https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

The Cryosphere Discussions

SAR Deep Learning Sea Ice Retrieval Trained with Airborne Laser Scanner Measurements from the MOSAiC Expedition

Karl Kortum[1, 2], Suman Singha[3, 4], Gunnar Spreen[2], Nils Hutter[5], Arttu Jutila[6, 5], and Christian Haas[5]

[1]German Aerospace Center (DLR)
[2]Institute of Environmental Physics, University of Bremen
[3]Danish Meteorological Institute (DMI)
[4]Department of Geography, University of Calgary
[5]Alfred-Wegener-Institut (AWI)
[6]Finnish Meteorological Institute (FMI), Helsinki, Finland)

Correspondence: karl.kortum@dlr.de

Abstract. Automated sea ice charting from Synthetic Aperture Radar (SAR) has been researched for more than a decade and still, we are not close to unlocking the full potential of automated solutions in terms of resolution and accuracy. The central complications arise from ground truth data not being readily available in the polar regions. In this paper, we build a dataset from 20 near coincident X-Band SAR acquisitions and as many Airborne Laser Scanner (ALS) measurements from the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC), between October and May. This dataset is then used to assess the accuracy and robustness of five machine learning based approaches, by deriving classes from the freeboard, surface roughness (standard deviation at $0.5m$ correlation length) and reflectance. It is shown that there is only a weak correlation of the radar backscatter and the sea ice topography. Accuracies between 40% and 69% percent and robustnesses between 68% and 85% give a realistic insight into modern classifiers' performance across a range of ice conditions over 8 months. It also marks the first time algorithms are trained entirely with labels from coincident measurements, allowing for a probabilistic class retrieval. The results show that segmentation models able to learn from the class distribution significantly perform pixel-wise classification approaches.

1 Introduction

Sea ice classification from remote sensing and especially SAR instruments have been used for monitoring the Arctic sea ice for multiple decades, with automation being proposed as early as the mid eighties by Fily and Rothrock (1986). However, even with the inception of advanced machine learning methods and modern data analysis, there does not yet exist a universally reliable classifier to retrieve sea ice classes from radar imagery. The potential for such a classifier is obvious: Humans are not able to match the speed and precision of an automated algorithm. Until now, however, this potential has yet to be fully unlocked; Human-generated ice charts (for an overview regard the World Meteorologial Organizations overview by JCOMM (2017)) are still dominant in operational usage, despite the considerable amount of research that has been focused on the subject. These products unfortunately can provide only coarse approximate labels of the sea ice. For cross-cutting research, a

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

The Cryosphere Discussions
Open Access
EGU

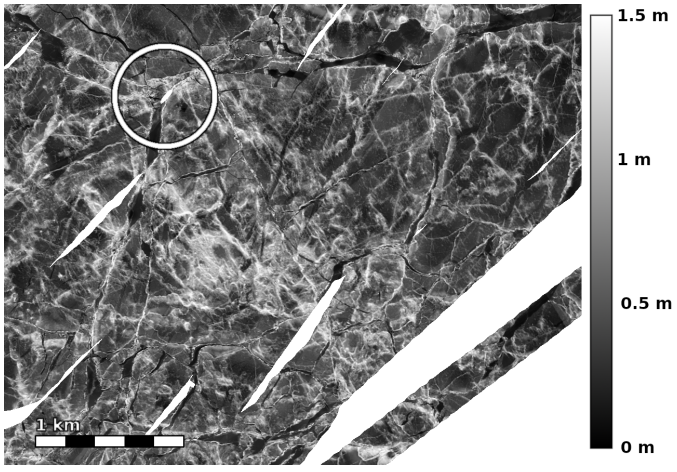


Figure 1. Section of ALS measured freeboard over the MOSAiC floe on April 8, 2020. RV Polarstern can be seen in the center of the red circle. Brighter values correspond to higher freeboard values whereas white areas indicate no data. The displayed freeboard range is 0 to 1.5 metres.

more detailed and higher resolution of classification would be preferable and should be possible given the spatial resolution of the SAR sensors. As a human analyst generating an ice chart has only limited time to annotate a SAR scene, such high-resolution labels are not contained in the ice charts. Leads, for example, hot spots of ocean and atmosphere interaction and thus of particular interest for the energy budgets, are generally not labelled in operational ice charting. At this point, with many different classifiers having been proposed and developed (e.g. Kwok et al. (1992); Soh and Tsatsoulis (1999); Hara et al. (1995); Karvonen (2004); Ressel et al. (2015); Doulgeris (2015); Johansson et al. (2020); Lohse et al. (2021)) one must ask the question why no meaningful direction has yet established itself in the ongoing research. The answer to the question - aside from the complexity of the subject - is twofold. Firstly and most important is the state of the data. Although we have a great wealth of satellite SAR acquisitions of the sea ice in diverse states and conditions, we lack the corresponding ground truth information. Secondly, the constantly varying and difficult-to-predict drift and deformation of sea ice makes it nearly impossible to image the same area of sea ice over longer time series to evaluate any proposed classifiers' robustness. The latter is particularly true for high-resolution imagery. These two shortcomings open this topic up to a plethora of different challenges because we have almost no way to test, iterate and improve sea ice retrieval algorithms in a structured manner. This stifles the rate at which progress in the field can be made or even recognised.

On a mission to fill gaps in our knowledge about the Arctic sea ice and its climatology, the MOSAiC expedition launched in the autumn of 2019 and the ship Polarstern spent a year adrift with the ice pack. Aboard, interdisciplinary teams of scientists worked to collect as many data as possible, which will help to further our understanding of one of our earth's most remote regions. With the mission came the unique opportunity to collect exactly the type of ground truth over a long time period, that is needed to test sea ice retrieval algorithms, with satellite-borne SAR data being acquired at the same time. An overview of the snow and ice related activities is given in Nicolaus et al. (2022).

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

Ice and snow transects from Itkin et al. (2021) or drilling hold the most detailed information of the underlying ice. Unfortunately, the spatial extents of these measurements are too sparse to be used for comparison with the satellite acquisitions. Aerial measurements taken from helicopters, such as the Airborne Laser Scanner (ALS) data products by Hutter et al. (2022a, b) being used in this approach (Fig. 1) provide information about the height of the snow and/or ice surface above the local sea level, i.e. freeboard, and surface reflectance at scales of kilometres to tens of kilometres. These data are therefore a prime candidate to extract ground truth information for ice classification based on roughness and thickness.

One prominent emerging method of segmenting image data are machine learning based approaches based on convolutional neural networs, such as published in Simonyan and Zisserman (2015); He et al. (2015); Liu et al. (2022); Ronneberger et al. (2015); Zhou et al. (2018, 2019). Advancements in the field of machine vision are being made at a rapid pace, able to leverage the improvements in chip design and the increasing amount of data that are being generated. The image-like properties of SAR acquisitions mean that this knowledge is transferable to the ice classification domain (e.g. Boulze et al. (2020); Ullah et al. (2021); Wang and Li (2021); Kortum et al. (2021, 2022)). Historically, this has been done with texture extraction and subsequent dense neural networks as in Ressel et al. (2016); Singha et al. (2018); Murashkin et al. (2018), pixel-wise classification using image classifiers based on convolutional neural networks as by Boulze et al. (2020); Ullah et al. (2021) and segmentation models that are able to segment an entire patch simultaneously like in Wang and Li (2021). In this study, we will use the unique opportunity provided by 20 instances of near-coincident ALS and SAR data over a period of 8 months to compare a variety of machine learning-based classification approaches in terms of classification accuracy and robustness on classes delineated directly from measurements. For the first time, we have accurate, high resolution sea ice topography measurements of freeboard and surface reflectance with high spatial overlap and low time differences between acquisitions to truly test the capability of retrieving freeboard and (above snow) surface roughness based sea ice classes from SAR data.

2 Methodology

2.1 The Data

The SAR component of the analysis is made up of TerraSAR-X X-band acquisitions in StripMap (SM) mode. The resulting scenes have a pixel spacing of 3.5 metres and a radiometric resolution of 16 bit. Both HH and VV bands are acquired by the satellite simultaneously. This configuration of polarisations has been shown to yield valuable information for ice classification in Ressel et al. (2016)Geldsetzer and Yackel (2009).

