ExtremeEarth

H2020 - 825258

Deliverable
D2.5
Evaluation report on Polar use case output products

Salman Khaleghian, Qiang Wang, Habib Ullah, Andrea Marinoni, Thomas Kræmer, Torbjørn Eltoft

31.12.2021

Status: Final
Scheduled Delivery Date: 31.12.2021

Executive Summary

ExtremeEarth aims to advance the state-of-the-art by developing ad-hoc distributed deep learning architectures tailored to the specific properties of Copernicus Sentinel satellite data. The information provided by the satellite data will be used for deriving information by focusing the attention on the following use cases:

- **Food Security Use Case**: aims at the assessment of water availability for fertilization and irrigation by combining a long time series of Sentinel 2 multispectral images with crop growth modelling to provide water availability. The deep learning architectures trained on Sentinel 2 data will generate crop type and crop boundaries maps;
- **Polar Use Case**: exploits a long time series of Sentinel 1 synthetic aperture radar (SAR) images to develop processing architectures and algorithms to cope with the extreme analytics and big data challenges associated with sea ice monitoring.

This document focuses on WP2 tasks related to the *Polar Use Case* scenario, with emphasis on the design and analysis of Deep Learning (DL) architectures for sea ice monitoring. The goals of the work were to exploit the free and open availability of SAR images from *Sentinel 1 A & B satellites* to explore the capabilities of DL architectures for sea ice classification and evaluate their potential of coping with the extreme analytics and big data challenges associated with the Copernicus data.

It is noted that *sea ice monitoring here* refers to *ice type classification*, which is about segmenting SAR sea ice scenes into areas labelled as a certain ice type. Two classification tasks have been explored: i) The two-class problem, which aims at separating between sea ice and open water, and ii) Multiple class-problem, which aims at classifying the scene into multiple ice types and water classes. This document presents a detailed evaluation of two scenarios.

The next section describes the results of the requirement analysis for the Polar Use Case. To put the work and its contribution into a wider context, Section 2 gives a brief *state-of-the-art* discussion of the research area and the challenges related to automatic sea ice classification from SAR. In Section 3, we describe the main activities carried out during the project. These are: i) the generation of the large training database; ii) the definition of the DL architectures studied; iii) the evaluation of the inference results by applying the previous DL architectures. Section 4 describes the training of the architectures and its implementation on Hopsworks and the PolarTEP. Moreover, Section 5 briefly summarizes some auxiliary investigations which have been conducted within the project. Finally, a summary of the project is concluded in terms of the objectives and user requirements in section 6.

Document Information

Contract Number	H2020 - 825258	**Acronym**	ExtremeEarth
Full title	From Copernicus Big Data to Extreme Earth Analytics		
Project URL	http://earthanalytics.eu/		
EU Project officer	Riku Leppänen		

Deliverable	**Number**	D2.5	**Name**	Evaluation report on Polar use case output products	
Task	**Number**	T2.3	**Name**	Deep learning architecture design and development	
Work package	**Number**	WP2			
Date of delivery	**Contractual**		M36	**Actual**	M36
Status	draft ☐ final ☒				
Nature	Prototype ☐ Report ☒				
Distribution Type	Public ☐ Restricted ☐ Consortium ☒				
Responsible Partner	UIT the Arctic University of Norway				
QA Partner	Met Norway				
Contact Person	Torbjørn Eltoft				
	Email	torbjorn.eltoft@uit.no	Phone: (+47) 95007345		

Project Information

This document is part of a research project funded by the IST Programme of the Commission of the European Communities as project number H2020-825258. The beneficiaries in this project are the following:

Partner	Acronym	Contact
National and Kapodistrian University of Athens Department of Informatics and Telecommunications (Coordinator)	UoA	Prof. Manolis Koubarakis National and Kapodistrian University of Athens Dept. of Informatics and Telecommunications Panepistimiopolis, Ilissia, GR-15784 Athens, Greece Email: koubarak@di.uoa.gr
VISTA Geowissenschaftliche Fernerkundung GmbH	VISTA	Heike Bach Email: bach@vista-geo.de
The Arctic University of Norway Department of Physics and Technology	UiT	Torbjørn Eltoft Email: torbjorn.eltoft@uit.no
University of Trento Department of Information Engineering and Computer Science	UNITN	Lorenzo Bruzzone Email: lorenzo.bruzzone@unitn.it

Royal Institute of Technology	KTH	Vladimir Vlassov Email: vladv@kth.se
National Center for Scientific Research - Demokritos	NCSR-D	Vangelis Karkaletsis Email: vangelis@iit.demokritos.gr
Deutsches Zentrum für Luft-und Raumfahrt e. V.	DLR	Corneliu Octavian Dumitru Email: corneliu.dumitru@dlr.de
Polar View Earth Observation Ltd.	PolarView	David Arthurs Email: david.arthurs@polarview.org

METEOROLOGISK INSTITUTT	METNO	Nick Hughes Email: nick.hughes@met.no
Logical Clocks AB	LC	Jim Dowling Email: jim@logicalclocks.com
United Kingdom Research and Innovation – British Antarctic Survey	UKRI-BAS	Andrew Fleming Email: ahf@bas.ac.uk

Table of Contents

1 User requirements **8**

2 Background, Challenges, and Workflow **9**

2.1 Background *9*

2.2 Challenges *10*

2.3 Workflow and research topics *13*

3 Training datasets and evaluation **14**

3.1 Training dataset generation *14*

3.2 Evaluation of patch-wise supervised classification strategies *15*

3.2.1 Patch-wise classification with VGG-16 15

3.2.2 Supervised patch-based learning of a 13-layers CNN with noise profiling 20

3.2.3 Multi-class inference results evaluation 25

3.3 Supervised pixel-wise sea-ice versus water classification *27*

3.3.1 Study area and training dataset generation 27

3.3.2 Model architecture and experiment setting 28

3.3.3 Results and evaluation 28

3.3.4 Conclusions 33

3.4 Semi-supervised learning *33*

3.4.1 Model design and experiment setup 33

3.4.2 Inference results evaluation 36

4 Implementation on Hopsworks and the PolarTEP **39**

5 Additional studies **40**

5.1 Distributed training *40*

5.2 Incidence angle dependency *42*

5.3 Polarimetry analysis *45*

6 Summary and conclusion **46**

References **48**

1 User requirements

Deliverable D5.1 “The Polar Use Case - User requirement analysis” analyzed the user requirements for the Polar Use Case and concluded that “*the types of satellite data needed varied with the type of user being supported: transportation and other tactical information needed high resolution Synthetic Aperture Radar (SAR) data, whilst climatological studies found low resolution Passive Microwave (PMW) data is sufficient.*” Mariners need information that can support navigation and safety in ice infested waters. In this regard, detailed and reliably accurate information about the location of the ice edge, location of leads and ridges, areas of (thin) first year ice and areas of multi-year ice are crucial information. Mariners also want to have short-term forecasts about the development of ice conditions in the area they are located and heading. However, forecast predictions have not been considered within the scope of the ExtremeEarth project.

Focusing on the mariners needs, the research on sea ice monitoring has addressed two classification problems:

- *Binary sea ice-versus-water classification* – High-resolution ice masks can give information about many of the parameters requested above, such as location of ice edge location, location of leads, and they may also be used to generate large-scale ice concentration maps.
- *Multi-class ice type classification* - This addresses the more general classification problem of creating multiple ice type maps, identifying first and multi-year ice, thin ice, deformed ice, ridges and leads.

Our main focus has been on the first task, i. e. binary ice-versus-water classification, and the work includes analysis of several deep learning architectures. However, the extension from binary classification to multi-class classification has also been addressed.

It is noted that for climate monitoring, low-resolution ice concentration maps can be generated from the high-resolution SAR maps or obtained from passive microwave radiometer observations. In fact, the combined use of both these two sensors would also be beneficial. Such data analysis has not been conducted in our study.

In the generation of a training database and the development of deep learning architectures for the Polar Use Case, we have used high-resolution mapping from SAR sensors, and the Sentinel 1 (S1) SAR data has been the *primary source of data*. However, auxiliary data from other high-resolution SAR satellites (TerraSAR-X, RADARSAT-2) and/or collocated optical data (Sentinel 2 & 3, Landsat) have been used to support the creation of the training database.

2 Background, Challenges, and Workflow

2.1 Background

National sea ice services have the responsibility to produce operational sea ice charts, often on a daily basis. They use Synthetic Aperture Radar (SAR) data as a primary source of information for the generation of these sea ice products. At present, none of the ice services has reported using *automatic classification algorithms* operationally throughout the year. All algorithms still require some form of human guidance. The ExtremeEarth project aims to explore if the recent developments in the Artificial Intelligence (AI) and Deep Learning (DL) domain can bring progress in sea ice monitoring from SAR, and advance ice charting towards more automated services.

Current satellite borne SAR systems operate in the microwave area of the electromagnetic (EM) spectrum, with carrier frequencies in the range 1 - 10 GHz. They are of invaluable importance because their signals are independent of light conditions and can support ice services around the clock, seven days a week, all year round. SARs are multi-channel systems in the sense that they can transmit and receive EM signals at different polarization states. This multi-dimensionality increases their information-providing capabilities.

On the other hand, the characteristics of SAR echoes from an ice surface are very sensitive to the imaging geometry and to properties of the sea ice and surface topography. So, even if SAR images can provide ice information, the dynamics and complexity of sea ice makes the SAR scenes difficult to interpret. This is especially the case during the summer season, when the snow on the ice melts and gets wet, and the radar signals have less penetration into the ice column. Summer is also the period of the year when vessel activity increases in the polar regions, and there is increased demand for accurate ice mapping. Hence, even if there is a substantial number of classification and segmentation algorithms around, none of them are sufficiently robust to support automatic operational sea ice charting, which is the reason why currently sea ice charts are made manually by analysts at the service centers. Automatization of the segmentation and classification processes will make sea ice map production more efficient, and the maps will be less subjective and probably more detailed than those made by the ice experts. Consequently, the effort to improve automatic segmentation and classification approaches for ice charting and monitoring is highly important. We briefly mention a few challenges that must be considered in this regard.

SAR scattering from sea ice: A SAR is a side-looking imaging radar. For mono-static systems, the imaging process relies on the transmitted signal being backscattered towards the sensor by surface scattering or volume scattering mechanisms. The strength of the backscattered signals depends on both *system* and *medium parameters*. The system parameters include frequency, polarization, incidence angle, and noise characteristics. The medium properties are the surface roughness associated with the upper ice surface or horizontal layers, or permittivity contrasts within the bulk of the medium. These contrasts are typically attributed to the brine or air being trapped between crystalline ice. Snow on the ice, and the properties of snow on the ice also affects the strength and characteristics of the radar backscatter.

In conclusion, SAR returns from sea ice are intrinsically complex and contain multiple ambiguities. Even the binary discrimination between sea ice and water, which in many cases is trivial, can be awfully difficult, because wind flow and turbulence over water may create signal patterns that look like sea ice. Hence, for sea ice classification with DL networks, generating training datasets that are large and diverse enough to cover all possible situations is a prerequisite. This is both a tedious and difficult task, but probably is equally important as finding an optimal network architecture.

2.2 Challenges

The European Union's Copernicus program now provides Sentinel-1 SAR products, free of charge and on a regular basis, for all users, with near-real time (NRT) access less than 3 hours for operational users. Sea ice monitoring in the Arctic primarily uses the Sentinel-1 Extra Wide (S-1 EW) mode, acquired with dual-polarization channels (HH+HV). The SAR data used for training and validation in the Extreme Earth project is S-1 EW Level-1 Ground Range Detected (GRD) products, which is available at a pixel-spacing of 40 x 40 m (ground resolution of approximately 80 x 80 m). The swath of a scene has a width of 400 km, corresponding to incidence angles (IA) ranging from 18.9 – 47^o. The swath is composed of 5 sub-swaths.

Incidence angle dependency: The IA dependent intensity gradient is a consequence of the side-looking geometry of spaceborne SAR systems where the local IA varies significantly from the near to the far range view. This strongly affects the resulting image intensity values and is a fundamental physical effect related to the radar illumination geometry and the scattering being dependent upon the physical properties of the sea ice surface and volume. The gradient is varying with ice type, and it is also different for sea ice and water. This is illustrated in Figure 2-2-1.

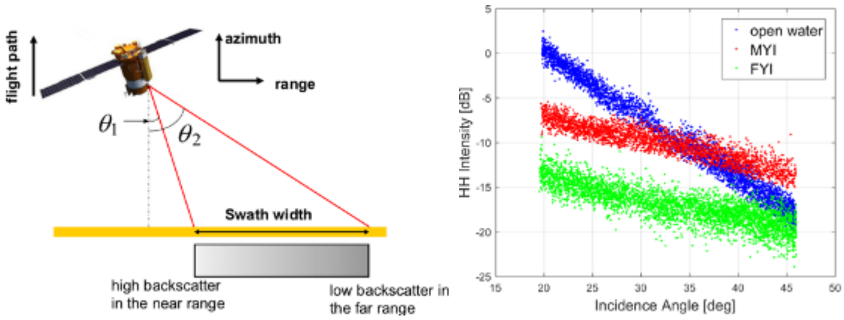


Figure 2-2-1: Basic incidence angle geometry and intensity variation for 3 classes: water, first-year ice (FYI) and multi-year ice (MYI) (Courtesy: Anthony Doulgeris).

<u>*Seasonal changes*</u>: The seasonality has a large impact on the dielectric properties of sea ice and its snow cover and, therefore, on the intensities of the backscattered signals. In winter, a sea ice scene is largely dominated by surface scattering from young ice and volume scattering from the bubble structure in the upper layer of multi-year ice (Ulaby et al., 1986). The onset of melt changes the characteristics and the dielectric properties of the sea ice and snow, resulting in dramatic changes in the radar return from a given ice type. The discrimination between various ice types gets more difficult as is illustrated in Figure 2-2-2. Figure 2-2-2 shows two S1 SAR scenes from the fast ice area at the east-coast of Greenland, before the onset of melting (left), and after (right).

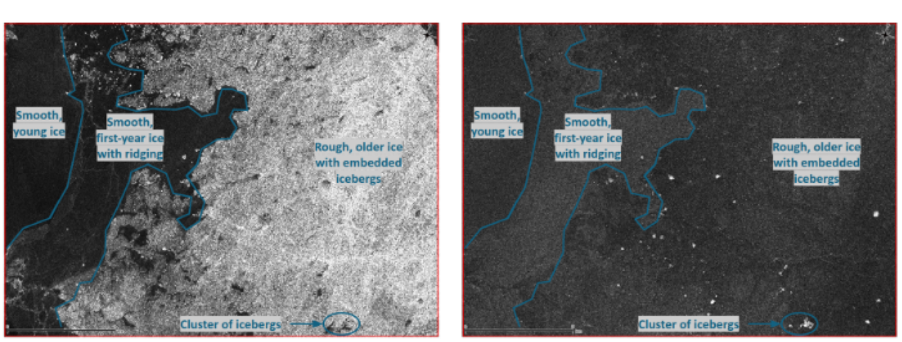


Figure 2-2-2: S1 SAR image of fast sea ice collected at pre-melt season (17.04.2018) in left panel, and melt-season (16.06.2018). (Courtesy: Nick Hughes)

Ambiguous scattering: Backscattering from sea ice and water surfaces is caused by roughness and dielectric properties of the target surfaces. For the ocean, the primary scattering is caused by small-scale roughness (on the order of the size of the wavelength of the radar signal), and these are modulated by the larger waves. The small-scale roughness is connected to the wind in such a way that high wind and turbulence will give high backscatter with texture associated to random directions of the wind turbulence. Over sea ice, the scattering is more complex. Ice surfaces may be level or deformed, they may be dry or wet, they may be snow covered or bare. And there are incidence angle dependencies. All these effects make the SAR returns ambiguous in such a way that very different ice types may give the same signal properties, and even the separation between open water and sea ice may be difficult. This is illustrated in Figure 2-2-3.

