Anonymous Referee #1

General comments

The manuscript by Park et al presents a machine learning technique to provide operational sea ice charts. The use of machine learning is interesting from the operational side as it enables processing of large volumes of data in a consistent manner. Such sea ice charts are of significant use for the both scientific community as well as the general community.
The manuscript is well structured, though would benefit from a spell and grammar check, e.g. Sentinel is misspelled several times. The authors claim that the method is free from subjective judgements though it requires the input of manually derived open water vs sea ice charts.

First of all, we would like to thank the reviewer for the positive evaluation and providing important comments. We hope that all your concerns will be cleared after reading our responses and modifications made to the manuscript. Please find below our answers (in green) and modifications (deleted in red and added in blue) to your comments/suggestions/questions.

Specific comments

The authors claim that by using operationally provide sea ice charts they are avoiding subjective decisions in the sea ice classification training data. To my knowledge the sea ice charts provided by the US National Ice Center are based on exactly such subjective decisions and are made to the best ability of the excellent sea ice experts working there. Please clarify how the training and validation data used here are not subjected to such decisions. The manually derived open water vs sea ice maps used as input training data could also be viewed as subjectively derived data.

As the reviewer pointed, ice charting cannot be free from subjective decisions but it requires best knowledges from credible experts like national/international ice services rather than from anonymous individuals. The aim of this study is to address such an issue; however, we admit that the manual selection of training/validation dataset causes another subjectivity issue. Regarding the concern on subjective judgement in selecting training dataset, the open water vs sea ice charts are not derived manually; the boundary between open water and sea ice is extracted from the electronic ice chart. The manual part is to check if the SAR image show high image contrast at the position of the extracted ice-water boundary. Although the visual inspection along the ice-water boundary overlaid on SAR image is also not completely free from subjective judgements, it requires much less effort and expertise compared to drawing manual ice chart from the SAR image itself. We apologize for this confusion; we omitted an essential description. In the revised manuscript, we added detailed explanations about the procedure on selecting training dataset.

[Section 2.2.6]
In order to automate image selection, the ice edges in SAR images need to be identified first. ~~To automate image selection for training, a good ice/water classifier for SAR image is needed.~~ Since even ~~such~~ an ~~simple binary~~ ice/water classifier has not been well developed yet for Sentinel-1, the image selection procedure has to be done manually in the beginning. However, once a classifier is generated with high accuracy, it can be used to automate the procedure, then the whole process in the proposed scheme will be fully automated. This is why the proposed algorithm is named "semi-" automated for now. Nevertheless, the manual selection is done by visual inspection of ice-water boundaries overlaid on SAR images. The ice-water boundary can be extracted easily from the reprojected ice chart by selecting the pixel borders of open water class. Then the SAR backscattering image contrasts across the ice-water boundaries are examined both in HH- and HV-polarization because the image contrast between ice-water is larger in HV and smooth level ice is better recognizable in HH.

It is unclear if only the sea ice parts of the images were incidence angle corrected using the sea ice estimated slope or if the whole image was corrected using the slopes derived for the sea ice part of the images. Please

clarify. If the incidence angle slope derived for the sea ice were used over the entire image how would this affect the open water areas of the image? How are the slopes presented here derived?

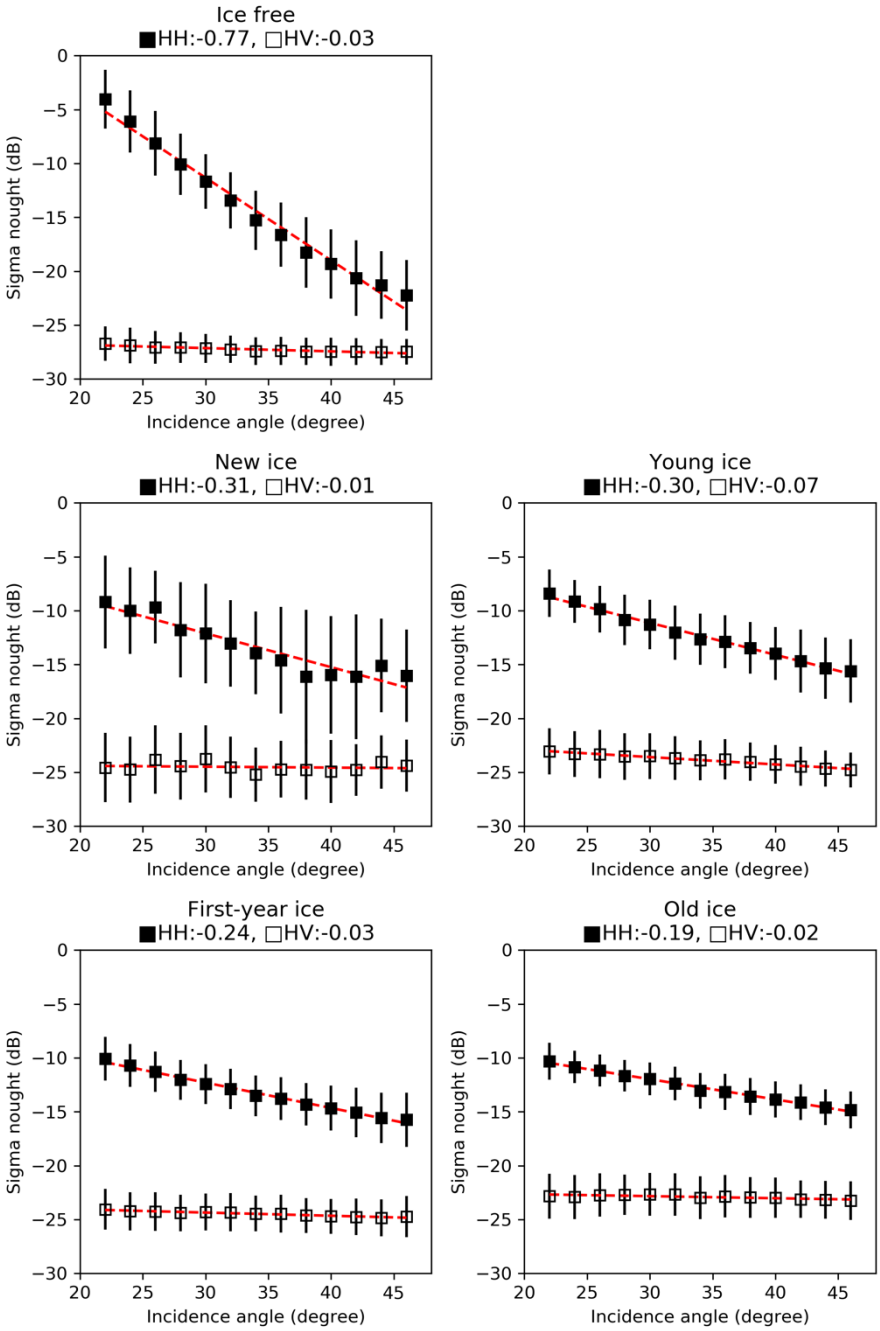
The whole image was corrected using the slopes derived for the sea ice regardless of the surface type. Since this correction is prior to the classification, sea ice type-specific correction cannot be made in this stage. Nevertheless, the bulk slope correction has known to be effective in literature (Zakhvatkina et al., 2013; Zakhvatkina et al., 2017). We added the following explanations and references to Section 2.2.3:

[Section 2.2.3]
Figure 4. shows two-dimensional histograms of incidence angle versus sigma nought for sea ice pixels in HH and HV polarization channels from Sentinel-1 data collected over sea ice and open water in the study area in winter 2018. From the Sentinel-1 dataset described in Section 2.1, sea ice pixels were extracted by using daily global sea ice edge products available from the EUMETSAT Ocean and Sea Ice Satellite Application Facilities (OSISAF). For HH polarization, the estimated slope was -0.200 dB/degree, which is slightly different from the estimation of the first-year ice (-0.24 dB/degree) in Mäkynen and Karvonen (2017) and in between the estimations for first-year ice (-0.22 dB/degree) and multi-year ice (-0.16 dB/degree) in Mahmud et al. (2018). For HV polarization, the estimated slope was only -0.025 dB/degree, which is much lower than the estimation in Mäkynen and Karvonen (2017), however, it is in line with the estimations from RADARSAT-2 (Leigh et al., 2014; Liu et al., 2015). We compensate for the incidence angle ~~dependency~~ dependence using the estimated slopes, ~~referencing~~ with respect to the nominal scene center angle of 34.5 degrees as reference. Although the incidence angle dependence changes with ice type and radar frequency (Mahmud et al., 2018), the compensation is done for all pixels in the image using a single value of mean slope because the ice types are not identified in this stage. Open water areas of the image are also affected; however, the correction is also beneficial since the incidence angle dependence for open water is stronger (-0.65 dB/degree for wind velocity of 5 m/s, computed from CMOD5 C-band geophysical model function in Hersbach et al., 2007), thus the corrected image has less incidence angle dependence.

[References]
Mahmud, M. S., Geldsetzer, T., Howell, S. E. L., Yackel, J. J., Nandan, V., and Scharien, R. K.: Incidence angle dependence of HH-polarized C- and L-band wintertime backscatter over Arctic sea ice, IEEE T. Geosci. Remote, 56(11), 6686-6698, doi:10.1109/TGRS.2018.2841343, 2018.
Hersbach, H., Stoffelen, A., and de Haan, S.: An improved C-band scatterometer ocean geophysical model function: CMOD5, J. Geophys. Res., 112, C03006, doi:10.1029/2006JC003743, 2007.

Mahmud et al 2018 showed that different sea ice types have different incidence angle dependent slopes. Have you considered using a sea ice type dependent slope factor? Moreover, how does the work by Mahmud et al 2018 fit in with the incidence angle dependencies presented here?

Typically, the slope for open water is higher than that for sea ice, thus the correction works in a way reducing the difference in sigma naught for open water as well. As the review pointed, different sea ice types have different incidence angle dependent slopes; however again, ice type-specific correction prior to ice type classification is controversial. Although estimating ice type dependent slope is not a part of this manuscript, we provide the values derived from the training/validation dataset for the review purpose only.



How do you define “good match” (P7 R3)? Temporal overlap? Spatial overlap? How was this manual selection of images performed? How were the open water vs sea ice charts, that are used as initial input into the classifier, derived?

Both temporal and spatial overlaps are important. Since the SAR image itself is potentially one of the sources for ice charting at the ice services, some images spatially match well with the shape of the polygons in ice chart. Temporal window length of 3 days from the publication date of ice chart was used for squeezing the number of images to make decisions of use/discard for training. We added a paragraph that explains processing details in the revised manuscript as follows:

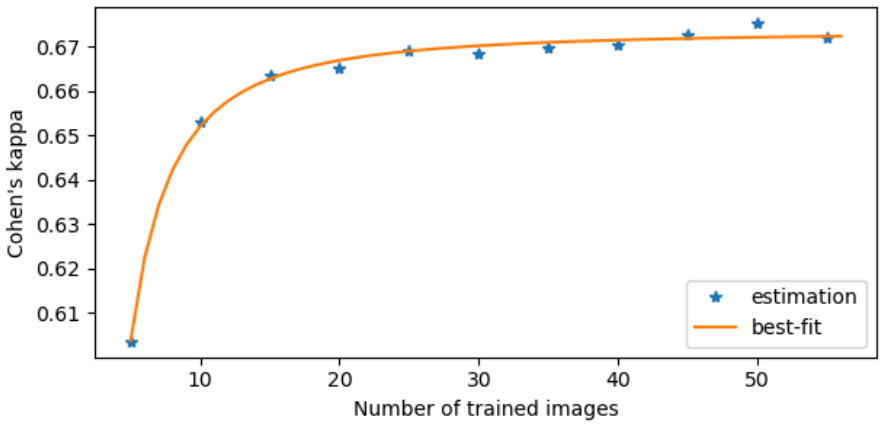
[Section 2.2.6]
To train an ice type classifier, a set of collocated SAR images and ice charts is required. After the preprocessing of the ice chart including reprojection into the SAR image geometry, only the samples with spatially and temporally good matches should be fed to the training phase. Image selection is trivial, but not easy to automate. Since the weekly ice chart is made partly based on the SAR images acquired in the past three3 days from the date of publication, the ice edges in some images match well with those in the ice chart. In order to automate image selection, the ice edges in SAR images need to be identified first. ~~To automate image selection for training, a good ice/water classifier for SAR image is needed.~~ Since even ~~such~~ an ~~simple binary~~ ice/water classifier has not been well developed yet for Sentinel-1, the image selection procedure has to be done manually in the beginning. However, once a classifier is generated with high accuracy, it can be used to automate the procedure, then the whole process in the proposed scheme will be fully automated. This is why the proposed algorithm is named "semi-" automated for now. Nevertheless, the manual selection is done by visual inspection of ice-water boundaries overlaid on SAR images. The ice-water boundary can be extracted easily from the reprojected ice chart by selecting the pixel borders of open water class. Then the SAR backscattering image contrasts across the ice-water boundaries are examined both in HH- and HV-polarization because the image contrast between ice-water is larger in HV and smooth level ice is better recognizable in HH.
After the image selection, the samples in the selected images are split randomly into training and test datasets with a ratio of 7:3. For the training dataset, further data selection is made by excluding the samples residing close to the polygon boundaries. This is to account for possible mismatch due to various reasons (e.g., ice drift, vector mapping error, image geocoding error, etc.). In this study, only the data from pixels more than 3 km away from the polygon boundaries was fed into the training process. Once the hyperparameter optimization is done, the RF classifier is trained for the training dataset. The trained classifier is then applied to the test dataset. For performance evaluation, we~~We~~ use confusion matrix and Cohen's kappa coefficient κ (Cohen, 1960), which measures the agreement between two raters (in this study, they are the trained classifier and the reference ice chart) with taking account of the possibility of the agreement occurring by chance ~~for performance evaluation~~. The validation is done in the same way but using a completely independent dataset. The 2018 data was used to run the training phase. Among 958 images in total, we selected 57 images of which ice edges match well with the collocated ice chart. From the selected images, 6.4 million samples covering open water and sea ice were divided into training and test dataset.

Are the open water areas separated from the sea ice areas at this stage of the classification process? Or are they classified at the same time as the sea ice types?

At this stage, no classifier is introduced. The open water is classified at the same time as the sea ice types in the later stage.

Given that one possible reason for the low accuracy in 2019 were stated to be insufficient training data (P9 R9-11), have you tested how the classification improves/remains the same if additional training data is added? Such an assessment would add strength to the accuracy of the method presented here.

We tested the changes in classification accuracy with varying number of trained images. As the results shown below indicates, the accuracy increases rapidly until adding 15 images, but the improvement was not great after adding 20-25 images.



By following the test result, we revised the manuscript as below.

[Section 3]
However, the accuracy decrease from 2018 data to 2019 data was ~~in~~ at a similar level to the case of the five~~5~~-class classification.~~, and t~~This could have been caused by ~~insufficient training of classifier and/or~~ inconsistent labeling in the reference ice chart.

Daily ice charts using Sentinel-1 data covering at least part of the study areas used here are provided by the Norway Ice Service. A comparison with their ice charts when they overlap, spatially and temporally, would have been beneficial and added strength to the accuracy assessment. Partially as it would have provided daily instead of weekly ice charts to compare to.

As far as we understand, the daily ice chart serviced by Met.no does not provide ice types; it provides ice concentration instead.

Would your method work also outside the winter season? Has it been tested for other seasons? A majority of the shipping industry is dependent on sea ice charts year-round and a consistent method employed year-round is therefore beneficial.

The method itself would work for other seasons if multiple classifiers are trained for each of the seasons. It is challenging to develop a single universal classifier that works for all seasons. Although we tried to make a classifier adapt to seasonal changes by including day of year as a feature (FC3), the result was not promising.

Day of year might not correspond to the same temperature, fluxes and weather regimes. Have you considered using a weather variant input parameter instead of day of year? Such a parameter might be more suitable to capture the seasonality within the scenes.

No, we didn't. As sea ice drifts continuously, weather variant information for each of the ice floes in the SAR image at the image acquisition time needs to be calculated and joined, and this may require complex and rigorous works. In order to simplify the problem, we used day of year only, but certainly other parameters need to be tested in the follow up paper. We are very interested in combining sea ice drift with ice type-specific texture changes in our future study. We added a discussion paragraph to the end of Section 3.

[Section 3]
Unfortunately, the proposed algorithm has several limitations. First of all, the variations in radar backscattering and its corresponding image textures due to seasonal changes were not properly captured. Although day-of-the-year was tested as a seasonality variable in the FC3 feature configuration, the result did not show any improvement. This is because day-of-the-year might not correspond to the same temperature, fluxes, and weather regimes.

How is the general accuracy derived? Is it a normalised average and does it account for the varying amounts of the different sea ice types? The overall accuracy of the sea ice classifier is only provided in the abstract and the conclusion. Please also provide it with the general results and discussion.