The ALS data from Hutter et al. (2022b, a) from 20 scenes (appendix A) between October 2019 and May 2020 are used to delineate sea ice classes. The data were acquired by flying a mow-the-lawn pattern over the ice. The resulting ALS grid has a geospatial resolution of 0.5 metres. For midwinter flights in high latitudes of >85°N, the post-processing of the helicopter INS/GPS data failed and ALS data processing was performed using a lower frequency real-time navigation solution with metre-scale undulations in GPS altitude that propagated to the surface elevation retrieved from the ALS. The undulations in the computed freeboard could be minimised using a correction calculated from swath-to-swath overlap. It should be noted that the local standard deviation of the freeboard is left intact by these processing artefacts and can still be used to derive a

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

parametrisation of the local surface roughness, where these undulations are present. An additional measurement aside from freeboard is the surface reflectance at the wavelength of the laser (1064 nm), which is useful to identify regions of young ice that have not yet been covered by snow. For the acquisitions with unphysical undulations in the freeboard measurement, freeboard was not used to delineate class labels. Instead, only classes which could be inferred from the surface roughness and reflectivity were used.

Colocation: For each ALS grid, the first step for co-locating with SAR data is to find the SAR acquisition that is closest to the ALS measurement time, whilst still having substantial spatial overlap. Then, by using the Polarstern ship to determine a common coordinate system, the two measurements are fused by assigning each ALS data point to the closest TSX pixel (see. Kortum et al. (2021); Hendricks (2019).) In the common coordinate system, this means that the two measurements are in the same TerraSAR-X grid cell relative to the ship. Because of the difference in resolutions ($0.5m$ ALS and $3.5m$ SAR), we obtain approximately 49 points of ALS measurements per SAR pixel. The freeboard and roughness are then computed as the respective mean and standard deviation of these points. Using the Polarstern as an origin of the common coordinate system is sensible, as we have accurate GPS positioning and heading to account for ice drift and rotation. The matching of the two products using this method was accurate to a couple of metres. To further improve the accuracy of colocation, a final translation and rotation was then determined manually. Afterwards, the features overlapped perfectly at (TerraSAR-X) pixel resolution. The accuracy of co-location is made possible by more than daily TerraSAR-X SAR acquisitions of the MOSAiC floe, which helps keep the time differences between satellite and helicopter measurements small.

Determining labels: We have categorised the measured sea ice into three classes. A label is given for each SAR pixel, for which ALS information is available. For ease of reference, we are giving them names in accordance with conventions, which are easier to contextualise. However, the exact definitions of the classes is given here. They are fully given by the ALS measurement. The three classes are: Open water and young ice (OW/YI), level first-year ice (LFYI) and deformed first-year and multiyear ice (DFYI/MYI). These classes we define as follows (see Fig. 2 for a visual aid):

- OW/YI: Ice whose reflectance (range corrected target echo amplitude) is significantly lower than that of the surrounding snow covered ice. Typically values around -7dB were used as a threshold value and adjusted manually if needed. Note, that finer separation here is not possible from the data alone, but from reports of scientists on the expedition we know that most ice in this class will have already formed a thin ice layer and entirely open water was very rare during the flights.

- LFYI: Snow-covered ice with a surface roughness (standard deviation of freeboard measurements at scales of the ALS grid ($0.5m$) calculated over one TSX pixel ($3.5m^2$)) of less than 1 centimetre or a freeboard value lower than the higher inflection point in the freeboard distribution (typically around 40 centimetres).

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

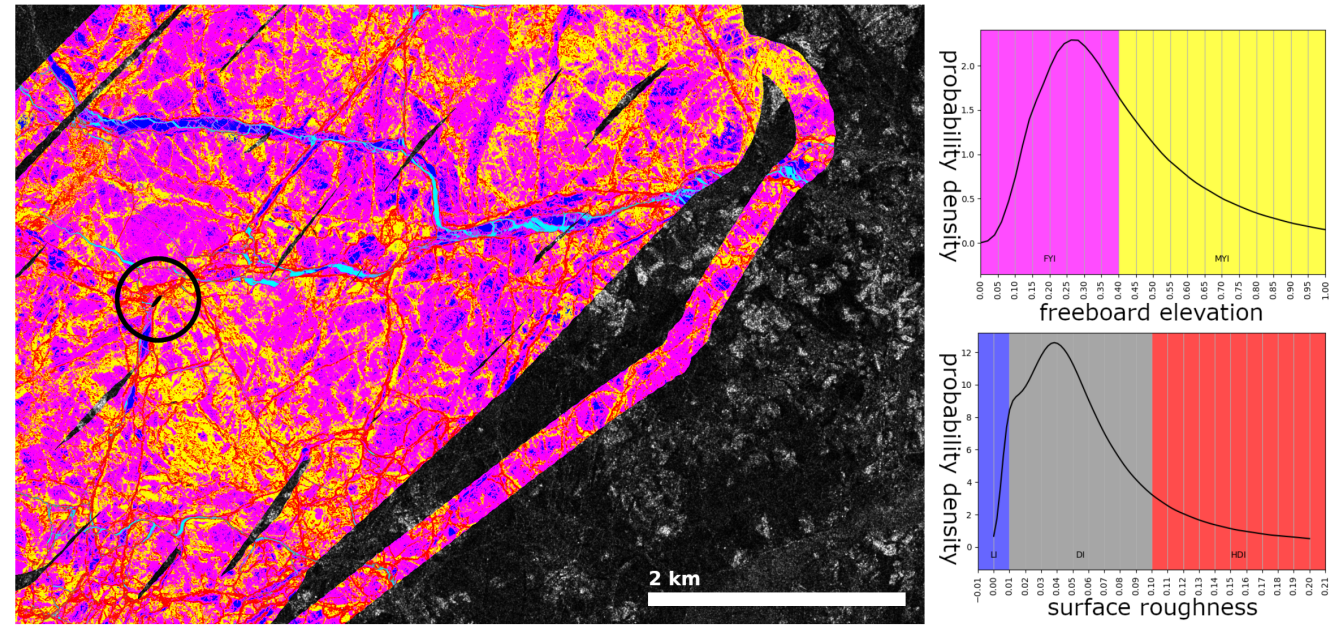


Figure 2. Derived labels from the ALS acquisition on the April 8, 2020 overlayed on the HH channel of the near-coincident SAR measurement (left) and estimated probability density functions from the distributions of freeboard and surface roughness (in this case this is the local standard deviation of the freeboard) (right). *Yellow* indicates ice with a higher freeboard than the high inflection point of the distribution. *Magenta* is ice with a lower free board than that. *Red* are areas with higher surface roughness than 10 cm. *Blue* areas are ice with surface roughness of less than 1 cm. *Cyan* areas have reflectivity indicating no snow cover (less than -7dB Echo Amplitude). For this study, yellow and red, as well as magenta and blue classes are combined. The grey background of the surface roughness distribution denotes the region that was not used to identify ice classes, as there was considerable mixing in this parameter region.

– DFYI/MYI: Snow covered ice with a surface roughness of more than 10 centimetres or a freeboard greater than the higher inflection point in the freeboard distribution.

Because these labels are entirely defined by measurements of the ice surface (Fig. 2), we can also infer the probabilities of belonging to each class, by assuming a gaussian distribution of ALS freeboard and reflection measurements at each TSX pixel. From the 49 ALS measurements, we compute the mean and standard deviation of the freeboard and can then compute the probabilities of lying below or above any freeboard thresholds by using the error function. Explicitly, we integrate the area under the curve of the gaussian distribution, above and below the threshold. Thus we obtain labels which give the probabilities of belonging to a certain class, rather than discrete classes. Assuming a gaussian distribution allows us to also infer uncertainties of the surface roughness.

