to predict SIC polygon boundaries in MIZs with mixed concentrations. Thus, inaccurate SIC boundaries may impact predictions for SOD and FLOE. 2) An ice chart polygon defines the same region for SIC, SOD, and FLOE. However, a larger homogeneous region for one parameter can be composed of multiple polygons, which may report mixed conditions for another (e.g., comparing SIC and SOD). Here, the separate decoders are able to reconstruct different representations based on common features extracted by the same encoder. Thus, task-specific decoders may be better at predicting homogeneous regions for one or two parameters, when the third is mixed.

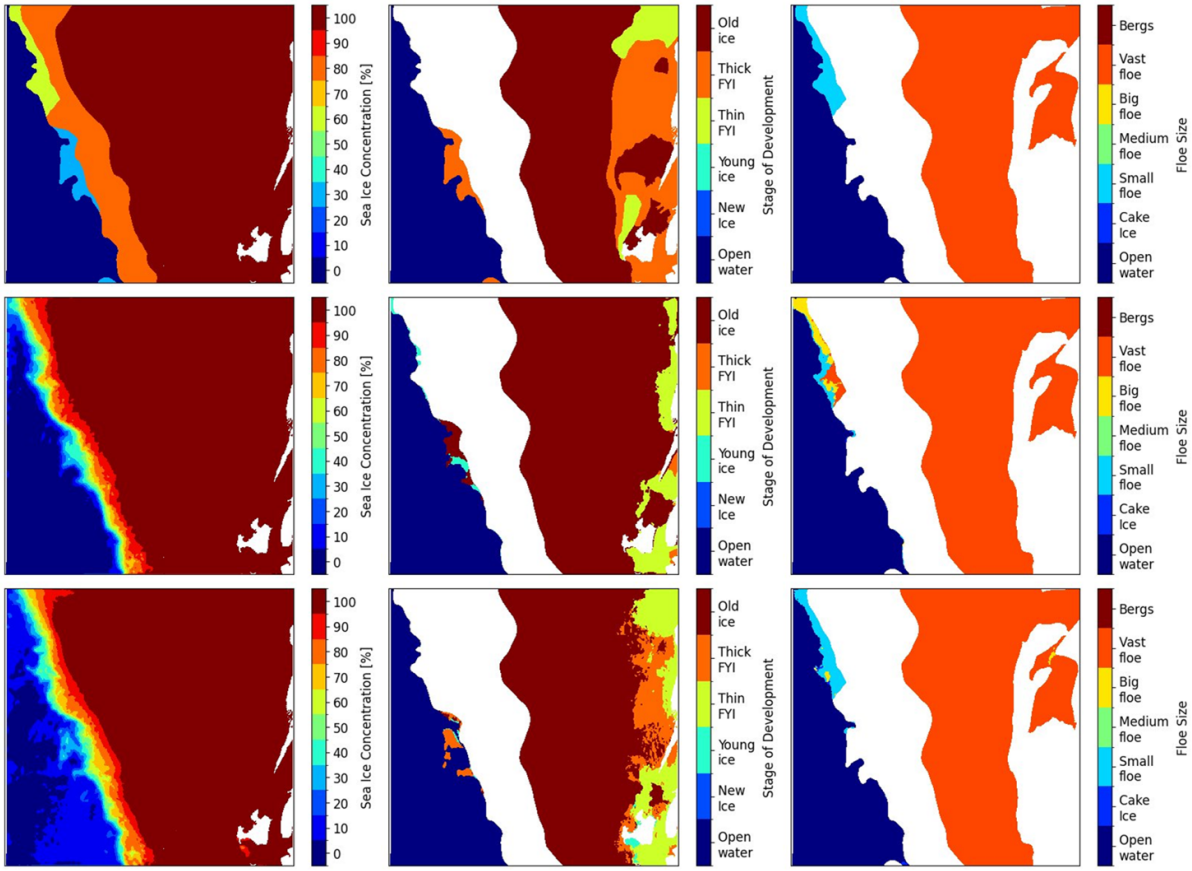


Figure 5.8: Summary of DEU-Net-V (middle row) and TDU-Net-V (bottom row) model predictions for 2021-03-01 off the west coast of Greenland. Predictions for SIC (left), SOD (middle) and FLOE (right) are compared to ice chart labels (top row).

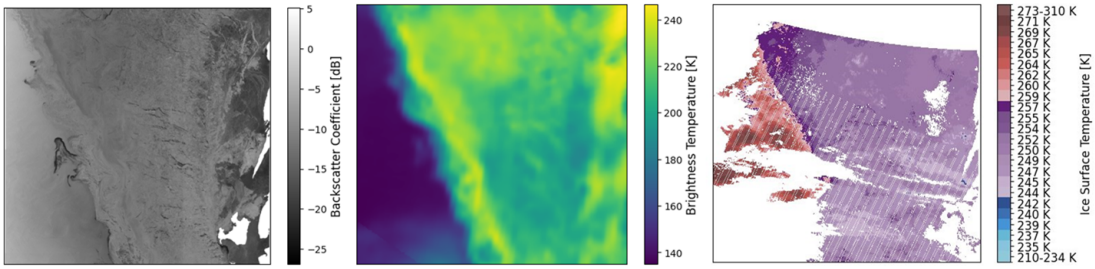


Figure 5.9: HH SAR (left), 36.5 GHZ H BT (middle), and IST (right) input channels for 2021-03-01. Note: the masked lines in the IST data are a visual artifact of the bow-tie effect. Affected scan line segments are assigned a fill value in the IST product [87].

2020-01-24

Returning to the scene off the coast of Labrador, Figure 5.10 demonstrates a case where the task-specific decoder performs worse than DEU-Net-V across all three parameters. While the model more closely approximates the polygons of 70 and 80% concentration that divide the region of consolidated ice, its performance across the shared open water class is considerably worse. Here, use of a single decoder reduces the impacts of ambiguous SAR and passive microwave data (Figure 5.7). Conversely, TDU-Net-V predicts different, incorrect regions of 20% concentration and young ice. Further, the model predicts open water as vast floe, which would not correlate with a region of 20% young ice.

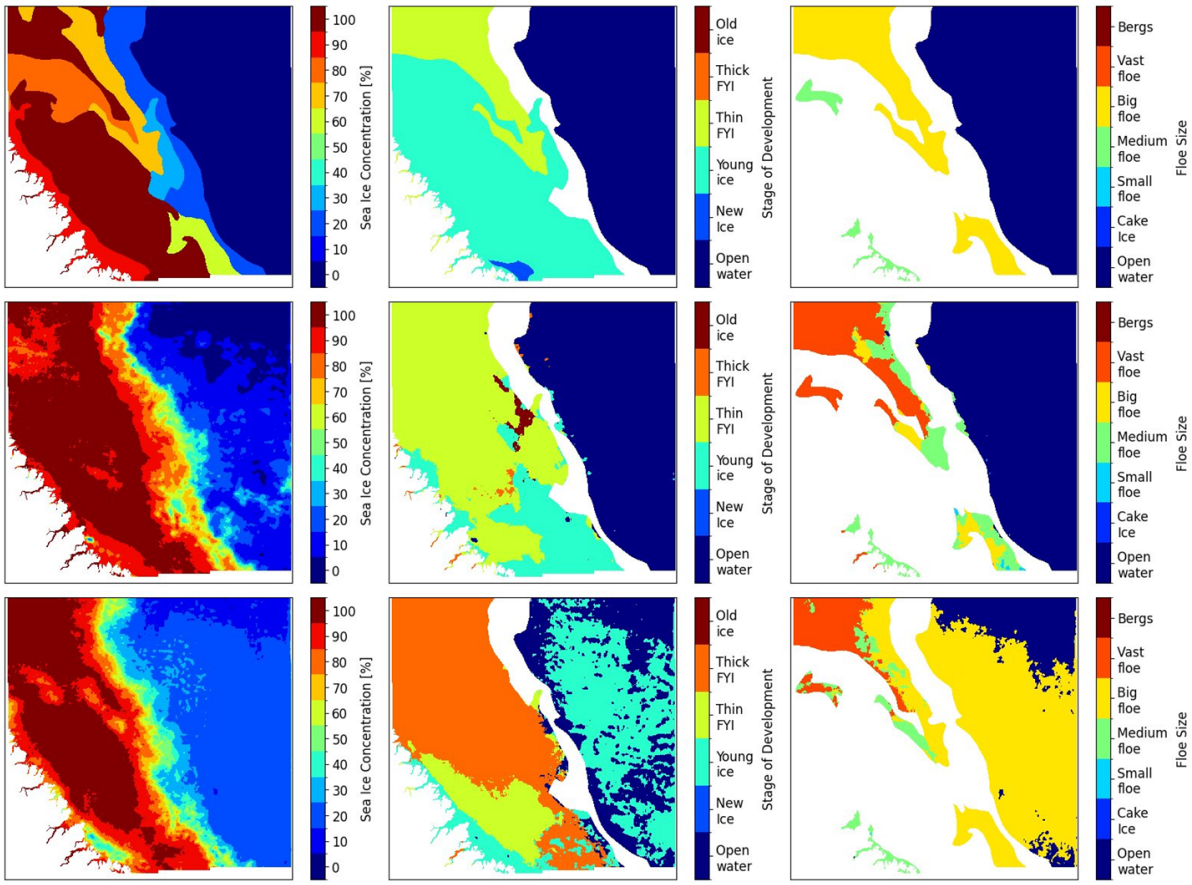


Figure 5.10: Summary of DEU-Net-V (middle row) and TDU-Net-V (bottom row) model predictions for 2020-01-24 off the coast of Labrador. Predictions for SIC (left), SOD (middle) and FLOE (right) are compared to ice chart labels (top row).

5.3 Improvements from DEU-Net-V

Finally, predictions from DEU-Net-V are compared to the base U-Net architecture, and the DEU-Net-V architecture with GNLL loss.

5.3.1 Improvements over Input-level Fusion

2019-01-19

Figure 5.11 is another scene off the coast of Labrador, Canada. Both DEU-Net-V and AutoICE incorrectly predict thin FYI in the homogeneous region of young ice, potentially influenced by SIC and FLOE predictions, which feature mixed conditions in this region (see Figure 5 in the appendix). Notably, low IST measurements result in DEU-Net-V correctly classifying the inland regions of thin FYI, which AutoICE misclassifies as open water (box a). Referring to the input channels, BT measurements are inaccurate near land boundaries, and the SAR imagery registers lower back scatter measurements, which may cause the classification error in the AutoICE model. Finally, neither model correctly predicts the top region of thin FYI (box b); referring to the IST data, this region is masked, causing the model to rely on SAR and passive microwave measurements.