To mitigate this, recent generations of SAR starting with RADARSAT-2 and now continuing with Sentinel-1, have included a dual-polarization capability. The cross-polarization channel is less affected by the open water roughness, and provides improved discrimination between ice and water. However, this channel is affected by a lower signal-to-noise (SNR) ratio.

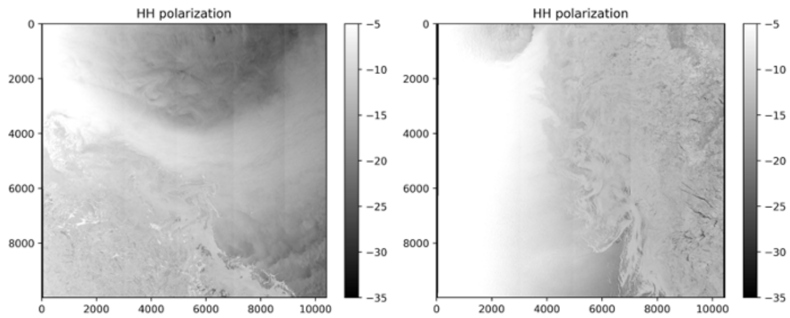


Figure 2-2-3: Two HH-polarization S1 SAR from the marginal ice zone.

Noise characteristics of Sentinel-1: SAR systems are affected by internal noise. This adds a structured but class independent intensity noise signal across the images, which is most prominent in the lower intensity HV or VH channels. This *noise floor* usually varies as a function of range (and thus IA) and is particularly pronounced with the five sub-swath assemblies of Terrain Observation with Progressive Scans SAR (TOPSAR) S1 EW imagery. The variable noise-floor may cause unwanted classification effects that are similar but separate from the IA dependency.

It has proved difficult to efficiently filter out this noise component, which is partly because the spatial correlation properties of the noise are very similar to the spatial properties of speckle, a well-known phenomenon associated with coherent radar imaging. Figure 2-2-4 provides an example of the noise pattern in a S1 EW-mode HV SAR scene (left), with the corresponding noise floor profile shown separately (right).

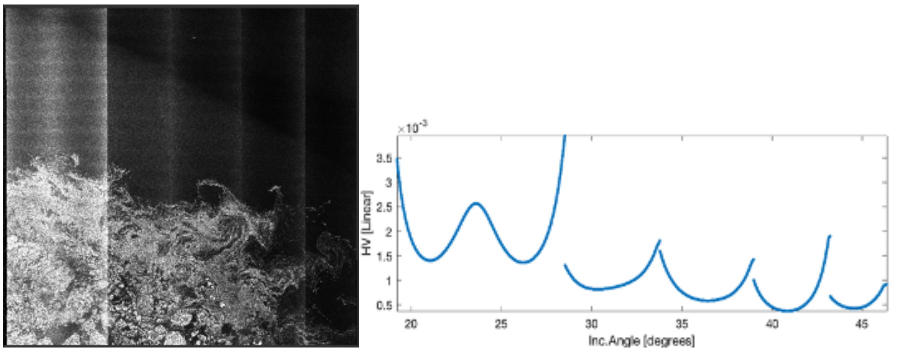


Figure 2-2-4: An example of a noise HV SAR image, with corresponding noise floor profile on the right.

2.3 Workflow and research topics

The development of DL architectures for sea ice mapping from SAR data involves several important work tasks. These include generation of labelled training data, pre-processing, architecture design, interpretation and evaluation. As discussed in the section above, there are also several additional challenges connected to the use of S1 SAR data which also need to be addressed, and some of them have implications for the architecture design including

- number of input layers,
- patch-wise or pixel-wise classification,
- number of output classes.

In addition, we need to address how to pre-processing the data, how to alleviate the noise problem and the scarcity of training data. Figure 2-3-1 summarizes the workflow and research tasks that have been addressed by UiT in the ExtremeEarth project.

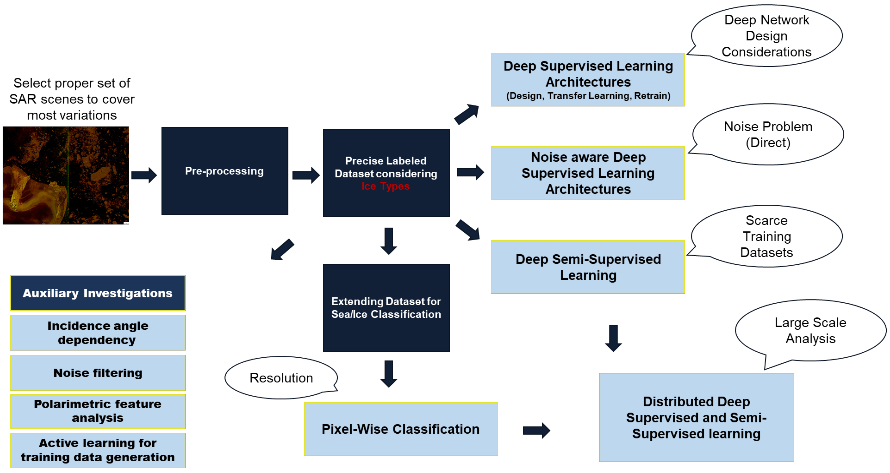


Figure 2-3-1: Illustratration of the workflow and research questions addressed by the UiT team.

3 Training datasets and evaluation

3.1 Training dataset generation

In the training and evaluation of the DL models, we have used several datasets.

<u>*UiT-dataset:*</u> This is a labelled training dataset generated from 31 S1 scenes acquired over ice-infested waters north of Svalbard having a resolution of 40m x 40m. Each image has a size of around 800 MB, summing up to 24.8 GB. We labelled polygons from these scenes (See details on the training data generation in D2.1) using co-registered optical images (S2 and S3) with small time-gaps to the S1 acquisitions. The polygons were carefully manually labelled into 6 classes (Lohse, 2020). The six classes were identified as; *open water, leads with water/newly formed ice, brash ice/pancake ice, thin first year ice, thick level first year ice, and thick deformed ice (first or multi-year ice)*. These classes have been qualitatively described and defined in D2.1. Nonetheless, to perform *water/sea ice* classification, we grouped the first two classes into one group, labelled '*water*', and gathered the other classes into a class labelled '*sea ice*'. Moreover, for each class and for a given set of patch sizes (see *Table 3-1-1*), we extracted all possible patches inside each polygon, with a stride of 10. Each patch input to the network contained HH, HV intensities, plus incidence angle.

Table 3-1-1: Number of samples in each class

Patch Size	Total	Ice	Sea
10x10	22,999	12,723	10,276
20x20	21,020	11,301	9,719
32x32	19,029	9,711	9,318
36x36	18,469	9,255	9,214
46x46	17,237	8,255	8,982

Met-dataset: For inference validation, i.e., to produce ice-water masks of unseen SAR scenes, we have used the "Met-dataset" from Danmarkshavn on the east coast of Greenland (Hughes and Amdal, 2020). This area was continuously monitored by key European Sentinel satellites during 2018. It has a wide variety of sea ice and iceberg conditions. The dataset consists of 12 days, approximately monthly, throughout the year, covering winter and spring, summer melt, and autumn freeze-up. The presence of extensive (land) fast ice in the area ensures that classifications can be applied to additional dates.

UiT-dataset 2: UiT has created a second dataset for the pixel-wise binary ice-water classification. This has been generated from 9 S1 SAR scenes over sea ice infested water in the Svalbard area. The images were classified into unlabeled classes using a segmentation algorithm, and thereafter carefully inspected by experts to assign the segments to the water and ice classes. The binary classified SAR scenes are subsequently used to randomly select an extensive number of patches of the required size to be used as training data for the pixel-wise classification network.

3.2 Evaluation of patch-wise supervised classification strategies

3.2.1 Patch-wise classification with VGG-16

In this subsection, we give a performance evaluation of the VGG-16 architecture for sea ice classification from SAR images under different training strategies. These strategies include:

(a) Training the VGG-16 network by *transfer learning*, where the pre-trained network is trained on the ImageNet dataset.

(b) Training the VGG-16 network from scratch.

(c) Training the VGG-16 network from scratch, with an augmented dataset.

(d) Training the modified VGG-16 network from scratch considering the augmented dataset with a patch-size equal to 32 × 32.

(e) Similar to (d) with a patch size of 20 × 20.

Based on these set-ups and the experimental analyses, we observed that the modified VGG-16 network trained from scratch with the augmented data with patch-size of 32× 32 provides the highest accuracies. We used the augmentation strategy of Buslaev and co-authors (Buslaev et al.,2020). According to this strategy, we perform horizontal flip, rotation with 90 degrees, blurring, and random changes to both brightness and contrast. The data augmentation technique aims to improve the robustness of the architecture by adding more variability to the training data and make the network more independent of changes in brightness and contrast. The overall dataset for different classes is indicated in Table 3-2-1-1.

Table 3-2-1-1: Number of samples in each class for augmented dataset. Smallest possible patch size for VGG16

Patch Size	Total	Ice	Sea
20x20	**263,541**	146,913	116,628
32x32	**238,059**	126,243	111,816

We present the classification results related to the VGG-16 network with different training approaches in Table 3-2-1-2. As can be seen, when the network is trained by transfer learning, the validation accuracy is equal to 97.9%, whereas when the same network is trained from scratch, the accuracy is 99.5%.

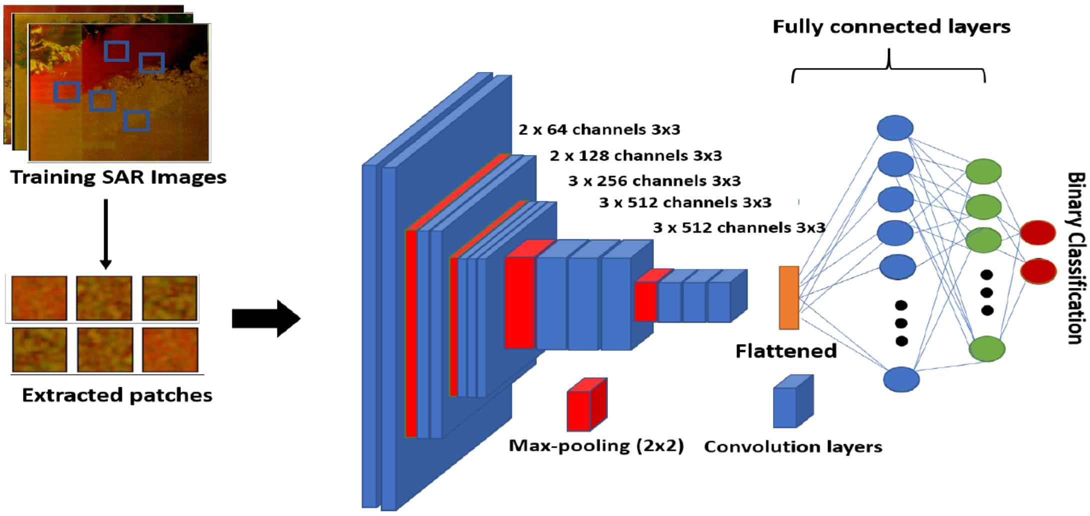


Figure 3-2-1-1: The overall architecture of the proposed method based on a modified version of VGG-16. We extract patches from these images and feed them to the network during the training process.

Table 3-2-1-2: Performance comparison from previous deliverable D2.3.

Training Strategies	Validation Accuracy	Resolution in Pixels	Resolution in Meters
VGG-16 transfer learning	97.9	32×32	1280
VGG-16 trained from scratch	99.5%	32×32	1280
VGG-16 trained from scratch with augmentation	99.79%	32×32	1280
VGG-16 Modified + augmentation	**99.89%**	32×32	1280
VGG-16 Modified + augmentation	99.30%	20×20	800

This leads us to conclude that in the case of sea ice classification from SAR data, training the network from scratch with an augmented dataset enables better adjustment and learning of the sea ice characteristics. Transfer learning, with pre-training on ImageNet data, which is fundamentally different from SAR data, does not allow the same adaptation to the data and poorer generalization.

The modified version of the VGG-16 model is obtained by reducing the number of max pooling layers. This also enables the size of the input patches to be reduced to 20 x 20. Moreover, by reducing the number of max pooling layers, the network better preserves the structure of the data and shows improved performance.

We also performed a comparison with three other reference models to show the stability and robustness of the modified VGG-16 model for sea ice classification. These reference models are MobileNetV2, RestNet50, and DenseNet110. The performance of our model in comparison with these reference models is presented in Figure 3-2-1-2 in the form of validation accuracy as a function of the number of epochs. As can be seen, our model presents higher and more consistent validation accuracy.