The derived labels from each scene are split into two mutually exclusive connected subsets. By connected we mean, that in all but edge cases pixels are neighbouring ones from the same subset. The training set is made up of 75% of labels whilst the test set consists of 20%. The remaining 5% of the data is used as a validation set. The validation data is used only to decide when

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

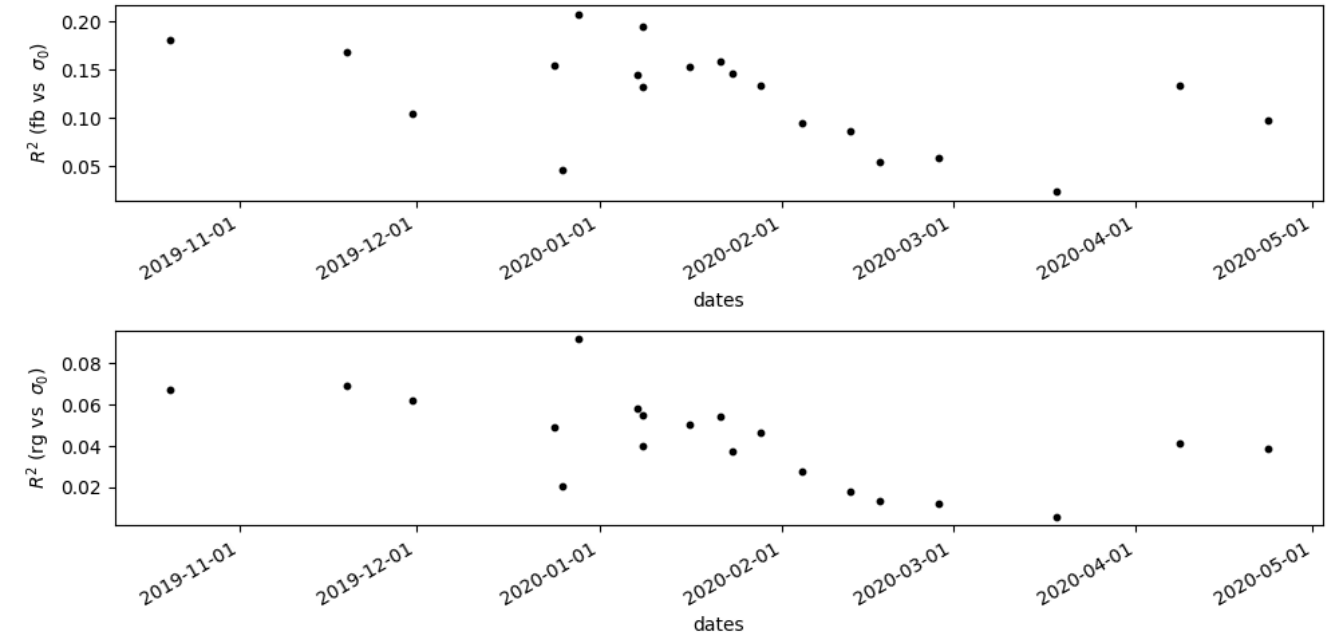


Figure 3. Evolution of correlations between freeboard or surface roughness and SAR backscatter over time. Note that the surface roughness is measured at the snow atmosphere interface and at correlation lenghts of 0.5 metres, whilst the SAR sensor is most sensitive to the ice snow interface and roughness at correlation lengths at the wavelength of the sensor, which is only 3.1 centimetres.

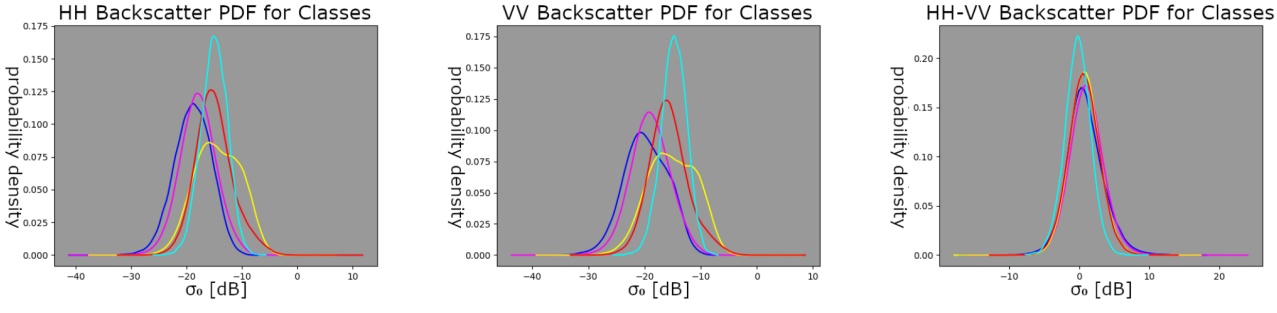


Figure 4. Approximate probability density functions for the sigma nought backscatter of each class across the different polarization configurations, for one flight on the 8th of April. Note that no two classes can be reliably separated using backscatter alone. Colours are as defined as in Fig. 2

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

to stop training. All subsets (test, training, validation) contain data from every scene. Imbalances of the classes were handled by balancing the dataset for pixel-wise classifiers and weighting the classes inversely to their frequency for the segmentation approaches, where an entire patch is segmented at once. Thus the training of the networks is set up so that performing equally well for each class yields the lowest loss.

In Figure 3 the correlation between backscatter and surface topography measurements is shown. It becomes evident immediately, that the backscatter characteristics alone are only very weakly correlated with the backscatter and thus separation using the backscatter alone would surely be futile. This is further underlined by looking at the backscatter distributions of the delineated classes from the flight on april 8th (fig. 4), where the correlations are relatively average in regards to all other flights. Here it is again obvious that the backscatter characteristics are not very valuable for class separation. Thus most of the information needed to classify accurately must come from contextual data.

2.2 Robustness

To test the robustness of each classifier, we will follow the same steps outlined in Kortum et al. (2022). In brief: Using the Polarstern as an origin, a 3km x 3km region around the ship is used as the robustness test set. This area has been identified in 162 TerraSAR-X SM scenes from different days. The robustness is then defined as the probability of each pixel being classified the same as in the previous and subsequent acquisitions (time between acquisitions is typically one day). Taking into account that the surface conditions are changing over time and that Polarstern was not perfectly stationary, this approximation of the robustness will serve only as a lower bound of the actual robustness of the classifier. In summary we are operating under the assumption that in a time period of two days, the percentage of ice that has changed class (e.g. through deformation) is significantly smaller than the percentage of ice that has remained in the same class. Note that this test is only sensible for the two solid ice classes and not for the OW/YI class, which is too dynamic on a daily timescale to be analysed in this manner. The robustness is first computed for the two classes and their average is used as an indicator for the network's robustness.

2.3 The Network Architectures

In this paper, we will compare five different architectures: two established image classifiers in the VGG16 developed by Simonyan and Zisserman (2015) (ice classification in e.g. Khaleghian et al. (2021b)) and the ConvNext network proposed by Liu et al. (2022) (an improvement over ResNet, used for SAR sea ice classification in e.g. Song et al. (2021)), a custom CNN (cCNN) pixel-wise classifier by Kortum et al. (2022) specifically designed for ice classification and two established segmentation models in the Unet by Ronneberger et al. (2015) (SAR sea ice classification in e.g. Nagi et al. (2021); Ren et al. (2022)) and Unet++ proposed in Zhou et al. (2018, 2019) (used in e.g. Murashkin and Frost (2021)). These first three (VGG16, ConvNext, cCNN) and last two (Unet, Unet++) models have one fundamental difference: Classification approaches (VGG16 etc.) are given a patch and are then asked to predict the class of the centre of the image. Segmentation approaches (Unet etc.) are tasked to produce a label for every pixel in the patch at the same time. For the Unet++ we chose to average over the features of the multiple output layers in the deep supervision part of the model. The exact specifications of all the models can be found in the appendix.

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

Model	Mean acc. [%]	Std. of acc. [%]	Mean KLD	Std. of KLD	Mean rb. [%]	Std. of rb. [%]
VGG16	40.52	7.83	0.8493	0.0458	79.95	5.82
ConvNext	45.12	3.17	0.872	0.0363	81.16	4.84
cCNN	47.89	3.74	0.7886	**0.0240**	68.52	18.81
Unet	**68.07**	**1.74**	**0.6032**	0.0406	**84.42**	1.78
Unet++	67.92	2.13	0.6249	0.0597	82.06	**1.36**

Table 1. Network performances on the independent test set after training. For brevity, we shortened accuracy to acc, standard deviation to std and robustness to rb. The means and standard deviations are computed from the 10 models in the population for every architecture. Best-in-category results are highlighted in bold font. Ten instances were trained for every model. The Unet and Unet++ architectures show significantly better performance than the rest.