The overall accuracy in the abstract and conclusion is an average of the accuracies for each of the classes, thus the varying amounts of the different sea ice types are accounted for. In the revised manuscript, we adopted Cohen's Kappa to support the performance evaluation.

[Section 2.2.6]
For performance evaluation, we~~We~~ use confusion matrix and Cohen's kappa coefficient κ (Cohen, 1960), which measures the agreement between two raters (in this study, they are the trained classifier and the reference ice chart) with taking account of the possibility of the agreement occurring by chance~~ for performance evaluation~~.

Is all level sea ice considered to have low backscatter? (P4 R7.)

No. They can have high backscatter in case frost flowers exist.

[Section 2.2.2]
For surfaces with low backscattering such as calm ~~ocean~~ open water and level sea ice without the presence of frost flowers on top, the effects from thermal noise contamination are visible not only in the backscattering image but also in some of the texture images (Park et al., 2019).

[Section 3]
This might be because the new ice has different types of recently formed ice including nilas, which is smooth but rafting can make rough features, and frost flowers, which introduces high surface roughness and volume scattering (Isleifson et al., 2014), thus the new ice can appear either featureless dark or complex bright in SAR image (Dierking, 2010). The large range in backscatter values makes it hard to define characteristic texture in the new ice patch. ~~there is no characteristic texture in the new ice patch; typically, they look just dark in SAR image.~~

How does the spatial resolution of 1 km affect the results? Have you tested using different spatial resolution sizes?

The spatial resolution was set as 1 km to meet with two conflicting requirements. The size of subwindow for texture calculation must be larger than the spatial scales of the major sea ice structures, e.g., cracks, ridges, rafting patterns, narrow leads, which are up to a few hundred meters. The resolution needs to be comparable with modern high-resolution sea ice products from SAR. The regional ice concentration map published by Met.no has also 1 km resolution.

How was the value of 3 km set? (P7 R14)

There is no specific reason but we assumed that positional disagreements can be up to this amount in case the same SAR image was also used in NIC ice charting.

Sea ice does not form bergy bits, the terminology used in Figure 3 of bergy water may therefore be misleading. Please change to a more appropriate term.

You are right. The “bergy water” label was removed from Figure 3.

Visual inspection of Figure 7 seems to indicate that the method struggles with SAR image edges and that the same sea ice on different side of an image edge is classified differently, e.g. top right corner where one size of the edge is young ice and on the other side there is first-year ice. Please comment. Is the First-year ice observed in Figure 8 to the east of the old ice an artefact of a beam problem for the method?

As the reviewer pointed, the proposed method partly fails in resulting consistent result when the same sea ice on different side of an image edge is classified. Regarding the artifacts in Figure 8, it is most likely due to the wind roughened surface. When the sea surface is roughened by wind, it creates texture and easily misclassified as ice.
https://www.star.nesdis.noaa.gov/sod/mecb/sar/AKDEMO_products/APL_winds/wind_images2/2019-02/S1A_ESA_2019_02_08_07_21_08_0602925668_003.77E_79.54N_HH_C5_GFS05CDF_wind.png
Regarding these limitations, we added a discussion paragraph to the end of Section 3.

[Section 3]
Unfortunately, the proposed algorithm has several limitations. First of all, the variations in radar backscattering and its corresponding image textures due to seasonal changes were not properly captured. Although day-of-year was tested as a seasonality variable in the FC3 feature configuration, the result did not show any improvement. This is because day-of-the-year might not correspond to the same temperature, fluxes, and weather regimes. Second, the proposed method struggles when the same sea ice is located on different marginal sides of SAR images because the incidence angle dependence could not be normalized perfectly. An example of such a failure can be seen along the image boundaries at 80N, 35E and 82.5N, 60E, approximately. Third, some artifacts were observed under an extreme marine condition. In the classified results in the bottom right panel of Figure 8, there is a misclassified FYI patch (yellow) in the open water area. According to the NOAA SAR wind image service, ANSWRS 2.0, the wind speed ranged from 17 to 21 m/s at the time of image acquisition heavily roughing the water surface.

What is your definition of New ice? For the ice types used here are you using the WMO definitions?

Yes, we follow the WMO definitions as the NIC ice chart does. In the WMO Sea Ice Nomenclature (WMO No. 259, volume 1 – Terminology and Codes), the definition of New ice is, "A general term for recently formed ice which includes *frazil ice*, *grease ice*, *slush* and *shuga*. These types of ice are composed of ice crystals which are only weakly frozen together (if at all) and have a definite form only while they are afloat."

In Figure 9 the sea ice type FYI thin is include in the sea ice classification results. Please clarify what thin FYI means. Why is this class not used throughout?

The label was wrong. It was just FYI without "thin". As the bottom panels (SAR results) are irrelevant in the context of the corresponding description, they were removed in the revised manuscript.

Consider adding something indicating the semi-automatic aspect of the manuscript in the title. And also indicate that the Sentinel-1 scenes used here only reflect the winter season.

We changed the title as "Classification of Winter Sea Ice Types in Sentinel-1 SAR images". The word "semi-automatic" and "semi-automated" appears several times in the revised manuscript including abstract.

Technical comments

In many places references to the appropriate work is missing, e.g. P1, R27, P4 R15-16 and P6 R9-10. Please carefully revise the manuscript to include references to earlier work.
Corrected.

The method is claimed to be semi-automatic in the body of the manuscript though the word semi- is left out of the abstract of the manuscript. Please correct.
Corrected.

P1 R22-24. Unclear sentence, consider rewriting.
Corrected.

P.1, R.14-15. Unclear sentence, consider rewriting.
Corrected.

Minor grammatical errors are present throughout the manuscript, e.g. P2 R5 : : : particularly in the cross-polarization: : : P2 R7 : : : considering the relative: : : p2 R31 : : :to train the classifier: : : P3 R15 To take advantage of the objective: : : Please carefully revise the English language throughout and pay particular attention to "the"
Corrected.

P2 R24 Region of interest -> region of study
Corrected.

P2 R25 coexit -> are found
Corrected.

P3 R2 National Ice Center (NIC) -> US National Ice Center
Corrected.

P3 R29. What does the precision of decimals mean here? That e.g. that the sea ice concentration can be 10.1%?
Reworded.

> Note that the ice concentration label in the SIGRID-3 format is assigned in an increment of 10%~~has precision of decimals~~.

P4 R5. What is strong noise?
Image contaminations as in Park et al., 2018

Consider using the commonly used term "open water" instead of "ocean" throughout. The same goes for ice free (IF) please use the much more commonly used open water (OW). MFYI could easily be confused with Multi-first-year ice. If MFYI is meant to indicate Mixed First Year Ice please change this or at least include this information in the header of the tables 4 and 5. In Table 4 and 5 are the MFYI meant to be used also on the predicted header?
Corrected.

P4 R12. Senitnel -> Sentinel
Corrected.

P5 R9 "Furthermore, further: : :" consider changing this.
Corrected.

P6 R10-12 you argue that when training dataset are prepared manually the sample size is usually less then 20 images. Please provide several references to support this statement.

> When training dataset is prepared by manual work (i.e., manual classification by human expert), the number of images is not large, usually less than 20 (e.g., 12 scenes in Zakhvatkina et al., 2013; 20 scenes in Leigh et al., 2013; 2 scenes in Liu et al., 2015; 4 scenes in Ressel et al., 2015).

P7 R6. Unclear sentence please revise.
Corrected.

P7 R19-21 appears to be a description of the method please move them to the methods section.
Following the reviewer's suggestion, they were moved to the end of the Section 2.

P8 R13-14. Substantial work studying young ice type has been carried out by e.g. Dierking 2010. Consider referencing such works. The young sea ice has a large range in backscatter values from very low to high and are also often subjected to frost flowers, see e.g. excellent work by Isleifson. Stating that young ice "typically just look dark in the SAR images" may therefore not be strictly true.
Revised.

> This might be because the new ice has different types of recently formed ice including nilas, which is smooth but rafting can make rough features, and frost flowers, which introduces high surface roughness and volume scattering (Isleifson et al., 2014), thus the new ice can appear either featureless dark or complex bright in SAR image (Dierking, 2010). The large range in backscatter values makes it hard to define characteristic texture in the new ice patch. ~~there is no characteristic texture in the new ice patch; typically, they look just dark in SAR image.~~

P9 R13-15. Unclear sentence, please revise.
Corrected.

P9 R16. What does “charted in” mean?

Reworded.

> The ~~type of the~~ same ice floe (red outline) is ~~annotated~~ classified differently in ~~the~~ two different ice charts (old ice on the left panel and first-year ice on the right panel)~~, while~~ although it looks almost the same in the corresponding SAR backscattering images. ~~Considering the ice edges in ice charts match well with those in the SAR backscattering images, thus the ices in the inner parts are also expected to be charted in the same time, i~~It should be noted that training with ice chart might have included ~~wrong~~ mislabeled small features ~~samples~~ even if the image selection based on ice edge matching was successful.

P9 R17. What does “included wrong samples” mean? What makes a sample wrong?

Reworded.

> It should be noted that training with ice chart might have included ~~wrong~~ mislabeled small features ~~samples~~ even if the image selection based on ice edge matching was successful.

P9 R19. When mentioning previous studies please provide references to these studies.

Added numbers from the references.

> Therefore, the lower classification accuracies compared to those in the previous studies (80% in Zakhvatkina et al., 2013; 91.7% in Liu et al., 2015; 87.2% in Aldenhoff et al., 2018), which used manually classified ice maps as training and validation reference, are expected.

P9 R31-32. Daily ice charts building on Sentinel-1 are provided by the Ice service at the Meteorological office in Norway, so it is certainly true that this can be done.

The ice chart from Met.no does not provide ice type but ice concentration.

Figure 4. What does the colorbar represent?

We accidentally omitted the title for the colorbar. It represents the number of subimages used for computing the histogram.

Figure 8. In Figure 3 and 7 the ice chart from 3 days later is used yet in Figure 8 the image from the same day is used. Please be consistent in which time interval is used for these weekly ice charts.

Since the reference ice chart is published weekly, the same NIC ice chart both in Figure 7 and 8 is supposed to be valid for both dates of Figure 7 and 8, but it is not as shown. The common time interval of 3 days in Figure 3 and 7 is just a coincidence. It can be one day or two days depending on the date of SAR image that used as source materials for ice charting at ice services. If the satellite images acquired two days prior to the publication of weekly ice chart, then the overall distribution of ice would represent the status of that time. Note that among the 57 images used for training, 42 and 35 percent of the images were acquired 2 and 3 days prior to the publication date of the corresponding weekly ice chart, respectively.

Anonymous Referee #2

General comments

In this paper the authors propose a classification method for determining ice types in Sentinel-1 SAR images. The structure and methodology are similar to those found in other studies, and overall the paper reads reasonably well. The study makes good use of recent published work by the authors on denoising Sentinel-1 SAR images, and improvements to the calculation of texture information in these images.

First of all, we would like to thank the reviewer for the positive evaluation and providing important comments. We hope that all your concerns will be cleared after reading our responses and modifications made to the manuscript. Please find below our answers (in green) and modifications (deleted in red and added in blue) to your comments/suggestions/questions.

Specific comments

To the best of my knowledge, this is the first study to examine the classification of Sentinel-1 SAR images for determination of sea ice types. These images are of great interest to the scientific and operational community. As the authors point out, the images are noisy, and the residual noise after the ESA correction is still significant. Certainly, the ability to classify ice types from such noisy images is of great use. However, I have difficulty following some of the claims made, in particular in the abstract and introduction. Primarily, I am not certain if it is clear to the authors that operational ice charts are generated manually, and contain significant bias and other possible errors of subjectivity. It is a little difficult to find information about this online, but the studies by Partington et al. (2003) and in the text by Johannessen et al. (2006) clearly state that the preparation of NIC charts (former reference) and AARI charts (latter reference) is through manual inspection of various sources of satellite imagery and other sources of data. Other studies (such as J. Karvonen, 2015) look at the accuracy of manual analyses by ice analysts. Training using a large volume of these charts would reduce operator-to-operator bias, but not the overall bias these charts are believed to contain since they are produced in the interest of marine safety. Based on this, the claim in the abstract and elsewhere that the use of ice charts allows training/testing data 'void of biased subjective decisions' should be revised.

Thank you very much for pointing an important issue. We revised the abstract and some parts in introduction as belows.

[Abstract]

A new Sentinel-1 image-based sea ice classification algorithm using a machine learning-based model trained in a semi-automated manner is proposed to support ~~automated~~ daily ice charting. Previous studies mostly rely on manual work in selecting training and validation data. We show that the ~~use of~~ readily available ice charts from ~~an~~ the operational ice services ~~allow to automate selection~~can reduce the number of manual works in preparation of large amounts of training/testing data. Furthermore, they reduce the ~~void of biased, subjective~~inconsistent decisions in the classification algorithm by indirectly exploiting the best ability of the sea ice experts working at the operational ice services.

[Section 1]

The use of a public ice chart as training and validation reference data may help in solving the validation problem and enabling automation. The preparation of a public ice chart is also through manual inspection of various sources of satellite imagery and other sources of data (Partington et al., 2003; Johannessen et al., 2006); however, training using a large volume of these charts would reduce operator-to-operator bias. The overall bias may exist since the public ice charts are produced in the interest of marine safety. Nevertheless, as the human interpretation available in the ice chart is currently considered as the best available information of sea ice (Karvonen et al., 2015), the best practice to make a sea ice type classifier is to train with the public ice chart so that the best knowledge of certified ice analysts is mimicked.

The 'novelty' of using ice charts in this way as training data should be clarified. These charts are fairly similar to the training data that was used for the sea ice type classification study by Zakhvatkina [2013], where homogeneous areas identified by trained ice analysts are used. Image analysis charts, which are very similar to

daily ice charts with the exceptions that they are based only on the SAR imagery, are used directly as training data in the study by Wang et al. [2017]. In that study the ice concentration information was used directly in the same manner as ice type in the present study (the available charts were mapped to the SAR image latitude and longitude), however it was ice concentration information that was used, not ice type. These similarities should be discussed.

They are now included in the introduction as below.

[Section 1]
In ~~most~~ many of the previous works on ice-water and/or sea ice classification (Soh and Tsatsoulis, 1999; Zakhvatkina et al., 2013; Leigh et al., 2014; Liu et al., 2015; Ressel et al., 2015; Zakhvatkina et al., 2017; Aldenhoff et al., 2018), the training and validation were done using manually produced ice maps. Although the authors claimed that the manual ice maps were drawn by ice experts, the selection of SAR scenes and interpretation can be ~~subjective~~inconsistent, and the number of samples ~~were~~was not enough to generalize the results because of the laborious manual work. Therefore, increasing objectivity is crucial, and automating the classification process is encouraged. The idea of training using SAR images and accompanying image analysis charts, which is a raw interpretation of SAR images by trained ice analysts working at operational ice services, were tested for sea ice concentration estimation by Wang et al. (2017); however, such image analysis charts are not accessible to the public.

Random forest classifiers are very popular at this time, and have been shown to be useful in many studies. To better motivate the present study, I suggest the authors compare their method to a multi-class random forest. In particular, the reference given for choosing the one-vs-all classification scheme as compared to multiclass problem is not closely related to the problem at hand. Did the authors try the multiclass method? Given that the motivation here is for operational implementation, it would be of interest to know if the mutliclass method performs similarly, and the computation time difference between the binary one-vs-all method and multiclass method.