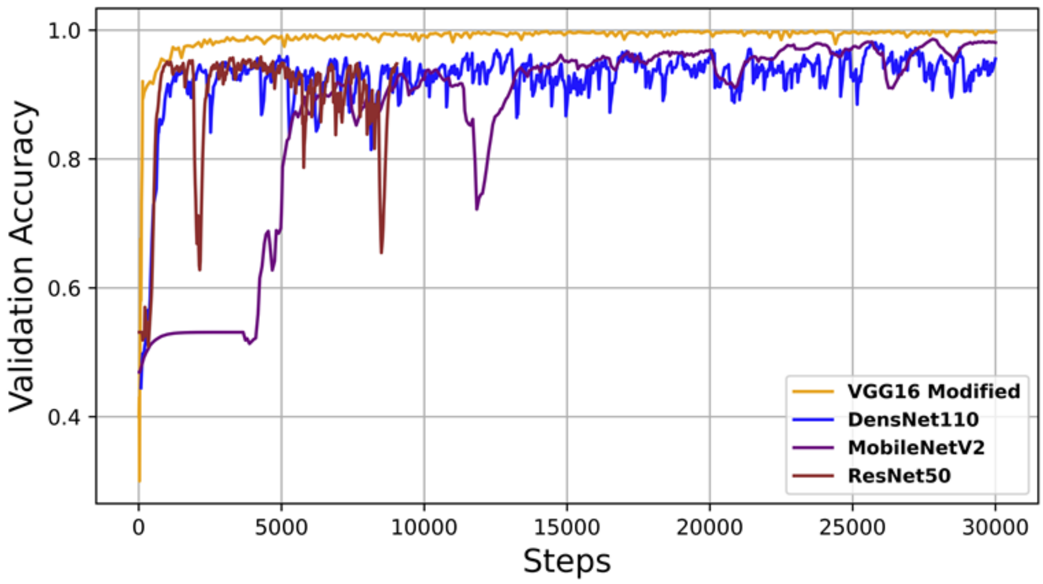


Figure 3-2-1-2: Validation accuracy

Inference evaluation of VGG-16

Figure 3-2-1-3 shows some inference results. The first row shows four input images from the area north of Svalbard. In the same figure, the patch-wise results of the ad hoc Convolutional Neural Network (CNN) are presented in the second row, the results of the VGG-16 model trained with transfer learning are presented in the third row, the results of the VGG-16 model retrained from scratch without the augmented data are presented in the fourth row, the results of the VGG-16 model retrained from scratch with the augmented data are presented in the fifth row, and the results of the modified VGG-16 model trained from scratch with augmented data are presented in the sixth row. Areas consisting of water and sea ice are annotated in blue and white, respectively. For better visualization, we applied a land mask to detect land areas, which is indicated as the black regions in the images. We zoom in on parts of some images to highlight specific details. The classification results obtained with ad hoc CNN (second row) are not satisfactory. The classified images are severely affected by the banding additive noise pattern, as can be clearly seen in columns two and three. The VGG-16 trained with transfer learning (third row) does not classify sea ice areas properly. In fact, open water and newly formed sea ice often have lower radar backscatter values in HV than HH channels. The HV intensity values are closer to the noise floor and therefore often have a lower signal-to-noise ratio, which explains the artifacts seen caused by noise patterns. In fact, this noise pattern may lead to problems during the interpretation of sea ice maps because the noise corrupts the true backscattered signal from the sea ice region.

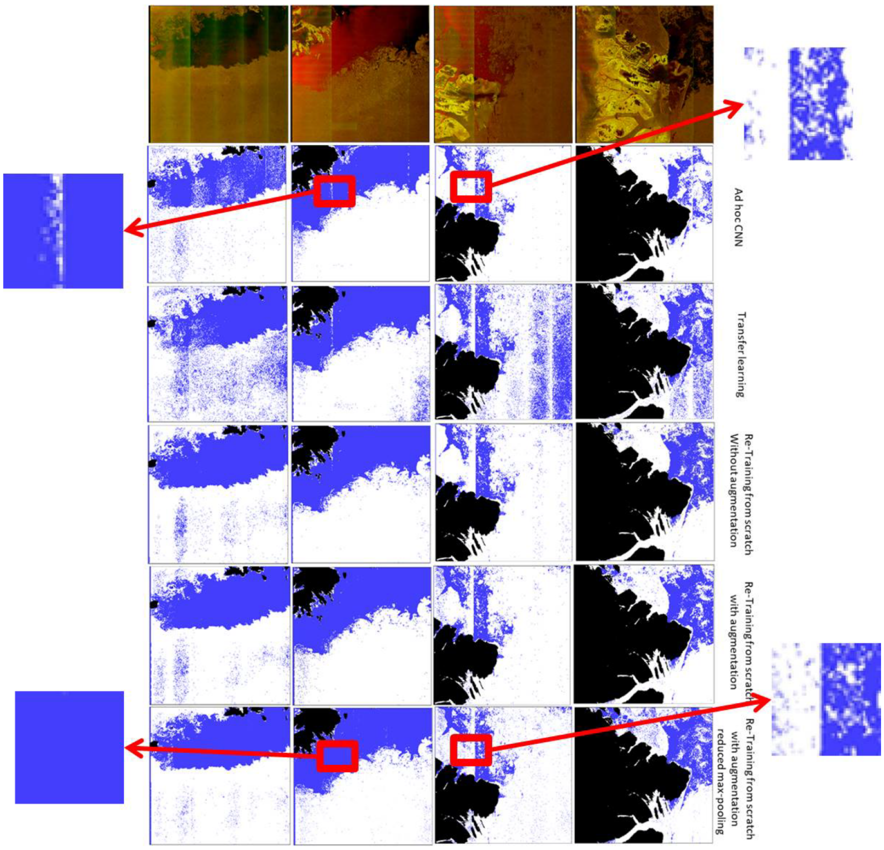


*Figure 3-2-1-3: Patch-wise results considering patch size equal to 32*32.* The first row presents the original images in two-bands. The second row presents results using ad-hoc CNN, the third row presents results using VGG-16 with transfer learning. The fourth row presents results using VGG-16 trained from scratch without augmentation. The fifth row presents results obtained using VGG-16 trained from scratch with augmented dataset. The sixth row presents the results of modified VGG-16 trained from scratch considering patch size equal to 20 × 20. Ice is annotated in white, and water is annotated in blue color. For better visualization, we applied a land mask on the obtained images annotated in black.

To further show the generalization performance of the CNN models for ice versus water classification, we also tested the models on images acquired from a different Arctic region, the area offshore of Danmarkshavn, East Greenland (76°46′ N, 18°40′ W), where the Norwegian Meteorological Institute provided vector polygon data representing manually interpreted sea ice areas (the Met-dataset). Figure 3-2-1-4 shows four of those images, corresponding to different months of the year. These include both the freezing and melting seasons and were then analyzed with the trained architectures. Figure 3-2-1-4 displays the classification results corresponding to the modified VGG network, trained from scratch with data augmentation, using patch sizes of 32 × 32 and 20 × 20.

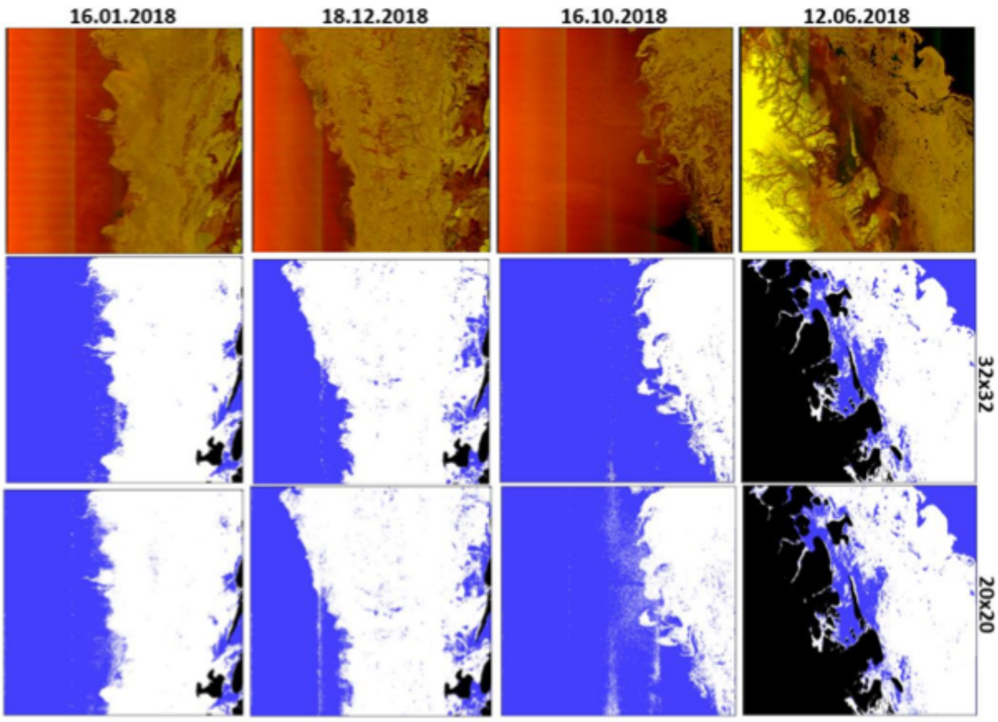


Figure 3-2-1-4: Binary classification maps with modified VGG16 from D2.3 considering 32x32 and 20x20 patches. Water is blue, Ice is white, and Black is a land mask added after classification.

As can be seen, the overall performance is good. It is also noticed that the results obtained with patch size equal to 32 × 32 are better than the results obtained with patch size equal to 20 × 20. The larger patch-size seems to be less affected by the noise and therefore we conclude that a patch size equal to 32 × 32 is a better choice for Sentinel-1 SAR images corrupted by additive noise. Overall, our experimental analysis shows that the VGG-16, when trained from scratch with augmented data, presents very good classification results when trained in a supervised mode.

3.2.2 Supervised patch-based learning of a 13-layers CNN with noise profiling

In this section, we evaluate the performance of the trained 13-layers CNN model shown in Figure 3-2-2-1 using different combinations of input features. In our experiments, four different compositions of the input patches are considered. Each approach uses a different combination of the calibrated Sigma0 horizontal-vertical (HV) and horizontal-horizontal (HH) polarization intensities, as well as IA and noise profile (NESZ) values:

1. HV + HH
2. HV + HH + IA
3. HV + HH + IA + NESZ
4. Filtered HV + HH + IA

As listed above, the experiments cover three different approaches to accounting for the noise: no correction (feature sets 1 and 2 above), noise as a separate feature (feature set 3) and noise filtering, which is the subtraction of the noise profile given in the meta data (feature set 4). Since the noise is most problematic for the HV channel, we only corrected the HV channel intensities.

The trends indicate that including information about the noise is valuable for the network both in terms of how quickly it converges and the achievable accuracies. Figure 3-2-2-2 shows the validation accuracies of the 13-layers CNN through the epochs for the different input combinations, and Table 3-2-2-1 shows the best accuracy achieved by the different compositions.

Comparing the validation accuracy of HH+HV with HH+HV+IA, we can see that the blue line (HH+HV+IA) tends to have higher accuracy than the orange line (HH+HV) in Figure 3-2-2-2. This is expected, as the backscatter of the classes varies with incidence angle with a different decay ratio for each class. Including IA as a feature allows the network to learn these dependencies. However, the improvement is not very large, which is surprising, as we know there is a significant difference in the dependency on incidence angle of the ice and water classes.

In contrast, when the network is trained using the noise corrected (filtered) HV intensities, it can better distinguish ocean from ice than with other combinations. The effect of correcting the noise is pronounced, leading to a much quicker convergence, in addition to converging to a higher accuracy. Subtracting the additive noise allows for easier visual interpretation, especially in the first swath, and it also allows the network to focus on learning the ocean and ice features, rather than trying to learn the noise. The extracted patches are then more distinguishable for the network, and the network can learn more quickly. However, when the noise profile is added as an input layer, the network can learn the relation between the signal and noise, and implicitly cope with noisy patches in the classification task. Therefore, the validation accuracy is slightly better than when filtered patches are used.

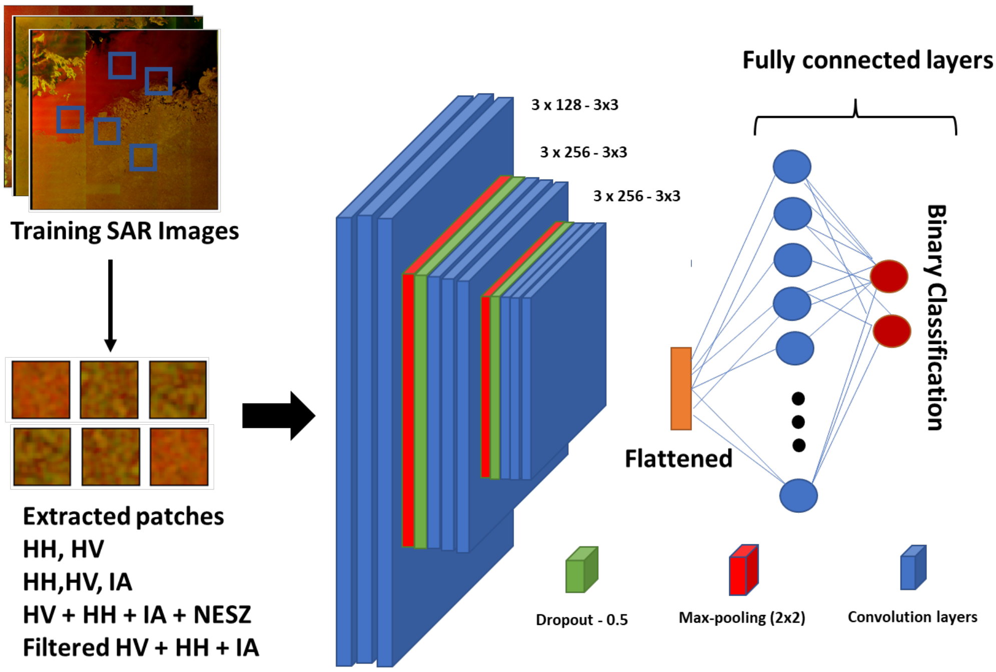


Figure 3-2-2-1: The overall architecture of the 13-layers CNN considering Noise-profile. We extract patches from the SAR images and feed them to the network during the training process.

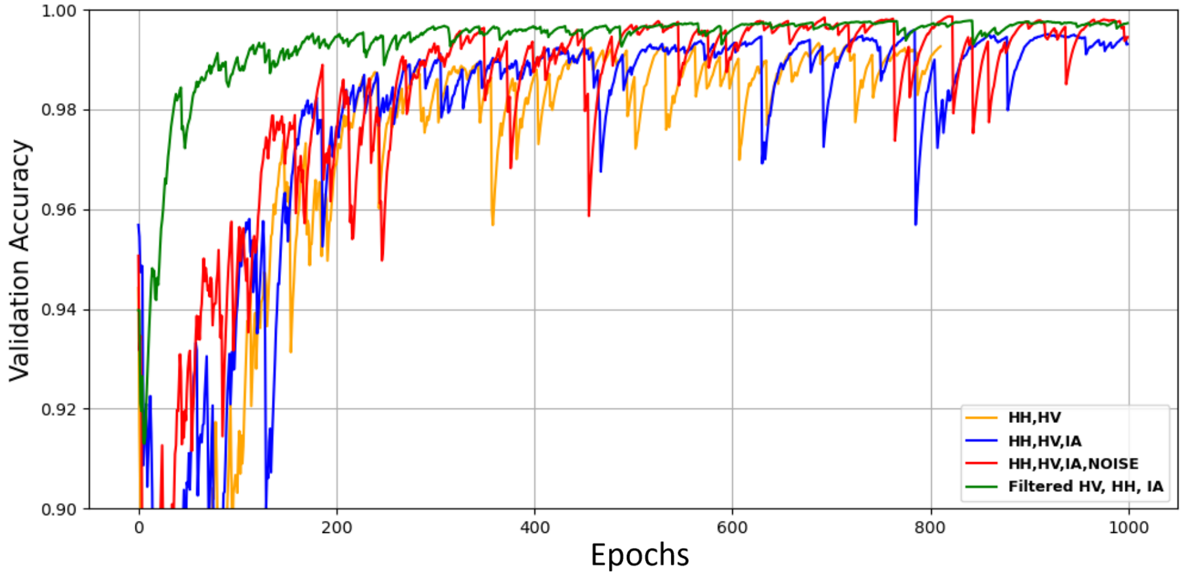


Figure 3-2-2-2: Validation accuracy for the 13-layers CNN, trained on four different combinations of input features.