2020-11-12

Figure 5.12 is a scene on the east coast of Greenland, characterized by largely consolidated, land-fast ice with interspersed marginal ice polygons, and a region of open water off the coast. Here, both single-encoder architectures largely misclassify the stage of development as new ice, whereas DEU-Net-V classifies the SOD as predominantly correct, old ice. Similarly, the dual-encoder architecture achieves a more accurate FLOE prediction. Referring to the input channels (Figure 5.13), there is minimal distinction between 90 and 100% concentration in the BT data, and vast and large floe in the SAR data. Further, the 90% concentration region between the peninsulas registers as lower back-scatter than the 90% concentration in other regions of the chart. IST measurements register higher temperatures across the big floe region, but this does not result in correct classification over the full region. DEU-Net-V overestimates 100% SIC; however, this appears to be learned in combination with the model's more correct old ice and vast/big floe classifications for SOD and FLOE. Conversely, AutoICE and AutoICE-V both overestimate 90% SIC. Finally, the DEU-Net-V's misclassification of open water as 10% ice aligns with the SAR back-scatter

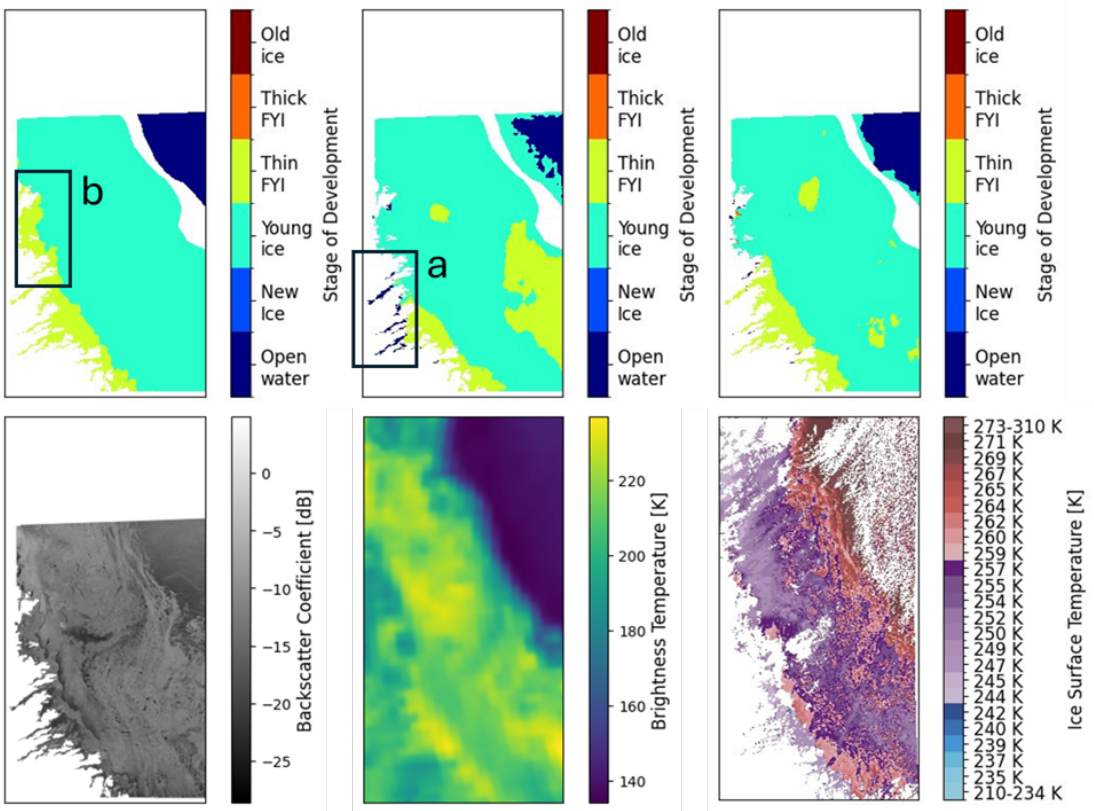


Figure 5.11: SOD model predictions for 2019-01-19 (top) off the coast of Labrador, comparing AutoICE (middle) and DEU-Net-V (right) to ice chart labels (left). HH SAR (left), 36.5 GHZ H BT (middle), and IST (right) input channels are presented in the bottom row.

readings in that region. This classification error may illustrate a case where learning SAR features in combination with IST would be beneficial. However, this error has a small impact overall because they are adjacent classes, and 10% SIC accounts for 0.11 and 0.19% of the train-validation and test sets, respectively.

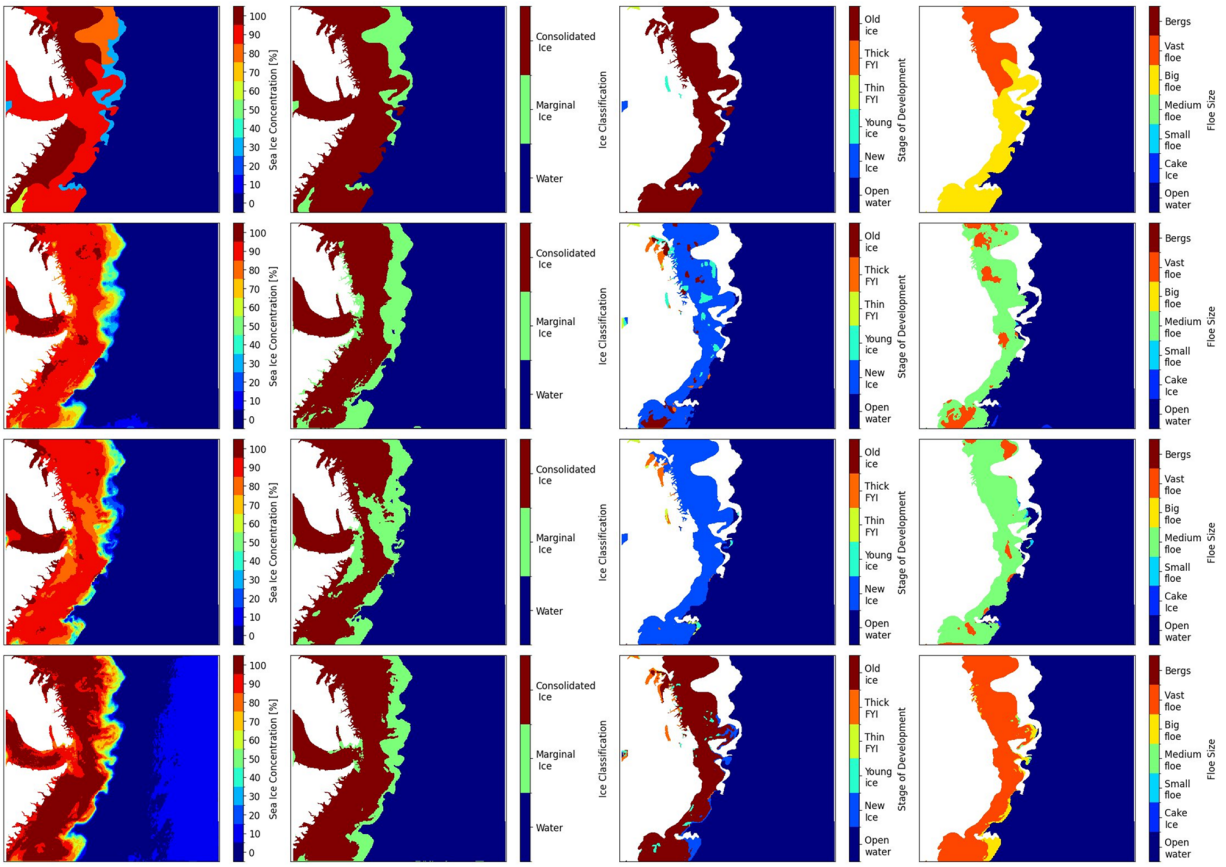


Figure 5.12: Summary of AutoICE (row 2), AutoICE-V (row 3) and DEU-Net-V (row 4) model predictions for 2020-11-12 on the east coast of Greenland. Predictions of SIC (left, column 1), SIC accuracy (column 2), SOD (column 3) and FLOE (column 4) are compared to ice chart labels (top, row 1).

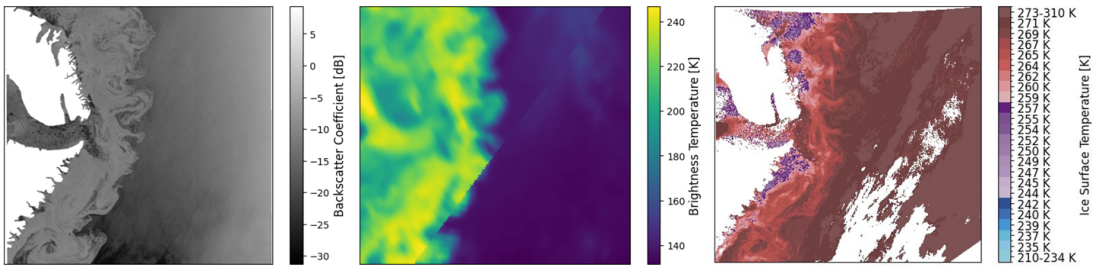


Figure 5.13: HH SAR (left), 36.5 GHZ H BT (middle), and IST (right) input channels for 2020-11-12.

5.3.2 Impacts of GNLL Loss

The impacts of implementing GNLL loss in place of MSE are assessed using the previously seen 2020-01-24 scene from the coast of Labrador (Figure 5.14). DEU-Net-V-GNLL more accurately predicts the shape of the MIZ, the polygons of 80 and 90% concentration, and the open water region. Referring to the variance map produced by the model (Figure 5.15), uncertainty is higher for predictions in the MIZ, and open water predictions improve where variance is minimized. This finding is expected, as ice chart polygons may vary the most from real-world pixel-level estimates for these concentrations. Specifically, consolidated ice and open water are characterized by the presence or absence of ice in a pixel. Conversely, marginal ice concentrations may differ at the pixel- and polygon-levels because polygons can contain multiple pixels with different conditions. Finally, DEU-Net-V-GNLL misclassifies consolidated ice as marginal ice along the land boundary (boxed). Referring back to the input channels (Figure 5.7), higher IST, lower BT, and higher SAR backscatter are observed in that region. This error is an example of how predicted variance can be used to assess model misclassifications; specifically, higher predictive uncertainty is reported in Figure 5.15, in line with the incorrect classification.

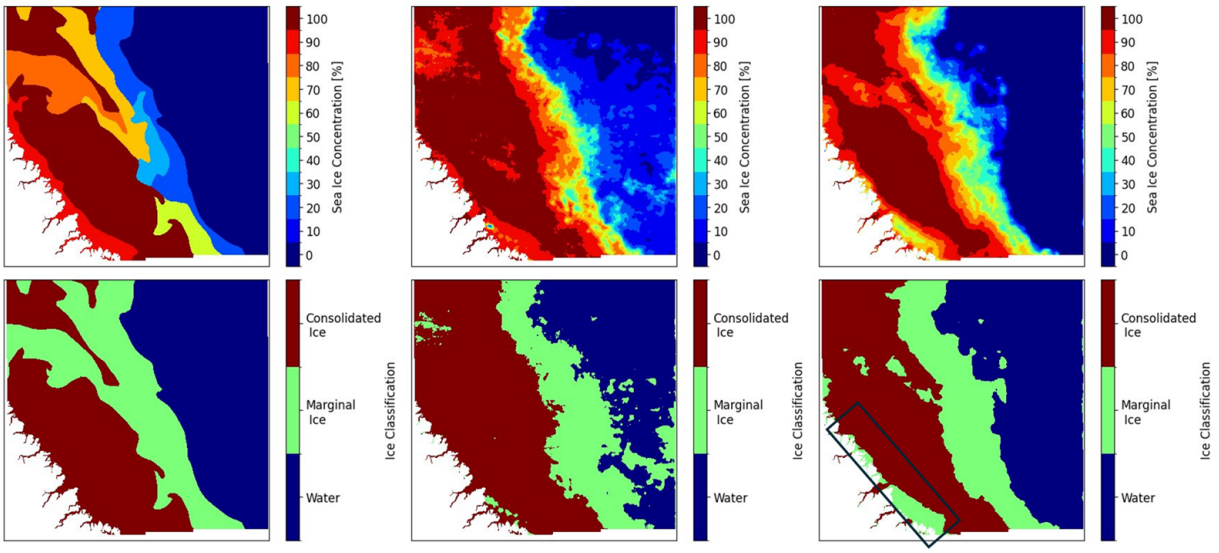


Figure 5.14: SIC (top) and SIC accuracy (bottom) model predictions for 2020-01-24 off the coast of Labrador. Predictions from DEU-Net-V (middle) and DEU-Net-V-GNLL (right) are compared to ice chart labels (left).