2.4 Training

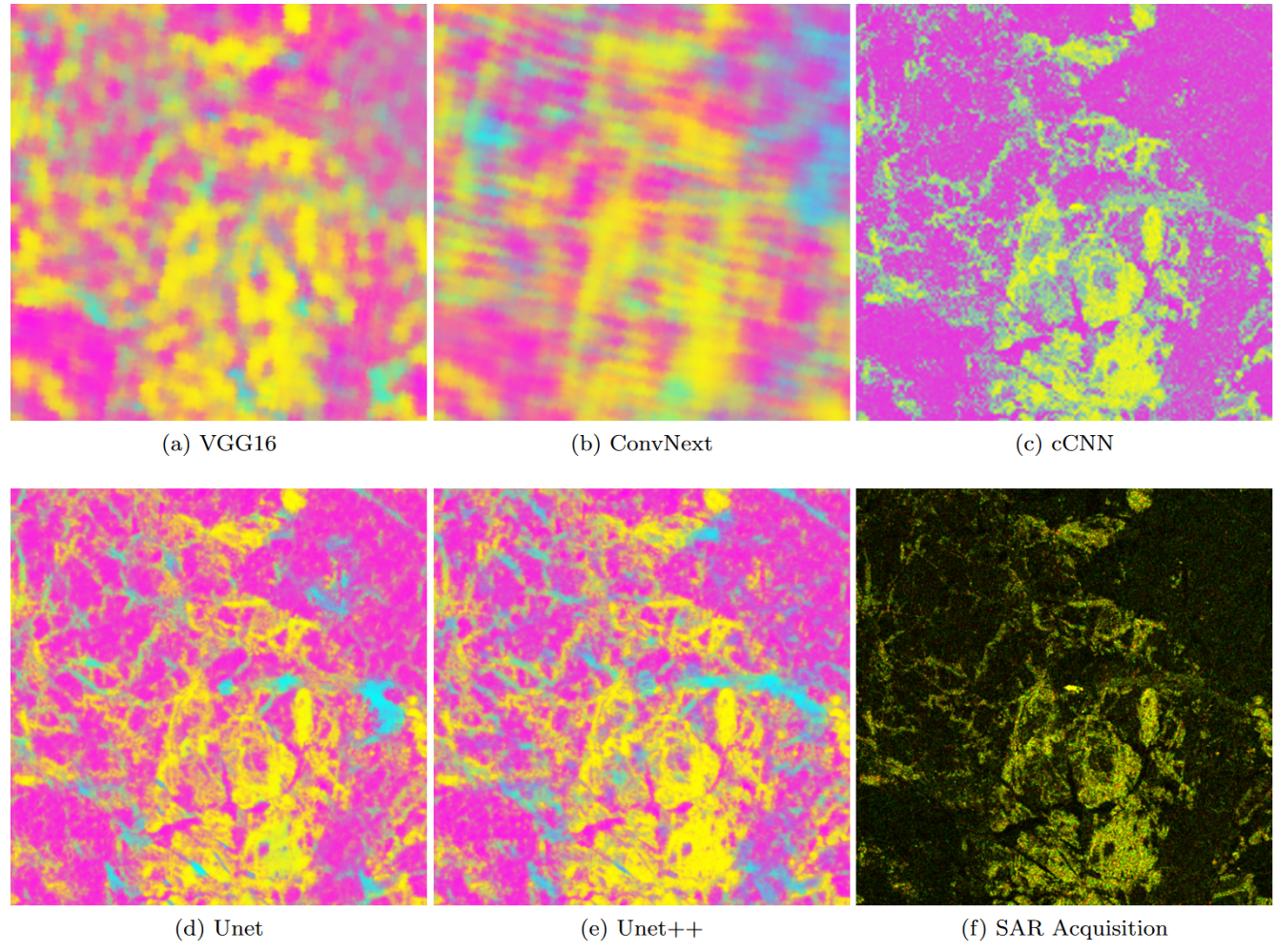
During training, the networks are tasked with minimising the Kullback Leibler Divergence (KLD) between the output and the label distributions. This allows us to fit the probabilities of each class occurring at each pixel, which we can infer from the ALS measurements. As this serves as a benchmark and comparison of these models concerning their applicability for sea ice retrieval, no further optimisations have taken place. For each of the model architectures, ten separate instances are trained. Training is stopped using the small independent validation set (5% of data). The model population allows some additional insight into the reliability of each architecture. The ingested SAR data are pre-processed by converting each band to sigma nought and then applying a logarithm. The incidence angle is provided in a third channel. The size of each patch to be classified is chosen to be 256x256 pixels, except for the cCNN which receives input patches at various scales (a 5x5, a 16x16 and a 64x64 pixel patch).

3 Results

The performance of different network architectures can be seen in table 1. They paint a clear picture of segmentation models' (Unet, Unet++) improvement over pixel-wise classification approaches. Of the pixel-wise classification approaches, the custom CNN classifier performed best, yet it was still significantly inferior to the segmentation models. We speculate that part of the reason for this is the high spatial resolution of the labels, as we get a label for every pixel from most of the ALS measurements. The pixel-wise classifiers cannot make use of any relationships between or spatial properties of labels, like shape, sparsity and correlations. This seems to be detrimental to their performance.

A more detailed analysis of the output of different models (Fig. 5) shows, how the VGG16 and ConvNext models struggle to relate all the information of the patch to only the classification of the central pixel, leading to a diffuse-looking classified scene. This seems most pronounced for the ConvNext model. A possible reason for this are the larger convolutional kernels

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023



(a) VGG16 (b) ConvNext (c) cCNN

(d) Unet (e) Unet++ (f) SAR Acquisition

Figure 5. Comparison of classifications from different models, randomly selected from the ten instances trained. Colours are the same as the classes discussed above, but the intensity is given by the predicted probabilities, so mixed colours can occur. This can be seen most easily in the cCNN classification (c). The scene was acquired over the Polarstern (center of the images) on January 14^{th}. The false colour composition consists of HH, VV and HH/VV channels, normalised with a $tanh$ function. The area shown is a 6 by 6 kilometre square.

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

(7x7 in contrast to 3x3) used in the architecture. The cCNN seems to struggle with using contextual data to separate rough ice and young ice. In general, the predicted probabilities at each pixel are higher in the non-dominant class, leading to a seemingly different colour palette in this visualisation. The Unet and Unet++ classifications are largely similar. Some difficulty in the separation of deformed and young ice signatures persists as can be seen in the mixing of yellow and cyan areas.

It is also worth pointing out that the very same cCNN and a VGG16 performed at accuracies around 85-95% on manual labels in Kortum et al. (2022), illustrating the difference between training and testing on quantitatively measured labels in contrast to human-generated annotations.

Whilst the mean KLD's are in accordance with the accuracies, the spread (std) of the KLD's across the model populations seems to be very similar across all models and there is no clear gap between segmentation and classification approaches. Overall, we cannot say that one model converges more reliably than another - as would be suggested by the accuracies alone. It is also apparent, that the cCNN does not perform well in the robustness scores on this dataset. This model is considerably smaller than the others (in terms of parameter count) and was heavily optimised using a different dataset, which seems to have come at a cost of flexibilty/generality of the architecture. The spread of robustnesses of the segmentation models seems to be considerably smaller that those of the generative models - additionally indicating these approaches are more reliable for ice classification from SAR.