Table 3-2-2-1: Best Validation accuracy of 13-layers CNN considering different channels composition.

	HH+HV	HH+HV+IA	HH+HV+IA+Noise	FilteredHV+HH+IA
Validation Accuracy	99.63	99.76	**99.95**	99.86

As shown in Table 3-2-2-1, the best classification accuracy was obtained by feeding in the NESZ profiles as a feature, allowing the network to figure out the appropriate scaling on its own. However, there is a significant downside to do it this way. As mentioned above, when introducing the NESZ, these profiles change with the processor version, meaning that we have to train separate models for different versions. This is not really feasible as it would require training data for all processor versions. Instead, it makes more sense to focus on feature set 4 (Filtered HV + HH + IA), as it is easier to try to adjust the noise scaling to make a consistent time series, than it is to create new training datasets for each version.

Inference results for the binary classification for the 13-layer network

To show the generalization performance of the 13-layer CNN model for the binary ice versus water classification, we tested the models on images acquired from the Danmarkshavn area offshore of Greenland (the Met-dataset). Figure 3-2-2-3 shows the inference results of four different images. Qualitatively, the ice vs. water classification looks reasonable for all images. What separates the cases is the classification performance in the noisy areas of the images and how well details within the ice pack are correctly classified.

For the first input image, and considering the noise profile as an input channel, the network can cope with the noise and produce a classification result that is almost free of noise artefacts (row 1, column 4). When the filtered patches are used (row 1, column 5), we can still see some noise effects on the water, showing that the European Space Agency (ESA) noise profile is not sufficient to completely remove the noise (this has also been shown by Nansen Environmental and Remote Sensing Center (NERSC) in several papers on thermal noise correction, see Park et al., 2018). This result shows that the model can successfully make use of the provided information about the noise floor profile. Since we have provided the ESA noise profile as an extra input layer, the model has not learned the profile itself, but rather an appropriate approximation. We have not yet run this experiment with the NERSC correction model, but the expectation is that with an improved noise profile, these two approaches of handling the noise should perform similarly.

The second row shows significant noise in large parts of the image, and clearly demonstrates the benefit of including information about the noise into the process. Whereas the HH, HV and HH, HV, IA classifications have noise artifacts along the sub-swath boundaries, the two experiments that actively consider the noise, provide much cleaner classification maps.

Rows 3 and 4 show quite similar images but demonstrate some remaining challenges that we have to look further into. The third row shows poor performance overall. In this case, the noise information does not seem to help. We can also see some differences in the inference results with and without IA. However, in this case, the quality of the inference results is almost the same. In the fourth row, we get better results with the filtered HV feature (column 5) than with the NESZ as a separate input feature (column 4). There are a few likely explanations for the poor performance, all of which relate to the ESA Instrument Processing Facility (IPF) version. First, the SAR images are captured a few years ago, when the noise profile only contained the range direction (left-to-right). Hence, no information was available in the azimuth (down-up) direction for the network to learn. Secondly, the training dataset may be generated by different ESA processor versions, and the noise profile is known to vary between versions. Thirdly, the noise profile was incorrectly scaled, compared to recent versions, meaning that the filtered HV cases are expected to be poor. Going forward, we will account for this, by training separately for different processor versions. Still, rows 3-4, columns 4-5 (with noise handling) show improved details in the classification map compared to columns 2-3. The vertical line in row 4, column 4 can be explained by the missing azimuth profile and is expected to improve when we include this information.

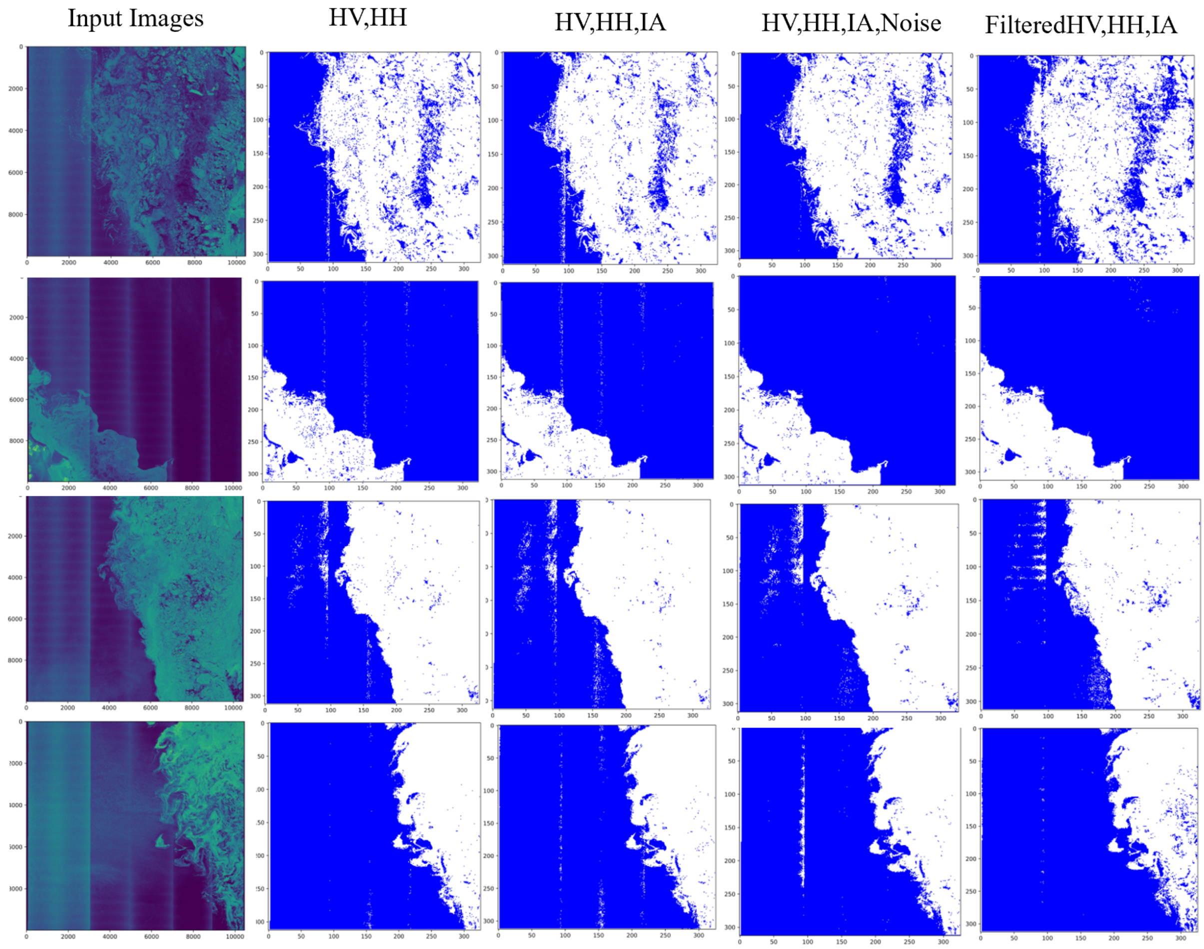


Figure 3-2-2-3: Inference results based on 13-layers CNN using different combinations of features. Blue indicates water and white indicates ice. The input images show the original HV intensity images without noise correction, where the five sub-swaths EW1-EW5 (left to right) are clearly visible.

3.2.3 Multi-class inference results evaluation

Ice type information is of high priority for ice charting and will be an important activity going forward. Following the binary sea ice-versus-water classification, we further extended our experimental analysis to the multi-class sea ice type classification case. We trained a model from scratch using 5 classes, including open water/leads with water, brash/pancake ice, young ice (YI), level first year ice (FYI), and old/deformed ice based on the UiT dataset. We considered the modified VGG-16 network and 13-layers CNN. The inference results are depicted in Figure 3-2-3-1 and Figure 3-2-3-2.

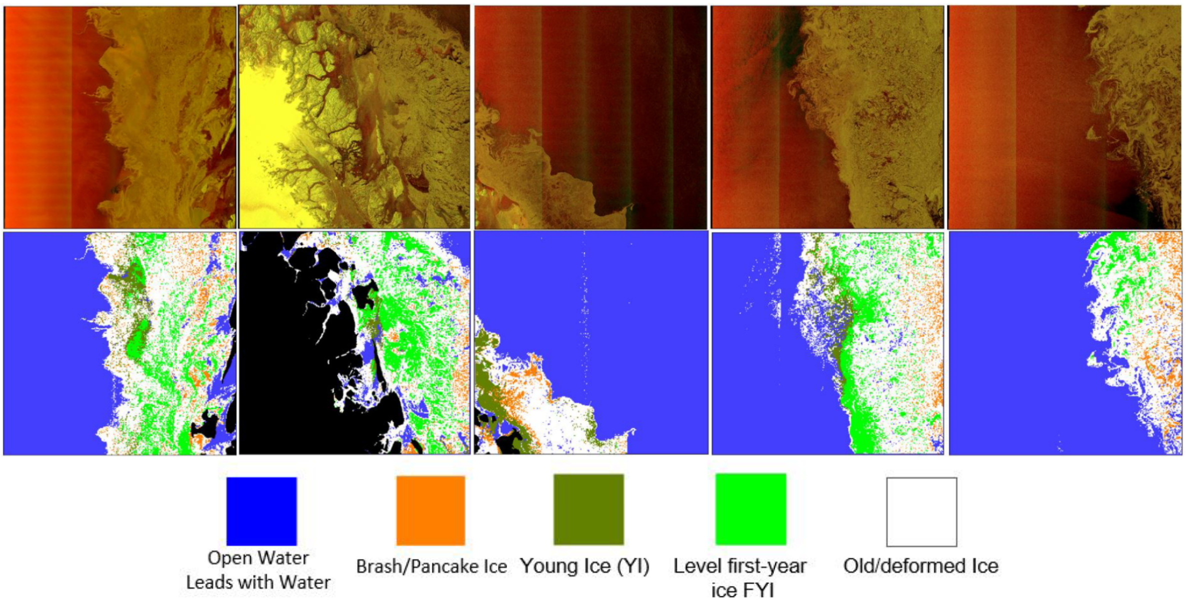


Figure 3-2-3-1: Inference result for multi-class classification based on modified version of VGG-16 considering HV, HH, IA channels composition. Top row is original HV intensity images. Bottom row is classification maps.

We would like to emphasize that the training dataset we use is scarce and unbalanced, with an unequal number of samples from the different ice types. This affects the classification performance, and the results presented in Figure 3-2-3-1 and Figure 3-2-3-2 are expected to be biased toward ice types where we have more samples than others. The effect of the imbalanced data can be seen for example in column 4, where brash/pancake ice is detected within the ice pack on the right-hand side. In general, brash/pancake ice is located at the edges towards open water, but is apparently here found in a dense ice area. However, brash ice can can be found inside the ice pack in leads between larger floes. The 13-layers CNN using filtered HV data produces quite clean results, especially in noisy areas. Despite this problem, the results indicate that the models are capable of distinguishing both different ice types as well as separating between ice and open water classes.

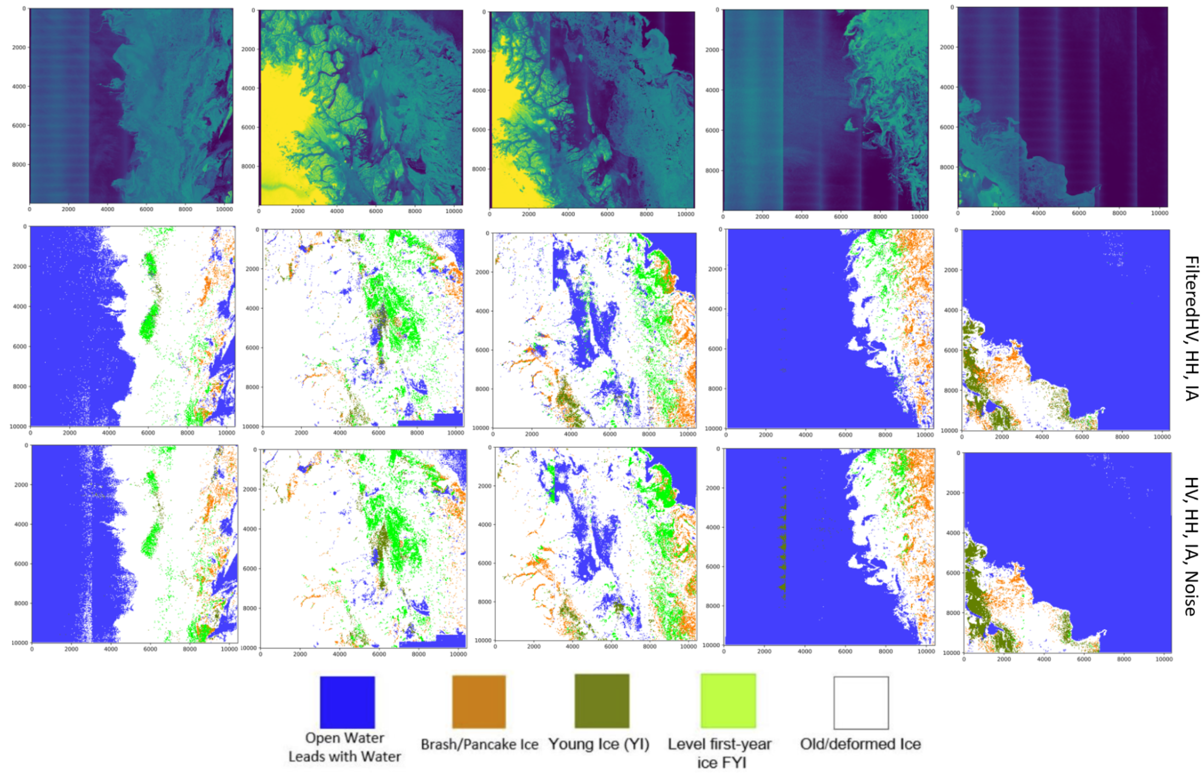


Figure 3-2-3-2: Inference result for multi-class classification based on 13 layers CNN considering Filtered HV, HH, IA channels composition.

3.3 Supervised pixel-wise sea-ice versus water classification

3.3.1 Study area and training dataset generation

The region between Greenland and Svalbard is selected as our area of interest and we have selected 9 scenes of the Extra Wide (EW) mode of Sentinel-1 over the time period January to March 2021 for training. To generate reliable data and label pairs, pixels at the boundaries between sub-swaths are removed and the noise floor correction methodology proposed by Nansen Environmental and Remote Sensing Center (https://github.com/nansencenter/sentinel1denoised) is implemented and applied. Subsequently, the images are reorganized into patches of size 224 x 224 to be fed into the model. The input layers are HH and HV intensities, as well as incidence angle (IA).

For label generation, we utilized a clustering method to generate a baseline cluster map trying out various settings of the clusters to minimize the errors. Based on this map, clusters are merged into ice or water. In this step, an expert is being consulted for assigning labels to the most ambiguous pixels. In the end, 5,713 and 1,424 patches of labeled data are generated for training and validation, respectively.