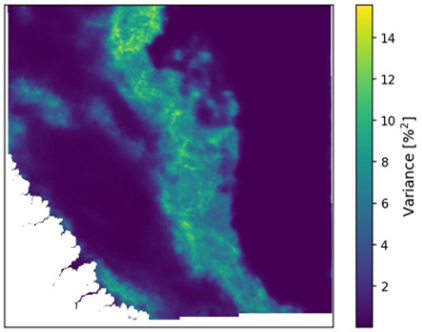


Figure 5.15: Variance map produced by DEU-Net-V-GNLL for 2020-01-24.

5.4 Impacts on Sea Ice Parameters

In the following section, the impacts of IST and the DEU-Net-V architecture are assessed, by parameter, using class-wise F1 scores. The baseline AutoICE architecture is compared to the best performing DEU-Net-V model, in addition to DEU-Net-V-GNLL. F1 scores are reported as the average over the six cross validation folds.

5.4.1 Sea Ice Concentration

The DEU-Net-V architectures see improvements over AutoICE in all but the 50% concentration class (Table 5.3). However, because the 10% through 80% classes represent between 0.19-4.2% of unmasked test set pixels, variation in F1 scores within the same magnitude is insignificant. Referring to Figure 5.16, which compares SIC confusion matrices for the test set, it is observed that all models struggle to predict the marginal ice classes. This finding agrees with results from the baseline AutoICE model using the full AI4Arctic dataset, although the confusion matrices differ based on dataset composition [88]. Overall, low class-wise F1 scores for these classes reflect the models' tendencies to misclassify lower marginal concentrations as any lower concentration, and mid-marginal concentrations (50-60%) as greater than 30% concentration. For concentrations of 70-90%, the models consistently predict these classes as consolidated ice, likely due to higher representation of these classes in the dataset. The distinction between classifications of low, mid, and high-marginal concentrations likely relate to relationships between SIC and the classification of SOD and FLOE ice types. All models consistently and correctly predict 100% ice concentration relative to the other classes. Additionally, the largest improvements from DEU-Net-V and DEU-Net-V-GNLL are seen in open water; this finding is

attributed to a reduction, relative to AutoICE, in the models predicting open water as low concentration ice.

Table 5.3: Summary of class-wise F1 scores for SIC, averaged over cross validation results for the test set.

SIC [%]	Train-Val Class Dist [%]	Test Class Dist [%]	Auto ICE F1 [%]	DEU-Net-V F1 [%]	DEU-Net-V-GNLL F1 [%]
0	21.92	24.65	69.13	82.98	**92.89**
10	0.11	0.19	1.42	1.47	**3.13**
20	1.11	2.13	12.69	**13.45**	9.19
30	2.11	1.26	9.37	**14.09**	8.16
40	0.72	0.51	5.08	6.99	**8.48**
50	1.58	0.29	**5.22**	4.04	4.66
60	0.70	1.34	10.40	11.53	**13.80**
70	1.85	2.82	5.49	7.33	**9.23**
80	5.21	4.20	21.18	**23.71**	23.59
90	10.03	12.84	32.85	**35.34**	32.58
100	54.66	49.78	84.45	84.18	**85.76**

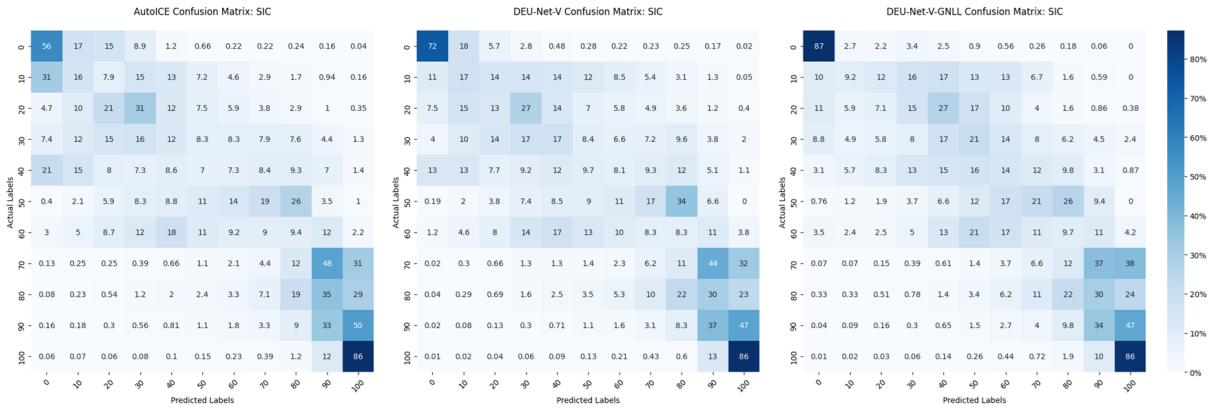


Figure 5.16: Confusion matrices for sea ice concentration: AutoICE (left), DEU-Net-V (middle), and DEU-Net-V-GNLL (right). Confusion matrices are the average of the test set across the six cross validation folds.

F1 scores and confusion matrices are also provided for sea ice concentration accuracy

in Table 5.4 and Figure 5.17, respectively. This representation more broadly assesses class-wise performance. Use of IST data and the DEU-Net-V architecture improves classification of open water and marginal ice, while all three models predict consolidated ice with comparable accuracy. Referring to Figure 5.17, improvements in predicting the open water and consolidated ice classes reduce misclassifications of marginal ice. Further, the models overwhelmingly misclassify marginal ice as consolidated ice, supporting the conclusion that their majority representation within the dataset skews predictions towards these classes. Comparing to AutoICE, reductions of 36% (DEU-Net-V) and 70% (DEU-Net-V-GNLL) demonstrate significant improvements in the classification of open-water regions where ambiguity in SAR and AMSR2 data impacts predictions.

Table 5.4: Summary of class-wise F1 scores for SIC accuracy, averaged over cross validation results for the test set.

Ice Type	**AutoICE F1 [%]**	**DEU-Net-V F1 [%]**	**DEU-Net-V-GNLL F1 [%]**
Open Water	69.15	83.01	**91.12**
Marginal Ice	42.11	50.50	**54.67**
Consolidated Ice	93.18	**93.67**	93.47

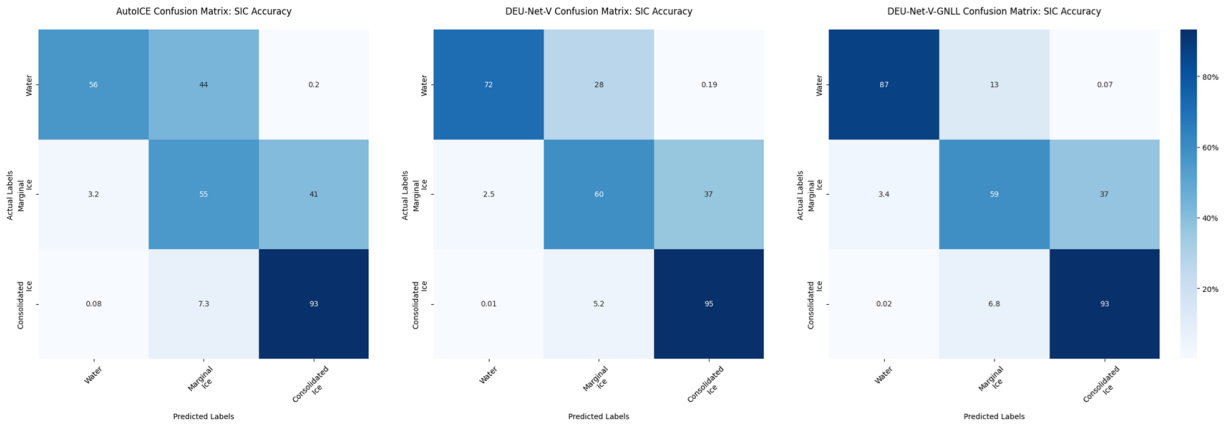


Figure 5.17: Confusion matrices for sea ice concentration accuracy: AutoICE (left), DEU-Net-V (middle), and DEU-Net-V-GNLL (right). Confusion matrices are the average of the test set across the six cross validation folds.

Finally, it must be understood that evaluation of SIC using class-wise assessments negatively impacts perceptions of the models' performances on intermediate classes. First, the

low total percentage of marginal class pixels within the train-validation dataset impacts a model's ability to learn these classes. Additionally, SIC estimation is treated as a regression problem, where concentrations are rounded to the nearest tenth. Thus, class-wise evaluation of this task unfairly penalizes incorrect predictions of neighbouring classes that the model was trained to forgive. This limitation is evident from the class-wise scores for sea ice concentration accuracy, where scores for the marginal ice class are higher than those calculated for the 11 discrete concentrations. Consequently, analysis within these broader, accepted categories presents a fairer assessment of model capabilities.

5.4.2 Stage of Development

Both dual-encoder architectures consistently improve class-wise F1 scores for SOD, as seen in Table 5.5. Referring to the test set confusion matrices (Figure 5.18), the most significant improvements are seen with open-water and old ice. In addition to ambiguous input channel measurements (e.g., from incidence angle, atmospheric moisture), this improvement encompasses cases such as Figure 5.11, where IST data is present and improves predictions for land-fast ice and tributaries. All three models are poor predictors of new ice, which is the least represented class. However, both dual-encoders reduce classification of new ice as open water, and increase correct classification of this ice type. Additionally, the models consistently misclassify thin FYI as thick FYI. This error may result from a combination of dataset representation and learned ice type relationships between SOD, SIC, and FLOE. Alternatively, this misclassification may reflect real-world conditions if both ice types are represented in a region but only the majority ice type (in this case thin FYI) is assigned as the polygon label. Lastly, Chen et al. [3] found that use of spatial-temporal encodings improved the accuracy of SOD predictions. VIIRS-AI4Arctic is a smaller dataset and sparser in terms of the encodings; thus, it would be useful to compare the models using a larger dataset, to determine if performance for the DEU-Net-V architecture remains comparable across certain classes (e.g., FYI).

Table 5.5: Summary of class-wise F1 scores for SOD, averaged over cross validation results for the test set.

SOD	Train-Val Class Dist [%]	Test Class Dist [%]	AutoICE F1 [%]	DEU-Net-V F1 [%]	DEU-Net-V-GNLL F1 [%]
Open Water	34.14	30.38	90.07	92.20	**94.51**
New Ice	2.16	0.41	0.53	5.17	**11.39**
Young Ice	12.29	16.51	61.30	**67.91**	67.50
Thin FYI	10.31	9.17	38.31	**39.74**	35.83
Thick FYI	32.89	33.44	77.98	**78.07**	72.34
Old Ice	8.22	10.10	79.07	79.99	**82.35**

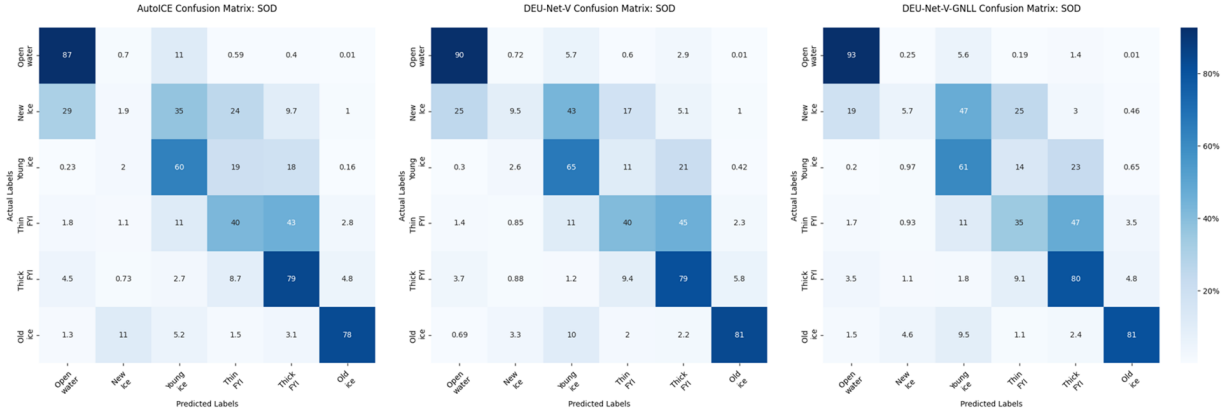


Figure 5.18: Confusion matrices for stage of development: AutoICE (left), DEU-Net-V (middle), and DEU-Net-V-GNLL (right). Confusion matrices are the average of the test set across the six cross validation folds.