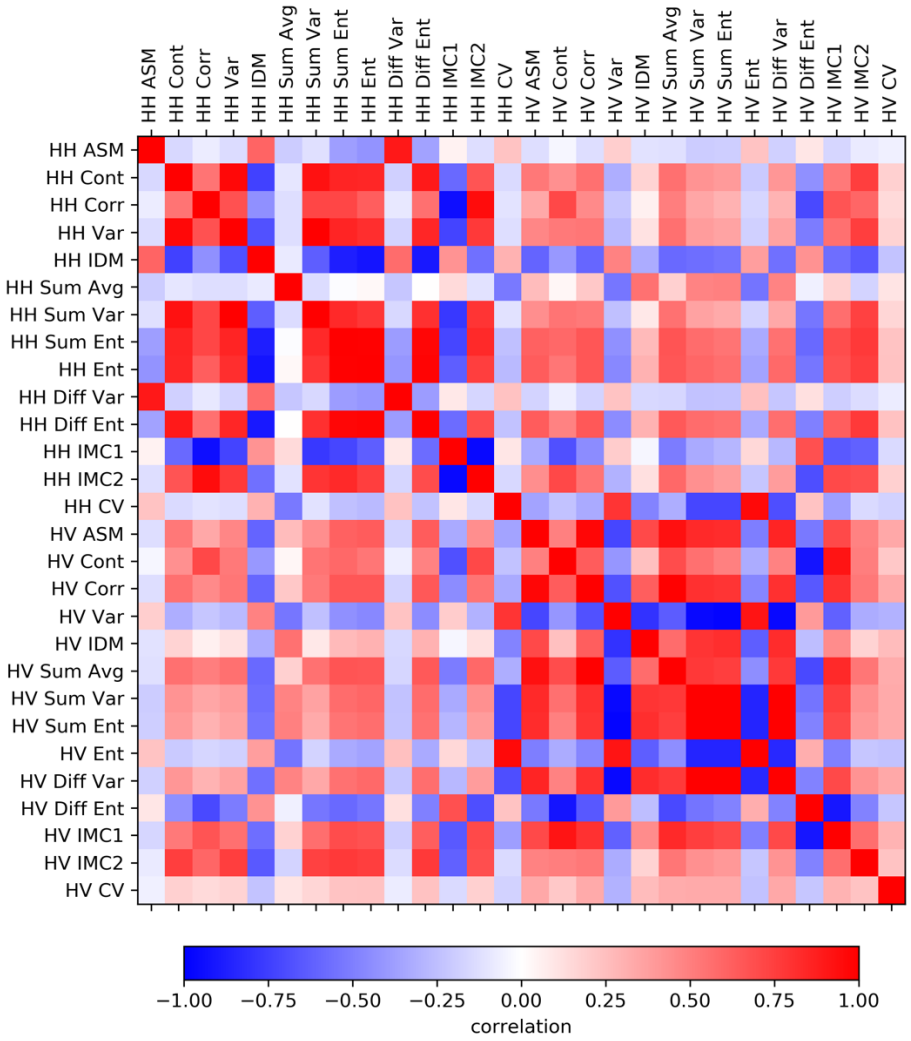
Yes, we tried the multi-class random forest as well. The main reason for using one-vs-all scheme is to check the difference in feature importance per each of the classes. As the multi-class random forest gives a single feature importance, it is impossible to see the differences among the classes. The performance of the one-vs-all binary approach were slightly better than that of the multi-class method as shown in the table below, and this is in line with the results in Adnan and Islam, 2015. However, the computation times of the multi-class method were 1/3 and 1/2 compared to those of the binary one-vs-all method for the cases of 5- and 3-class.

Feature configurations	FC1		FC2		FC3	
Number of classes	5 classes	3 classes	5 classes	3 classes	5 classes	3 classes
Overall accuracy (multi-class)	58.5	86.2	58.0	86.1	57.4	75.5
Overall accuracy (One-vs-all)	58.8	86.2	58.4	86.7	54.6	75.8
Cohen's kappa (multi-class)	0.66	0.79	0.66	0.79	0.52	0.53
Cohen's kappa (One-vs-all)	0.67	0.80	0.67	0.80	0.49	0.53

Adnan, M. N., and Islam, M. Z., One-Vs-All Binarization Technique in the Context of Random Forest, Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges (Belgium), 22-24 April 2015.

I have a similar question regarding the use of all Haralick texture features. How long did it take to calculate these features over the 64 grey levels used here? Are all features needed, or is it not relevant (in the sense that the additional time required and change in accuracy is not significant).

There are some high correlations between the features as shown in the heatmap below. For example, ASM and Diff Var, Cont and Var, Sum Ent and Ent have highly correlated each other. Removing some of them may not lead to significant decrease in prediction accuracy, but the computational efficiency is out of the scope of this study. The computation time for extracting Haralick texture features per image is approximately five minutes in the given conditions (64 grey levels, 25 x 25 pixels of subwindow size) with an Intel i7 quad-core processor.



If the main contribution is to be the classifier itself, then a more careful examination of the method should be carried out. It would also be very interesting to see how the denoising methods they have developed lead to improved ice type classification. I am not sure if that would be difficult. Without this information, others are likely to attempt ice type classification without following rigorous denoising procedures. With this information, this piece of work could be a much stronger contribution to the sea ice community.

We conducted an additional test by following your suggestion, and the results are added to the revised manuscript. As shown in the table below, the textural denoising led to improved accuracies for all the classes except New ice.

[Section 3]
To see how the denoising step in Section 2.2.2 led to improvements in the classification accuracies, the same training and evaluation were conducted for the same dataset without applying the textural noise correction (Table 4). In both FC1 and FC2, the accuracies improved for young ice (+8.2-9.8%) and first-year ice (+9.2-11.6%) which were most pronounced compared to those for open water (+1.7%) and old ice (+1.2-1.7%). On the contrary, a small accuracy decrease was observed for new ice (-2.8-4.7%). Nevertheless, the improvement in kappa (+0.05) demonstrates a clear improvement in the overall classification result.

Table 4: Changes in classification accuracies before and after applying textural denoising

class	case								
	FC1			FC2			FC3		
	Thermal denoising only	Textural denoising applied	difference	Thermal denoising only	Textural denoising applied	difference	Thermal denoising only	Textural denoising applied	difference
OW	88.4	90.1	+1.7	88.9	90.6	+1.7	88.0	85.4	-2.6
NI	30.2	28.0	-2.8	27.7	23.0	-4.7	31.8	23.9	-7.9
YI	34.9	44.7	+9.8	36.2	44.6	+8.2	43.4	51.5	+8.1
FYI	29.3	38.9	+9.6	30.4	42.0	+11.6	38.0	47.0	+9.0
OI	91.5	92.7	+1.2	90.3	91.7	+1.4	75.2	66.3	-8.9
kappa	0.62	0.67	+0.05	0.62	0.67	+0.05	0.54	0.49	-0.05

In the end, it is found that the classification accuracies are higher when considering only three classes, first-year ice, muliti-year ice and open water. Could the authors add a little discussion to the conclusions as to if they envision a three-class or five-class operational implementation? If it is three-class, would they recommend using ice types from another sensor as training data? Some discussion on how the method is expected to work for other times of year should also be included.

We added relevant discussions to Section 3 and 4.

[Section 4]
Based on the results, we envisage that three-class ice type classification from SAR imagery would be useful for making a global sea ice type product like EUMETSAT OSI-403-C (Aaboe et al., 2014) with higher spatial resolution.

[Section 3]
Unfortunately, the proposed algorithm has several limitations. First of all, the variations in radar backscattering and its corresponding image textures due to seasonal changes were not properly captured. Although day-of-year was tested as a seasonality variable in the FC3 feature configuration, the result did not show any improvement. This is because day-of-the-year might not correspond to the same temperature, fluxes, and weather regimes.

page 6 - line 10 - Can the authors explain what they mean here by a 'sparse dataset' and why a dataset used for ice/water and ice types from SAR imagery would be considered a 'sparse dataset'? I am not sure I follow this line of reasoning.

The sentence was reworded as belows.

In the literatures about sea ice classification, the SVM was used often because by nature it works relatively well even when the~~for sparse~~ number of ~~dataset~~datasets are small. When training dataset is prepared by manual work (i.e., manual classification by human expert), the number of images is not large, usually less than 20 (e.g., 12 scenes in Zakhvatkina et al., 2013; 20 scenes in Leigh et al., 2013; 2 scenes in Liu et al., 2015; 4 scenes in Ressel et al., 2015).

page 6 - line 32 - Why is the 'Richard's curve' chosen over a typical curve fit? Do the authors obtain more robust or interpretable results using this method? Please provide more context.

We added more explanations.

> Classification scores with values ranging from 0 (worst performance) to 1 (best performance) are evaluated for each node of the grid and are interpolated between the nodes by curve fitting. The~~the~~ Richards's~~s~~ Curve (Richards, 1959) was used as the fit model because it allows easy estimation of the model's maximum value.

page 7 - If I understand correctly, the authors manually selected 57 image (or do the authors mean scenes here?) for training and testing from a set consisting of 958 images (or again is this scenes)? Can they say something about these 57? Are they from similar geographic regions? times of year? specific features? Going through 958 images manually to choose a training data set is not automated. Using an ice type product generated in an automated manner from another sensor (for example open water/FYI/MYI from passive microwave data or scatterometer data) could provide an automated workflow.

If you mean the image-subimage things, it is SCENES here. As in Section 2.1, a total of 958 scenes were acquired, and the selected 57 scene are from various geographic region within the study area and various time of year. Using ice type product from passive microwave data or scatterometer data cannot help the image selection procedure due to the large difference in spatial resolution of them and SAR. Regarding the automation issue, we clarify throughout the revised manuscript that the developed algorithm is "semi-" automated.

> **[Section 2.2.6]**
> In order to automate image selection, the ice edges in SAR images need to be identified first. ~~To automate image selection for training, a good ice/water classifier for SAR image is needed.~~ Since even ~~such~~ an ~~simple binary~~ ice/water classifier has not been well developed yet for Sentinel-1, the image selection procedure has to be done manually in the beginning. However, once a classifier is generated with high accuracy, it can be used to automate the procedure, then the whole process in the proposed scheme will be fully automated. This is why the proposed algorithm is named "semi-" automated for now. Nevertheless, the manual selection is done by visual inspection of ice-water boundaries overlaid on SAR images. The ice-water boundary can be extracted easily from the reprojected ice chart by selecting the pixel borders of open water class. Then the SAR backscattering image contrasts across the ice-water boundaries are examined both in HH- and HV-polarization because the image contrast between ice-water is larger in HV and smooth level ice is better recognizable in HH.

page 8 and Figure 6 - What method was used to determine the feature importance score and why was this method chosen? How is this score calculated?

> **[Section 2.3]**
> For each sub-classifier, each ~~of the~~ texture feature~~s~~ has a different weight in decision making. The fraction of the samples that each texture feature contributes can be used to compute the relative importance of the features, and the averaged estimates of them over several randomized trees serve as an indicator of feature importance (Louppe, 2014).
>
> **[References]**
> Louppe, G.: Understanding random forests: From theory to practice, PhD Thesis, U. of Liege, 2014.

Technical comments

1) abstract - overall accuracies vs. overall accuracy - Be consistent in your use of plurals here

Corrected.

2) abstract - In what way would this work support automated ice charting? Were the authors thinking that fewer operational (manual) charts would need to be produced? Clarification of this point would be helpful.

Revised.

> A new Sentinel-1 image-based sea ice classification algorithm using a machine learning-based model trained in a semi-automated manner is proposed to support ~~automated~~ daily ice charting.

3) page 1 - line 12 - 'In most of the previous works...'... please provide a few references in this sentence to the works you have in mind.

References added.

> In ~~most~~ many of the previous works on ice-water and/or sea ice classification (Soh and Tsatsoulis, 1999; Zakhvatkina et al., 2013; Leigh et al., 2014; Liu et al., 2015; Ressel et al., 2015; Zakhvatkina et al., 2017; Aldenhoff et al., 2018), the training and validation were done using manually produced ice maps.

4) page 1 - lines 12-15 and lines 20. I reiterate my earlier point. Ice charts are generated manually by trained analysts. Although they are available in the public domain, and this means using these charts directly relieves the individual designing the classification algorithm from the 'laborious' and possibly biased process of manually choosing training and testing data, it does not enable an automated workflow.

Revised.

> A new Sentinel-1 image-based sea ice classification algorithm using a machine learning-based model trained in a semi-automated manner is proposed to support ~~automated~~ daily ice charting. Previous studies mostly rely on manual work in selecting training and validation data. We show that the ~~use of~~ readily available ice charts from ~~an~~ the operational ice services ~~allow to automate selection~~can reduce the number of manual works in preparation of large amounts of training/testing data. Furthermore, they reduce the ~~void of biased, subjective~~inconsistent decisions in the classification algorithm by indirectly exploiting the best ability of the sea ice experts working at the operational ice services.

5) page 1 - line 20 - Again, ice charts are generated by humans. They contain human error. They are often produced under a strong time constraint, and in the interest of marine safety (the latter point meaning they likely contain bias to ensure safety).

Revised.

> The use of a public ice chart as training and validation reference data may help in solving the validation problem and enabling automation. The preparation of a public ice chart is also through manual inspection of various sources of satellite imagery and other sources of data (Partington et al., 2003; Johannessen et al., 2006); however, training using a large volume of these charts would reduce operator-to-operator bias. The overall bias may exist since the public ice charts are produced in the interest of marine safety. Nevertheless, as the human interpretation available in the ice chart is currently considered as the best available information of sea ice (Karvonen et al., 2015), the best practice to make a sea ice type classifier is to train with the public ice chart so that the best knowledge of certified ice analysts is mimicked.

6) page 3 - line 25 - 'ice edge determined from AMSR-E' - an ice edge cannot be determined from AMSR-E without using an algorithm. Which algorithm was used? Please revise.

Revised.

> Heinrichs et al. (2006) reported that the ice edge determined from ~~the~~ AMSR-E passive microwave radiometer~~, which is a passive microwave radiometer,~~ data using the isoline of 15% concentration matches best the ice edge determined from RADARSAT-1 SAR data using visual inspection~~, which is a C-band HH-polarization SAR~~.

7) page 3 - lines 28-29 - I don't know what the authors mean by 'has a precision of decimals'.

Revised.

> Note that the ice concentration label in the SIGRID-3 format is assigned in an increment of 10%~~has precision of decimals~~.

8) page 3 - line 26 - Similar comment regarding the ice edge determined from SAR - a methodology must have been used to get this ice edge. Was it visual inspection, or another method? Please revise.

Revised.

Heinrichs et al. (2006) reported that the ice edge determined from ~~the~~ AMSR-E passive microwave radiometer~~, which is a passive microwave radiometer,~~ data using the isoline of 15% concentration matches best the ice edge determined from RADARSAT-1 SAR data using visual inspection~~, which is a C-band HH-polarization SAR~~.

9) page 4 - line 3 - better than what?

Revised.

Comparing the original SoD in~~In~~ the top left panel with the processed SoD in the bottom left panel, it is clear that the ice edge of the processed SoD match better with the SAR backscattering images.

10) page 5 - line 2 - wording is not specific - many of the previously developed methods - methods for what? These references are a mixture of ice/water and ice type classification studies. These two tasks are different from the perspective of a computer algorithm. Also, a reference to Shokr [1991] should be included.

Revised and added reference.

Like many of the previously developed sea ice type classification methods (Shokr, 1991; Barber and LeDrew, 1991; Soh and Tsatsoulis, 1999; Deng and Clausi, 2005; Zakhvatkina et al., 2013; Leigh et al., 2014; Liu et al., 2015~~; Karvonen, 2017; Zakhvatkina et al., 2013, 2017~~), the proposed approach starts from gray level co-occurrence matrices (GLCM) calculation.

11) page 5 - lines 9-11 - I don't understand this sentence, what is averaged for multiple distances, and what is the normalized GLCM?

Normalized GLCM is the GLCM divided by the sum of all elements, representing probability of co-occurrence. As there are multiple normalized GLCMs, one for each of the co-occurrence distances, the averaged values were used to reduce the dimensionality of the data to analyze.

12) page 5 - line 5 - direction? or should this be orientation?

Revised. It should be orientation in the context.

13) page 5 - line 12 - The term spatial resolution is not clear. Some authors consider this the scales that are resolved. It may be better to state that the spacing between the GLCM texture feature windows is 1km? (or please reword if I am not interpreting this point correctly).

Revised.

In this study, we set *w* as 25 so that the grid spacing of the result of texture analysis~~spatial resolution~~ is 1 km.

14) page 5 - It would be nice to have the Haralick features listed in a table, and to provide a brief rationale for including all of them in the study. Information as to how long it took to calculate these features using 64 grey levels for their set of imagery is also important.

As the usefulness of GLCM-based texture features for sea ice classification has been demonstrated in literature (Shokr, 1991; Soh and Tsatsoulis, 1999; Deng and Clausi, 2005; Zakhvatkina et al., 2013; Leigh et al., 2014; Liu et al., 2015) and the Haralick features include most of them, it might be not necessary to list all the features in the manuscript. The computation time for extracting Haralick texture features per image is approximately five minutes in the given conditions (64 grey levels, 25 x 25 pixels of subwindow size) with an Intel i7 quad-core processor.

15) page 5 - the number of Haralick features is referred to inconsistently as 13 on line 7 and 26 on line 25. The 26 is likely just accounting for the two polarizations, but the two should be referred to in a consistent manner. Similarly on page 7 lines 22-23, please use either 'Haralick texture features' or 'texture features' consistently when describing the three classifiers.

Revised.

In addition to ~~the~~ 1326 ~~26~~ Haralick features, the coefficient of variation (CV) which is reported as a useful feature for ice-water discrimination (Keller et al., 2017) is included. The CV is defined as follows:
…
We trained three RF classifiers with different feature configurations: i) FC1: ~~texture features from~~ Haralick texture features and CV, ii) FC2: Haralick texture features, CV, and ~~texture features and~~ incidence angle, iii) FC3: Haralick texture features, CV~~texture features~~, incidence angle, and ~~day of year~~day-of-the-year.

16) page 6 - line 16 'they are' - what are 'they'? Is this the number of operations?

Revised.