3.3.2 Model architecture and experiment setting

For the pixel-wise sea-ice classification, we have reviewed several classical semantic segmentation methods including FCN-8, FCN-16 (Long et al., 2015), deeplabv3 (Chen et al., 2017), as well as the UNet (Ronneberger et al., 2015). After evaluating the performance of the architectures on referenced data, we decided to use *the UNet* for designing our model architecture. On this basis, the number of layers as well as the number of features of each layer, were further explored. Then, the final model architecture was chosen. It consists of 3 layers of down-sampling and 3 layers of up-sampling. We have denoted this model “the reduced UNet”, and its detailed architecture is depicted in Figure 3-3-2-1.

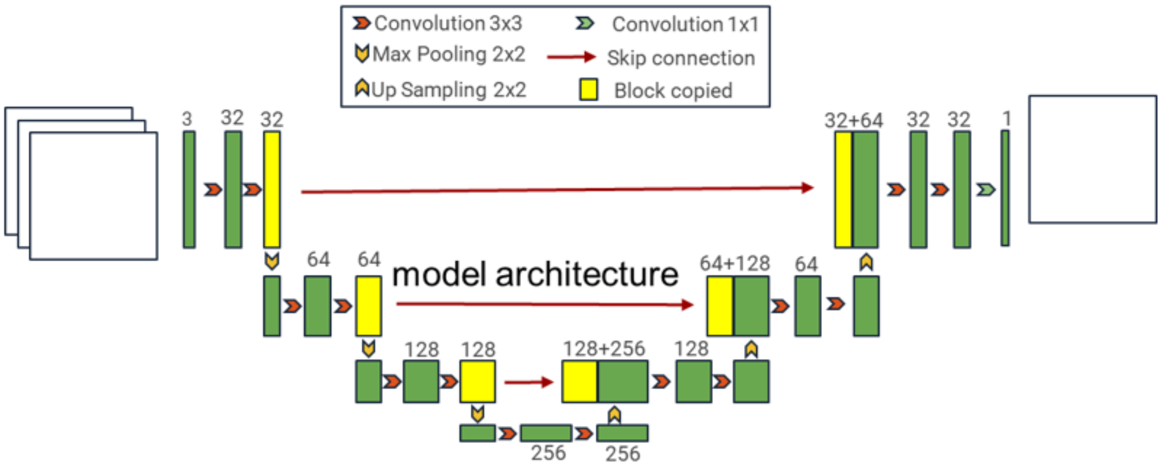


Figure 3-3-2-1: Model architecture of reduced UNet

During the training process, optimization method of Stochastic Gradient Descent (SGD) was utilized. Moreover, early stopping was considered at epoch n if the validation error of epoch *n* did not decrease with more than 0.001 for 7 consecutive epochs, i.e. from n+1 to n+7. The batch size is32, learning rate is 0.001. The training and validation procedure was carried out on a NVIDIA Quadro RTX 5000 Graphics Card.

3.3.3 Results and evaluation

We explored the model performance for the following feature combination: i) HH, HV, and IA; ii) HH and IA; iii) HV and IA. The corresponding accuracies and losses for training and validation are shown in Figure 3-3-3-1 and in Table 3-3-3-1, respectively. It can be seen from the figure that after 20 epochs, the accuracy for the 3-layer input model converges to the value of 91.39%, with a training loss of 0.048, when using HH, HV and IA as the input features.

The training procedure takes roughly 20 minutes, and the model size storing the parameters is 4.13 MB.

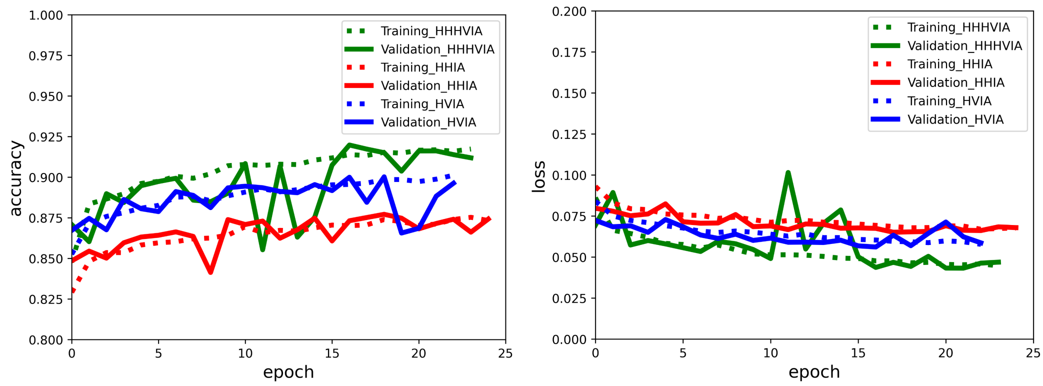


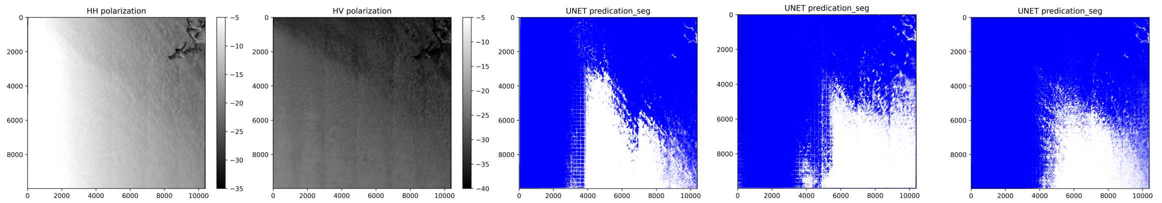
Figure 3-3-3-1: Training and validation accuracies (left) and loss (right) versus epochs.

Table 3-3-3-1: Accuracy and loss for training and validation for different features selections

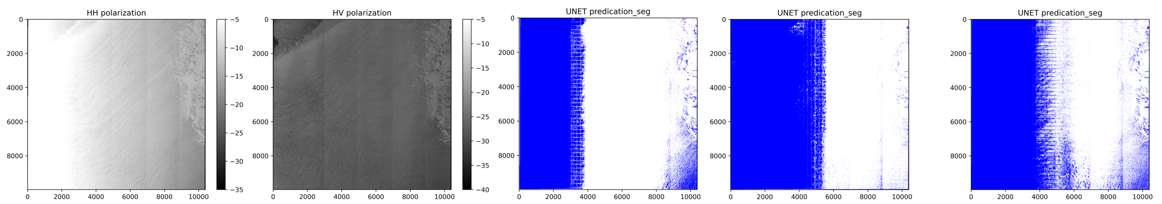
Features	accuracy (-)		loss (-)	
	training	validation	training	validation
HH + IA	87.06%	87.52%	0.069	0.065
HV + IA	89.53%	89.17%	0.061	0.057
HH + HV + IA	91.39%	91.98%	0.048	0.044

To further evaluate the model, we performed an inference study with the *reduced UNet* considering various scenarios of different percentage of ice and water in the scene, as well as the relative range location of ice and water. In addition, the chosen scenes illustrate how ambiguities in the SAR intensities of ice and water may arise as a consequence of varying wind conditions over the ocean. The inference processing of one scene takes about 20 seconds, depending on the size of the scene. Overall, we performed inference on 23 scenes, of which we include the results of 12 of them below. The Figures 3-3-3-2 to 3-3-3-7 show from left to right: the images of HH and HV intensities, and the inference results for input compositions HH & IA, HV & IA, and HH & HV & IA, respectively. The identification numbers of the 12 scenes are indicated.

HH	HV	inference (HH+IA)	inference (HV+IA)	inference (HH+HV+IA)

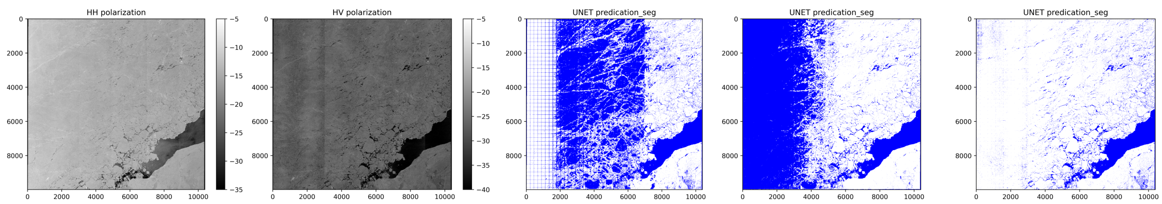


S1A_EW_GRDM_1SDH_20200112T070525_20200112T070625_030764_038727_C5F3

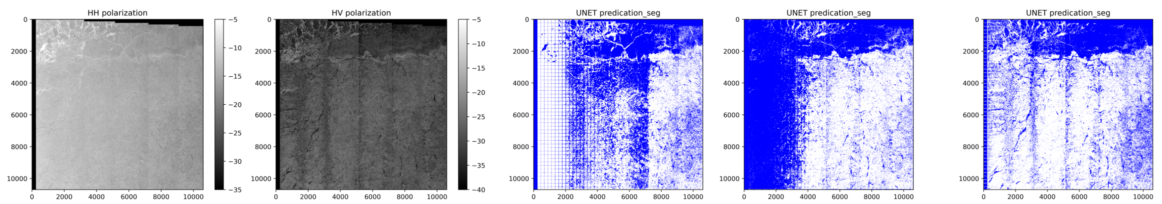


S1A_EW_GRDM_1SDH_20200205T070524_20200205T070624_031114_03936B_4DC6

Figure 3-3-3-2: Scenes dominated by water.



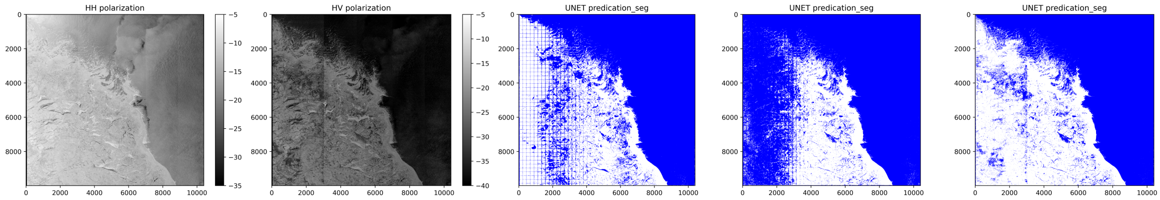
S1A_EW_GRDM_1SDH_20210604T202343_20210604T202443_038195_0481F7_E982



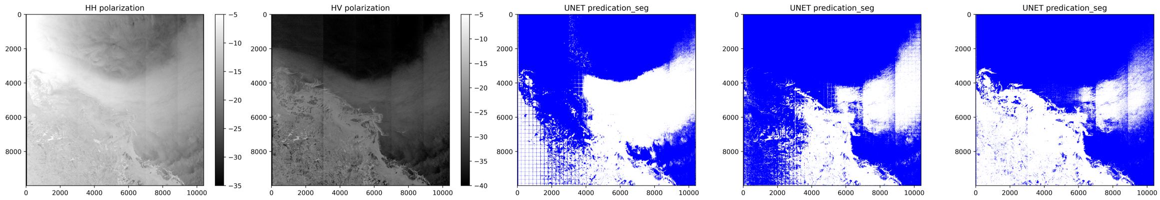
S1B_EW_GRDM_1SDH_20200128T121225_20200128T121329_020017_025DF2_AF51

Figure 3-3-3-3: Scenes dominated by ice.

HH	HV	inference (HH+IA)	inference (HV+IA)	inference (HH+HV+IA)

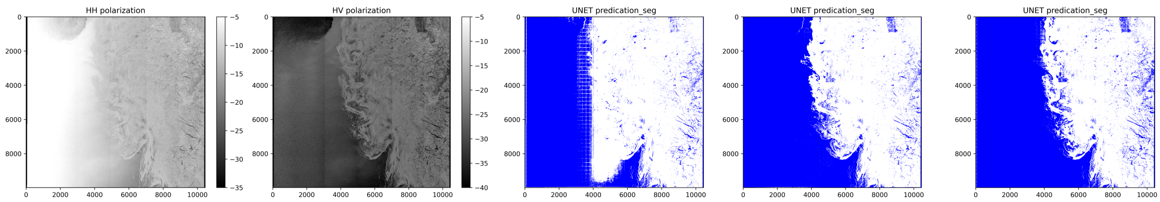


S1B_EW_GRDM_1SDH_20201217T170625_20201217T170725_024745_02F173_36F7

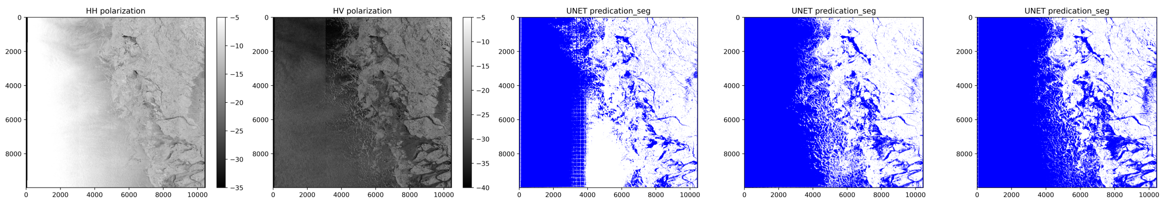


S1B_EW_GRDM_1SDH_20200209T170617_20200209T170717_020195_0263BC_B84

Figure 3-3-3-4: Scenes with ice in the near-range and water in the far range.



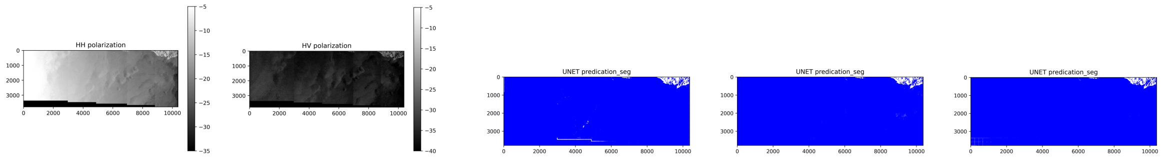
S1B_EW_GRDM_1SDH_20201109T073739_20201109T073839_024185_02DF97_51C2



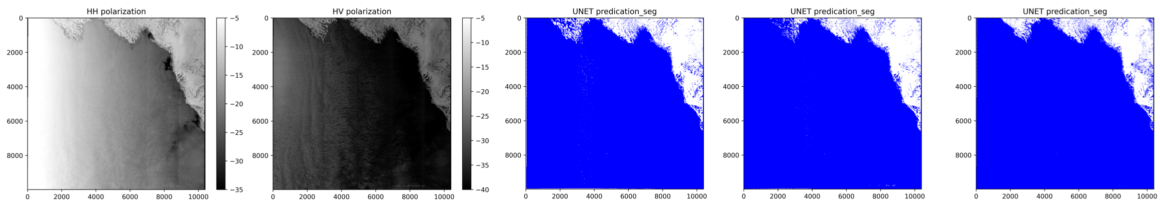
S1A_EW_GRDM_1SDH_20200315T073001_20200315T073101_031683_03A71B_38

Figure 3-3-3-5: Scenes with water in the near range and ice in the far-range.