5.4.3 Floe Size

The DEU-Net-V architecture consistently improves class-wise F1 scores for open water and medium and big floes (Table 5.6), resulting in the performance improvements seen in Table 5.1. Additionally, accuracy is comparable for vast floes. Cake ice is not represented in the test set, but is present in scenes from the train-validation set. Thus, no scores are reported for this class. AutoICE outperforms both dual-encoder models in small floe detection; this result is attributed to a combination of class distribution, polygon location, and architecture. Specifically, this class is underrepresented in the test set. DEU-Net-V is slightly more likely to classify small floe as medium floe, which is frequently an adjacent polygon, and errors in underrepresented classes have a greater impact on class-wise scores. Notably, FLOE classification errors often result from the model under- or over-estimating boundaries between adjacent polygons (e.g., in Figure 5.8, open water and small floe are adjacent), resulting in higher classification of non-adjacent classes, as seen in Figure 5.19. Additionally, in the case of small floes, learning the salient features of solely IST data may impact predictions, because the floe size (20-100 m) is smaller than the resolution of the IST product.

While DEU-Net-V-GNLL is the only model to correctly classify bergs, this result is discounted. The two test scenes containing berg labels are shown in Figure 5.20; these regions are boxed in the corresponding IST data, and this data is either mostly or entirely masked. Consequently, the limited presence of this class in the dataset (0.06% train-validation and 0.13% test), combined with the models being trained on randomized patches, suggests that the correct or incorrect classification of bergs is not representative of the models learning these conditions.

Table 5.6: Summary of class-wise F1 scores for FLOE, averaged over cross validation results for the test set.

FLOE	**Train-Val Class Dist [%]**	**Test Class Dist [%]**	**AutoICE F1 [%]**	**DEU-Net-V F1 [%]**	**DEU-Net-V-GNLL F1 [%]**
Open Water	31.86	34.63	88.62	**95.84**	95.59
Cake Ice	0.44	0.00	0.00	0.00	0.00
Small Floe	2.36	3.05	**20.24**	10.18	12.19
Medium Floe	7.37	10.06	69.55	72.37	**73.54**
Big Floe	7.94	7.95	29.61	**42.07**	39.55
Vast Floe	49.97	44.19	92.55	**92.78**	92.74
Bergs	0.06	0.13	0.00	0.00	**3.36**

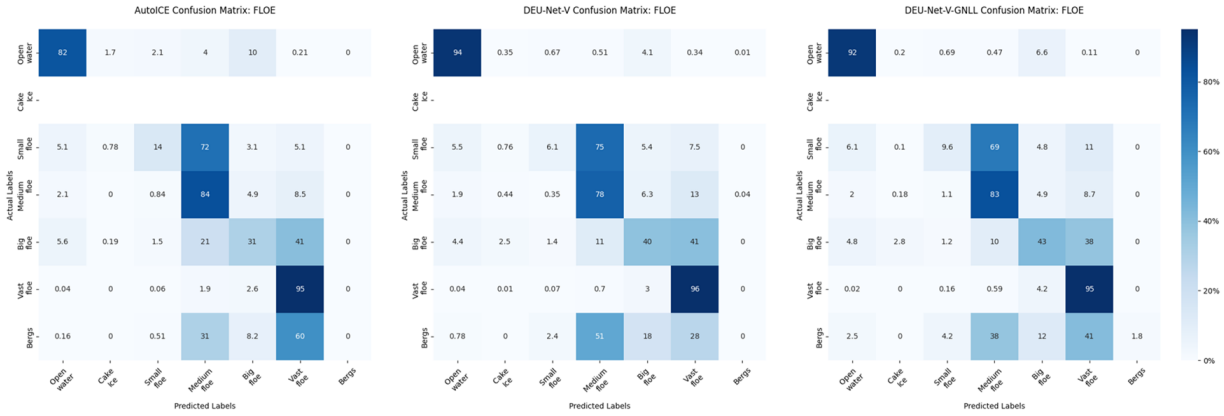


Figure 5.19: Confusion matrices for floe size: AutoICE (left), DEU-Net-V (middle), and DEU-Net-V-GNLL (right). Confusion matrices are the average of the test set across the six cross validation folds.

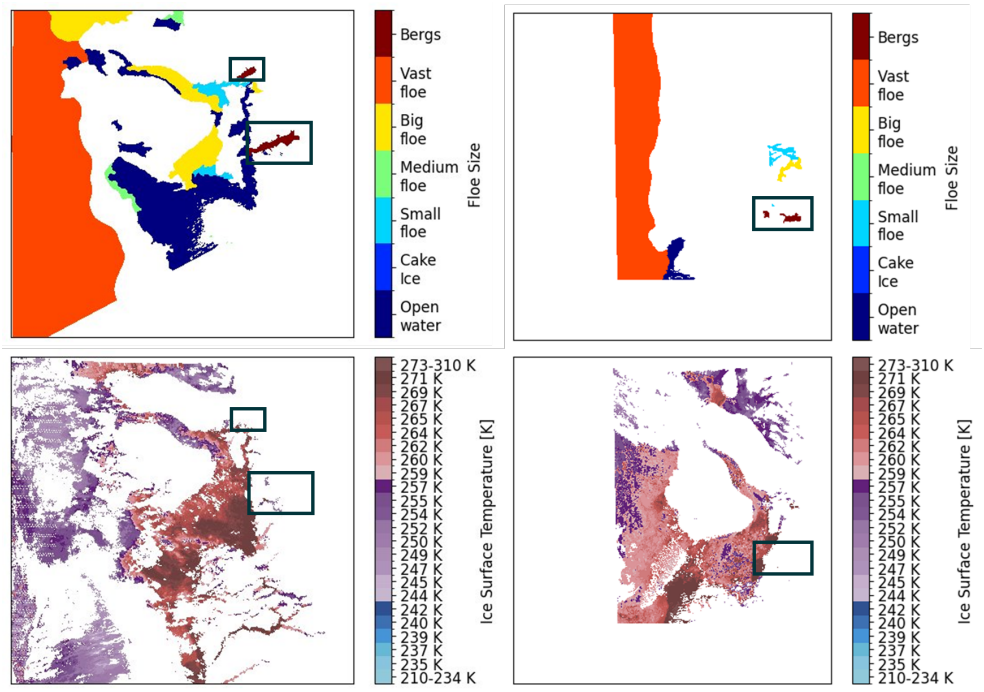


Figure 5.20: Comparison of the two scenes containing berg data: 2021-04-19 (left) and 2019-01-29 (right) on the west coast of Greenland. IST data (bottom) is compared to ice charts (top), demonstrating minimal-to-no IST data in regions containing bergs.

5.5 Impacts of Novel Evaluation Methods

In addition to the metrics used in evaluating models during the AutoICE Challenge (Section 4.4.1), this study introduced sea ice concentration accuracy and uncertainty quantification using Gaussian negative log likelihood loss. The following sections discuss the impacts of introducing these methods.

5.5.1 Sea Ice Concentration Accuracy

SIC accuracy provides a fairer evaluation metric by allowing for misclassifications within the marginal ice class while appropriately penalizing errors in open water, marginal ice, and consolidated ice classifications. Relative to open-water and consolidated ice, higher rates of misclassification for marginal ice (Section 5.4.1) are driven by a combination of low class representation within the dataset, and the task's treatment as a regression problem, which limits penalties for the model predicting adjacent classes. Further, variability in mid-range concentration estimates by ice analysts [53], and use of polygon labels at the pixel-level, may both contribute. Ultimately, SIC accuracy presents another means by which to evaluate and understand model predictions, which benefits marginal ice classes. Class-wise evaluation of marginal concentrations provides limited meaningful insight into model performance, while the R^2 score fails to penalize adjacent classes that belong to different categories. E.g., 10% and 20% are adjacent concentrations, but belong to the open-water and marginal ice categories, respectively.

While R^2 remains a valuable metric for comparing models on a shared dataset, SIC accuracy is found to reflect larger classification errors between the three concentration categories. For example, SIC accuracy in Figure 5.2 reflects the models' tendencies to over-report marginal ice, while Figure 5.6 highlights DEU-Net-SV's misclassification of open water as low concentration ice. Segmenting SIC maps into these three categories provides quick and effective visual analysis; however, the more significant impact of this metric is in reflecting misclassification outside of a label's category. Specifically, comparing R^2 and SIC accuracy for the test set (Table 5.1), SIC accuracy reports lower model performance, which is attributed to misclassification of open water as low concentration ice. Conversely, the validation set (Table 5.2) reports a higher accuracy; from analysis of the train-validation set, this is attributed to evaluation over a larger dataset, featuring a smaller percentage of scenes where input channel data results in misclassification of open water. Further, the metric allows for greater forgiveness of classification errors within the marginal and consolidated ice categories, relative to R^2. This forgiveness may be useful for considering

real-world, pixel-level conditions. Specifically, ice chart polygons are assigned a label based on the majority ice type, and overestimation of concentration is often linked to older/vaster SOD and FLOE. Thus, the ice type associated with the dominant concentration may skew SIC estimates to be higher than the total concentration, which is the sum of partial concentrations.

Ultimately, SIC accuracy is a useful metric for evaluation, which should be used in combination with the R^2 score. As long as the goal remains for the model to predict ice chart labels, the SIC task should involve training on labels for the 11 discrete concentrations, with a loss function that acknowledges relative distance between classes. Further, if models are compared using R^2 across studies, then the loss function must still penalize classification errors within those broader concentration categories. However, these thresholds for open-water, marginal ice, and consolidated ice are widely accepted ([50, 86, 51]). Thus, further assessment of model predictions using these categories can highlight common classification errors for an architecture, which are relevant when ice analysts consider adopting automated mapping approaches for safety- or science-critical applications (e.g., navigation or climate monitoring).

5.5.2 Uncertainty Quantification

Use of GNLL loss achieves comparable predictions to the DEU-Net-V architecture using MSE (Tables 5.1 and 5.2), meaning that inclusion of uncertainty quantification neither improves nor reduces model accuracy overall. Further, DEU-Net-V-GNLL demonstrates consistent improvements in classifying open water (and to a lesser degree, consolidated ice) across the three sea ice parameters. Figure 5.21 summarizes predicted variance maps for three additional test set scenes. Similarly to Figure 5.15, it is observed that uncertainty increases along the MIZ and ice edge, in agreement with previous studies [44]. Greater variance is expected in marginal ice regions because pixel-level conditions are expected to vary the most from ice chart labels. Thus, quantifying uncertainty may be useful for analysis of model predictions, which may be less accurate relative to the ice chart labels, but more accurate relative to real-world conditions. Alternatively, when deploying automated mapping approaches, variance maps may be useful for flagging scenes or regions that require additional verification by ice analysts.