For the SVM, ~~they are~~the number of operations is $O(n^2p + n^3)$ and $O(n_{sv}p)$ for training and prediction while for RF, $O(n^2pn_{tr})$ and $O(n_{tr}p)$, respectively, where n is the number of samples, p is the number of features, n_{sv} for the number of support vectors, and n_{tr} for the number of trees.

17) page 7 - If a binary ice/water classifier is 'simple' (line 6), why are the authors starting with ice type classification? I suggest this be reworded.

Revised.

Since even ~~such~~ an ~~simple binary~~ ice/water classifier has not been well developed yet for Sentinel-1, the image selection procedure has to be done manually in the beginning.

18) page 8 - lines 20-21 - The sentence starting with, 'Since the training and test datasets were extracted from the same…' I find a little out of place. With this placement, it seems like it is trying to account for the results from FC2 and FC3. It might be better to start this one with 'When the evaluation is carried out with the 2018 data, the training and test datasets….'

Revised.

19) page 9 - line 32 - 'capturing' should be 'to capture'

Corrected.

20) page 9 - line 32 -'more details' - as compared to what?

Corrected.

21) Figure 3 - Could the authors provide some information in the text as to what the map of partial concentration is (top right). Is this the partial concentration of the dominant ice type for the given polygon?

Revised.

22) Figures 3,7,8 and 9 should have geolocation data provided.

Added geolocation grid to Figure 7 and Figure 8. In Figure 3 and Figure 9, the geolocation information may be irrelevant for understanding the contents.

23) all numbers less than ten should be written out in words, eg., 3 -> three

Corrected

References

A comparison between high-resolution EO-based and ice analyst-assigned sea ice concentrations, Juha Karvonen and others, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(4):1-9, 2015.
Evaluation of second-order texture parameters for sea ice classification from radar images, Mohammed E. Shokr, Journal of Geophysical Research, 96(C6),10,625-10640, 1991.
Late twentieth century Northern Hemisphere sea-ice record from the U.S. National Ice Center ice charts, Kim Partington, Tom Flynn, Doug Lamb, Cheryl Bertoia and Kyle Dedrick, Journal of Geophysical Research, 108(C11), doi:10.1029/2002JC001623, 2003.
Remote sensing of sea ice in the Northern Sea route: Studies and applications, Ola M. Johannessen and others, Springer Science and Business Media, 472 pages, 2006.

References above are now in the reference list of the revised manuscript.

Anonymous Referee #3

First of all, we would like to thank the reviewer for the providing important comments. We hope you reconsider your decision based on our responses and modifications made to the manuscript. Please find below our answers (in green) and modifications (deleted in red and added in blue) to your comments/suggestions/questions.

General comments

The manuscript by Park et al presents Random Forest based classifier (Python Scikit-Learn) for sea ice type classification from dual pol (HH-HV) Sentinel-1 images which were collected during the winter period only. Training dataset were collected from the National Ice Center weekly ice chart and the classification algorithm exploits standard GLCM features along with some additional features. Since the launch of Sentinel-1 SAR sensors (a+b), it continuously monitoring Arctic Sea ice with high spatial and temporal resolution and an automated sea ice type classification product from high resolution SAR is highly desirable and relevant for the sea ice monitoring community. Having said that, there are several limiting factors which are preventing researchers to come up with a robust SAR based sea ice type classification scheme, (1) Backscatter variation due to varying incidence angle, along with sensor specific noise related issues (2) Backscatter variation due to seasonal changes (winter–melt- early summer - summer). In this manuscript authors performed a denoising technique which was developed by authors previously (Park et al., 2018), and a standard linear incidence angle correction (Section 2.2.3). It is important to note here that different sea ice types have different incidence angle dependency. Moreover, if the incidence angle correction was applied over all classes (i.e. including open water), it will most likely not contribute to the robustness of the classifier. Backscatter variation due to seasonal changes is completely ignored in the presented manuscript. As authors aimed to develop an operational system, in my opinion authors cannot ignore this major issue completely.

First of all, we would like to thank the reviewer for the providing important comments. Please find below our answers (in green) and modifications (deleted in red and added in blue) to your comments and suggestions.

- The use of Sentinel-1 in ice type classification is highly demanded for sea ice monitoring community but has not going well because of the two reasons that the reviewer pointed. We partly solved the problem of radar backscatter variation due to sensor noise issue, and this is one of the main significances of this study.

- The fact that different sea ice types have different incidence angle dependency is well known, but it is not true that applying the incidence angle correction over all classes will not contribute to the robustness of the classifier. Typically, the slope for open water is higher than that for sea ice, thus the correction works in a way reducing the backscatter variation for open water as well. Moreover, ice type-specific correction prior to ice type classification is controversial.

- As the reviewer pointed, backscatter variation due to seasonal changes is important for operational ice charting. Since the developed algorithm was tested for winter season only, we changed the title as "Classification of **Winter** Sea Ice Types in Sentinel-1 SAR images" and the limitation related to the variations in radar backscattering and its corresponding image textures due to seasonal changes is added to the end of Section 3 in the revised manuscript.

- The only variable that we introduced to capture the seasonality, day of year, might not correspond to the same temperature, fluxes and weather regimes. Weather variant input parameters may be more suitable. However, as sea ice drifts continuously, weather variant information for each of the ice floes in the SAR image at the image acquisition time needs to be calculated and joined, and this may require complex and rigorous works. In order to simplify the problem, we used day of year only, but certainly other parameters need to be tested in the follow up paper. We are very interested in combining sea ice drift with ice type-specific texture changes in our future study.

Specific comments

The proposed classification scheme is based on Python Scikit-Learn library and GLCM features, this kind of classification scheme is well known and published several times for different frequency bands. Therefore the current manuscript is very limited in terms of innovation. What I find slightly different is use of weekly ice chart for training data generation.

Our manuscript includes several significance and innovations:

i) This is the first study to examine the classification of Sentinel-1 SAR images for determination of sea ice types.

ii) This study demonstrates the ability to classify ice types from the noisy Sentinel-1 image by adopting our previous development, textural denoising method, and in the revised manuscript, we show how the denoising methods they have developed lead to improved ice type classification.

[Section 3]

To see how the denoising step in Section 2.2.2 led to improvements in the classification accuracies, the same training and evaluation were conducted for the same dataset without applying the textural noise correction (Table 4). In both FC1 and FC2, the accuracies improved for young ice (+8.2-9.8%) and first-year ice (+9.2-11.6%) which were most pronounced compared to those for open water (+1.7%) and old ice (+1.2-1.7%). On the contrary, a small accuracy decrease was observed for new ice (-2.8-4.7%). Nevertheless, the improvement in kappa (+0.05) demonstrates a clear improvement in the overall classification result.

Table 4: Changes in classification accuracies before and after applying textural denoising

class	case								
	FC1			FC2			FC3		
	Thermal denoising only	Textural denoising applied	difference	Thermal denoising only	Textural denoising applied	difference	Thermal denoising only	Textural denoising applied	difference
OW	88.4	90.1	+1.7	88.9	90.6	+1.7	88.0	85.4	-2.6
NI	30.2	28.0	-2.8	27.7	23.0	-4.7	31.8	23.9	-7.9
YI	34.9	44.7	+9.8	36.2	44.6	+8.2	43.4	51.5	+8.1
FYI	29.3	38.9	+9.6	30.4	42.0	+11.6	38.0	47.0	+9.0
OI	91.5	92.7	+1.2	90.3	91.7	+1.4	75.2	66.3	-8.9
kappa	0.62	0.67	+0.05	0.62	0.67	+0.05	0.54	0.49	-0.05

iii) the use of public ice chart for training is new, and the difference from the conventional approaches were discussed to the introduction section.

[Section 1]

In ~~most~~ many of the previous works on ice-water and/or sea ice classification (Soh and Tsatsoulis, 1999; Zakhvatkina et al., 2013; Leigh et al., 2014; Liu et al., 2015; Ressel et al., 2015; Zakhvatkina et al., 2017; Aldenhoff et al., 2018), the training and validation were done using manually produced ice maps. Although the authors claimed that the manual ice maps were drawn by ice experts, the selection of SAR scenes and interpretation can be ~~subjective~~inconsistent, and the number of samples ~~were~~was not enough to generalize the results because of the laborious manual work. Therefore, increasing objectivity is crucial, and automating the classification process is encouraged. The idea of training using SAR images and accompanying image analysis charts, which is a raw interpretation of SAR images by trained ice analysts working at operational ice services, were tested for sea ice concentration estimation by Wang et al. (2017); however, such image analysis charts are not accessible to the public.

However I am also concerned about the automated training data generation as it is not clear which images were used to generate the ice chart by NIC and there is a high probability that that the ice chart polygons will not match that Sentinel-1 mosaics ice types. Hence there is a high risk that the classifiers were trained with wrong training data. Authors mentioned that 57 images were selected manually where only 'ice egdes' match well with

the ice chart. In my opinion, this is also a manual selection of training data which authors criticized in the introduction section.

We criticized conventional approach for two reasons: selecting training/testing data by anonymous human ice expert can be subjective, and preparing a large volume of such dataset is laborious. If public ice charts are used for training and validation, these two issues are partly resolved. Although the public ice chart is also made by human expert thus it cannot be completely free from subjective decisions, training using a large volume of these charts would reduce operator-to-operator bias. Another advantage of the use of public ice chart is that the best knowledge of certified ice analysts rather than anonymous expert is to be mimicked through the machine learning process. Regarding the manual selection of training data, the visual inspection along the ice edges overlaid on SAR image is also not automated and this is why the proposed method is "semi-" automated algorithm, but it requires much less effort and expertise compared to drawing manual ice chart from the SAR image itself.

[Section 2.2.6]

In order to automate image selection, the ice edges in SAR images need to be identified first. ~~To automate image selection for training, a good ice/water classifier for SAR image is needed.~~ Since even ~~such~~ an ~~simple binary~~ ice/water classifier has not been well developed yet for Sentinel-1, the image selection procedure has to be done manually in the beginning. However, once a classifier is generated with high accuracy, it can be used to automate the procedure, then the whole process in the proposed scheme will be fully automated. This is why the proposed algorithm is named "semi-" automated for now. Nevertheless, the manual selection is done by visual inspection of ice-water boundaries overlaid on SAR images. The ice-water boundary can be extracted easily from the reprojected ice chart by selecting the pixel borders of open water class. Then the SAR backscattering image contrasts across the ice-water boundaries are examined both in HH- and HV-polarization because the image contrast between ice-water is larger in HV and smooth level ice is better recognizable in HH.

The selection (and definitions) of ice types for SAR based sea ice classification scheme is crucial. The 5 class classifier has some classes which might be very close to each other in terms of backscatter and texture. This might be the main reason for significantly low classification accuracy. I would recommend the authors to restrict the classifier for 4 classes (Open Water, Young Ice, FYI and old ice).

As the reviewer pointed, the five classes do not fit very well to the characteristic radar signatures from sea ice. We have tested for the four classes and the results are as below.

		Predicted											
		OW (open water)			NYI (New + Young ice)			FYI (First-year ice)			OI (old ice)		
	Case	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3
Actual	OW	**91.9**	**92.1**	**87.3**	2.0	2.3	6.0	6.1	5.7	6.7	0.0	0.0	0.0
	NYI	11.2	10.2	10.3	**47.7**	**49.1**	**54.8**	32.7	34.2	30.2	8.4	6.6	4.6
	FYI	5.5	4.6	27.8	26.8	27.4	27.8	**41.0**	**43.5**	**48.1**	26.7	24.5	19.1
	OI	0.6	0.5	8.1	3.5	3.8	8.1	3.1	4.0	25.9	**92.8**	**91.6**	**65.5**

Compared to the results of five-class scheme in Table 3, the accuracy increases for merged young ice (new ice and young ice) were 3.0-4.5%. However, the misclassification among young ice and first-year ice was still significant.

A seasonal assessment of the classification scheme is missing. It the most important issue to address and without this assessment it would not be reasonable to claim the scheme to be either operational or innovative.

As the reviewer pointed, we did not conduct seasonal assessment. Since the developed algorithm was tested for winter season only, we changed the title as "Classification of Winter Sea Ice Types in Sentinel-1 SAR images" and the limitation related to the variations in radar backscattering and its corresponding image textures due to seasonal changes is added to the end of Section 3 in the revised manuscript.

Due to above mentioned major issues I didn't listed any technical corrections. I would kindly invite the authors to address this issues in future. Due to the lack of innovation and failed to address the basic issues, at current stage I can only recommend the manuscript to be rejected despite it is well within the scope of The Cryosphere.

We kindly ask the reviewer to reconsider his/her decision based on our responses and modifications made to the manuscript.

Classification of Winter Sea Ice Types in Sentinel-1 SAR images

Jeong-Won Park[1,2], Anton A. Korosov[1], Mohamed Babiker[1], Joong-Sun Won[3], Morten W. Hansen[1,4], Hyun-cheol Kim[2]

[1]Ocean and Sea Ice Remote Sensing Group, Nansen Environmental and Remote Sensing Center, Bergen, 5006, Norway
[2]Department of Earth System Sciences, Yonsei University, Seoul, 03722, South Korea
[2]Unit of Arctic Sea Ice Prediction, Korea Polar Research Institute, Incheon, 21990, South Korea
[3]Department of Earth System Sciences, Yonsei University, Seoul, 03722, South Korea
[4]Department of Remote Sensing and Data Management, Norwegian Meteorolotical Institute, 0371 Oslo, Norway

Correspondence to: Jeong-Won Park (orepaku.andoid@gmail.comjeong-won.park@kopri.re.kr)

Abstract. A new Sentinel-1 image-based sea ice classification algorithm using a machine learning-based model trained in a semi-automated manner is proposed to support automated daily ice charting. Previous studies mostly rely on manual work in selecting training and validation data. We show that the use of readily available ice charts from an the operational ice services allow to automate selectioncan reduce the number of manual works in preparation of large amounts of training/testing data. Furthermore, they reduce the void of biased, subjectiveinconsistent decisions in the classification algorithm by indirectly exploiting the best ability of the sea ice experts working at the operational ice services. The proposed scheme has two phases: training and operational. Both phases start from the removal of thermal, scalloping, and textural noise from Sentinel-1 data and calculation of gray level co-occurrence matrix and Haralick texture features in a sliding window. In the training phase, the weekly ice charts are reprojected into the SAR image geometry. A random forest classifier is trained with the texture features on input and labels from the rasterized ice charts on output. Then, the trained classifier is directly applied to the texture features from Sentinel-1 images in an operational manner. Test results from the datasets spanning two winter seasons dataset acquired over the Fram Strait and the Barents Sea area showed that the classifier is capable of retrieving three3 generalized cover types (ice freeopen water, integrated mixed first-year ice, old ice) with an overall accuracyies of 85% and five5 cover types (ice freeopen water, new ice, young ice, first-year ice, old ice) with an accuracy of 58%. The errors are attributed both to incorrect manual classification on the ice charts and to the semi-automated algorithm. Finally, We we demonstrate the potential for near-real time service of the ice map using type classification through an example of ice maps made from daily mosaiced Sentinel-1 images.

1 Introduction

Wide swath SAR observation from several spaceborne SAR missions (RADARSAT-1, 1995-2013; Envisat ASAR, 2002-2012; ALOS-1 PALSAR, 2006-2011; RADARSAT-2, 2007-present; Sentinel-1, 2014-present) played an important role in studying global ocean and ice-covered polar Polar regionRegions. The Sentinel-1 constellation (1A and 1B) is producing dual-polarization observation data with the largest Arctic coverage and the highest temporal resolution ever. The cross-polarization

is known to be more sensitive to the difference in scattering from sea ice and open water than the co-polarization (Scheuchl et al., 2004), and the combination of HH- and HV-polarizations has been widely used for ice edge detection and ice type classification. However, most of the recent ice classification algorithms were developed using RADARSAT-2 ScanSAR ~~images~~ (Leigh et al., 2014; Liu et al., 2015; Zakhvatkina et al., 2017) which has different sensor characteristics from Sentinel-1 TOPSAR, and the use of Sentinel-1 for the same purpose is very limited in literature. The main drawback of applying existing algorithms to Sentinel-1 TOPSAR data is the relatively high level of thermal noise contamination and its propagation to image textures.

For ~~a~~ proper use of dense time-series of Earth observations using SAR sensors, radiometric properties must be ~~well~~ well-calibrated. Thermal noise is often neglected in many ~~applications~~ cases, but is impacting seriously the utility of dual-polarization SAR data. Sentinel-1 TOPSAR image intensity is particularly disturbed by the thermal noise ~~particularly~~ in the cross-polarization channel. Although the European Space Agency (ESA) provides calibrated noise vectors for noise power subtraction, residual noise contribution is still significant considering the relatively narrow backscattering distribution of the cross-polarization channel. In our previous study (Park et al. 2018), a new denoising method with azimuth de-scalloping, noise scaling, and inter-swath power balancing was developed and showed improved performance in various SAR intensity-based applications. Furthermore, when it came to texture-based image classification, we suggested a correction method for textural noise (Park et al., 2019) which distorts local statistics thus degrades texture information in the Sentinel-1 TOPSAR images.