HH	HV	inference (HH+IA)	inference (HV+IA)	inference (HH+HV+IA)

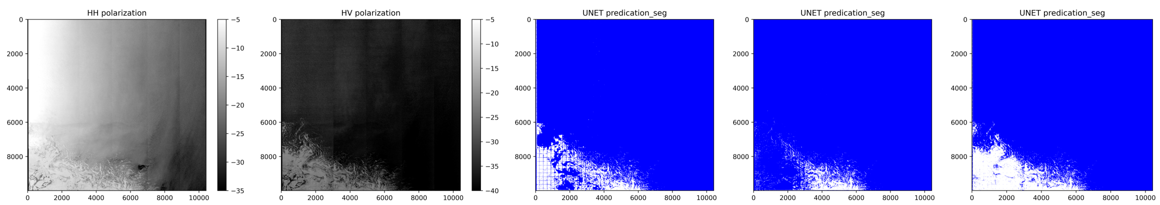


S1B_EW_GRDM_1SDH_20210501T042851_20210501T042914_026706_0330A5_734F

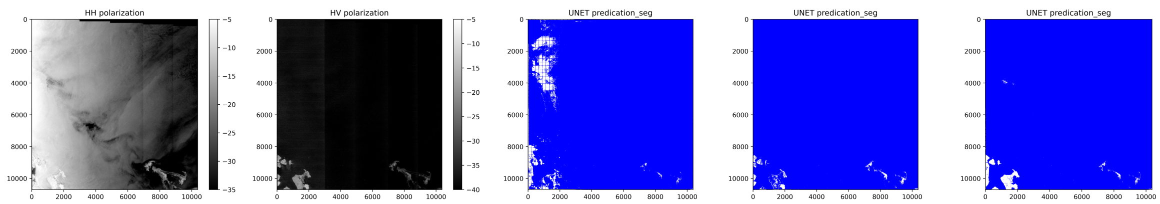


S1B_EW_GRDM_1SDH_20210415T050122_20210415T050222_026473_03290E_110E

Figure 3-3-3-6: Scenes dominated by water, with ice mostly in the far range.



S1B_EW_GRDM_1SDH_20201119T173823_20201119T173923_024337_02E45F_A9CC



S1B_EW_GRDM_1SDH_20210827T174654_20210827T174758_028435_03648F_057E

Figure 3-3-3-7: Scenes dominated by water, with ice mostly in the near range.

3.3.4 Conclusions

Based on the analysis performed using the reduced UNet presented in Figure 3-3-2-1. for pixel-wise sea ice-versus-water classification, we have made the following conclusions:

i) In general, the performance is good. The network is capable of distinguishing between sea ice and water.

ii) The best performance is obtained when the input layers contain HH and HV intensities and the incidence angle.

iii) The ice edge is quite well captured.

iv) The inference results clearly show that noise is still a problem, even after state-of-the-art filtering. This is particularly a problem for sub-swath one, which is most heavily contaminated.

v) It seems like the network may misclassify isolated water as ice or vice versa when the surroundings are dominated with pixels from the other class.

All in all, we conclude that the capabilities of the reduced UNet is very promising. We think that the problems we have identified can be remedied with more accurate training dataset, covering larger variations in the imaging scenario. Our conclusions are currently quite qualitative. Future expansion into quantitative analysis of the performance will be pursued.

3.4 Semi-supervised learning

3.4.1 Model design and experiment setup

To address the problem of a limited amount of labeled training data, we have investigated several methods of label propagation using semi-supervised learning approaches. These are methods that have proved themselves efficient in applications such as object detection (Tan et al., 2020), information protection (Yan et al., 2019), and natural language processing (Otter et al., 2020), and are designed to handle the situation where one wants to combine a small amount of labeled data with a large amount of unlabeled data. Our investigations have included experiments with the *MixMatch* method, with classification based on "*Wide ResNet*", as well as an own-developed *Teacher–Student* architecture, where pseudo-labels for the unlabeled data are generated based on label propagation using the feature space structure.

The MixMatch method is an interesting method since it shows very high accuracy in standard computer vision datasets. However, for our application, this model suffers from convergence problems and fluctuates through the training process. MixMatch has several important hyperparameters which should be tuned to optimize the training process; the estimation of hyperparameters is complicated, and training is very time consuming. Therefore, we need more investigation on this method to tune and customize it for our specific applications.

Our main focus, however, has been on a new Teacher–Student based label propagation model using semi-supervised deep learning, referred to as SSL-LP. The 13-layer network presented in the supervised case is used both for the Teacher and the Student models. In our approach, the Teacher model is trained in a two-step procedure. In the first step, the model is trained in a supervised manner using the labeled data. Subsequently, both the labeled and unlabeled samples are fed to the trained Teacher model and based on an evaluation of the feature space embedding, pseudo-labels for the unlabeled data are generated through a label propagation procedure (Khaleghian et al., 2021b). The training methodology is depicted in the Figure 3-4-1-1.

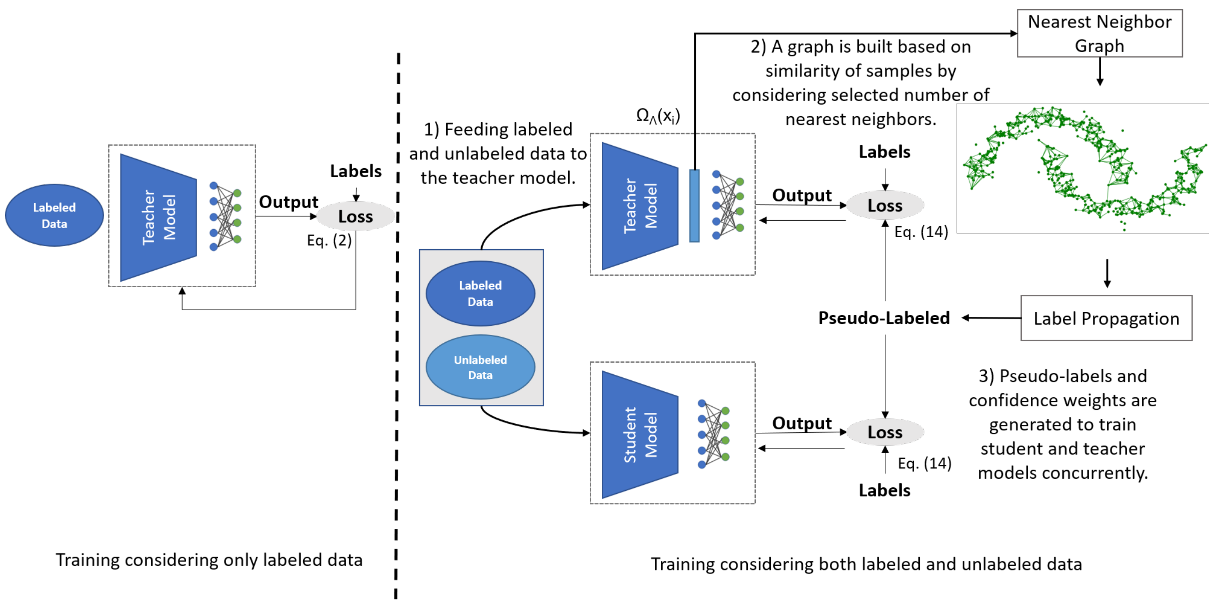


Figure 3-4-1-1: TSLP-SSL training methodology.

We have two models namely Teacher and Student. The Teacher model is trained on labeled data during the first stage and then both models are trained on labeled and unlabeled data during the second stage of the training.

With the aforementioned methodology, we have carried out the following experiments.

Experiment 1: We evaluated the validation accuracy of the proposed method considering different numbers of labeled data from the UiT dataset. In this case, the rest of the UiT dataset makes up the unlabeled data. Table 3-4-1-1 shows the obtained validation accuracy based on the different numbers of labeled data. Note that for validation, we extracted ice and water patches from polygons drawn in the MET dataset generated by Norwegian Meteorological Institute over the Danmarkshavn region. The classification results are compared with the fully supervised network presented above, using the same number of labeled data. As can be seen, the proposed method can improve the accuracy by being able to extract relevant information from unlabeled data. Table 3-4-1-2 shows other metrics including precision, recall and F1 score. These metrics also show the same behavior.

Table 3-4-1-1: Validation accuracy of proposed method considering different numbers of labeled samples

	15	30	40	60	100	500	1000
Fully supervised	39.60	52	55.67	70.35	88.72	91.50	92.06
semi-GANs	-	-	71.96	88.14	89.23	90.5	90.04
MixMatch	-	-	-	86.28	85.88	88.19	91.02
LP-SSL	55.62	75.34	75.23	89.97	90.42	**91.73**	91.24
TSLP-SSL	**88.03**	**86.96**	**90.87**	**90.47**	**91.21**	91.07	**91.94**

Table 3-4-1-2: Average of precision, recall and fi-score for different numbers of labelled data and unlabelled data from training dataset.

	15			30			40			60			100			500			1000		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
Fully supervised	.5623	.5981	.3645	.6064	.7100	.4948	.5991	.7016	.5123	.6259	.7452	.6163	.7812	.6977	.7283	.8860	.8036	.8336	.8955	.8154	.8480
semi-GANs	-	-	-	-	-	-	.4278	.4295	.4286	.7951	.6709	.7086	.7914	.7192	.7474	.8957	.7602	.8059	.8506	.7825	.8103
MixMatch	-	-	-	-	-	-	-	-	-	**.9160**	.5263	.5041	.9137	.5108	.4740	.8804	.6848	.7319	.9089	.7665	.8143
LP-SSL	.5963	.6922	.4921	.6392	**.7474**	.6493	.6611	**.8038**	.6674	.7971	**.7738**	.7847	.8041	.7949	.7990	**.9154**	.7866	.8324	**.9074**	.7751	.8211
TSLP-SSL	**.7591**	**.7927**	**.7741**	**.7314**	.7345	**.7329**	**.8299**	.7640	**.7914**	.8599	.7546	**.7925**	**.8452**	.7609	**.7990**	.8674	**.8070**	**.8326**	.9007	**.8062**	**.8432**

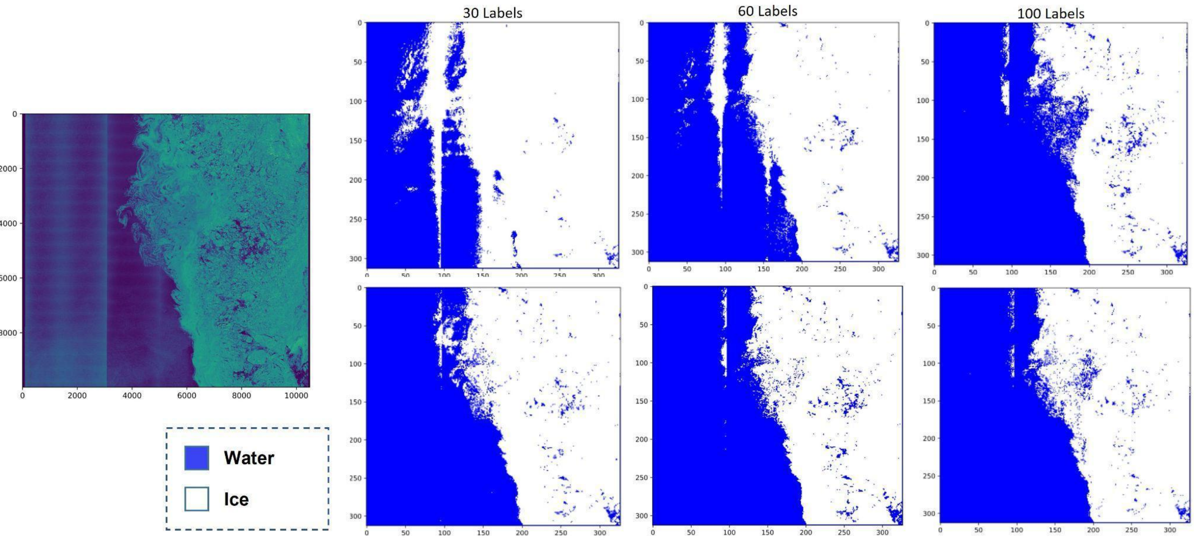


Figure 3-4-1-2: Inference results. We present qualitative results of a single input image. The first row depicts the results considering supervised deep learning and the second row depicts the results using our proposed TSLP-SSL model.

Experiment 2: We also did an experiment where the unlabeled data consisted of 5000 patches extracted randomly from the MET dataset. In this case, we did not actually know the labels of extracted patches. We trained our model using all the samples in the UiT set as labeled data. Figure 3-4-1-2 shows some inference results comparing the classification result using this trained model with the result using the 13-layers CNN with supervised learning (Khaleghian et al., 2021b). As can be seen, the initial results of the semi-supervised learning approach were quite promising, and we will conduct further research to investigate the full potential of this methodology.

Table 3-4-1-3: Validation accuracy, average precision, average recall and average f1-score considering additional real unlabeled data.

	Acc.	Pre.	Rec.	F1
Fully supervised	92.06	0.8955	0.8154	0.8480
semi-GANs	89.22	0.8730	0.7264	0.7716
MixMatch	89.55	0.8788	0.7345	0.7800
LP-SSL	91.19	0.9243	0.7626	0.8144
TSLP-SSL	**92.93**	**0.9291**	**0.8182**	**0.8606**

3.4.2 Inference results evaluation

We also present some inference results using four different images from the Danmarkshavn data provided by MET Norway, considering the student model trained on labeled datasets and extended with unlabeled data (Experiment 2). In Figure 3-4-2-1, the left column depicts the original SAR images, the middle column presents the inference results obtained with the supervised learning model, and the last column shows the results obtained with our proposed TSLP-SSL method. Water is highlighted in blue color and ice is highlighted in white color. These inference results again show the capability of our proposed semi-supervised method in using the information of unlabeled data.

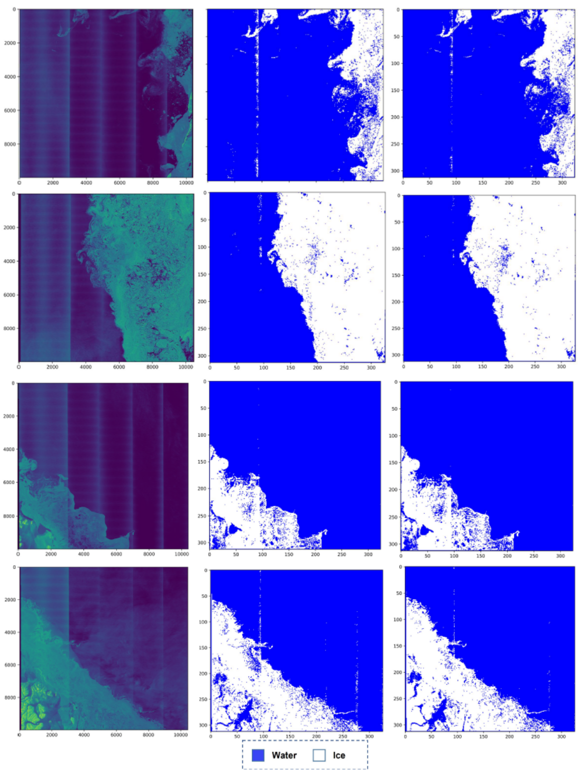


Figure 3-4-2-1: (Left) HV intensity image. (Middle) inference result of supervised model. (Right) inference result of semi-supervised model.