Additionally, low variance is reported over regions of homogeneous concentrations. In combination with improved classification accuracy, this finding suggests that as the model learns to minimize predictive uncertainty, it consistently improves predictions for lower variance, homogeneous regions. While this study presents positive preliminary findings on

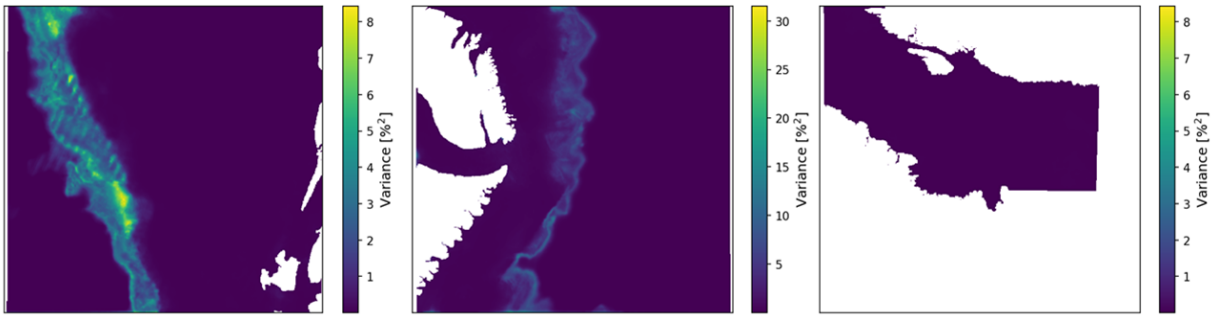


Figure 5.21: Variance maps produce by DEU-Net-V-GNLL for 2021-03-01 (left), 2021-04-19 (middle), and 2018-12-31 (right).

incorporating uncertainty quantification in multi-task deep learning architectures, further work is required to assess the impacts of IST data on label uncertainty. Specifically, variance estimates should be compared to AutoICE incorporating GNLL loss in order to distinguish between the impacts of the loss function and IST data.

Chapter 6

Conclusion

6.1 Summary

This thesis investigated the use of thermal-infrared data, specifically ice surface temperature, for improving sea ice parameter predictions in regions where SAR and passive microwave data are challenging to interpret. Chapter 3 discusses the design and co-registration of a VIIRS IST dataset that can be used in conjunction with the AI4Arctic dataset. 84 AI4Arctic scenes acquired from January 2018 to December 2021 and encompassing November through May are found to contain up to 40% cloud-free IST data in their associated VIIRS imagery. This work uses established criteria to select compatible VIIRS scenes, and demonstrates that sufficient coverage and overpass windows exist between VIIRS, Sentinel-1, and AMSR2 to expand the use of this product with existing instruments. While the dataset captures examples of freeze-up and break-up within the same regions, the exclusion of scenes from June to October impacts the ability to assess IST data under late-season melt conditions and early freeze-up. Further, the exclusion of these months highlights the limitation of using a passive product that cannot register accurate measurements in cloud-contaminated pixels.

Chapter 4 introduced three series of U-Net architectures (AutoICE, DEU-Net, and TDU-Net) for multi-task estimation of sea ice concentration, stage of development, and floe size. Models fusing IST with AI4Arctic channels at the input- and feature-levels are compared to the winning AI4EO AutoICE Challenge architecture, establishing baseline performance on the VIIRS-AI4Arctic dataset. AutoICE-V highlights the impacts of using IST data with the baseline architecture. DEU-Net-SV explores the impacts of features learned at the same resolution, while DEU-Net-V establishes that IST features learned

independently from AI4Arctic channels reduce classification errors. Finally, TDU-Net-V investigates how task-specific decoders affect model predictions. This chapter also introduces SIC accuracy as a method for evaluating model predictions, and presents the implementation of Gaussian negative log likelihood loss in the DEU-Net-V architecture as a means of quantifying aleatoric uncertainty.

In Chapter 5, models are assessed using a combination of evaluation metrics, class-wise F1 scores, and visual analysis of predicted sea ice maps. Overall, it is found that learning the salient features of IST data using the DEU-Net-V architecture reduces ambiguity introduced by SAR and passive microwave features. Across all models, IST data in combination with brightness temperature reduces the impacts of ambiguous SAR textures. Further, IST measurements consistently improve the classification of open water where incidence angle impacts backscatter measurements (SAR) and atmospheric moisture or wind increase brightness temperature readings (BT).

Comparing the dual-encoder architectures, complimentary features learned between SAR and IST data result in higher rates of misclassification due to instances where IST data is masked and SAR is ambiguous. Thus, learning complimentary SAR and PM features, where BT effectively distinguishes between ice and water, improves predictions. Concurrently, salient features learned from IST data are able to reduce SAR and PM ambiguities where data is present. Overall, masked data and sensor resolution (relative to small floes and smaller tributaries) constrain the potential for IST to improve accuracy. Finally, task-specific decoders are found to improve predictions that are unique to an individual parameter (e.g. mixed SOD conditions but consistent SIC and FLOE). However, this architecture struggles with detection of the shared open-water class. While all models report variability in this classification, TDU-Net-V predicts incongruent conditions across the three parameters, demonstrating that multi-task classification using a shared decoder improves contextual awareness.

Finally, incorporation of GNLL loss in the DEU-Net-V architecture is found to achieve comparable accuracy and improve predictions in regions with homogeneous conditions. Variance maps report the highest uncertainty at the MIZ and ice edge, providing a valuable method of quantifying uncertainty for safety-critical or climatological applications. Further, this heightened uncertainty may represent differences between ice chart labels and pixel-level conditions in regions of marginal concentration. Specific to AI4Arctic ice chart labels, use of majority ice types ($\geq$65% of a polygon) and simplification of ice type categories both reduce the accuracy of pixel-level labels. Further, all models misclassify marginal concentrations, which is attributed to low class representation, pixel-level variability, and the task's treatment as a regression problem. Consequently, a combination of uncertainty quantification and evaluation using the SIC accuracy metric is found to yield

more relevant findings on model performance. Ultimately, SIC accuracy is found to be a useful metric, which, when used in combination with R^2, highlights misclassifications of open water, marginal ice, and consolidated ice.

6.2 Future Work

6.2.1 Expanding on Current Architectures

Previous research reports similar variability in model predictions for the open-water class, which should be consistent across SIC, SOD, and FLOE [89]. Building on the DEU-Net-V architecture, use of a second decoder, which accomplishes an additional ice-water classification task, is one potential method for improving model accuracy. A dedicated decoder would allow the model to focus on distinguishing between the two classes. Further, in learning shared features that are useful for both ice-water classification and sea ice parameter estimation, the model may learn better feature representations that improve model accuracy and generalization. Alternatively, the second decoder could classify open-water, marginal ice, and consolidated ice, in line with the SIC accuracy classes. In addition to improving ice-water classification, this approach has the potential to reduce misclassification of marginal ice as consolidated ice.

Selection of AI4Arctic scenes where IST data contains less than 60% cloud cover restricts dataset size and spatial-temporal coverage, in addition to excluding certain conditions (e.g., melt). This work concludes that IST data does improve model predictions; thus, future work could increase the allowed threshold for cloud cover, with model evaluation acknowledging that IST cannot improve predictions where no data is present. Because DEU-Net-V uses a separate encoder for IST features, it is possible that a different architecture might improve generalization when the percentage of IST data in a scene is limited.

6.2.2 Further Assessing Uncertainty Quantification

This study explored the use of Gaussian negative log-likelihood loss in a positive yet limited context. Existing literature critiques GNLL loss for producing suboptimal regression predictions due to overconfident variance estimates and premature convergence on poor initial predictions [90]. Although similar behaviour was not observed in this work, it is noted that multiple authors have proposed alternative loss formulations to address these issues [90, 91].

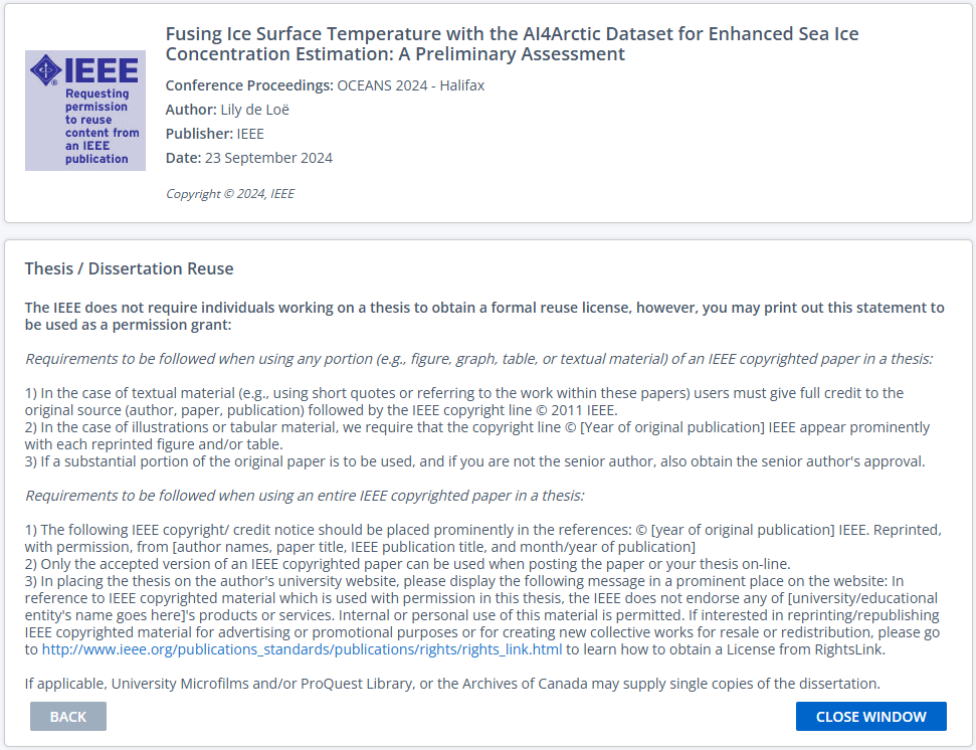
This study only applied GNLL loss to the best-performing model. Thus, use of GNLL loss in other architectures, such as AutoICE, may yield suboptimal predictions, which would complicate comparisons. Future research should strengthen this study's preliminary conclusions by comparing variance estimates across architectures (AutoICE and DEU-Net-V) and examining IST's impact on model uncertainty (AutoICE and AutoICE-V). This investigation should also explore improved approaches and loss functions for estimating aleatoric uncertainty.

6.2.3 Leveraging Next-Generation Thermal-Infrared Instruments

Ice surface temperature data has been shown to improve sea ice parameter predictions, although its effectiveness is limited by instrument resolution and atmospheric effects. Upcoming satellite missions offer promising advancements in thermal-infrared products, which may reduce these limitations. For example, the TRISHNA satellite, set to launch in 2026, will provide a spatial resolution of 60 meters and global revisit period of three days [92]. Higher-resolution thermal-infrared data could enhance sea ice parameter estimation near land boundaries and within inland features, such as tributaries. Further, improved spatial resolution may also increase data retrieval within cloudy regions. Although these products will not overlap with the AI4Arctic dataset, they could complement existing algorithms for estimating SIC, such as those by Ludwig et al. [28] and Konig et al. [29]. Additionally, combining higher-resolution products with SAR imagery can refine pixel-level labels of ice conditions, enabling advancements beyond ice chart labels and supporting the exploration of semi-supervised or transfer learning approaches [93, 65]. Ultimately, the benefits of accurate, high-resolution, automated sea ice mapping can be realized through emerging instruments and efficient data retrieval systems. This study advances sea ice parameter estimation with current products, and supports investment in future instruments.