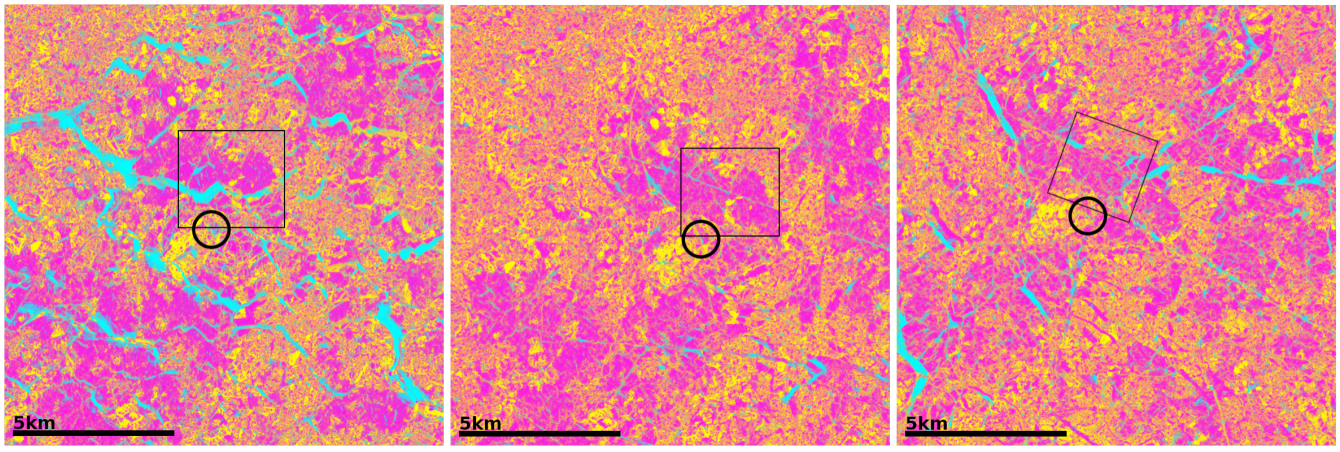
The classifications (e.g. in Fig. 6) show a very plausible set of results, that align with the observations of members on board the expedition. The fine labels at high resolution seem to have transcended into a similarly detailed classification map. The examples in Fig 7 also illustrate a general increase of deformation in the first year ice: The magenta FYI area close to Polarstern, marked by a black square in figure 6, is getting progressively more deformed as time progresses (detail in Fig. 7). The areas most prone to error seem to be the OW/YI classifications. This is to be expected as they are naturally the most sparse in the training data set. Additionally, they are very dynamic, which leads to extremely diverse backscatter properties that can be exhibited, in turn making them more difficult to classify.

We also observe decreasing correlation of backscatter and surface topography variables from the onset of the expedition until early April - particularly during January and February (3), where the MOSAiC expedition was met by numerous storms. Some of the decorrelation can be accounted for because of snow accumulation and redistribution, but it is difficult to quantify this phenomenon. However, since this trend is broken in April it seems that whatever drove this decorrelation is revertable and therefore changes in the snow are a more plausible explanation than ice deformation.

4 Discussion

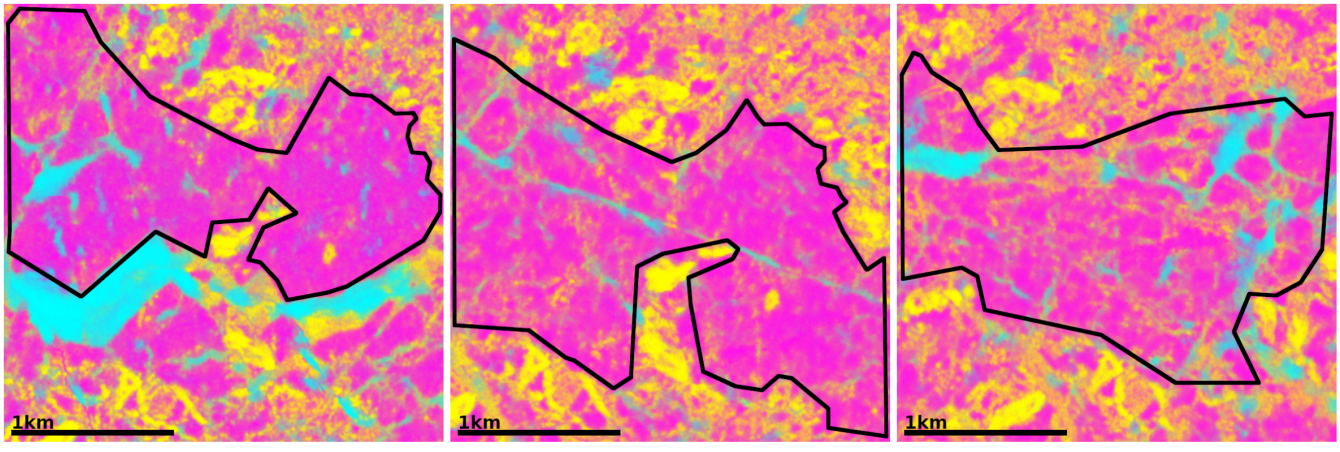
The top models in our investigation perform at around 68% accuracy on the test data set (Tab. 1). The segmentation models predictions are approximately 20% more accurate than the classification models. The only concrete difference between these models is that the segmentation approaches can learn from the distribution of labels, which appears to be highly important. Even the highest accuracies measured here are considerably lower than what many author's report for algorithms trained with

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023



(a) 22nd of November 20219 (b) 13th of January 2020 (c) 25th of March 2020

Figure 6. Collection of classified subscenes (Unet, pixel spacing = 3.5m) including the MOSAiC floe, after a storm (a), in calm conditions with some shearing indications (b) and with some breakup of the ice cover visible (c). The Polarstern location is indicated by the black circle. The DFYI/MYI class probability is displayed in yellow, the LFYI probability in magenta and the OW/YI probability is cyan. The black square marks the area shown at full resolution below (Fig. 7)



(a) 22nd of November 2019 - DFYI: 3.1% (b) 13th of January 2020 - DFYI: 5.6% (c) 25th of March 2020 - DFYI: 13.5%

Figure 7. Full resolution excerpt from the scenes show in Fig. 6. The classified images reflect the increased deformation of the first year ice area over time accurately, as the DFYI occurence rises. The DFYI fraction is computed inside of the black border. In the first scene some misclassification of the open lead (cyan) as older, deformed ice (yellow) is seen (outside of the area we are computing the DFYI fraction in) - this is a common issue in SAR sea ice retrieval as the backscatter can become very similar.

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

human-made labels. To understand these discrepancies, we will discuss the main differences between these measured labels and human annotation.

The measured labels used in this study have some underlying difficulties, because we do not know the snow depth and density, we do not know how strong the correlation of freeboard and ice-thickness is and cannot eliminate this error. Also the reflectance used to disseminate young ice and open water is based purely on the coverage of the surface being snow free and thus not directly correlated with ice age: if thin ice has formed the atmospheric conditions will dictate whether or not snow has gathered on top or if the bare ice is visible to the sensor. Thus the quality of labels could still be improved on, if more information were available.

To assess the impact of the individual thresholds (e.g. the location of the inflection point in the freeboard distribution) we also evaluated the top-performing unet architecture on the same dataset, but excluding points near the thresholds. To to this we did not consider labels, where the certainty of the most probable class was lower than 90%. For example regions with local standard deviation of approximately 3cm, that means that points within 6cm of the thresholds are not considered and the exact value of the thresholds have little bearing on the data considered. In case of the test dataset, these data points account for 24.1% of all data. Under these circumstances, the average accuracy of the unet model is 72.5% which is an increase of only 4.18% although 24.1% of the least certain labels where removed. Thus we can conclude that the exact location of the thresholds had only marginal impact on model performance, lending increased confidence that the model performances are representative of performance evaluated against ground truth.

For comparison with human annotation/ ice chart we must mention the resolution. In our case, every individual pixel gets its own class and there is no semantic grouping of pixels into the same class based on proximity or likeness. This is a stark contrast to ice charts, where the labels are made up of only few polygons per scene. Even when not training from such ice charts, humans generating training data for algorithms at high resolution generally limit themselves to areas which they can confidently identify. Not much can be said about the correctness of these labels per se, but one should keep in mind that in these instances, the accuracy achieved by the classifier is constrained to those easy-to-identify regions and are therefore not representative of the classifier's performance on the whole. Because of the size of SAR acquistions obtaining labels at pixel resolution from human annotation is not feasible. The great advantage that labels from measurements have is that they are truly indicative of performance on the entire scope of ice conditions in the scene (every pixel is labelled, thus there is no selection bias). Only by holding the testing of our high resolution retrieval algorithms to this standard can we show with certainty when an improved method of classification is developed, but of course to do so we are lacking available data sets.