We also evaluated the proposed method on a number of different images from the MET Norway dataset. Figure 3-4-2-2 shows the inference results of these images. The annotated red rectangles indicate areas of misclassification. We mainly have two issues. We have some noise effects on the boundary between the first and the second sub-swaths. However, we believe the noise removal method developed by NERSC (see above) can lead to better classification results. We also see some misclassification of young ice and ice covered by water areas.

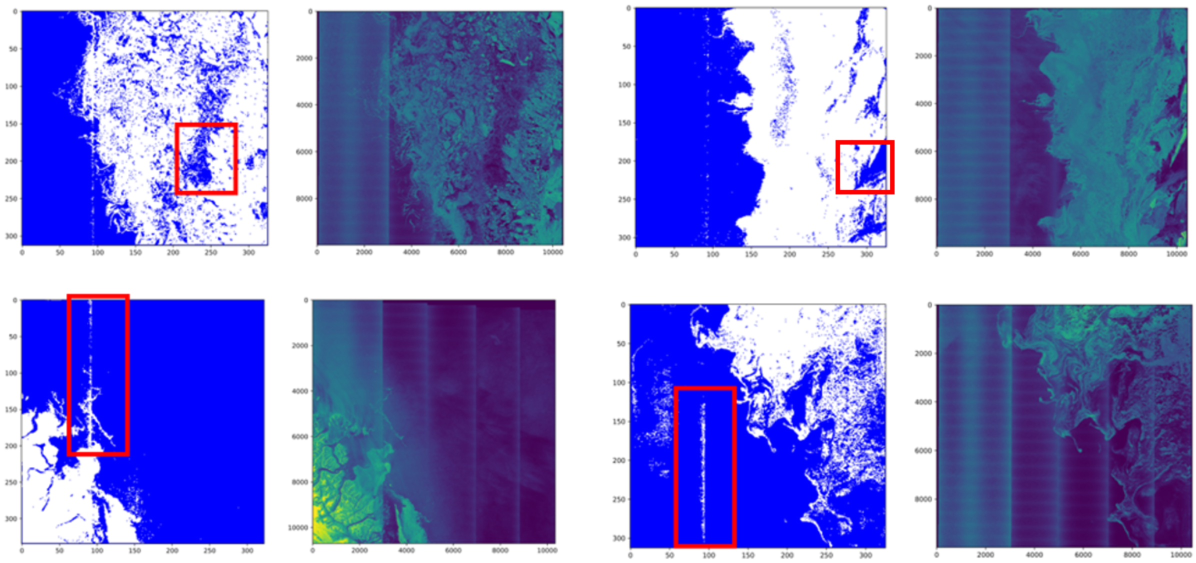


Figure 3-4-2-2: Inference results using more images from MET Norway dataset. We annotated the challenging areas in these images.

4 Implementation on Hopsworks and the PolarTEP

Hopsworks is utilized for distributed training to test our model in an operational phase. Figure 4-1. part a (lower part) shows the overall architecture, in which Hopsworks is used for the data processing and distributed training. We used the Hopsworks capabilities to read the training data and train it on several spark executors which hold TensorFlow codes. The training dataset associated with the water – ice classification task was converted to the TFrecord file format and stored on HopsFS and then used by TensorFlow for distributed training. For distributed training, we considered different training strategies, including MirroredStrategy and MultiWorkerMirroredStrategy. We have reported our experiments in Deliverable 2.3.

PolarTEP pre-processes the data, dividing a full satellite scene into smaller blocks. Then PolarTEP classifies each block using the Hopsworks model serving through the exposed REST APs. Next, PolarTEP reassembles the result into a full classification map. This was the most natural way to set up the inference step, given the capabilities of Hopsworks. In fact, Hopsworks provides a nice way to work with model training and serving in Python or PySpark. For the inference phase, we considered two different implementations, which are indicated in part a and b of Figure 4-1. In the first implementation, a, inference is performed on Hopsworks. In this case, patches are generated and sent to Hopsworks to be classified. For implementation b, inference is performed on PolarTEP. In this case, the trained models are transferred to PolarTEP and PolarTEP is responsible for the pre-processing and inference.

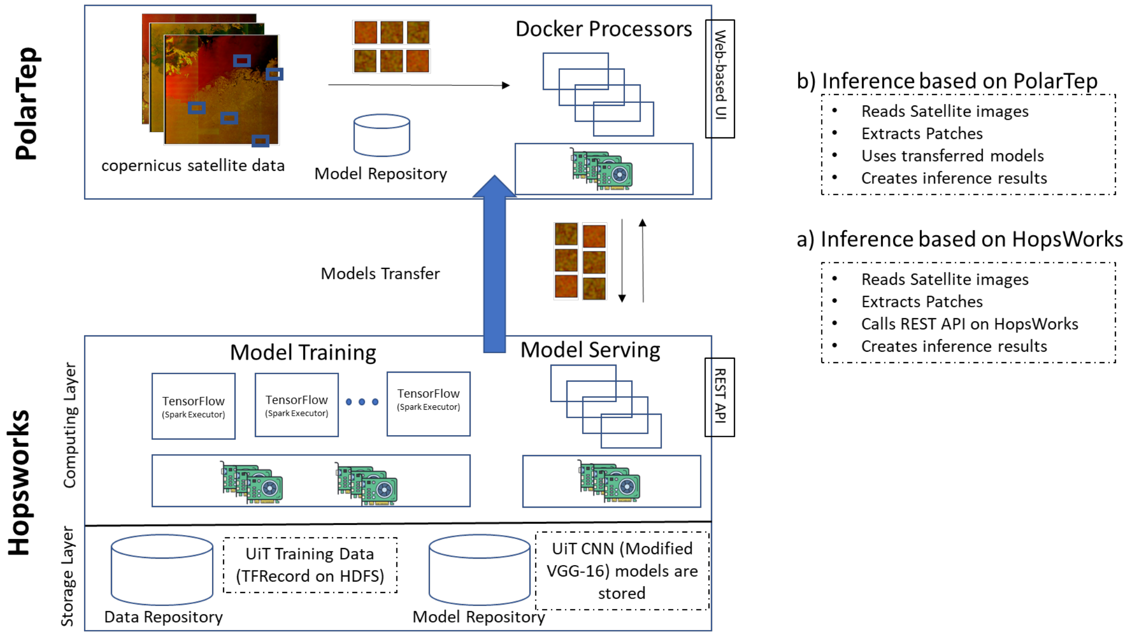


Figure 4-1: Overall architecture of UiT implementations using Hopsworks and PolarTEP.

In response to the need for custom data processing, in the second implementation, the trained model is transferred to PolarTEP in an off-line way. In PolarTEP, by using a workflow like the scheme based on Docker container technology, we can accomplish pre-processing and inference at the same place and close to the Copernicus data using PolarTEP on the CreoDIAS platform. In this implementation, the trained modes are transferred in offline mode by creating a Docker image. More details are found in deliverable D5.5.

5 Additional studies

5.1 Distributed training

Deep Neural Network (DNN) is rapidly taking over a variety of aspects in our daily lives. However, DNNs' rise into prominence is tightly coupled to the available computational power. A single machine is often not capable of finishing a training task in a desired time frame. It is more significantly important when we have a large amount of training data. Especially in semi-supervised learning where we can increase unlabeled data used in the training process easily. Therefore, to have scalable computing to analyze large amounts of data is essential. In Table 5-1-1, we report our training time considering different semi-supervised methods. Although our training dataset is not so big, we can observe high training time based on these methods. It clearly shows the need for scalable deep learning algorithms for processing large datasets.

Table 5-1-1: Training time for different methods using 60 labeled data and 10000 unlabeled data on the UiT training dataset on a single GPU (Quadro RTX 5000 16GB).

Methods	Batch Size	Training time
LP-DeepSS (WideResNet)	100	1h-2 h
LP-DeepSS (CNN)	100	2h-3h
TSLP-SSL (CNN)	100	4h-6h
MixMatch (WideRestNet)	100	8h-10h
SGANs (CNN)	64 (best)	45m-1h

We investigated a distributed training based on knowledge distillation called co-distillation (Anil, 2018). This method is interesting because, 1) it is related to our proposed method in supervised and semi-supervised and we can apply this method to extend our proposed method especially TSLP-SSL to distributed training, 2) this method provides more parallelism while reducing the communication cost, and 3) it can be used on commodity hardware in contrast to other approaches that need high speed and low latency networks.

Anil et al. (2018) proposed an online variant of distillation that is called co-distillation. In contrast to the two-step distillation, when two or more models co-distill, there is only one phase. Co-distillation trains n copies of a model in parallel and starts distillation early in the training process using an additional distillation term to the loss function. Co-distillation is an elegant alternative approach to distributed training with reduced communication overhead. Rather than synchronizing models to have the same weights, co-distillation seeks to share information by having the models represent the same function (i.e., input-output mapping).

Based on co-distillation, we proposed a method called feature-based co-distillation. We take advantage of including learned feature space by CNNs to perform more effective distillation. The overall architecture of this method is shown in Figure 5-1-1.

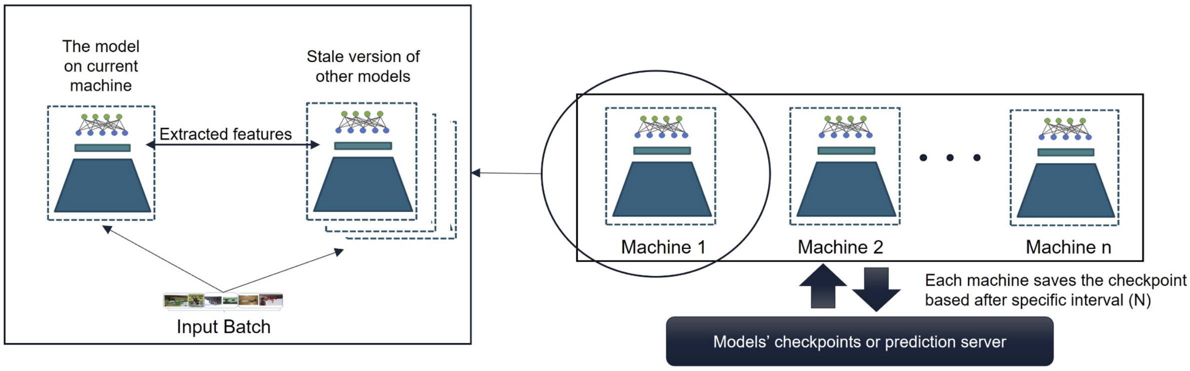


Figure 5-1-1: Overall architecture of feature-based co-distillation (FBC).

Our method can achieve the same accuracy of co-distillation with fewer epochs. Figure 5-1-2 shows the validation accuracies of our proposed method, co-distillation, and AllReduce algorithms on two machines with 3 GPUs on each. In this experiment, we use early stopping when co-distillation and our method reach the highest accuracy of AllReduce. In this experiment, we considered a high update interval (9000 steps) and we can see a jump in training loss when we start distillation in the training process.

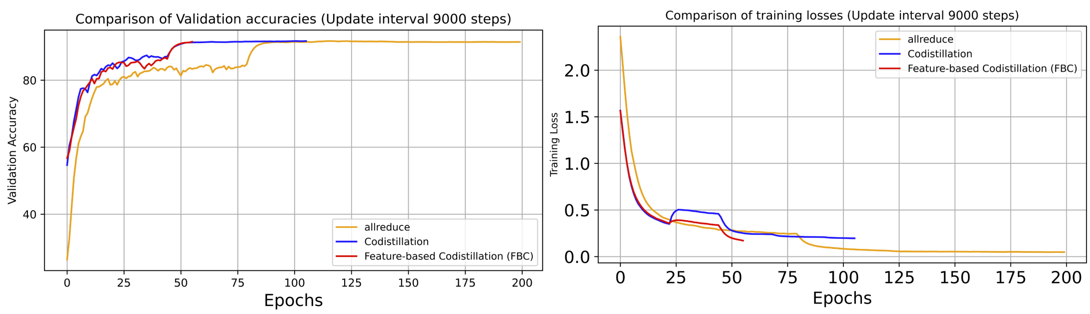


Figure 5-1-2: Validation accuracy and training loss of AllReduce, Co-distillation, and our proposed method.

We are working to publish our method for scalable deep supervised learning. However, we will continue to research on extending our distributed training method for semi-supervised learning (especially our proposed method TSLP-SSL) in the future.

5.2 Incidence angle dependency

Another side study at UiT addresses statistical sea ice classification with ice-type dependent incidence angle dependency integrated into the classification algorithm. As can be seen from

Figure 2-2-1, open water has a steeper intensity decay with incidence angle than multi-year and first-year ice. Detailed studies indicate that the incidence angle decay also varies with ice-type. The common pre-processing method, and currently the best solution, is known as *incidence angle correction* and applies a range-dependent image scaling to the whole image. However, this must be performed with a single common rate, since the classes are not yet known, and is therefore not optimized for all classes. A single slope correction may suffice when limited to classes with similar slopes, such as first-year and multi-year ice only in the Canadian Archipelago (Mahmud et al., 2018), but not for more general cases that can include small water areas such as leads. As intensities of various objects can be affected differently with incidence angle owing to the geometrical interaction between microwaves and image targets, Cristea and co-authors proposed an unsupervised method to distinguish water, oil slicks and ice (Cristea et al., 2020). For this method, a multivariate Gaussian Mixture model is employed to fit the HH and HV intensities, where the mean value of each component decays linearly with incidence angle. In the end, the appropriate number of clusters, which correspond to the Gaussian functions, are being estimated through goodness-of-fit tests. This physical based clustering algorithm may give a better confidence for labeling and eventually can aid the pixel-wise classification. To further investigate the benefit, we performed training and validation on the same images we selected in Section 3.3, with the same setting. It turns out that the early stopping happens before epoch 20, which indicates that the model converges more quickly than the model using the basic clustering method, without physical information provided. Furthermore, the accuracy for training and validation are 94.4% and 94.5%, respectively, which is slightly better than the achieved performance by using our training labels used in Section 3.3. Meanwhile, the loss for using these labels is 0.032 and 0.031 for training and validation, which is also better, compared with our findings in Table 3-3-3-1. The detailed training and validation accuracy as well as loss as function of epochs are shown in Figure 5-2-1.

Furthermore, we also carried out the inference procedure using the model trained by using this segmentation method, which is shown in Figure 5-2-2. In the figure, the first and second columns indicate the intensity over HH and HV polarization, while the third and fourth columns represent the inference obtained by the model trained using the clustering method without (Section 3.3) and with (this section) physical information about incidence angle considered.