Letters of Copyright Permission

Fusing Ice Surface Temperature with the AI4Arctic Dataset for Enhanced Sea Ice Concentration Estimation: A Preliminary Assessment [1]:



Fusing Ice Surface Temperature with the AI4Arctic Dataset for Enhanced Sea Ice Concentration Estimation: A Preliminary Assessment

Conference Proceedings: OCEANS 2024 - Halifax

Author: Lily de Loë

Publisher: IEEE

Date: 23 September 2024

Copyright © 2024, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

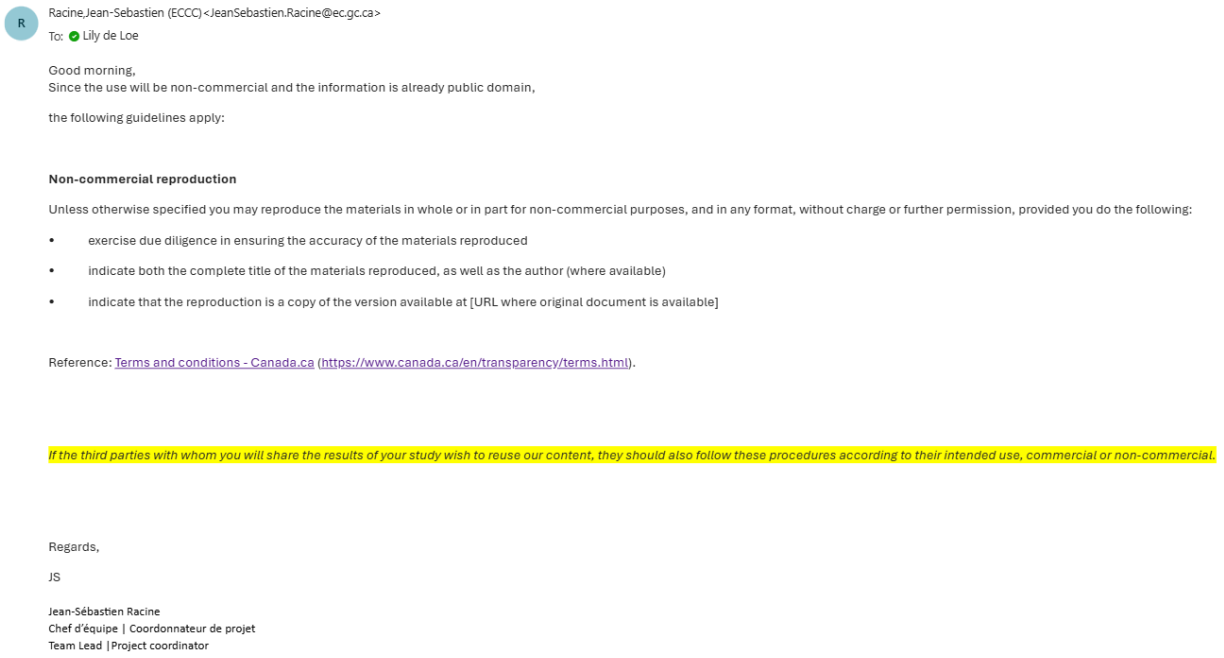
Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

Daily Ice Chart colour WMO CT - Davis Strait - WIS33CT - 2025/01/12, adapted from [52]:



Racine,Jean-Sebastien (ECCC)<JeanSebastien.Racine@ec.gc.ca>
To: Lily de Loe

Good morning,
Since the use will be non-commercial and the information is already public domain,

the following guidelines apply:

Non-commercial reproduction

Unless otherwise specified you may reproduce the materials in whole or in part for non-commercial purposes, and in any format, without charge or further permission, provided you do the following:

- exercise due diligence in ensuring the accuracy of the materials reproduced
- indicate both the complete title of the materials reproduced, as well as the author (where available)
- indicate that the reproduction is a copy of the version available at [URL where original document is available]

Reference: Terms and conditions - Canada.ca (https://www.canada.ca/en/transparency/terms.html).

If the third parties with whom you will share the results of your study wish to reuse our content, they should also follow these procedures according to their intended use, commercial or non-commercial.

Regards,

JS

Jean-Sébastien Racine
Chef d'équipe | Coordonnateur de projet
Team Lead |Project coordinator

References

[1] L. de Loë, L. Xu, and K. A. Scott, "Fusing ice surface temperature with the AI4Arctic Dataset for enhanced sea ice concentration estimation: A preliminary assessment," in *OCEANS 2024-Halifax*, pp. 1–9, IEEE, 2024.

[2] J. Buus-Hinkler, T. Wulf, A. Stokholm, A. Korosov, R. Saldo, L. Pedersen, D. Arthurs, R. Solberg, N. Longépé, and M. Brandt Kreiner, "Ready-To-Train AI4Arctic Sea Ice Challenge Dataset," 2022.

[3] X. Chen, M. Patel, F. J. Pena Cantu, J. Park, J. Noa Turnes, L. Xu, K. A. Scott, and D. A. Clausi, "MMSeaIce: Multi-task mapping of sea ice parameters from AI4Arctic Sea Ice Challenge Dataset," *The Cryosphere*, vol. 18, no. 4, pp. 1621–1632, 2024.

[4] European Space Agency, "ENVISAT satellite: Advanced Synthetic Aperture Radar (ASAR) data," 2007.

[5] Copernicus; European Space Agency, "Sentinel 1A C-band Synthetic Aperture Radar (SAR): Extra Wide (EW) mode Ground Range Detected (GRD) Medium Resolution (MR) Level 1 data," 2015.

[6] European Space Agency (ESA), "Sentinel-3 Sea and Land Surface Temperature Radiometer (SLSTR) Level-1B Radiances and Brightness Temperature Data." https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-3-slstr, 2016. Copernicus Programme.

[7] K. A. Kilpatrick, G. P. Podesta, and R. Evans, "Overview of the NOAA/NASA Advanced Very High Resolution Radiometer Pathfinder algorithm for sea surface temperature and associated matchup database," *Journal of Geophysical Research*, vol. 106, no. C5, pp. 9179–9197, 2001.

[8] D. K. Hall and G. A. Riggs., "MODIS/Aqua Sea Ice Extent 5-Min L2 Swath 1km, Version 61," 2021.

[9] M. G. Tschudi, D. K. Riggs, Hall, and M. O. Román., "VIIRS/NPP Ice Surface Temperature 6-Min L2 Swath 750m, Version 1," 2017.

[10] R. Kwok, A. A. Petty, G. Cunningham, T. Markus, D. Hancock, A. Ivanoff, J. Wimert, M. Bagnardi, N. Kurtz, and ICESat-2 Science Team, "ATLAS/ICESat-2 L3A Sea Ice Height, Version 6," 2023.

[11] Canadian Space Agency, "RADARSAT-1 data products specifications," Technical Report RSI-GS-026, Canadian Space Agency, Canada, Aug. 2004.

[12] Maxar Technologies Ltd., "RADARSAT-2 product description," Sept. 2018. Issue 1/14.

[13] J. C. Comiso, "AMSR-E/AMSR2 Unified L3 Daily 12.5 km Brightness Temperatures, Sea Ice Concentration, Motion & Snow Depth Polar Grids, Version 1," 2018.

[14] W. Meier, A. Petty, S. Hendricks, L. Kaleschke, D. Divine, S. Farrell, S. Gerland, D. Perovich, R. Ricker, X. Tian-Kunze, and M. Webster, "Sea ice: Arctic report card 2023," tech. rep., National Oceanic and Atmospheric Administration, 2023.

[15] G. Boutin, T. Williams, P. Rampal, E. Olason, and C. Lique, "Impact of wave-induced sea ice fragmentation on sea ice dynamics in the MIZ," tech. rep., Copernicus GmbH, 2020.

[16] N. Melia, K. Haines, and E. Hawkins, "Sea ice decline and 21st century trans-Arctic shipping routes," *Geophysical Research Letters*, vol. 43, no. 18, pp. 9720–9728, 2016.

[17] A. W. Bateson, D. L. Feltham, D. Schröder, L. Hosekova, J. K. Ridley, and Y. Aksenov, "Impact of sea ice floe size distribution on seasonal fragmentation and melt of Arctic sea ice," *The Cryosphere*, vol. 14, no. 2, pp. 403–428, 2020.

[18] M. Rantanen, A. Y. Karpechko, A. Lipponen, K. Nordling, O. Hyvärinen, and R. Kimmo et al., "The Arctic has warmed nearly four times faster than the globe since 1979.," *Communications Earth Environment*, vol. 3, 2022.

[19] Y. Liu, J. R. Key, X. Wang, and M. Tschudi, "Multidecadal Arctic sea ice thickness and volume derived from ice age," *The Cryosphere*, vol. 14, p. 1325–1345, 2020.

[20] L. Beaulieu, A. Arreak, R. Holwell, S. Dicker, O. Qamanirq, L. Moorman, K. Wilson, R. Segal, S. Crichton, and T. Bell, "Indigenous self-determination in cryospheric science: The Inuit-led Sikumik Qaujimajjuti ("tools to know how the ice is") program in Inuit Nunangat, Canada," *Frontiers in Earth Science*, vol. 11, 2023.

[21] J. Eamer, G. Donaldson, A. Gaston, K. Kosobokova, K. Lárusson, I. Melnikov, J. Reist, E. Richardson, L. Staples, and C. von Quillfeldt, "Life linked to ice: A guide to sea-ice-associated biodiversity in this time of rapid change," Tech. Rep. 10, Conservation of Arctic Flora and Fauna, Iceland, 2013.

[22] N. Asadi, P. Lamontagne, M. King, M. Richard, and K. A. Scott, "Probabilistic spatiotemporal seasonal sea ice presence forecasting using sequence-to-sequence learning and ERA5 data in the Hudson Bay region," *The Cryosphere*, vol. 16, no. 9, pp. 3753–3773, 2022.

[23] J. D. Ford, W. A. Gough, G. J. Laidler, J. MacDonald, C. Irngaut, and K. Qrunnut, "Sea ice, climate change, and community vulnerability in northern Foxe Basin, Canada," *Climate Research*, vol. 38, no. 2, pp. 137–154, 2009.

[24] A. R. Stokholm, J. Buus-Hinkler, T. Wulf, A. Korosov, R. Saldo, and L. T. Pedersen et al., "The AutoICE Challenge," *EGUsphere*, vol. 2023, pp. 1–28, 2023.

[25] M. Leppäranta and J. E. Lewis, "Observations of ice surface temperature and thickness in the Baltic Sea," *International Journal of Remote Sensing*, vol. 28, no. 17, pp. 3963–3977, 2007.

[26] T. Zeng, L. Shi, M. Mäkynen, and J. Zhang, "Sea ice thickness analyses for the Bohai Sea using MODIS thermal infrared imagery," *Acta Oceanologica Sinica*, vol. 35, pp. 96–104, may 2016.

[27] C. Melsheimer, , and G. Spreen, "AMSR2 ASI sea ice concentration data, Arctic, version 5.4 (NetCDF) (July 2012 - December 2018)," 2019.

[28] V. Ludwig, G. Spreen, and L. T. Pedersen, "Evaluation of a new merged sea-ice concentration dataset at 1 km resolution from thermal infrared and passive microwave satellite data in the Arctic," *Remote Sensing*, vol. 12, no. 19, 2020.

[29] C. König, T. König, S. Singha, A. Frost, and S. Jacobsen, "Combined use of space borne optical and SAR data to improve knowledge about sea ice for shipping," *Remote Sensing*, vol. 13, no. 23, 2021.