In most data-driven approaches to classification, the performance of the classifier is limited by the quality of the labels. Therefore, one should be careful when using manually labelled data, such as ice charts, as ground truth. These practices are common in the current research - as not many other sources of labels are available. However, the potential is much greater than that. The great challenge of course remains, that high-resolution measurements are very sparse.

Because the MOSAiC mission provided us with an unmatched opportunity for training and testing algorithms with measured labels over a long time period, this study has made obvious that there is considerable room for improvement even with modern deep learning algorithms. It needs to be mentioned, that due to the spatial constraint to the area near the MOSAiC floe, the

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

training dataset does not capture the full extent of possible winter ice conditions in the Arctic, thus we cannot expect the classifier to perform equally well on a pan-Arctic scale. Instances of OW/YI are very sparse and their entire span of possible conditions and consequent radar response is not covered well by data. Since a better in-situ dataset is probably not going to emerge in the near future, it is clear that measured labels alone are not enough to train a stable algorithm that can deal with the full span of ice conditions. It seems that to achieve this, one would need to leverage a great number of scenes without labels. Semi-supervised and self-supervised approaches come to mind. Some first examples of their development exist for optical data by HAN et al. (2019), ice and open water discrimination from SAR in Li et al. (2015); Khaleghian et al. (2021a) and for sea ice classes from SAR in Imber (2022).

5 Conclusions

The MOSAiC expedition enabled the generation of a large dataset (ca. 20 million data points) of SAR acquisitions and appropriate labels delineated from in-situ laser scanning measurements. It has become clear that both the freeboard and the above snow surface roughness (at correlation lengths of 50 cm) are only weakly correlated with X-Band SAR backscatter, with average R^2 values of 0.124 and 0.043 respectively. We have shown that deep-learning segmentation approaches such as the Unet can approximate these labels from the SAR measurement at accuracies around 68%. We thus measured the performance of modern network architectures on a representative set of labels for the first time. From the performances of the different models, we can conclude that the segmentation approaches advantage of having access to the distribution of labels is crucial (20% accuracy) to the performance. It is notable that these label distributions at the scale of the measurement resolution are not contained in ice charts or human annotations, which suggests that classifying accurately at the resolution of the measurement when trained on human-annotated labels is improbable. As a more comprehensive dataset than created here is unlikely to be acquired in the near future, newly developed classifiers aiming at classification at the resolution of the sensor will need to find some way to gain access to the spatial ice type distributions to be successful.

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023


The Cryosphere Discussions

Appendix A: List of Helicopter Flights

1	20191020_01_PS122-1_2-167
2	20191119_01_PS122-1_8-23
3	20191130_01_PS122-1_9-98
4	20191224_01_PS122-2_17-98
5	20191225_01_PS122-2_17-99
6	20191228_01_PS122-2_17-101
7	20200107_01_PS122-2_19-44
8	20200108_01_PS122-2_19-46
9	20200108_03_PS122-2_19-52
10	20200116_01_PS122-2_20-52
11	20200121_01_PS122-2_21-41
12	20200123_02_PS122-2_21-78
13	20200128_01_PS122-2_22-16
14	20200204_01_PS122-2_23-14
15	20200212_01_PS122-2_24-31
16	20200217_02_PS122-2_25-8
17	20200227_01_PS122-3_29-49
18	20200318_01_PS122-3_32-42
19	20200408_01_PS122-3_35-49
20	20200423_01_PS122-3_37-63

Table A1. List of the 20 helicopter flights used in this research. Data is published in Hutter et al. (2022a).

Appendix B: Network Architectures

We briefly present the network architectures used in this investigation. We make use of the following conventions to keep the figures concise. FCX is short for a fully connected layer with X neurons. ConvX x Y denotes a 2D convolutional layer with filter sizes X and number of filters Y. Unless otherwise specified the convolutional layers have stride 1. If a layer has multiple inputs, they are concatenated before being parsed to the layer.

Author contributions. Karl Kortum - Conceptualization, Formal Analysis, Investigation, Methodology, Writing - original draft.
Suman Singha - Conceptualization, Data curation, Funding acquisition, Project administration, Supervision, Writing - review and editing.
Gunnar Spreen - Funding acqusition, Supervision, Writing - review and editing.
Nils Hutter - Data curation, Writing - review and editing.
Arttu Jutila - Data curation, Writing - review and editing.
Christian Haas - Funding acqusition, Supervision, Writing - review and editing.

Competing interests. Some authors are members of the editorial board of The Cryosphere. The peer-review process was guided by an independent editor, and the authors have also no other competing interests to declare.

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

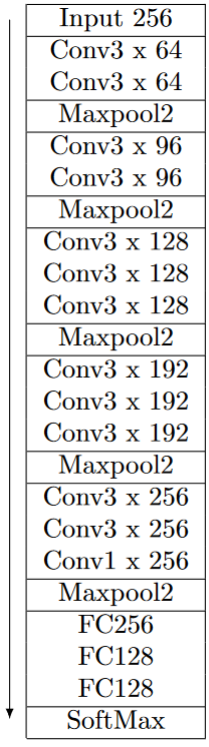


Table B1. VGG16 architecture as used in the paper. Published in Simonyan and Zisserman (2015). The ReLU activation is used throughout the network. The padding is set to 'same'.

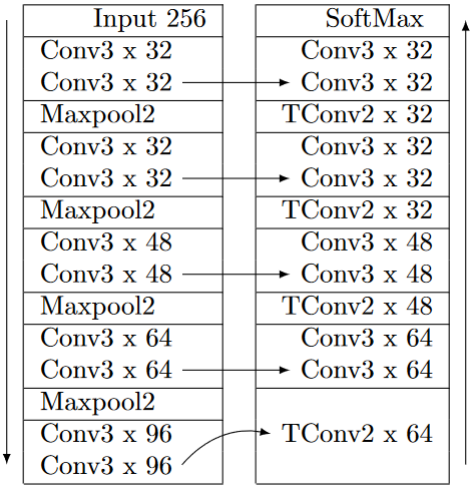


Table B2. The Unet architecture as used in this paper and published in Ronneberger et al. (2015). The ReLU activation is used throughout the network and the padding is set to 'same' where applicable.

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

Input 256
Conv4 (stride 4) x 96
ConvNx x 96 ConvNx x 96 ConvNx x 96
Conv2 (st 2) x 96
ConvNx x 192 ConvNx x 192 ConvNx x 192
Conv2 (st 2) x 192
ConvNx x 384 ConvNx x 384 ConvNx x 384 ConvNx x 384 ConvNx x 384 ConvNx x 384 ConvNx x 384 ConvNx x 384 ConvNx x 384
Conv2 (st 2) x 384
ConvNx x 768 ConvNx x 768 ConvNx x 768
GlobalAvgPool2D LayerNorm
SoftMax

where:

ConvNx x n =

Layer	Activation
Input	
Conv7 x n	Layer Norm
Conv1 x (4 n)	GELU
Conv1 x n	
Output	

Table B3. The ConvNext-T architecture used in this paper. Developed in Liu et al. (2022).

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

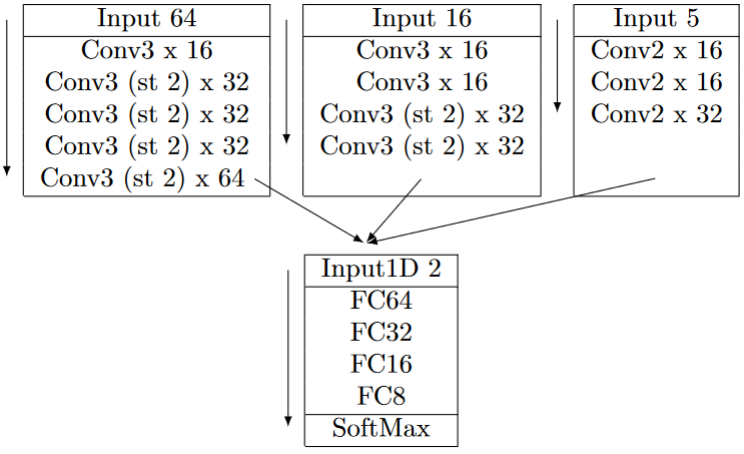


Table B4. The custom CNN architecture from Kortum et al. (2022) used in this paper. The inputs at different scales are flattened and concatenated before being output to the fully connected layers. Leaky ReLU is used for activation and padding is set to 'valid'. The 16x16 pixel input is downscaled from the original scene by factor 5 and the 64x64 pixel input is a square cutout that is rescaled so that the width of the entire scene is 64 pixels. The 1D input contains the relative coordinates of the pixel in the 64x64 pixel input.