In ~~most~~ many of the previous works on ice-water and/or sea ice classification (Soh and Tsatsoulis, 1999; Zakhvatkina et al., 2013; Leigh et al., 2014; Liu et al., 2015; Ressel et al., 2015; Zakhvatkina et al., 2017; Aldenhoff et al., 2018), the training and validation were done using manually produced ice maps. Although the authors claimed that the manual ice maps were drawn by ice experts, the selection of SAR scenes and interpretation can be ~~subjective~~inconsistent, and the number of samples ~~were~~ was not enough to generalize the results because of the laborious manual work. Therefore, increasing objectivity is crucial, and automating the classification process is encouraged. The idea of training using SAR images and accompanying image analysis charts, which is a raw interpretation of SAR images by trained ice analysts working at operational ice services, were tested for sea ice concentration estimation by Wang et al. (2017); however, such image analysis charts are not accessible to the public.

The use of a public ice chart as training and validation reference data may help in solving the validation problem and enabling automation. The preparation of a public ice chart is also through manual inspection of various sources of satellite imagery and other sources of data (Partington et al., 2003; Johannessen et al., 2006); however, training using a large volume of these charts would reduce operator-to-operator bias. The overall bias may exist since the public ice charts are produced in the interest of marine safety. Nevertheless, as the human interpretation available in the ice chart is currently considered as the best available information of sea ice (Karvonen et al., 2015), the best practice to make a sea ice type classifier is to train with the public ice chart so that the best knowledge of certified ice analysts is mimicked.

In this work, we present a semi-automated Sentinel-1 image-based sea ice classification algorithm which takes ~~an~~ advantage of our denoising method. The noise corrected dual-polarization images are processed into image textures that capture sea ice

features in various spatial scales, and they are used for supervised classification with a random forest classifier by relating with ice charts published by operational ice services. The use of ice charts has dual purposes: semi-automatization of classifier training, and minimization of human error.

2 Data and methods

2.1 Study area and used data

The region of study ~~interest~~ for developing and testing the proposed algorithm is the Fram Strait and the Barents Sea including a part of the Arctic Ocean (10ºW-70ºE, 75ºN-85ºN) as shown in Figure 1. Various sea ice types ~~coexist~~ are found in this area due to the intensive export of multi-year ice through the Fram Strait (Smedsrud et al., 2017), and development of young and first-year ice between Svalbard and Franz Josef Land.

Sentinel-1 ~~images~~ TOPSAR data in Extended Wide-swath (EW) mode acquired in two winter seasons ~~for 2 years~~ (Dec. 2017 – Mar. 2018 and Dec. 2018 – Mar. 2019) were collected from the Copernicus Open Access Hub (https://scihub.copernicus.eu). The number of daily image acquisitions covering the study area ranges from 6 to 10 depending on the orbits. The images from the first year (hereafter called 2018 data) ~~is~~ are used to train the classifier and those from the second year (~~hearafter~~ hereafter called 2019 data) ~~is~~ are used for validation.

The ice charts covering the same periods were collected. There are two ice services that publish weekly ice charts with Pan-Arctic coverage: U. S. National Ice Center (NIC) of the United States of America, and Arctic and Antarctic Research Institute (AARI) of Russia. Although the accuracies are known to be comparable (Pastusiak, 2016) to each other, there is no partial ice concentration information in the AARI ice chart. In this study, we use the ice charts downloaded from the NIC website (https://www.natice.noaa.gov/Main_Products.htm).

2.2 Methods

Figure 2 shows the flow of the semi-automated ice classification scheme that we propose. It is divided into two phases: training and operational. Both phases start from the removal of thermal noise from Sentinel-1 data (Section 2.2.2), incidence angle calibration (Section 2.2.3), and calculating texture features (Section 2.2.4). The training phase (shown ~~by~~ in gray ~~on~~ in Figure 2) continues with preprocessing and collocation of the ice charts with the Sentinel-1 data (Section 2.2.1) and machine learning step (Section 2.2.5 and 2.2.6). The operational phase uses the classifier of which developed during the training phase for processing texture features that were computed from the input SAR data and for generating ice charts. Detailed explanations for each step~~s~~ are given in the following subsections.

2.2.1 Ice chart preprocessing

To take ~~the~~ advantage of the objective identification of the ice type and to develop a semi~~n~~-~~-~~automated processing scheme, the proposed algorithm uses electronic ice charts published by international ice chart services. The electronic ice chart follows

SIGRID-3 format (JCOMM, 2014a), which is based on a vector format called shapefile (ESRI, 1998). The first step is to reproject the ice chart into the geometry of each SAR image. Although an accurate reprojection needs several pieces of information such as orbit, look angle, topographic height, etc., our interest is in the sea ice where the topographic difference does not exceed more than few meters, hence the reprojection of coordinates of ice chart polygons is done with Geospatial Data Abstraction Library (GDAL; GDAL/OGR contributors, 2019) using a simple 3rd order polynomial fitted using the ground control points information from the Sentinel-1 product-included auxiliary data.

After the reprojection, the following three layers are extracted: total ice concentration (CT), partial ice concentration of each ice type (CP), and stage of development (SoD). CT is important because areas with low CT can be ~~seen like water~~misinterpreted as open ocean in a SAR image. Heinrichs et al. (2006) reported that the ice edge determined from ~~the~~ AMSR-E passive microwave radiometer~~, which is a passive microwave radiometer,~~ data using the isoline of 15% concentration matches best the ice edge determined from RADARSAT-1 SAR data using visual inspection~~, which is a C-band HH-polarization SAR~~. After the visual comparison of many SAR ~~backscattering~~ images and the corresponding reprojected ice charts, we set a threshold of 20% for CT to discard water-like pixels. Note that the ice concentration label in the SIGRID-3 format is assigned in an increment of 10%~~has precision of decimals~~. CP is also important ~~to~~ in finding the dominant ice type in the given polygons. SoD is a so-called ice type. It is challenging to differentiate ice types using SAR data only, thus we ~~simplified~~ merged the SoDs ~~by merging~~ into simple five classes: ~~ice free~~open water, new ice, young ice, first-year ice, and old ice. ~~Bergy water is treated as open water since its ice concentration (by definition, less than 10%) is below the threshold that we set (i.e., 20%).~~

Figure 3 demonstrates an example of the ice chart preprocessing explained above with the colors following the WMO nomenclature (JCOMM, 2014b). Comparing the original SoD in~~In~~ the top left panel with the processed SoD in the bottom left panel, it is clear that the ice edge of the processed SoD match better with the SAR backscattering images.

2.2.2 Denoising of Sentinel-1 imagery

Sentinel-1 cross-polarization images suffer from strong noise of which are originated from combined effects of the relatively low signal-to-noise ratio of the sensor system and insufficient noise vector information in the Extra Wide-swath mode Level-1 product (Park et al., 2018). For surfaces with low backscattering such as calm ~~ocean~~ open water and level sea ice without the presence of frost flowers on top, the effects from thermal noise contamination are visible not only in the backscattering image but also in some of the texture images (Park et al., 2019). The authors have developed an efficient method for textural denoising which is essential for the preprocessing of Sentinel-1 TOPSAR dual polarization products. Denoising ensures beam-normalized texture properties for all subswaths, which helps seamless mosaic of multi-pass images regardless of the satellite orbit and image acquisition geometry. By following the methods developed in Park et al. (2018, 2019), each of the ~~Senitnel-1~~Sentinel-1 images was denoised before further processes are applied.

2.2.3 Incidence angle correction

It is well known that there is a strong incidence angle ~~dependency~~ dependence in the SAR backscattering intensity for ~~ocean~~ open water and sea ice surface (Mäkynen et al., 2002; Mäkynen and Karvonen, 2017). For wide-swath SAR system like Sentinel-1 TOPSAR, varying backscatter intensity confuses image interpretation. The quasi-linear slopes in the plane of incidence angle versus sigma nought in decibel scale for typical first-year ice are reported as -0.24 and -0.16 dB/degree for HH- and HV-polarization, respectively (Mäkynen and Karvonen, 2017). To normalize the backscattering intensity for all swath range, these slopes ~~are~~ were compensated for or used as input layer in ~~several~~several ice classification algorithms in the literature (Liu et al., 2015; Zakhvatkina et al., 2013, 2017; Karvonen, 2014, 2017; Aldenhoff et al., 2018). Although the angular dependency is not a system-dependent variable but is governed by physical characteristics of the backscattered surface, the numbers need to be reassessed because the estimations of Mäkynen and Karvonen (2017) might have been affected by the residual thermal noise which used to be very strong before the ESA has updated the noise removal scheme in 2018 (Miranda, 2018).

Figure 4~~.~~ shows two-dimensional histograms of incidence angle versus sigma nought for sea ice pixels in HH and HV polarization channels from Sentinel-1 data collected over sea ice and open water in the study area in winter 2018. From the Sentinel-1 dataset described in Section 2.1, sea ice pixels were extracted by using daily global sea ice edge products available from the EUMETSAT Ocean and Sea Ice Satellite Application Facilities (OSISAF). For HH polarization, the estimated slope was -0.200 dB/degree, which is slightly different from the estimation of the first-year ice (-0.24 dB/degree) in Mäkynen and Karvonen (2017) and in between the estimations for first-year ice (-0.22 dB/degree) and multi-year ice (-0.16 dB/degree) in Mahmud et al. (2018). For HV polarization, the estimated slope was only -0.025 dB/degree, which is much lower than the estimation in Mäkynen and Karvonen (2017), however, it is in line with the estimations from RADARSAT-2 (Leigh et al., 2014; Liu et al., 2015). We compensate for the incidence angle ~~dependency~~ dependence using the estimated slopes~~, referencing~~ with respect to the nominal scene center angle of 34.5 degrees as reference. Although the incidence angle dependence changes with ice type and radar frequency (Mahmud et al., 2018), the compensation is done for all pixels in the image using a single value of mean slope because the ice types are not identified in this stage. Open water areas of the image are also affected; however, the correction is also beneficial since the incidence angle dependence for open water is stronger (-0.65 dB/degree for wind velocity of 5 m/s, computed from CMOD5 C-band geophysical model function in Hersbach et al., 2007), thus the corrected image has less incidence angle dependence.

2.2.4 Texture feature computation

Like many of the previously developed sea ice type classification methods (Shokr, 1991; Barber and LeDrew, 1991; Soh and Tsatsoulis, 1999; Deng and Clausi, 2005; Zakhvatkina et al., 2013; Leigh et al., 2014; Liu et al., 2015~~; Karvonen, 2017; Zakhvatkina et al., 2013, 2017~~), the proposed approach starts from gray level co-occurrence matrices (GLCM) calculation. The GLCM is a four-dimensional matrix $P(i, j, d, a)$ calculated from ~~the~~ two grey tones of reference pixel i, and its neighbor

j, with co-occurrence distance d, and ~~direction~~ orientation a. Haralick et al. (1973) ~~has~~ have introduced a set of GLCM-based texture features called Haralick features, and ~~the usefulness of it~~its practicality has been reported in ~~several literatures~~numerous research. Since ~~the~~ 13 Haralick features can be calculated for each of the two-dimensional slices $P(i,j)$ for multiple d and a, the maximum number of texture features is to be as $2 \times 13 \times d \times a = 26da$, where 2 is for accounting dual polarization. It is common to take the directional average for 0°, 45°, 90°, and 135° to reduce GLCM dimensionality. ~~Furthermore, f~~Further averaging for multiple distances (1 to $w/2$ where w is the size of subwindow for GLCM computation) is taken after computing normalized GLCM. The spatial resolution of the texture features is the pixel spacing of Sentinel-1 EW-mode GRDM image (40 m) multiplied by w. In this study, we set w as 25 so that the grid spacing of the result of texture analysis~~spatial resolution~~ is 1 km.

An important factor that influences the computed texture features is the number of gray levels, L. Considering the radiometric stability of Sentinel-1 EW mode (0.32 dB; Miranda, 2018) and the range of sigma nought for various ice types (-31 to 0 dB for HH, -32 to -7 dB for HV; e~~E~~stimated from ~~the~~ Figure 4 after incidence angle correction), the number of gray level should be sufficiently large enough to capture their actual differences in sigma nought values. The optimal quantization level can be calculated using the ratio of sigma nought range to radiometric resolution as follows~~.~~:

For HH, $$\frac{(0\ dB)-(-31\ dB)}{0.32\ dB} = 96.875 \quad (1)$$

For HV, $$\frac{(-7\ dB)-(-32\ dB)}{0.32\ dB} = 78.125 \quad (2)$$

Since L should be sufficiently large to take the full advantage of system capability and yet, the computation cost should not be too expensive, in this study, we set L as 64, which is the closest power of 2 to the resulting numbers from the equations above. In addition to ~~the~~ 13~~26~~ Haralick features, the coefficient of variation (CV) which is reported as a useful feature for ice-water discrimination (Keller et al., 2017) is included. The CV is defined as follows:

$$CV = \sigma/\mu \quad (3)$$

where σ and μ are the standard deviation and mean of the samples in a given subwindow. Since CV can be computed for each polarization images, the number of texture features for Sentinel-1 dual-~~-~~polarization ~~product~~ data is extended to 28. ~~Other features can be added are incidence angle and day of year~~Incidence angle and day-of-the-year can also be added. The former is adopted to account for possible residuals from the angular dependency correction while the latter is to account for seasonal variability. Although these two are not any type of ~~image~~ textures, they can be used as input features for image classification ~~as input features~~. Note that it is important to have each ice type spatially and temporally even distributions ~~for each of the ice types~~ if these two additional features are included~~,~~; otherwise, the trained classifier will ~~give~~ result a biased prediction. The effects of including these extra features will be tested and discussed in later sections.

2.2.5 Machine learning classifier

Since there are hundreds of algorithms in the field of machine learning (ML) and each of the different algorithms has its own pros and cons, it is not easy to compare their performances and decide what to use. ~~In~~ Fernández-Delgado et al. (2014) evaluated that~~,~~ the Random Forest (RF; Ho, 1998) was ~~evaluated as~~ the best classifier ~~over~~ for various types of datasets~~,~~ with slight difference from~~but the difference with the second best,~~ Support Vector Machine (SVM; Cortes and Vapnik, 1995)~~, was not statistically significant~~. In the literatures about sea ice classification, the SVM was used often because by nature it works relatively well even when the~~for sparse~~ number of ~~dataset~~datasets are small. When training dataset is prepared by manual work (i.e., manual classification by human expert), the number of images is not large, usually less than 20 (e.g., 12 scenes in Zakhvatkina et al., 2013; 20 scenes in Leigh et al., 2013; 2 scenes in Liu et al., 2015; 4 scenes in Ressel et al., 2015). However, the number can increase with less effort when the readily available ice charts are used as training references~~, many more images become available with less effort,~~. ~~thus~~ Besides, there is no need to rely on additional manual work prone to contamination by biased decisions. The RF has two practical advantages when processing a large number of datasets. First, the RF is ~~scale~~ scale-invariant~~. It can use the data as is, while the SVM~~ and does not require~~s~~ preprocessing of the datasets whereas the SVM requires scaling and normalization~~ for the input features~~. Second, the computational complexity ~~for~~ of the RF is lower than that of the SVM. For the SVM, the number of operations ~~they are~~ is $O(n^2p + n^3)$ and $O(n_{sv}p)$ for training and prediction while for RF, $O(n^2pn_{tr})$ and $O(n_{tr}p)$, respectively, where n is the number of samples, p is the number of features, n_{sv} for the number of support vectors, and n_{tr} for the number of trees. Considering the practical requirements of fast processing for near-real time ice charting services, the RF can be a reasonable solution. We use the RF with the Python Scikit-Learn implementation (Pedregosa et al., 2011).

We split the RF classifier into several binary classifiers using a one-vs-all scheme (Anand et al., 1995). Although the standard RF algorithm can inherently deal with a multiclass problem, the one-vs-all binarization to the RF results in better accuracy with smaller forest sizes than the standard RF (~~Ramírez et al., 2018~~Adnan and Islam, 2015).

Three hyper-parameters of the ~~random forest~~RF classifier were tuned: number of trees (N_T), maximum tree depth (D), and maximum number of features (N_F). Usually, with the higher N_T and D, the model better fits to the data. However, increasing forest size can slow down the training process considerably, and more importantly, it can cause overfitting. Therefore, it is important to tune these hyperparameters adequately~~,~~ so that the processing time and performance are in balance. To determine the best values of the hyperparameters, a grid search with five~~5~~-fold cross-validation (Kohavi, 1995) is used. The grid (all possible combinations of N_T, D, and N_F values) is set in a logarithmic scale (~~see values in~~ Table 1) because the performance change with hyperparameter is typically in a logarithmic scale. Classification scores with values ranging from 0 (worst performance) to 1 (best performance) are evaluated for each node of the grid and are interpolated between the nodes by curve fitting. The~~the~~ Richards'~~s~~ Curve (Richards, 1959) was used as the fit model because it allows easy estimation of the model's maximum value. The optimal values for N_T, D, and N_F are selected based on the saturation of score increment, difference between training and testing score, and computational load considerations.