[30] A. S. Komarov, A. Caya, L. Pogson, and M. Buehner, "Assimilation of RCM data in the Canadian ice concentration analysis system," *Remote Sensing of Environment*, vol. 306, p. 114113, 2024.

[31] Environment Canada, "Manual of Standard Procedures for Observing and Reporting Ice Conditions (MANICE)," 2005. Issuing authority: Assistant Deputy Minister, Meteorological Service of Canada.

[32] D. Malmgren-Hansen, L. T. Pedersen, A. A. Nielsen, M. B. Kreiner, R. Saldo, H. Skriver, J. Lavelle, J. Buus-Hinkler, and K. H. Krane, "A convolutional neural network architecture for Sentinel-1 and AMSR2 data fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 1890–1902, 2021.

[33] H. Boulze, A. Korosov, and J. Brajard, "Classification of sea ice types in Sentinel-1 SAR data using convolutional neural networks," *Remote Sensing*, vol. 12, no. 13, 2020.

[34] K. Radhakrishnan, K. A. Scott, and D. A. Clausi, "Sea ice concentration estimation: Using passive microwave and SAR data with a U-Net and curriculum learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 5339–5351, 2021.

[35] Y. Wang and X. Li, "Arctic sea ice cover data from spaceborne synthetic aperture radar by deep learning," *Earth System Science Data*, vol. 13, no. 6, pp. 2723–2742, 2021.

[36] A. Stokholm, T. Wulf, A. Kucik, R. Saldo, J. Buus-Hinkler, and S. M. Hvidegaard, "AI4SeaIce: toward solving ambiguous SAR textures in convolutional neural networks for automatic sea ice concentration charting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[37] M. Rogers, M. Fox, A. Fleming, L. van Zeeland, J. Wilkinson, and J. Hosking, "Sea ice detection using concurrent multispectral and synthetic aperture radar imagery," *Remote Sensing of Environment*, vol. 305, p. 114073, 2024.

[38] S. C. Deodoro, R. de Andrade Moral, R. Fealy, T. McCarthy, and R. Fealy, "Using the surface scattering mechanism from dual-pol SAR data to estimate topsoil particle-sizefractions," *International Journal of Applied Earth Observation and Geoinformation*, vol. 128, p. 103742, 2024.

[39] G. J. Macdonald, R. K. Scharien, K. Duncan, S. L. Farrell, P. Rezania, and A. Tavri, "Arctic sea ice topography information from RADARSAT Constellation Mission (RCM) synthetic aperture radar (SAR) backscatter," *Geophysical Research Letters*, vol. 51, no. 4, p. e2023GL107261, 2024.

[40] L. Pogson, T. Geldsetzer, M. Buehner, T. Carrieres, M. Ross, and K. A. Scott, "Collecting empirically derived SAR characteristic values over one year of sea ice environments for use in data assimilation," *Monthly Weather Review*, vol. 145, no. 1, pp. 323 – 334, 2017.

[41] J. Chi, H. cheol Kim, S. Lee, and M. M. Crawford, "Deep learning based retrieval algorithm for Arctic sea ice concentration from AMSR2 passive microwave and MODIS optical data," *Remote Sensing of Environment*, vol. 231, p. 111204, 2019.

[42] T. Feng, X. Liu, and R. Li, "Super-resolution-aided sea ice concentration estimation from AMSR2 images by encoder–decoder networks with atrous convolution," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 962–973, 2023.

[43] Q. Liu, R. Zhang, Y. Wang, H. Yan, J. Xu, and Y. Guo, "Short-term forecasting of sea ice thickness based on PredRNN++," *Journal of Physics: Conference Series*, vol. 2486, p. 012017, may 2023.

[44] X. Chen, R. Valencia, A. Soleymani, and K. A. Scott, "Predicting sea ice concentration with uncertainty quantification using passive microwave and reanalysis data: A case study in Baffin Bay," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[45] W. N. Meier and J. S. Stewart, "Assessing the potential of enhanced resolution gridded passive microwave brightness temperatures for retrieval of sea ice parameters," *Remote Sensing*, vol. 12, no. 16, 2020.

[46] N. Ivanova, L. T. Pedersen, R. T. Tonboe, S. Kern, G. Heygster, T. Lavergne, A. Sørensen, R. Saldo, G. Dybkjær, L. Brucker, and M. Shokr, "Inter-comparison and evaluation of sea ice algorithms: Towards further identification of challenges and optimal approach using passive microwave observations," *The Cryosphere*, vol. 9, no. 5, pp. 1797–1817, 2015.

[47] K. Naoki, J. Ukita, F. Nishio, M. Nakayama, J. C. Comiso, and A. Gasiewski, "Thin sea ice thickness as inferred from passive microwave and in situ observations," *Journal of Geophysical Research: Oceans*, vol. 113, no. C2, 2008.

[48] M. A. Tschudi, G. A. Riggs, D. K. Hall, and M. O. Román, "Suomi-NPP VIIRS Ice Surface Temperature algorithm theoretical basis document (ATBD)," tech. rep., NASA Goddard Space Flight Center, Greenbelt, MD, 2016.

[49] J. C. Comiso, W. N. Meier, and R. Gersten, "Variability and trends in the Arctic Sea ice cover: Results from different techniques," *Journal of Geophysical Research: Oceans*, vol. 122, no. 8, pp. 6883–6900, 2017.

[50] C. Strong, D. Foster, E. Cherkaev, I. Eisenman, and K. M. Golden, "On the definition of marginal ice zone width," *Journal of Atmospheric and Oceanic Technology*, vol. 34, no. 7, pp. 1565 – 1584, 2017.

[51] A. W. Bateson, D. L. Feltham, D. Schröder, Y. Wang, B. Hwang, J. K. Ridley, and Y. Aksenov, "Sea ice floe size: Its impact on pan-Arctic and local ice mass and required model complexity," *The Cryosphere*, vol. 16, no. 6, pp. 2565–2593, 2022.

[52] Government of Canada, "Latest ice conditions." https://www.canada.ca/en/environment-climate-change/services/ice-forecasts-observations/latest-conditions.html, 2025. Accessed: 2025-01-24.

[53] A. Cheng, B. Casati, A. Tivy, T. Zagon, J.-F. Lemieux, and L. B. Tremblay, "Accuracy and inter-analyst agreement of visually estimated sea ice concentrations in Canadian Ice Service ice charts using single-polarization RADARSAT-2," *The Cryosphere*, vol. 14, no. 4, pp. 1289–1310, 2020.

[54] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[55] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[56] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, and O. Al-Shamma et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions.," *Journal of Big Data*, vol. 8, no. 53, 2021.

[57] L. Wang, K. A. Scott, L. Xu, and D. A. Clausi, "Sea ice concentration estimation during melt from dual-pol SAR scenes using deep convolutional neural networks: A case study," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4524–4533, 2016.

[58] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015.

[59] A. Kucik and A. Stokholm, "AI4SeaIce: Comparing loss representations for SAR sea ice concentration charting," in *ICLR Workshop on AI for Earth Science*, 2022.

[60] A. Kucik and A. Stokholm, "AI4SeaIce: selecting loss functions for automated SAR sea ice concentration charting," *Scientific Reports*, vol. 13, 2023.

[61] A. Stokholm, A. Kucik, N. Longépé, and S. M. Hvidegaard, "AI4SeaIce: Task separation and multistage inference CNNs for automatic sea ice concentration charting," *EGUsphere*, vol. 2023, pp. 1–25, 2023.

[62] A. Mahbod, G. Schaefer, G. Dorffner, S. Hatamikia, R. Ecker, and I. Ellinger, "A dual decoder U-Net-based model for nuclei instance segmentation in hematoxylin and eosin-stained histological images," *Frontiers in medicine*, vol. 9, 2022.

[63] L. M. Pereira, A. Salazar, and L. Vergara, "A comparative study on recent automatic data fusion methods," *Computers*, vol. 13, no. 1, 2024.

[64] L. Wang, K. A. Scott, and D. A. Clausi, "Improved sea ice concentration estimation through fusing classified SAR imagery and AMSR-E data," *Canadian Journal of Remote Sensing*, vol. 42, no. 1, pp. 41–52, 2016.

[65] Y. Liu, Z. Huang, and J. Han, "Aleatoric uncertainty embedded transfer learning for sea-ice classification in SAR images," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4980–4983, 2022.

[66] X. Chen, K. A. Scott, L. Xu, M. Jiang, Y. Fang, and D. A. Clausi, "Uncertainty-incorporated ice and open water detection on dual-polarized SAR sea ice imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[67] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, "Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation," *Computational Statistics Data Analysis*, vol. 142, p. 106816, 2020.

[68] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," 2017.

[69] K. Gupta, A. Mukhopadhyay, D. G. Babb, D. G. Barber, and J. K. Ehn, "Landfast sea ice in Hudson Bay and James Bay: Annual cycle, variability and trends, 2000–2019," *Elementa: Science of the Anthropocene*, vol. 10, p. 00073, May 2022.

[70] T. J. Ballinger, G. W. K. Moore, Y. Garcia-Quintana, P. G. Myers, A. A. Imrit, D. Topál, and W. N. Meier, "Abrupt northern Baffin Bay autumn warming and sea-ice loss since the turn of the twenty-first century," *Geophysical Research Letters*, vol. 49, no. 21, p. e2022GL101472, 2022.

[71] C. Derksen, D. Burgess, C. Duguay, S. Howell, L. Mudryk, S. Smith, C. Thackeray, and M. Kirchmeier-Young, "Canada's changing climate report," tech. rep., Environment and Climate Change Canada, 2019.

[72] A. Proshutinsky, R. H. Bourke, and F. A. McLaughlin, "The role of the Beaufort Gyre in Arctic climate variability: Seasonal to decadal climate scales," *Geophysical Research Letters*, vol. 29, no. 23, pp. 15–1–15–4, 2002.

[73] R. Mottram, S. B. Simonsen, S. Høyer Svendsen, V. R. Barletta, L. Sandberg Sørensen, T. Nagler, J. Wuite, A. Groh, M. Horwath, J. Rosier, A. Solgaard, C. S. Hvidberg, and R. Forsberg, "An integrated view of Greenland ice sheet mass changes based on models and satellite observations," *Remote Sensing*, vol. 11, no. 12, 2019.

[74] A. Korosov, D. Demchev, N. Miranda, N. Franceschi, and J.-W. Park, "Thermal denoising of cross-polarized Sentinel-1 data in interferometric and extra wide swath modes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[75] S. Dominicus, A. K. Mishra, and C. Rautenbach, *New Methodologies for Understanding Radar Data*. Institution of Engineering and Technology, 2021.

[76] C. Soriot, C. Prigent, C. Jimenez, and F. Frappart, "Arctic sea ice thickness estimation from passive microwave satellite observations between 1.4 and 36 GHz," *Earth and Space Science*, vol. 10, no. 2, p. e2022EA002542, 2023.

[77] S. Andersen, R. Tonboe, L. Kaleschke, G. Heygster, and L. T. Pedersen, "Intercomparison of passive microwave sea ice concentration retrievals over the high-concentration Arctic sea ice," *Journal of Geophysical Research: Oceans*, vol. 112, no. C8, 2007.

[78] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, and J. Muñoz-Sabater et al., "The ERA5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[79] J. Key, J. A. Maslanik, T. Papakyriakou, M. C. Serreze, and A. J. Schweiger, "On the validation of satellite-derived sea ice surface temperature," *Arctic*, vol. 47, no. 3, pp. 280–287, 1994.

[80] Y. Yu, D. A. Rothrock, and R. W. Lindsay, "Accuracy of sea ice temperature derived from the advanced very high resolution radiometer," *Journal of Geophysical Research: Oceans*, vol. 100, no. C3, pp. 4525–4532, 1995.