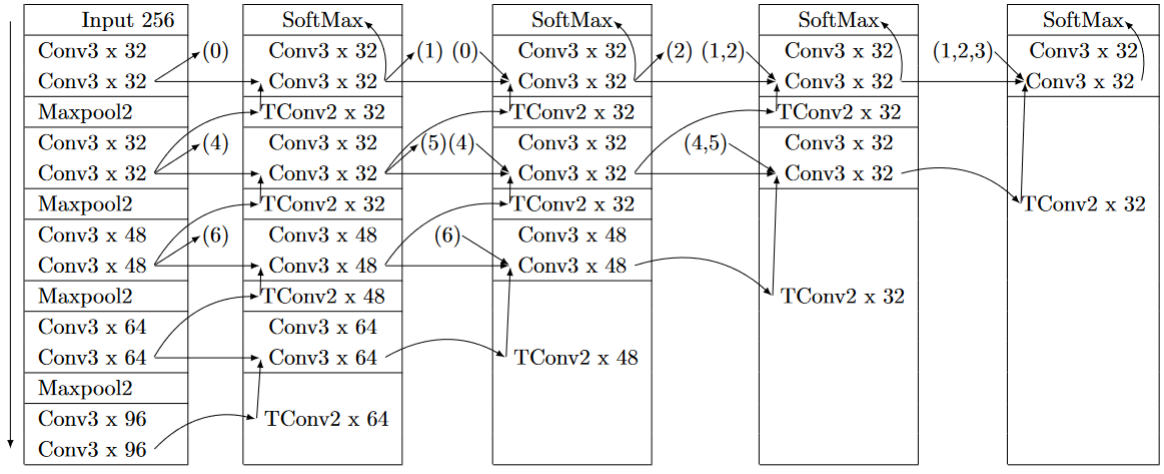


Table B5. Unet++ architecture used in this paper, published in Zhou et al. (2018, 2019). Note that the left column is identical to the downwards convolution side of the regular Unet and the lowest rows from left to right form the upwards side of the Unet. The Unet++ then uses extra layers in between to extend the architecture. All layers within a cell are considered to be a block, so they are all executed before parsing the output to the next block. All layers marked 'Softmax' are averaged before the final linear layer and the softmax are applied. ReLU is used as the activation function throughout and the padding is set to 'same'.

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

275 *Acknowledgements.* This study was funded by Deutsche Forschungsgemeinschaft (DFG) under project name 'MOSAiCmicrowaveRS' (SI 2564/1-1 and SP 1128/8-1).

Data used in this manuscript was produced as part of the international Multidisciplinary drifting Observatory for the Study of the Arctic Climate (MOSAiC) with the tag MOSAiC20192020 and Project_ID: AWI_PS122_00..

We thank all persons involved in the expedition of the Research Vessel Polarstern during MOSAiC in 2019-2020 as listed in Nixdorf et al.
280 (2021) Nixdorf et al. (2021).

TerraSAR-X images used in this study were acquired using the TerraSAR-X AO OCE3562_4 (PI: Suman Singha).

German Federal Ministry of Education and Research (BMBF) project IceSense—Remote Sensing of the Seasonal Evolution of Climate-relevant Sea Ice Properties (03F0866A).

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

References

Boulze, H., Korosov, A., and Brajard, J.: Classification of Sea Ice Types in Sentinel-1 SAR Data Using Convolutional Neural Networks, Remote Sensing, 12, https://doi.org/10.3390/rs12132165, 2020.

Doulgeris, A. P.: An Automatic $\mathcal{U}$-Distribution and Markov Random Field Segmentation Algorithm for PalSAR Images, IEEE Transactions on Geoscience and Remote Sensing, 53, 1819–1827, https://doi.org/10.1109/TGRS.2014.2349575, 2015.

Fily, M. and Rothrock, D. A.: Extracting Sea Ice Data from Satellite SAR Imagery, IEEE Transactions on Geoscience and Remote Sensing, GE-24, 849–854, https://doi.org/10.1109/TGRS.1986.289699, 1986.

Geldsetzer, T. and Yackel, J. J.: Sea ice type and open water discrimination using dual co-polarized C-band SAR, Canadian Journal of Remote Sensing, 35, 73–84, https://doi.org/10.5589/m08-075, 2009.

HAN, Y., ZHAO, Y., ZHANG, Y., WANG, J., YANG, S., HONG, Z., and CAO, S.: A Cooperative Framework Based on Active and Semi-supervised Learning for Sea Ice Classification using EO-1 Hyperion Data, TRANSACTIONS OF THE JAPAN SOCIETY FOR AERONAUTICAL AND SPACE SCIENCES, 62, 318–330, https://doi.org/10.2322/tjsass.62.318, 2019.

Hara, Y., Atkins, R., Shin, R., Kong, J. A., Yueh, S., and Kwok, R.: Application of neural networks for sea ice classification in polarimetric SAR images, IEEE Transactions on Geoscience and Remote Sensing, 33, 740–748, https://doi.org/10.1109/36.387589, 1995.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, https://doi.org/10.48550/ARXIV.1512.03385, 2015.

Hendricks, S.: Ice Drift - Transformation of GPS positions into a translating and rotating coordinate reference system, https://gitlab.awi.de/floenavi-crs/icedrift, 2019.

Hutter, N., Hendricks, S., Jutila, A., Birnbaum, G., von Albedyll, L., Ricker, R., and Haas, C.: Merged Grids of Sea-Ice or snow freeboard from helicopter-borne laser scanner during the MOSAiC Expedition, version 1., Pangea, https://doi.org/https://doi.pangaea.de/10.1594/PANGAEA.950896, 2022a.

Hutter, N., Hendricks, S., Jutila, A., Birnbaum, G., von Albedyll, L., Ricker, R., and Haas, C.: Digital elevation models of the sea-ice surface from airborne laser scanning during MOSAiC., Scientific Data (in review), 2022b.

Imber, J.: Generative Network For Semi-supervised Sea Ice Classification, https://doi.org/10.36227/techrxiv.21081136.v1, 2022.

Itkin, P., Webster, M., Hendricks, S., Oggier, M., Jaggi, M., Ricker, R., Arndt, S., Divine, D. V., von Albedyll, L., Raphael, I., Rohde, J., and Liston, G. E.: Magnaprobe snow and melt pond depth measurements from the 2019-2020 MOSAiC expedition, https://doi.org/10.1594/PANGAEA.937781, 2021.

JCOMM: Sea-Ice Information Services in the World, World Meteorological Organization, https://doi.org/10.25607/OBP-1325, 2017.

Johansson, A. M., Malnes, E., Gerland, S., Cristea, A., Doulgeris, A., Divine, D., Pavlova, O., and Lauknes, T. R.: Consistent ice and open water classification combining historical synthetic aperture radar satellite images from ERS-1/2, Envisat ASAR, RADARSAT-2 and Sentinel-1A/B, Annals of Glaciology, 61, 1–11, https://doi.org/10.1017/aog.2019.52, 2020.

Karvonen, J.: Baltic Sea ice SAR segmentation and classification using modified pulse-coupled neural networks, IEEE Transactions on Geoscience and Remote Sensing, 42, 1566–1574, https://doi.org/10.1109/TGRS.2004.828179, 2004.

Khaleghian, S., Ullah, H., Kræmer, T., Eltoft, T., and Marinoni, A.: Deep Semisupervised Teacher–Student Model Based on Label Propagation for Sea Ice Classification, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 10 761–10 772, https://doi.org/10.1109/JSTARS.2021.3119485, 2021a.