2.2.6 Training and validation

To train an ice type classifier, a set of collocated SAR images and ice charts is required. After the preprocessing of the ice chart including reprojection into the SAR image geometry, only the samples with spatially and temporally good matches should be fed to the training phase. Image selection is trivial, but not easy to automate. Since the weekly ice chart is made partly based on the SAR images acquired in the past three~~3~~ days from the date of publication, the ice edges in some images match well with those in the ice chart.

In order to automate image selection, the ice edges in SAR images need to be identified first. ~~To automate image selection for training, a good ice/water classifier for SAR image is needed.~~ Since even ~~such~~ an ~~simple binary~~ ice/water classifier has not been well developed yet for Sentinel-1, the image selection procedure has to be done manually in the beginning. However, once a classifier is generated with high accuracy, it can be used to automate the procedure, then the whole process in the proposed scheme will be fully automated. This is why the proposed algorithm is named "semi-" automated for now. Nevertheless, the manual selection is done by visual inspection of ice-water boundaries overlaid on SAR images. The ice-water boundary can be extracted easily from the reprojected ice chart by selecting the pixel borders of open water class. Then the SAR backscattering image contrasts across the ice-water boundaries are examined both in HH- and HV-polarization because the image contrast between ice-water is larger in HV and smooth level ice is better recognizable in HH.

After the image selection, the samples in the selected images are split randomly into training and test datasets with a ratio of 7:3. For the training dataset, further data selection is made by excluding the samples residing close to the polygon boundaries. This is to account for possible mismatch due to various reasons (e.g., ice drift, vector mapping error, image geocoding error, etc.). In this study, only the data from pixels more than 3 km away from the polygon boundaries was fed into the training process. Once the hyperparameter optimization is done, the RF classifier is trained for the training dataset. The trained classifier is then applied to the test dataset. For performance evaluation, we~~We~~ use confusion matrix and Cohen's kappa coefficient κ (Cohen, 1960), which measures the agreement between two raters (in this study, they are the trained classifier and the reference ice chart) with taking account of the possibility of the agreement occurring by chance ~~for performance evaluation~~. The validation is done in the same way but using a completely independent dataset. The 2018 data was used to run the training phase. Among 958 images in total, we selected 57 images of which ice edges match well with the collocated ice chart. From the selected images, 6.4 million samples covering open water and sea ice were divided into training and test dataset.

3 Results and discussion

~~The 2018 data was used to run the training phase. Among 958 images in total, we manually selected 57 images of which ice edges match well with the collocated ice chart. From the selected images, 6.4 million pixels covering open water and sea ice were divided into training and test dataset.~~ We trained three RF classifiers with different feature configurations: i) FC1: ~~texture features from~~ Haralick texture features and CV, ii) FC2: Haralick texture features, CV, and ~~texture features and~~ incidence angle, iii) FC3: Haralick texture features, CV~~texture features~~, incidence angle, and ~~day of year~~day-of-the-year.

As expected, the classification score increases with the number of trees (crosses on Figure 5, upper panel) and ~~the~~ Richards~~'s~~ curve (dashed line) fits well to the observations (RMSE=2.3×10^{-4}). The optimal N_T value is selected where the score increment per tree (i.e., local slope) becomes less than 0.001 (i.e., accuracy increase of 0.1%) and constitutes 11 trees thus keeping the forest size small. The scores also increase with the maximum tree depth (crosses on Figure 5, middle panel) but ~~the~~ Richards~~'s~~ curve (dashed line) doesn't fit so well (RMSE=3.6×10^{-3}) and cannot be used for finding the optimal D value. This can be explained by overfitting of the classifier and illustrated by the difference between training and testing scores (Figure 5, lower panel): small difference between the scores (for $D \leq 8$) indicate similar performance on training and testing datasets, while large difference (for $D > 8$) indicate that testing dataset is processed with worse results. The optimal D value is therefore selected where the score difference become higher than 0.03 and constitutes 8 levels. The optimal value of the number of features (N_F) was selected using the same criterion as for N_T and the value constitutes 10 features. As a result, the optimal hyperparameters of the number of trees, the maximum tree depth, and the number of features were ~~of the hyper parameter estimation,~~ 11, 8, and 10 ~~were selected as the optimal values for the number of trees, the maximum tree depth, and the number of features~~, respectively.

The trained five~~5~~-class classifier consists of five~~5~~ binary sub-classifiers, each of them is used for discriminating one specific class from the others. For each sub-classifier, each ~~of the~~ texture feature~~s~~ has a different weight in decision making. The fraction of the samples that each texture feature contributes can be used to compute the relative importance of the features, and the averaged estimates of them over several randomized trees serve as an indicator of feature importance (Louppe, 2014). The feature importance ~~for~~ of the sub-classifiers is presented in Figure 6. The ~~Overall~~ overall pattern shows that the features of HV polarization play~~s~~ a more important role than those of HH polarization. For HH polarization, the sum average, which is equal to the mean backscattering intensity in each subwindow, was the prominent feature. For HV polarization, however, variance- and entropy-related features were more important. The classifiers for ~~ice free~~open water and old ice have more strong dependencies on HV polarization than others. This is understandable because the main radar scattering mechanisms for those two types are strongly characterized by the portion of volume scattering: low for calm water and high for dry ice with low salinity (old ice). The classifier for new ice has a distinctive pattern ~~so~~ that the sum averages in both polarizations are much more important than other features. This might be because the new ice has different types of recently formed ice including nilas, which is smooth but rafting can make rough features, and frost flowers, which introduces high surface roughness and volume scattering (Isleifson et al., 2014), thus the new ice can appear either featureless dark or complex bright in SAR image (Dierking, 2010). The large range in backscatter values makes it hard to define characteristic texture in the new ice patch. ~~there is no characteristic texture in the new ice patch; typically, they look just dark in SAR image.~~

The confusion matrix for testing the trained classifier with the test dataset (2018 data) is shown is in Table 2. ~~The three~~ Three cases with different feature configurations (FC1-FC3) were tested. The accuracies for ~~ice free~~open water and old ice were higher than 85%~~;~~, however, those for young ice and first-year ice were around 60%. The mean difference between the results of FC1 and FC2 was only 1.6%, ~~which indicates~~ indicating that residual angular dependency was negligible after the incidence angle correction ~~was insignificant~~. However, the accuracy significantly improved from ~~when comparing the results of~~ FC2 ~~and~~

to FC3, ~~there were notable accuracy improvements,~~ especially ~~for~~ with new ice (24.5%). κ for FC1, FC2, and FC3 were 0.70, 0.71, and 0.77, respectively. It should be noted that the evaluation of the 2018 data was carried out with the input dataset that was used for training. Thus, the test and training data share the same ice conditions as well as spatio-temporal coverage. As a result, κ ~~Since the training and test datasets were extracted from the same selected images, thus sharing the same overall ice conditions and spatial/temporal coverage, there~~ might ~~be~~ contain correlation which is not preferable for proper evaluation. Table 3 shows the confusion matrix for validation results from the 2019 data of which the accuracy of ~~. Comparing to the results from 2018 data, the high accuracies for ice free~~open water and old ice ~~were~~ was ~~maintained in~~at a similar level, compared to the 2018 data. ~~, while~~ Meanwhile, ~~those for~~for the accuracy of new ice, young ice, and first-year ice ~~were~~ decreased considerably. The differences between the results of FC1 and FC2 were insignificant~~, while~~whereas there were notable accuracy ~~decreases~~ degrades from FC2 to FC3~~, which is the opposite to the result in Table 2. This means~~. This result is opposite to the 2018 data inferring that the training with FC3 was overfitted~~,~~ and the ~~day of year~~day-of-the-year may not ~~help ice type classification if the training does not cover the whole seasonal cycle.~~correspond to the temperature, air-sea fluxes, or weather regimes. κ for FC1, FC2, and FC3 with the 2019 data were 0.67, 0.67, and 0.49, respectively.

To see how the denoising step in Section 2.2.2 led to improvements in the classification accuracies, the same training and evaluation were conducted for the same dataset without applying the textural noise correction (Table 4). In both FC1 and FC2, the accuracies improved for young ice (+8.2-9.8%) and first-year ice (+9.2-11.6%) which were most pronounced compared to those for open water (+1.7%) and old ice (+1.2-1.7%). On the contrary, a small accuracy decrease was observed for new ice (-2.8-4.7%). Nevertheless, the improvement in kappa (+0.05) demonstrates a clear improvement in the overall classification result.

Figure 7 shows a daily mosaic of Sentinel-1 SAR images over the study area and the classified ice map. For comparison, the NIC weekly ice chart is also displayed. Despite the SAR images had been acquired three days before the ice chart was published, ~~The~~ the ice edges of the ice chart match well with the SAR mosaic in most parts ~~probably~~ because the same SAR data ~~was~~ were used. ~~when the ice chart was made, although the SAR images had been acquired 3 days before the ice chart was published.~~ In ~~Overall~~overall, the discriminations between ice and non-ice, old ice and other ice types, and detection of new ice patches look reasonable. However, some young ice patches, for example the ice patches between the Svalbard archipelago, are misclassified as the first-year ice~~, and vice versa~~. Figure 8 shows another daily mosaic made by the images acquired on the same day of the ice chart publication. ~~Comparing~~ Considering notable ice drift in the backscattering images in Figure 7 and ~~Figure~~ 8~~, there was notable ice drift. The~~the SAR-based ice classification results in both figures look consistent, well in line with the ice drift. Although the weekly ice chart is supposed to represent the averaged ice status in the past few days, the actual ice distribution on the actual date of ~~weekly ice chart~~the publication can be largely different ~~from that in the ice chart~~. This example shows a clear potential of near-real time service of ice type classification.

To cope with the ambiguous classification for the ice types with low accuracy, we conducted a test with the further simplification of ice types by combining the new ice, young ice, and first-year ice into the "~~integrated~~mixed" first-year ice, and then training new classifiers. Table 5~~4~~ and ~~Table 5~~ 6 show the confusion matrices for the three~~3~~-class classifiers. κ for

FC1, FC2, and FC3 were 0.84, 0.86, and 0.92 in 2018 data, and 0.80, 0.80, and 0.53 in 2019 data, respectively. The dramatic increase in the accuracy ~~for~~ of the ~~integrated~~ mixed first-year ice ~~class~~ indicates that the misclassification for the new ice, young ice, and first-year ice ~~were~~ was mostly among themselves. However, the accuracy decrease from 2018 data to 2019 data was ~~in~~ at a similar level to the case of the five~~5~~-class classification. ~~, and t~~This could have been caused by ~~insufficient training of classifier and/or~~ inconsistent labeling in the reference ice chart.

Figure 9 shows an example of the inconsistent labeling in the reference ice chart. The SoDs from the NIC ice charts are superimposed on the Sentinel-1 backscattering images. The ~~type of the~~ same ice floe (red outline) is ~~annotated~~ classified differently in ~~the~~ two different ice charts (old ice on the left panel and first-year ice on the right panel)~~, while~~ although it looks almost the same in the corresponding SAR backscattering images. ~~Considering the ice edges in ice charts match well with those in the SAR backscattering images, thus the ices in the inner parts are also expected to be charted in the same time, i~~It should be noted that training with ice chart might have included ~~wrong~~ mislabeled small features ~~samples~~ even if the image selection based on ice edge matching was successful. Furthermore, the boundaries between different ice types in the ice chart are normally not as precise as those in the SAR image-based classification results. Therefore, the lower classification accuracies compared to those in the previous studies (80% in Zakhvatkina et al., 2013; 91.7% in Liu et al., 2015; 87.2% in Aldenhoff et al., 2018), which used manually classified ice maps as training and validation reference, are expected.

Unfortunately, the proposed algorithm has several limitations. First of all, the variations in radar backscattering and its corresponding image textures due to seasonal changes were not properly captured. Although day-of-the-year was tested as a seasonality variable in the FC3 feature configuration, the result did not show any improvement. This is because day-of-the-year might not correspond to the same temperature, fluxes, and weather regimes. Second, the proposed method struggles when the same sea ice is located on different marginal sides of SAR images because the incidence angle dependence could not be normalized perfectly. An example of such a failure can be seen along the image boundaries at 80N, 35E and 82.5N, 60E, approximately. Third, some artifacts were observed under an extreme marine condition. In the classified results in the bottom right panel of Figure 8, there is a misclassified FYI patch (yellow) in the open water area. According to the NOAA SAR wind image service, ANSWRS 2.0, the wind speed ranged from 17 to 21 m/s at the time of image acquisition heavily roughing the water surface.

4 Conclusion

A new semi-automated SAR-based sea ice type classification scheme was proposed in this study. For the first time several ice types can be successfully identified on Sentinel-1 SAR imagery. The main technological innovation is two-fold: i) minimized manual work in the preparation of training and validation reference data and ~~,~~ii) more objective evaluation of the SAR-based sea ice type classifier. A conventional approach for selecting training/testing data by anonymous~~a~~ human ice expert is undesirable not only because it is laborious, but also due to subjectivity and lack of standardization in the assessment of the automated classifier. Therefore, the performance from different literature sources cannot be intercompared directly.

Test results from the datasets of two winter seasons ~~dataset~~ acquired over the Fram Strait and the Barents Sea area showed overall accuracies of 85% and 58% and κ of 0.80 and 0.67 for the three~~3~~-class and five~~5~~-class ice type classifiers, respectively. These are slightly lower than the numbers in the previous studies, and the errors are attributed not only to the automated algorithm but also to the inconsistency of the ice charts and the high level of their generalization. Based on the results, we envisage that three-class ice type classification from SAR imagery would be useful for making a global sea ice type product like EUMETSAT OSI-403-C (Aaboe et al., 2014) with higher spatial resolution. The proposed approach importantly showed that a daily ice type mapping from the Sentinel-1 data is feasible~~,~~ and ~~it~~ can help ~~capturing~~ capture ~~more~~ details ~~in~~ of short-term changes in the stage of sea ice development. Based on the achieved results, we believe that the proposed approach may be efficiently used for operational ice charting services for supporting navigation in the Arctic.

Code/Data availability

Not applicable

Author contributions

JP and AK formulated the research plan, JP and AK developed the algorithm, JP implemented the algorithm and performed the data processing, JP, AK, MB, JW, MH, and HK carried out the analyses, and JP wrote the paper.

Competing interests

The authors declare that they have no conflict of interests.

~~Acknowledgements~~Financial support

This work was supported by the French Service Hydrographique et Océanographique de la Marine (SHOM) under SHOM-ImpSIM Project~~,~~ 111222, ~~and~~ the Research Council of Norway and the Russian Foundation for Basic Research under NORRUSS Project 243608, SONARC~~.~~, and the Korea Polar Research Institute under grant number PE20080.

References

Aaboe, S., Breivik, L.-A., and Eastwood, S.: Improvement of OSI SAF product of sea ice edge and sea ice type. EUMETSAT Meteorological Satellite Conference, Geneva (Switzerland), 22-26 September 2014.

Adnan, M. N., and Islam, M. Z., One-Vs-All Binarization Technique in the Context of Random Forest, Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges (Belgium), 22-24 April 2015.

Aldenhoff, W., Heuzé, C., and Eriksson, L.: Comparison of ice/water classification in Fram Strait from C- and L-band SAR imagery. Ann. Glaciol., 59(76pt2), 112-123. doi:10.1017/aog.2018.7, 2018.

Anand, R., Mehrotra, K., Mohan, C. K., and Ranka, S.: Efficient classification for multiclass problems using modular neural networks, IEEE T. Neural Networ., 6(1), 117-124, doi:10.1109/72.363444, 1995.

Barber, D. G. and LeDrew, E. F.: SAR sea ice discrimination using texture statistics: A multivariate approach. Photogramm. E. Rem. S., 57(4), 385-395, 1991.

Cohen, J.: A coefficient of agreement for nominal scales, Educ. Psychol. Meas., 20(1), 37–46, doi:10.1177/001316446002000104, 1960.

Cortes, C. and Vapnik, V.: Support-vector networks, Mach. Learn., 20(3), 273-297, doi: 10.1007/BF00994018, 1995.

Deng, H. and Clausi, D. A.: Unsupervised segmentation of synthetic aperture radar sea ice imagery using a novel Markov random field model," IEEE T. Geosci. Remote, 43(3), 528-538, doi:10.1109/TGRS.2004.839589, 2005.