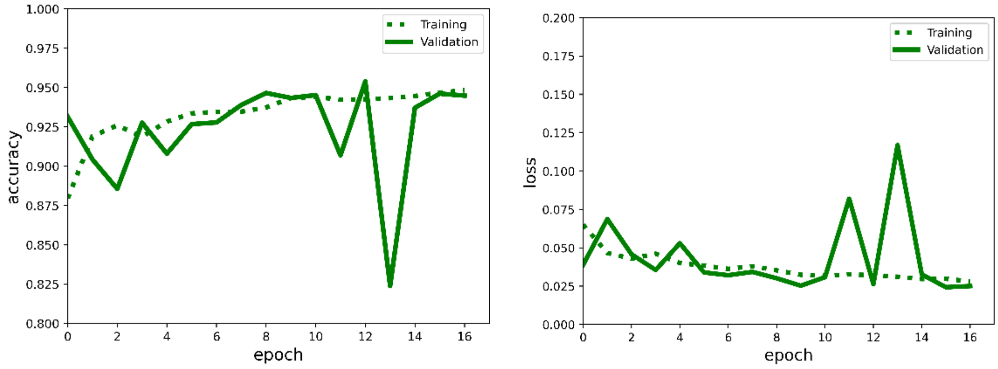
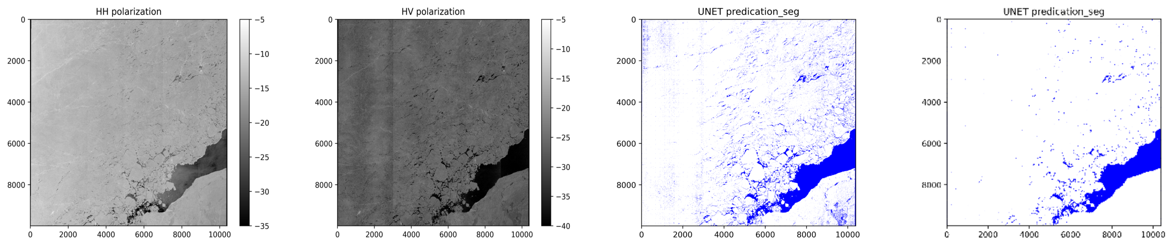
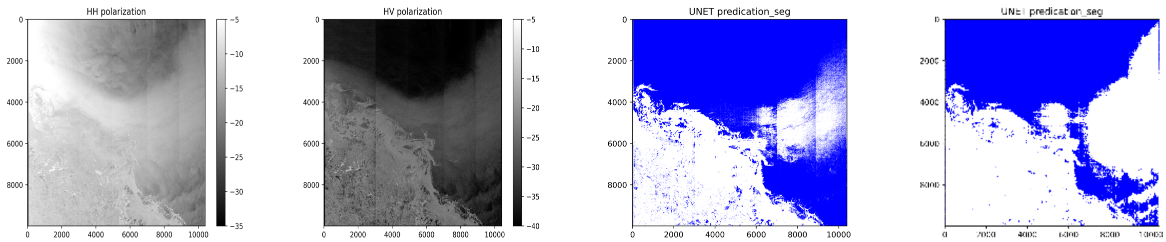


Figure 5-2-1: Training and validation accuracies (left) and loss (right) versus epochs.

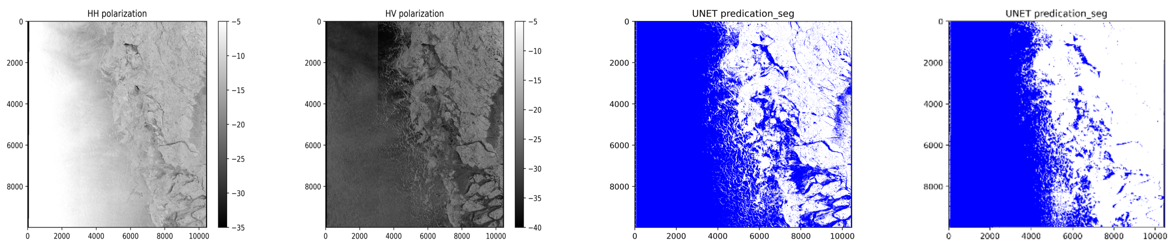
HH	HV	inference (without physical information)	inference (with physical information)



S1A_EW_GRDM_1SDH_20210604T202343_20210604T202443_038195_0481F7_E982



S1B_EW_GRDM_1SDH_20200209T170617_20200209T170717_020195_0263BC_B84



S1A_EW_GRDM_1SDH_20200315T073001_20200315T073101_031683_03A71B_3

Figure 5-2-2: Inference by using physical based labeling for training compared, with the labelling without physical constraints

As can be seen, by considering the physical information about the incidence angle effect, the inference over ice is much better than the inference captured by the original method. However, more misclassifications are also found for using physical information, especially when the wind is over the open water. This indicates that the inclusion of windy scenarios over sea-ice is demanding for obtaining good inference.

5.3 Polarimetry analysis

The ambiguity of SAR scattering from sea ice is a significant obstacle to automatic sea ice classification. In order to mediate this problem, several approaches to extract more surface information from the SAR signals have been addressed. One way to reduce ambiguities is by adding more polarimetric features to the input dataset. A full-polarimetric SAR system transmits and receives both linear horizontal (H) and vertical (V) polarized electromagnetic waves and is hence providing measurements in four polarization channels (quad-pol). These are generally referred to as the HH, VV, HV and VH channels. It is well known that a full polarimetric SAR system enables the extraction of scattering information through so-called polarimetric target decompositions. However, there are methods to derive similar scattering information from dual-polarimetric data. At UiT, this direction has been pursued with the goal of understanding the additional value of these extra features with regard to the interpretation of sea ice SAR scenes. More specifically, we have studied the capabilities of some recently introduced features defined through their geodesic distance (GD) measures from canonical targets (Debanshu et al., 2020) in providing sea ice information. As an example, Figure 5-3-1 illustrates the ice-versus-water distinguishing power of two GD parameters derived from a Radarsat-2 scene, where the image to the left is a Pauli image generated from the full-polarimetric scene, the next two are two GD-parameters generated from a quad-pol scene, and the two panels to the right are the same features generated from only the HH & HV channels. As can be seen, the dual-pol based parameters preserve the difference between ice and water. A drawback of the method is that it relies on *single-look complex* data, which is why the features have not yet been tested in a DL classification architecture.

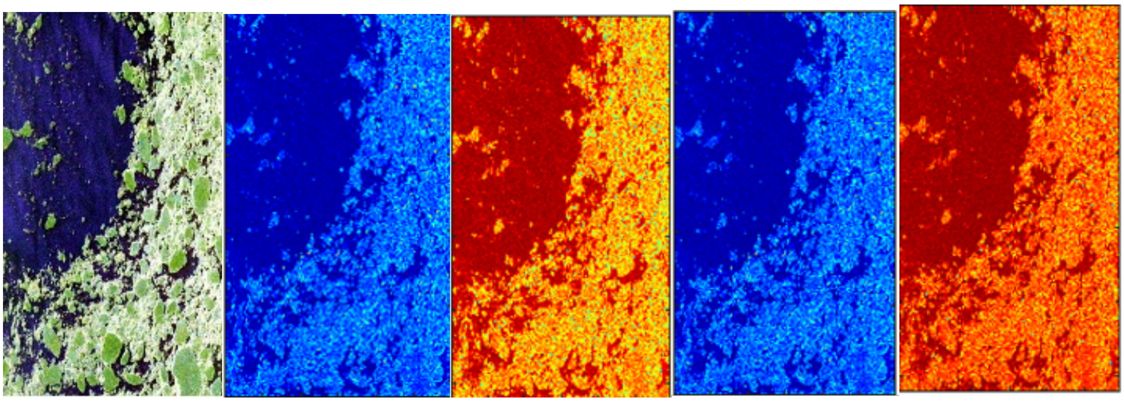


Figure 5-3-1: Comparison of GD parameters generated from quad- and dual-pol RS-2 data. Left panel: Pauli image showing water as dark blue and sea ice as green/white. The next two panels: Quad-pol generated GD parameters. Two right-most panels: same GD parameters generated from dual-pol data.

6 Summary and conclusion

The user requirements for the Polar Use Case concluded that: "*the types of satellite data needed varied with the type of user being supported: transportation and other tactical information needed high resolution synthetic aperture radar (SAR) data, whilst climatological studies found low resolution passive microwave (PMW) data sufficient.*" The former users, i.e., the ship captains operating in ice-infested waters, need high-resolution information, on the orders of meters, whereas climatologists can do well with low resolution data. When studying and developing deep learning architectures, we have had both these requirements in mind. We have studied both low-resolution architectures, which are based on patch-wise SAR-inputs and as well as patch-wise outputs, and high-resolution models, which produce pixel-wise output classification maps.

In this document, we briefly discuss some of the intrinsic challenges of sea ice classification based on Sentinel-1 EW mode SAR images. We focus on binary sea ice-versus-water classification but include a section on multi-type classification. It is argued that ice versus water separation is the natural first step, and the task of performing ice-type maps will come next.

The patch-wise architectures were at the outset relying on common structures, which had shown good performance in the computer vision application area. Hence, the VGG-16 network was selected as the base model. We investigated various ways of training the model, including transfer learning approaches, and found that due to the specific properties of SAR data, the primary data source for the polar use case, the network should be trained from scratch using reliable training data in order to perform optimally.

In the ExtremeEarth project, we learnt that for sea ice mapping, reliable training data in fact is a critical factor in order for DL models to perform well. For the sea ice classification case, the generation of enough reliable training data is probably more critical, than which architecture is used. In order to alleviate the scarcity of training data, we extended the study of patch-wise classification to also include semi-supervised learning and introduced *a new model*. We demonstrated that our semi-supervised model was in fact able to extract useful label information from un-labelled samples and overall, was able to perform well for the ice-versus-water classification task.

In the pixel-wise case, after testing out several alternatives, we ended up using the UNet architecture. In this case, we also generated a new training dataset using an *ad-hoc active learning strategy* based on standard K-means clustering. The current deliverable documents that ice-water separation may sometimes be easy, but, because of often varying wind conditions over the oceans, very challenging cases may also occur. Our trained reduced UNet shows promising potential, with some cases of excellent performance.

In conclusion, the lessons learnt from the ExtremeEarth project can be summarized as follows:

- DL algorithms show promising performance, but they need extensive amounts of reliable training data to perform well.
- The generation of training data is a time-consuming and labor-costly process, which needs to involve analyst experts with profound experience in interpreting SAR images.
- Semi-supervised approaches may alleviate the lack of large labeled data sets for training, by taking advantage of huge amounts of un-labeled data.
- For pixel-wise classification, the reduced UNet seems to be a suitable architecture. However, the appropriate size of the network and the depth in terms of feature layers at each stage, need to be further explored.

Our experience from the ExtremeEarth project is supporting the experience that many other researchers have made when it comes to automatic sea ice classification from SAR. The problem is *not trivial* due to the complexity of the SAR imaging, the complexity of scattering from the sea ice medium, and due to the changes of the SAR signal properties as a function of seasonal changes. The DL algorithms definitely have proved themselves to be valuable tools that will be helpful in the future, but the algorithm development needs to be supported by understanding of the physics of the problem.

References

Anil R., G. Pereyra, A. Passos, R. Ormandi, G. E. Dahl, and G. E. Hinton, "Large scale distributed neural network training through online distillation," arXiv preprint arXiv:1804.03235, 2018.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A. A. "Albumentations: fast and flexible image augmentations." Information, 11 (2), 2020.

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H, "Rethinking atrous convolution for semantic image segmentation. " arXiv preprint arXiv:1706.05587, 2017.

Cristea, A., Van Houtte, J., & Doulgeris, A. P, "Integrating incidence angle dependencies into the clustering-based segmentation of SAR images. " IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 2925-2939, 2020.

Debanshu, R., E. Pottier, A. Bhattacharya, A. F. Frerey, "A PolSAR Scattering Power Factorization Framework and Novel Roll-Invariant Parameters Based Unsupervised Classification Scheme Using a Geodesic Distance", IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no 9, 2020.

Gao Y., F. Gao, J. Dong, S. Wang, "Transferred deep learning for sea ice change detection from synthetic-aperture radar images." *IEEE Geoscience and Remote Sensing Letters.* vol. *16*, 2019.

Han Y., Y. Gao, Y. Zhang, J. Wang, S. Yang, "Hyperspectral Sea Ice Image Classification Based on the Spectral-Spatial-Joint Feature with Deep Learning." *Remote Sensing. 11*, 2019.

Huang G., Z. Liu, L. Van Der Maaten, K. Q. Weinberger, "Densely connected convolutional networks." In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.

Hughes, NN. and F. Amdal (2021). ExtremeEarth Polar Use Case Training Data (2.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4683174

Khaleghian S., H. Ullah, T. Kræmer, N. Hughes, T. Eltoft, and A. Marinoni, "Sea ice classification of SAR imagery based on convolution neural networks." Remote Sensing, vol. 13, no. 9, 2021a.

Khaleghian S., H. Ullah, T. Kræmer, N. Hughes, T. Eltoft, and A. Marinoni, "Deep Semi-Supervised Teacher–Student Modelbased on Label Propagation for Sea Ice Classification." Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2021b).

Kruk R., M. C. Fuller, A. S. Komarov, D. Isleifson, I. Jeffrey, "Proof of Concept for Sea Ice Stage of Development Classification Using Deep Learning." *Remote Sensing. 12*, 2020.

Lohse, J.; Doulgeris, A.P.; Dierking, W, “Mapping sea-ice types from Sentinel-1 considering the surface-type dependent effect of incidence angle.“ Annals of Glaciology 2020, p. 1–11. doi:10.1017/aog.2020.45.

Long, J., Shelhamer, E., & Darrell, T, “Fully convolutional networks for semantic segmentation.“ In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440), 2015.

Mahmud, M. S., Geldsetzer, T., Howell, S. E., Yackel, J. J., Nandan, V., & Scharien, R. K, “Incidence angle dependence of HH-polarized C-and L-band wintertime backscatter over Arctic sea ice.“ IEEE Transactions on Geoscience and Remote Sensing, 56(11), 6686-6698,2018.

Otter D. W., J. R. Medina, and J. K. Kalita, “A survey of the usages of deep learning for natural language processing,” IEEE Transactions on Neural Networks and Learning Systems, vol. 32, 2020

Park J., A. A. Korosov, M. Babiker, S. Sandven and J. Won, "Efficient Thermal Noise Removal for Sentinel-1 TOPSAR Cross-Polarization Channel." IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 3, pp. 1555-1565, March 2018.

Ronneberger, O., Fischer, P., & Brox, T, "U-net: Convolutional networks for biomedical image segmentation." In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham, 2015.

Tan M., R. Pang, and Q. V. Le, “EfficientDet: Scalable and efficient object detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

Ukaby, F. T., Moore, R. K., Fung, A. K.: “Microwave Remote Sensing. Active and Passive”, Artech House Inc., 1986.

Wang L., K. A. Scott, L. Xu, 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, 2016.

Wang L., K. Scott, D. Clausi, “Sea ice concentration estimation during freeze-up from SAR imagery using a convolutional neural network.” Remote Sensing. vol. 9, 2017.

Yan X., B. Cui, Y. Xu, P. Shi, and Z. Wang, “A method of information protection for collaborative deep learning under GAN model attack,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, 2019.