Khaleghian, S., Ullah, H., Kræmer, T., Hughes, N., Eltoft, T., and Marinoni, A.: Sea Ice Classification of SAR Imagery Based on Convolution Neural Networks, Remote Sensing, 13, https://www.mdpi.com/2072-4292/13/9/1734, 2021b.

https://doi.org/10.5194/tc-2023-72
Preprint. Discussion started: 3 July 2023

Kortum, K., Singha, S., Spreen, G., and Hendricks, S.: Automating Sea Ice Characterisation from X-Band SAR with Co-Located Airborne Laser Scanner Data Obtained During The MOSAiC Expedition, in: Geoscience and Remote Sensing Symposium, 2021 IEEE International, 2021.

Kortum, K., Singha, S., and Spreen, G.: Robust Multiseasonal Ice Classification From High-Resolution X-Band SAR, IEEE Transactions on Geoscience and Remote Sensing, 60, 1–12, https://doi.org/10.1109/TGRS.2022.3144731, 2022.

Kwok, R., Rignot, E., Holt, B., and Onstott, R.: Identification of sea ice types in spaceborne synthetic aperture radar data, Journal of Geophysical Research: Oceans, 97, 2391–2402, https://doi.org/https://doi.org/10.1029/91JC02652, 1992.

Li, F., Clausi, D. A., Wang, L., and Xu, L.: A semi-supervised approach for ice-water classification using dual-polarization SAR satellite imagery, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 28–35, https://doi.org/10.1109/CVPRW.2015.7301380, 2015.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S.: A ConvNet for the 2020s, https://doi.org/10.48550/ARXIV.2201.03545, 2022.

Lohse, J., Doulgeris, A., and Dierking, W.: Incident Angle Dependence of Sentinel-1 Texture Features for Sea Ice Classification, Remote Sensing, 13, https://doi.org/10.3390/rs13040552, 2021.

Murashkin, D. and Frost, A.: Arctic Sea ICE Mapping Using Sentinel-1 SAR Scenes with a Convolutional Neural Network, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, pp. 5660–5663, https://doi.org/10.1109/IGARSS47720.2021.9553206, 2021.

Murashkin, D., Spreen, G., Huntemann, M., and Dierking, W.: Method for detection of leads from Sentinel-1 SAR images, Annals of Glaciology, 59, 1–13, https://doi.org/10.1017/aog.2018.6, 2018.

Nagi, A. S., Kumar, D., Sola, D., and Scott, K. A.: RUF: Effective Sea Ice Floe Segmentation Using End-to-End RES-UNET-CRF with Dual Loss, Remote Sensing, 13, https://doi.org/10.3390/rs13132460, 2021.

Nicolaus, M., Perovich, D. K., Spreen, G., Granskog, M. A., von Albedyll, L., Angelopoulos, M., Anhaus, P., Arndt, S., Belter, H. J., Bessonov, V., Birnbaum, G., Brauchle, J., Calmer, R., Cardellach, E., Cheng, B., Clemens-Sewall, D., Dadic, R., Damm, E., de Boer, G., Demir, O., Dethloff, K., Divine, D. V., Fong, A. A., Fons, S., Frey, M. M., Fuchs, N., Gabarró, C., Gerland, S., Goessling, H. F., Gradinger, R., Haapala, J., Haas, C., Hamilton, J., Hannula, H.-R., Hendricks, S., Herber, A., Heuzé, C., Hoppmann, M., Høyland, K. V., Huntemann, M., Hutchings, J. K., Hwang, B., Itkin, P., Jacobi, H.-W., Jaggi, M., Jutila, A., Kaleschke, L., Katlein, C., Kolabutin, N., Krampe, D., Kristensen, S. S., Krumpen, T., Kurtz, N., Lampert, A., Lange, B. A., Lei, R., Light, B., Linhardt, F., Liston, G. E., Loose, B., Macfarlane, A. R., Mahmud, M., Matero, I. O., Maus, S., Morgenstern, A., Naderpour, R., Nandan, V., Niubom, A., Oggier, M., Oppelt, N., Pätzold, F., Perron, C., Petrovsky, T., Pirazzini, R., Polashenski, C., Rabe, B., Raphael, I. A., Regnery, J., Rex, M., Ricker, R., Riemann-Campe, K., Rinke, A., Rohde, J., Salganik, E., Scharien, R. K., Schiller, M., Schneebeli, M., Semmling, M., Shimanchuk, E., Shupe, M. D., Smith, M. M., Smolyanitsky, V., Sokolov, V., Stanton, T., Stroeve, J., Thielke, L., Timofeeva, A., Tonboe, R. T., Tavri, A., Tsamados, M., Wagner, D. N., Watkins, D., Webster, M., and Wendisch, M.: Overview of the MOSAiC expedition: Snow and sea ice, Elementa: Science of the Anthropocene, 10, https://doi.org/10.1525/elementa.2021.000046, 2022.

Nixdorf, U., Dethloff, K., Rex, M., Shupe, M., Sommerfeld, A., Perovich, D. K., Nicolaus, M., Heuzé, C., Rabe, B., Loose, B., Damm, E., Gradinger, R., Fong, A., Maslowski, W., Rinke, A., Kwok, R., Spreen, G., Wendisch, M., Herber, A., Hirsekorn, M., Mohaupt, V., Frickenhaus, S., Immerz, A., Weiss-Tuider, K., König, B., Mengedoht, D., Regnery, J., Gerchow, P., Ransby, D., Krumpen, T., Morgenstern, A., Haas, C., Kanzow, T., Rack, F. R., Saitzev, V., Sokolov, V., Makarov, A., Schwarze, S., Wunderlich, T., Wurr, K., and Boetius, A.: MOSAiC Extended Acknowledgement, https://doi.org/10.5281/zenodo.5541624, 2021.

Ren, Y., Li, X., Yang, X., and Xu, H.: Development of a Dual-Attention U-Net Model for Sea Ice and Open Water Classification on SAR Images, IEEE Geoscience and Remote Sensing Letters, 19, 1–5, https://doi.org/10.1109/LGRS.2021.3058049, 2022.

Ressel, R., Frost, A., and Lehner, S.: A Neural Network-Based Classification for Sea Ice Types on X-Band SAR Images, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8, 1–9, https://doi.org/10.1109/JSTARS.2015.2436993, 2015.

Ressel, R., Singha, S., Lehner, S., Rösel, A., and Spreen, G.: Investigation into Different Polarimetric Features for Sea Ice Classification Using X-Band Synthetic Aperture Radar, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9, 3131–3143, https://doi.org/10.1109/JSTARS.2016.2539501, 2016.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, https://doi.org/10.48550/ARXIV.1505.04597, 2015.

Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, https://arxiv.org/abs/1409.1556, 2015.

Singha, S., Johansson, M., Hughes, N., Hvidegaard, S. M., and Skourup, H.: Arctic Sea Ice Characterization Using Spaceborne Fully Polarimetric L-, C-, and X-Band SAR With Validation by Airborne Measurements, IEEE Transactions on Geoscience and Remote Sensing, 56, 3715–3734, https://doi.org/10.1109/TGRS.2018.2809504, 2018.

Soh, L.-K. and Tsatsoulis, C.: Texture Analysis of SAR Sea Ice Imagery using Gray Level Co-occurrence Matrices, IEEE Transactions on Geoscience and Remote Sensing, 37, 780 – 795, https://doi.org/10.1109/36.752194, 1999.

Song, W., Li, M., Gao, W., Huang, D., Ma, Z., Liotta, A., and Perra, C.: Automatic Sea-Ice Classification of SAR Images Based on Spatial and Temporal Features Learning, IEEE Transactions on Geoscience and Remote Sensing, 59, 9887–9901, https://doi.org/10.1109/TGRS.2020.3049031, 2021.

Ullah, H., Khaleghian, S., Kromer, T., Eltoft, T., and Marinoni, A.: A Noise-Aware Deep Learning Model for Sea Ice Classification Based on Sentinel-1 Sar Imagery, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, pp. 816–819, https://doi.org/10.1109/IGARSS47720.2021.9553971, 2021.

Wang, Y.-R. and Li, X.-M.: Arctic sea ice cover data from spaceborne synthetic aperture radar by deep learning, Earth System Science Data, 13, 2723–2742, https://doi.org/10.5194/essd-13-2723-2021, 2021.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J.: Unet++: A Nested U-Net Architecture for Medical Image Segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11, Springer, 2018.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J.: UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation, IEEE Transactions on Medical Imaging, 2019.