Dierking, W.: Mapping of different sea ice regimes using images from Sentinel-1 and ALOS synthetic aperture radar, IEEE T. Geosci. Remote, 48(3), 1045-1058, doi:10.1109/TGRS.2009.2031806, 2010.

ESRI (Environmental Systems Research Institute, Inc.): ESRI Shapefile Technical Description, An ESRI White Paper, 1998. Available at: http://downloads.esri.com/support/whitepapers/mo_/shapefile.pdf

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems?, J. Mach. Learn. Res., 15, 3133-3181, 2014.

GDAL/OGR contributors: GDAL/OGR Geospatial Data Abstraction software Library, Open Source Geospatial Foundation, 2019. URL https://gdal.org

Haralick, R. M., Shanmugam, K., Dinstein, I.: Textural features for image classification, IEEE T. SYST. MAN. CY.-S., SMC-3(6), 610-621, doi:10.1109/TSMC.1973.4309314, 1973.

Hersbach, H., Stoffelen, A., and de Haan, S.: An improved C-band scatterometer ocean geophysical model function: CMOD5, J. Geophys. Res., 112, C03006, doi:10.1029/2006JC003743, 2007.

Heinrichs, J. F., Cavalieri, D. J., and Markus, T.: Assessment of the AMSR-E sea ice concentration product at the ice edge using RADARSAT-1 and MODIS imagery, IEEE T. Geosci. Remote, 44(11), 3070-3080, doi: 10.1109/TGRS.2006.880622, 2006.

Ho, T. K.: The random subspace method for constructing decision forests, IEEE T. Pattern Anal., 20(8), 832-844, doi:10.1109/34.7096011998, 1998.

Isleifson, D., Galley, R. J., Barber, D. G., Landy, J. C., Komarov, A. S., and Shafai, L.: A study on the C-band polarimetric scattering and physical characteristics of frost flowers on experimental sea ice," IEEE T. Geosci. Remote, 52(3), 1787-1798, doi:10.1109/TGRS.2013.2255060, 2014.

JCOMM (Joint WMO-IOC Technical Commission for Oceanography and Marine Meteorology): Ice chart colour code standard, JCOMM Technical Report No. 24, Tech. rep., World Meteorological Organization, Geneva, Switzerland, 2014a.

JCOMM (Joint WMO-IOC Technical Commission for Oceanography and Marine Meteorology): SIGRID-3: a vector archive format for sea ice georeferenced information and data, JCOMM Technical Report No. 23, Tech. rep., World Meteorological Organization, Geneva, Switzerland, 2014b.

Johannessen, O.M., Alexandrov, V., Frolov, I.Y., Sandven, S., Pettersson, L.H., Bobylev, L.P., Kloster, K., Smirnov, V.G., Mironov, Y.U., and Babich, N.G.: Remote sensing of sea ice in the Northern Sea route: Studies and applications, Springer, Berlin, Heidelberg, doi:10.1007/978-3-540-48840-8, 2006.

Karvonen, J.: A sea ice concentration estimation algorithm utilizing radiometer and SAR data. The Cryosphere, 8, 1639-1650, doi:10.5194/tc-8-1639-2014, 2014.

Karvonen, J., Vainio, J., Marnela, M., Eriksson, P., and Niskanen, T.: A comparison between high-resolution EO-based and ice analyst-assigned sea ice concentrations, IEEE J. Sel. Top. Appl., 8(4), 1799-1807, 2015.

Karvonen, J.: Baltic sea ice concentration estimation using Sentinel-1 SAR and AMSR2 microwave radiometer data, IEEE T. Geosci. Remote, 55(5), 2871-2883, doi:10.1109/TGRS.2017.2655567, 2017.

Keller, M. R., Gifford, C. M., Walton, W. C., and Winstead, N. S.: Ice analysis based on active and passive radar images, U.S. Patent 9652674 B2, May 16, 2017. Available at: https://patents.google.com/patent/US9652674~~Keller, M.: Active/passive dual polarization sea ice detection, NISAR Applications Workshop: Sea Ice, Maryland, USA, 23 Jun. 2017. Available at: https://www.star.nesdis.noaa.gov/sod/mecb/sar/NISAR_Sea_Ice_Workshop/Keller_et_al_23_June_2017_Active_Passive_Sea_Ice_Detection.pptx~~

Kohavi, R.: A Study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of the 14th international joint conference on Artificial intelligence, Montreal, Canada, 20-25 August,1995, 2, 1137-1143, 1995.

Leigh, S., Wang, Z., and Clausi, D. A.: Automated ice-water classification using dual polarization SAR satellite imagery, IEEE T. Geosci. Remote, 52(9), 5529–5539, doi:10.1109/TGRS.2013.2290231, 2014.

Liu, H., Guo, H., and Zhang, L.: SVM-based sea ice classification using textural features and concentration from RADARSAT-2 dual-pol ScanSAR data, IEEE J. Sel. Top. Appl., 8(4), 1601–1613, doi:10.1109/JSTARS.2014.2365215, 2015.

Louppe, G.: Understanding random forests: From theory to practice, PhD Thesis, U. of Liege, 2014.

Mahmud, M. S., Geldsetzer, T., Howell, S. E. L., Yackel, J. J., Nandan, V., and Scharien, R. K.: Incidence angle dependence of HH-polarized C- and L-band wintertime backscatter over Arctic sea ice, IEEE T. Geosci. Remote, 56(11), 6686-6698, doi: 10.1109/TGRS.2018.2841343, 2018.

Mäkynen, M. and Karvonen, J.: Incidence angle dependence of first-year sea ice backscattering coefficient in Sentinel-1 SAR imagery over the Kara Sea, IEEE T. Geosci. Remote, 55(11), 6170-6181, doi:10.1109/TGRS.2017.2721981, 2017.

Mäkynen, M. P., Manninen, A. T., Simila, M. H., Karvonen, J. A., and Hallikainen, M. T.: Incidence angle dependence of the statistical properties of C-band HH-polarization backscattering signatures of the Baltic Sea ice, IEEE T. Geosci. Remote, 40(12), 2593-2605, doi:10.1109/TGRS.2002.806991, 2002.

Miranda, N.: S-1 constellation product performance status, SeaSAR 2018, Frascati, Italy, 7-10 May 2018. Available at: http://seasar2018.esa.int/files/presentation216.pdf

Park, J.-W., Korosov, A. A., Babiker, M., Sandven, S., and Won, J.-S.: Efficient thermal noise removal for Sentinel-1 TOPSAR cross-polarization channel, IEEE T. Geosci. Remote, 56(3), 1555–1565, doi:10.1109/TGRS.2017.2765248, 2018.

Park, J.-W., Won, J.-S., Korosov, A. A., Babiker, M., and Miranda, N.: Textural noise correction for Sentinel-1 TOPSAR cross-polarization channel images, IEEE T. Geosci. Remote, 57(6), 4040-4049, doi:10.1109/TGRS.2018.2889381, 2019.

Partington, K., Flynn, T., Lamb, D., Bertoia, C., and Dedrick, K: Late twentieth century Northern Hemisphere sea-ice record from the U.S. National Ice Center ice charts, J. Geophys. Res., 108(C11), doi:10.1029/2002JC001623, 2003.

Pastusiak, T.: Accuracy of sea ice data from remote sensing methods, its impact on safe speed determination and planning of voyage in ice-covered areas, International Journal on Marine Navigation and Safety of Sea Transportation, 10(2), 229-248, doi:10.12716/1001.10.02.06, 2016.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12, 2825-2830, doi:10.1016/j.patcog.2011.04.006, 2011.

~~Ramírez, J., Górriz, J. M., Ortiz, A., Martínez-Murcia, F. J., Segovia, F., Salas-Gonzalez, D., Castillo-Barnes, D., Illán, I. A., and Puntonet, C. G.: Ensemble of random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares, J. Neurosci. Meth., 302, 47-57, doi: 10.1016/j.jneumeth.2017.12.005, 2018.~~

Ressel, R., Frost, A., and Lehner, S.: A neural network-based classification for sea ice types on X-band SAR images, IEEE J. Sel. Top. Appl., 8(7), 3672-3680, doi:10.1109/JSTARS.2015.2436993, 2015.

Richards, F. J.: A flexible growth function for empirical use, J. Exp. Bot., 10(29), 290-300, doi: 10.1093/jxb/10.2.290, 1959.

Scheuchl, B., Flett, D., Caves, R., and Cumming, I.: Potential of RADARSAT-2 data for operational sea ice monitoring, Can. J. Remote Sens., 30(3), 448-471, 2004.

Shokr, M. E.: Evaluation of second-order texture parameters for sea ice classification from radar images, J. Geophys. Res.-Oceans, 96(C6), 10625-10640, 1991.

Smedsrud, L. H., Halvorsen, M. H., Stroeve, J. C., Zhang, R., and Kloster, K.: Fram Strait sea ice export variability and September Arctic sea ice extent over the last 80 years, The Cryosphere, 11, 65-79, doi:10.5194/tc-11-65-2017, 2017.

Soh, L.-K. and Tsatsoulis, C.: Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices, IEEE T. Geosci. Remote, 37(2), 780-795, doi:10.1109/36.752194, 1999.

Wang, L., Scott, K. A., and Clausi, D. A., Sea ice concentration estimation during freeze-up from SAR imagery using a convolutional neural network, Remote Sens-Basel, 9(5), doi:10.3390/rs9050408, 2017.

Zakhvatkina, N. Y., Alexandrov, V. Y., Johannessen, O. M., Sandven, S., and Frolov, I. Y.: Classification of sea ice types in ENVISAT synthetic aperture radar images, IEEE T. Geosci. Remote, 51(5), 2587-2600, doi:10.1109/TGRS.2012.2212445, 2013.

Zakhvatkina, N., Korosov, A., Muckenhuber, S., Sandven, S., and Babiker, M.: Operational algorithm for ice-water classification on dual-polarized RADARSAT-2 images, The Cryosphere, 11, 33-46, doi:10.5194/tc-11-33-2017, 2017.

Figures

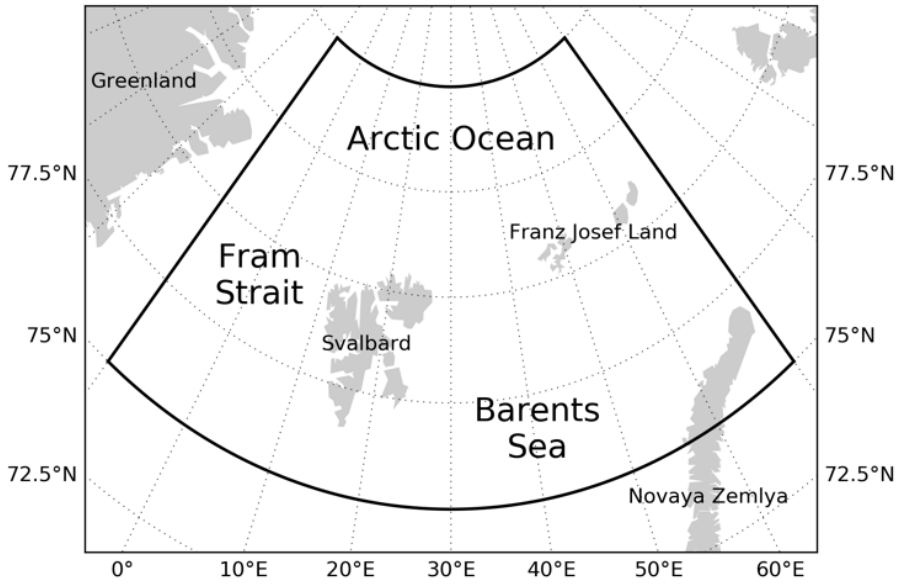


Figure 1: Study area.

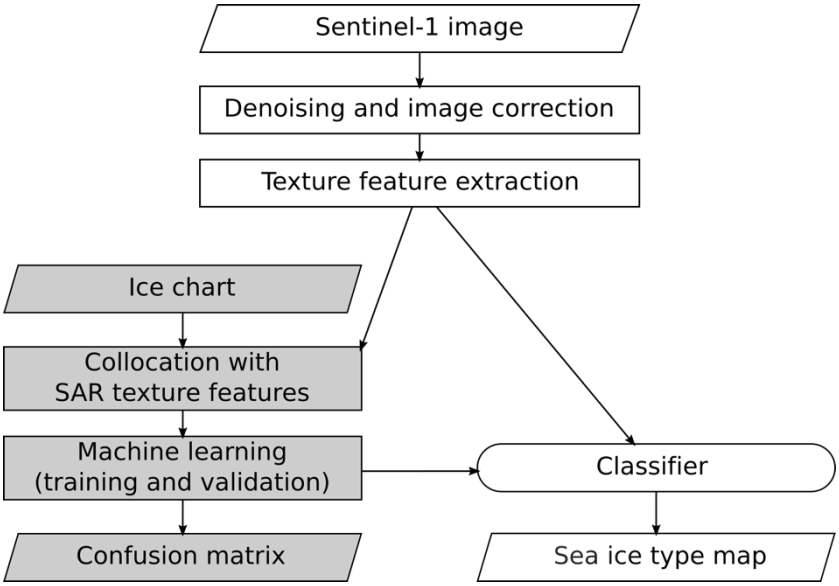
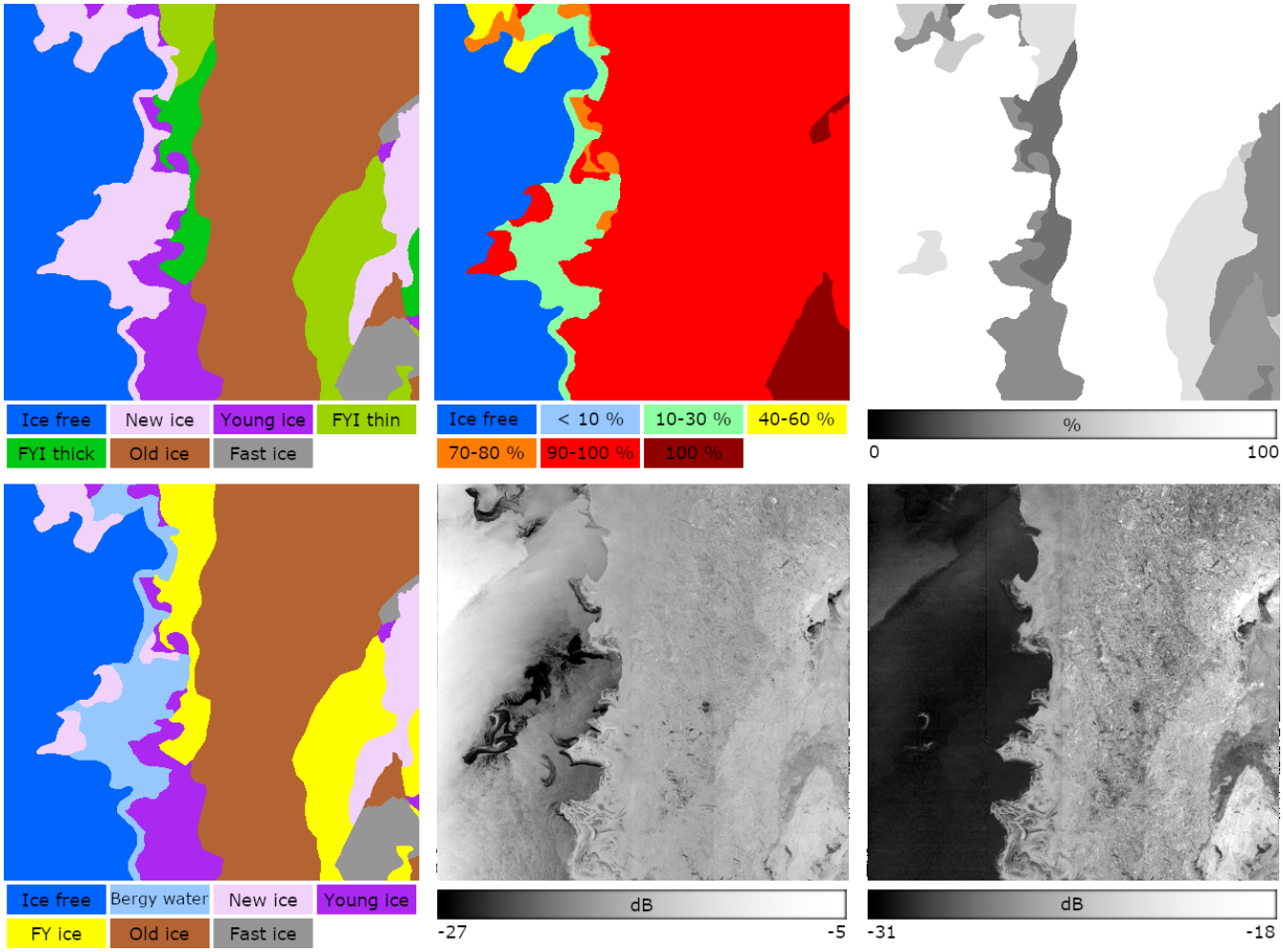


Figure 2: Processing flowchart of the proposed algorithm. The ~~Gray~~ gray color shows the training phase.