[81] J. R. Key, J. B. Collins, C. Fowler, and R. S. Stone, "High-latitude surface temperature estimates from thermal satellite data," *Remote Sensing of Environment*, vol. 61, no. 2, pp. 302–309, 1997.

[82] J. Key and Y. Liu, "Algorithm theoretical basis document: Ice surface temperature, ice concentration, and ice cover," tech. rep., NOAA NESDIS Center for Satellite Applications and Research, 2019.

[83] R. Wolfe, G. Lin, P. Zhang, J. Dellomo, and B. Tan, "VIIRS/NPP Moderate Resolution 6-Min L1B Swath 750m, Version 2," 2022.

[84] Y. Liu, J. Key, M. Tschudi, R. Dworak, R. Mahoney, and D. Baldwin, "Validation of the Suomi NPP VIIRS Ice Surface Temperature Environmental Data Record," *Remote Sensing*, vol. 7, no. 12, pp. 17258–17271, 2015.

[85] R. A. Frey, S. A. Ackerman, R. E. Holz, S. Dutcher, and Z. Griffith, "The Continuity MODIS-VIIRS Cloud Mask," *Remote Sensing*, vol. 12, no. 20, 2020.

[86] C. Horvat, E. Blanchard-Wrigglesworth, and A. Petty, "Observing waves in sea ice with ICESat-2," *Geophysical Research Letters*, vol. 47, no. 10, 2020.

[87] C. Cao, X. Xiong, R. Wolfe, F. DeLuccia, Q. Liu, and S. Blonski et al., "Visible Infrared Imaging Radiometer Suite (VIIRS) sensor data record (SDR) user's guide," tech. rep., National Oceanic and Atmospheric Administration (NOAA), Washington, DC, 2013.

[88] X. Chen, F. J. P. Cantu, M. Patel, L. Xu, N. C. Brubacher, K. A. Scott, and D. A. Clausi, "A comparative study of data input selection for deep learning-based automated sea ice mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 131, p. 103920, 2024.

[89] F. J. Pena Cantu, X. Chen, Y. Liu, K. Kanani, J. Park, J. N. Tunes, L. Xu, K. A. Scott, and D. A. Clausi, "A hierarchical multitask U-Net for automated sea ice mapping from AI4Arctic Sea Ice Challenge Dataset," in *OCEANS 2023 - MTS/IEEE U.S. Gulf Coast*, pp. 1–7, 2023.

[90] M. Seitzer, A. Tavakoli, D. Antic, and G. Martius, "On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks," 2022.

[91] N. Skafte, M. Jø rgensen, and S. r. Hauberg, "Reliable training and estimation of variance networks," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[92] T. H. Vidal, P. Gamet, A. Olioso, and F. Jacob, "Optimizing TRISHNA TIR channels configuration for improved land surface temperature and emissivity measurements," *Remote Sensing of Environment*, vol. 272, p. 112939, 2022.

[93] M. Jiang, X. Chen, L. Xu, and D. A. Clausi, "Semi-supervised sea ice classification of SAR imagery based on graph convolutional network," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1031–1034, 2022.

Appendix

.1 Results and Analysis Additional Figures

Pixel Class Distributions

Figures 1- 3 demonstrate the variation in pixel class distribution that can be found between folds in the train-validation dataset. This variation results in classification differences from models trained on different folds, which contributes to the standard deviations reported in Table 5.1.

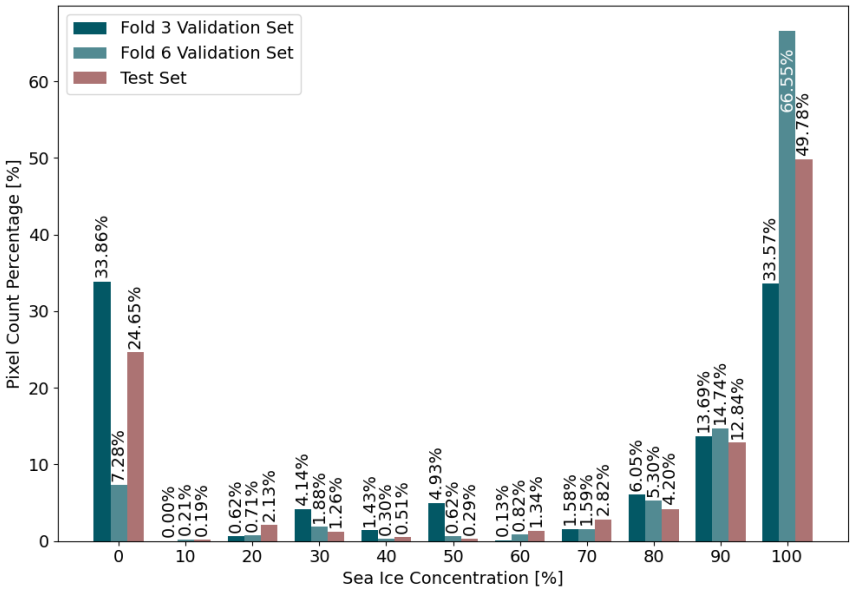


Figure 1: Comparison of the sea ice concentration class distributions for folds 3 and 6.

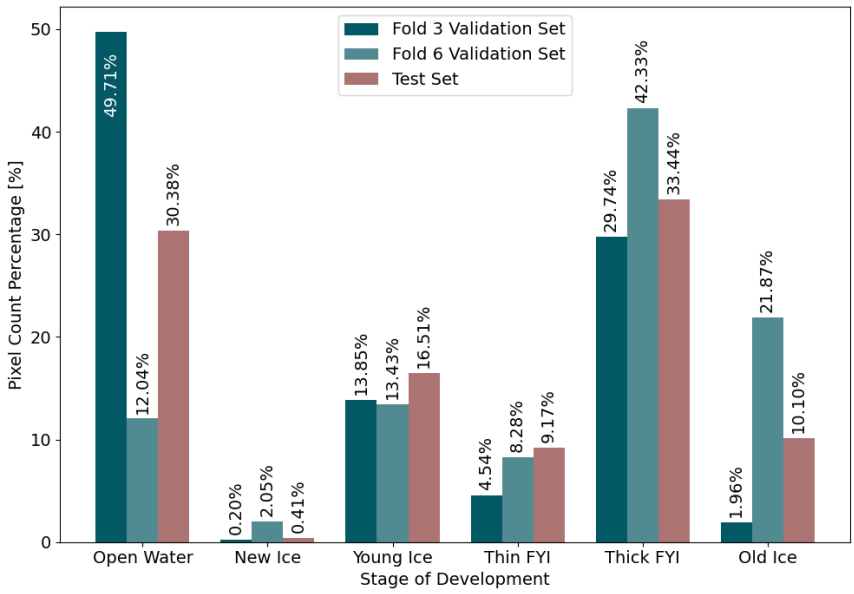


Figure 2: Comparison of the stage of development class distributions for folds 3 and 6.

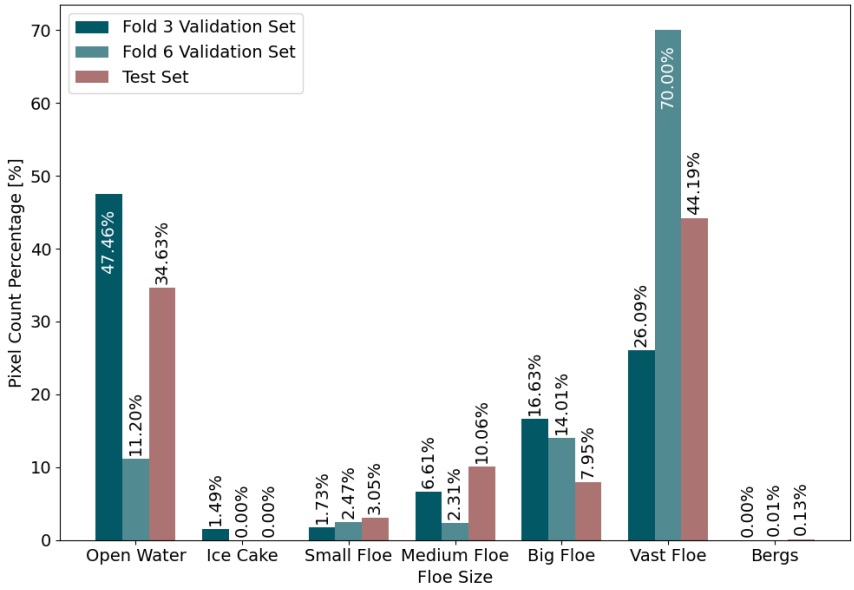


Figure 3: Comparison of floe size class distributions for folds 3 and 6.

2018-12-31 Supplementary Scene

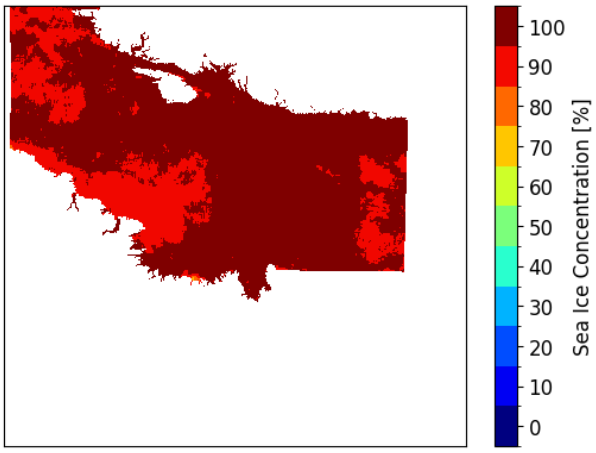


Figure 4: SIC prediction from DEU-Net-V for 2018-12-31 in the Hudson Strait.

2019-01-19 Ice Charts

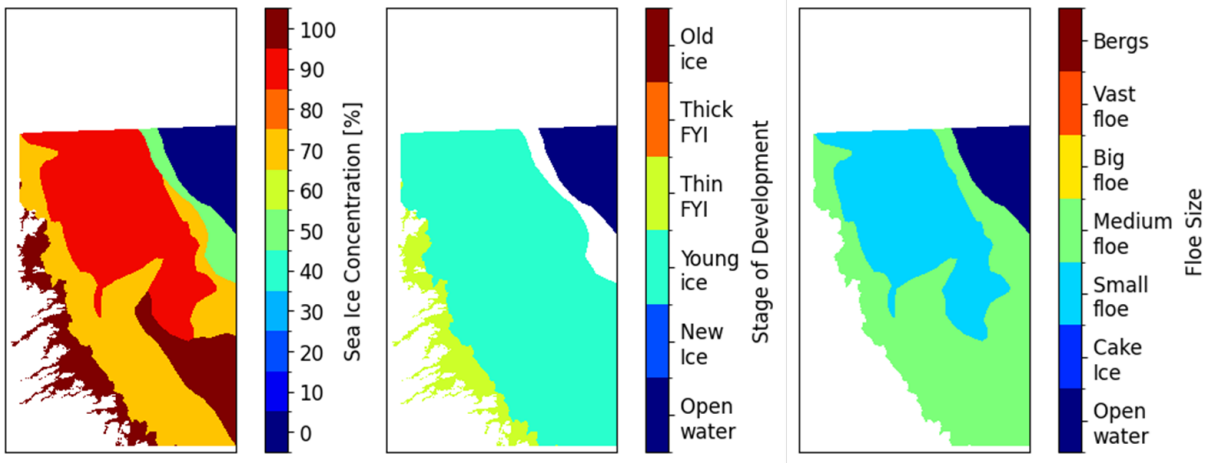


Figure 5: Comparison of SIC, SOD, and FLOE ice charts for 2019-01-19 off the coast of Labrador.