Ice free
New ice
Young ice
FYI thin
FYI thick
Old ice
Fast ice
Ice free
< 10 %
10-30 %
40-60 %
70-80 %
90-100 %
100 %
%
0
100
Ice free
Bergy water
New ice
Young ice
FY ice
Old ice
Fast ice
dB
-27
-5
dB
-31
-18

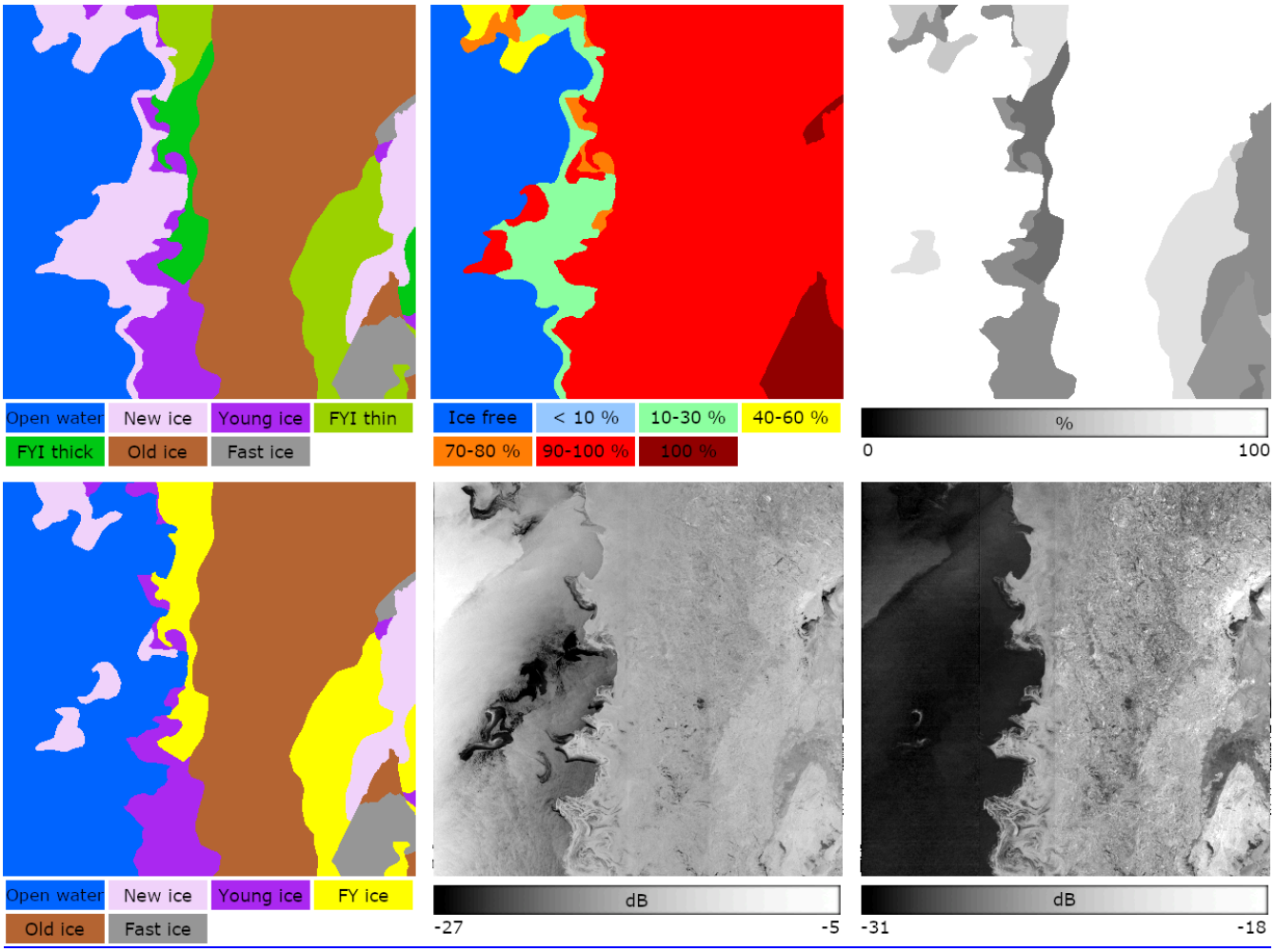
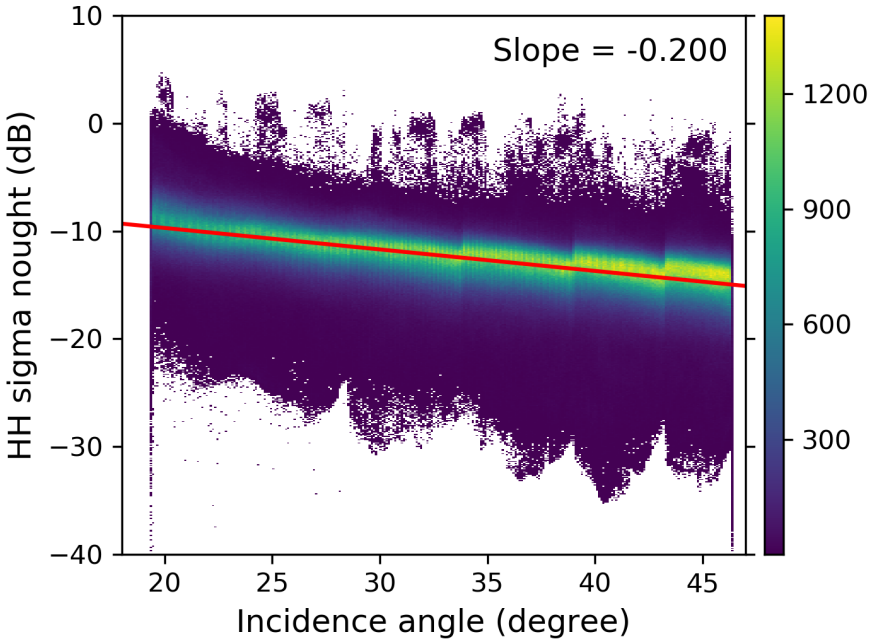
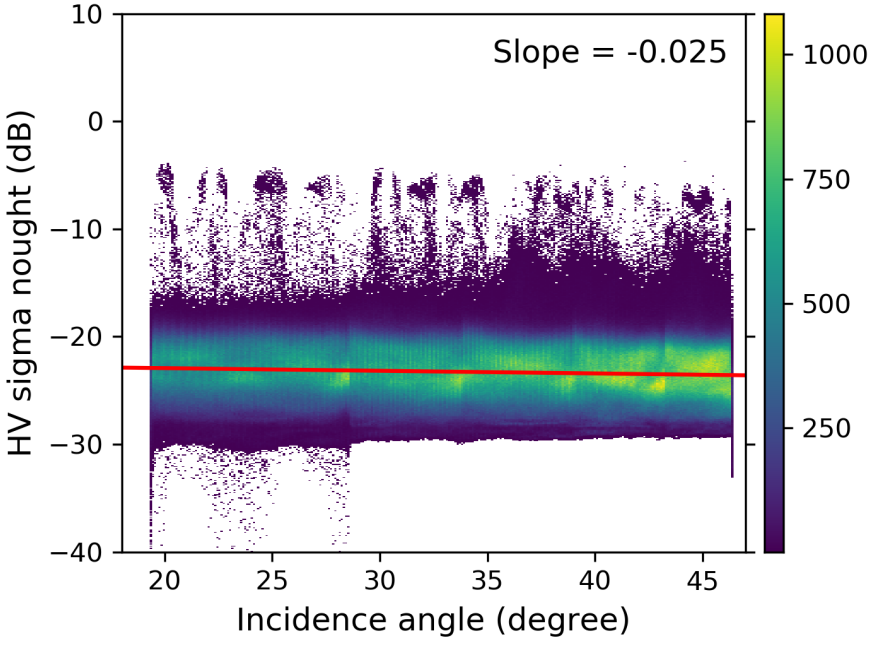


Figure 3: An example of the ice chart preprocessing. From the ice chart, stage of development (SoD; Top left), ice concentration (CT; Top center), and partial concentration of the dominant ice type (CP; Top right) maps are extracted. Then, some of the different SoDs are merged (e.g., thin and thick first-year ices are merged into a single label as first-year ice), and the area with low ice concentration is labeled as ~~bergy ice~~open water. The processed map of SoD (Bottom left) is ~~used to~~ related with ~~texture~~ textural features extracted from HH and HV polarization images (Bottom center and bottom right). Note that the NIC ice chart which was published on January 25, 2018, and the Sentinel-1 product S1B_EW_GRDM_1SDH_20180122T075237_20180122T075337_009281_010A4D_65AA acquired over the Fram Strait were used in this example.


Slope = -0.200
HH sigma nought (dB)
Incidence angle (degree)
10
0
−10
−20
−30
−40
20
25
30
35
40
45
1200
900
600
300


Slope = -0.025
HV sigma nought (dB)
Incidence angle (degree)
10
0
−10
−20
−30
−40
20
25
30
35
40
45
1000
750
500
250

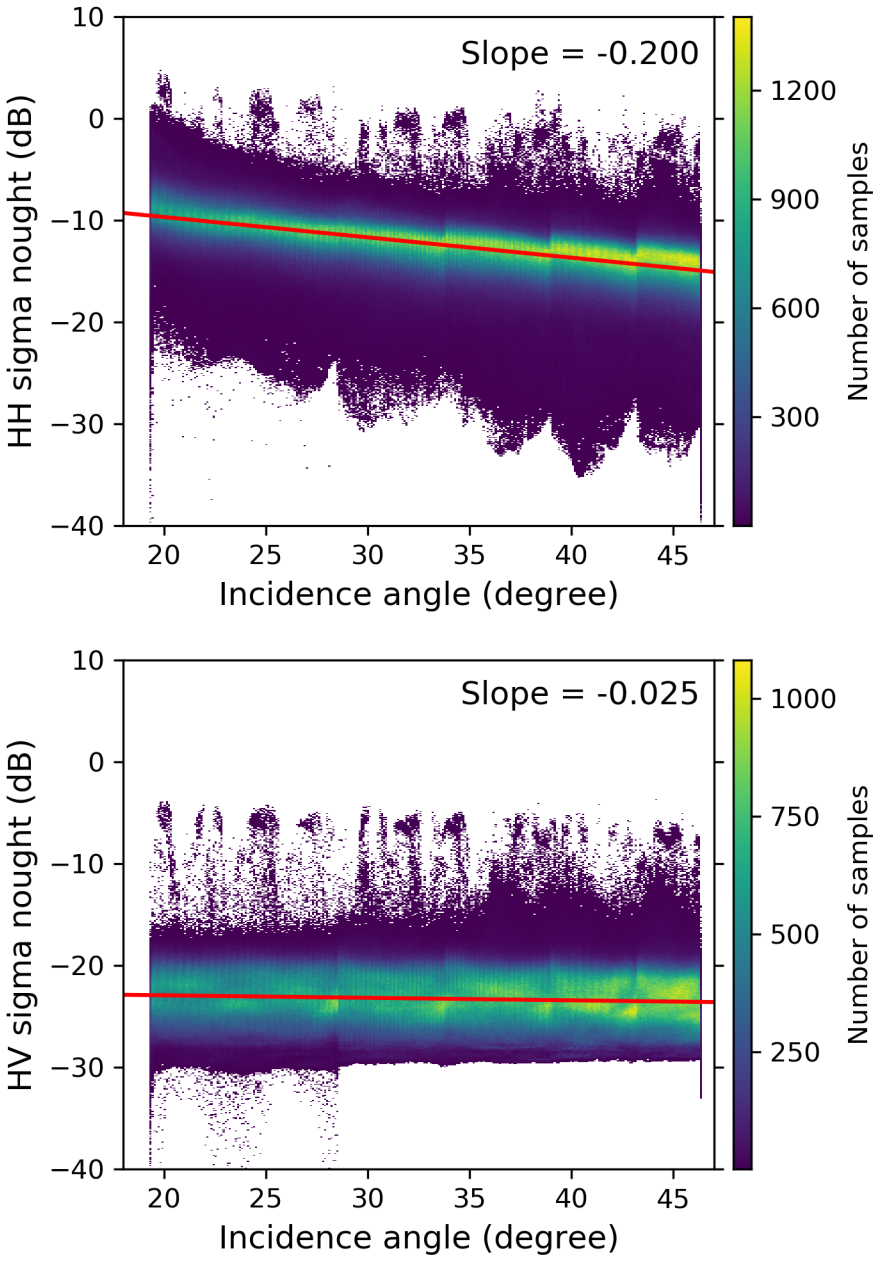


Figure 4: Two-dimensional histograms of incidence angle versus sigma nought for HH (top) and HV (bottom) polarization channels. Pixels covering various types of sea ice were merged so that the averaged property can be estimated. The best fit linear trends are shown with red lines.

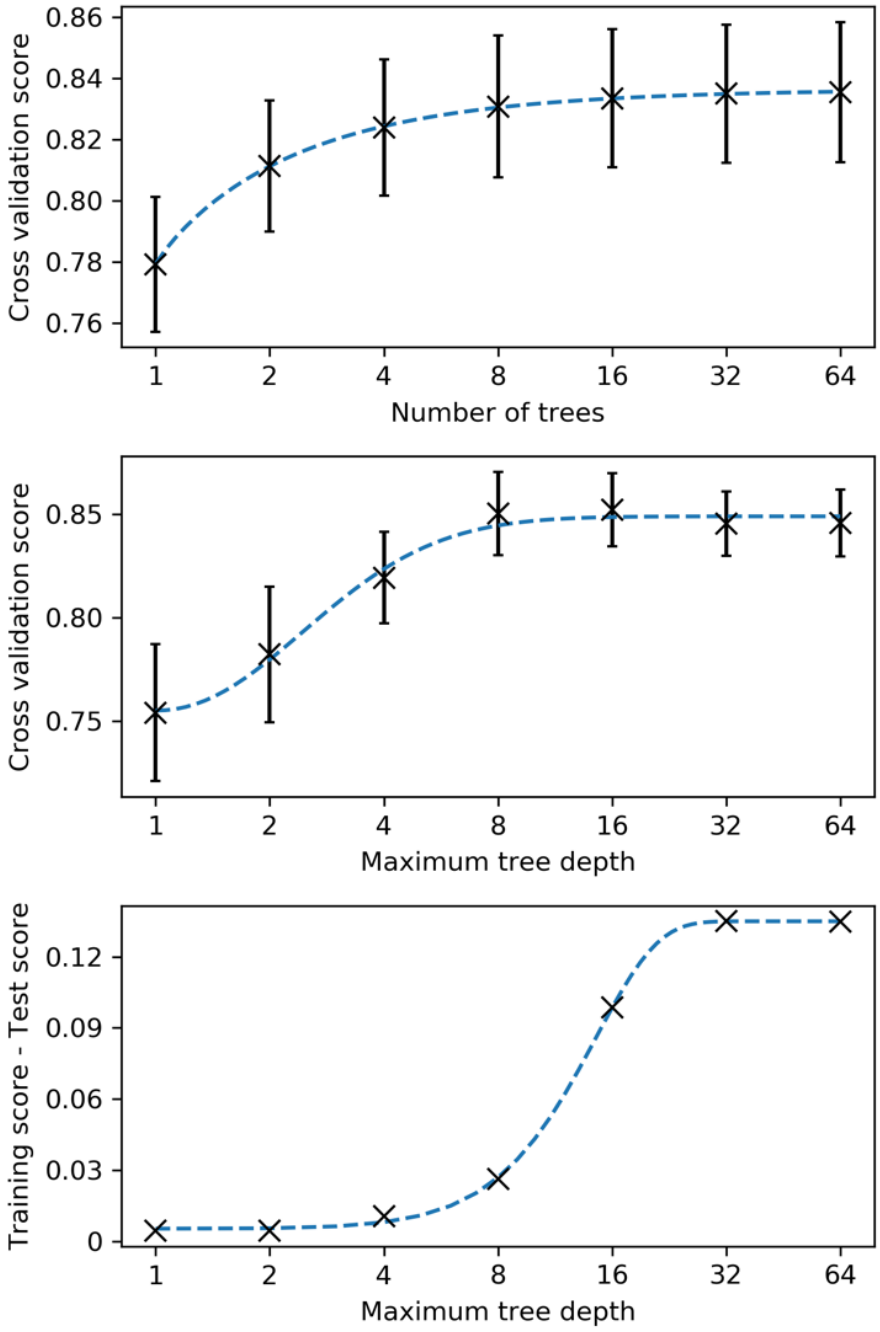


Figure 5: Hyperparameter optimization using grid search results (cross). Dashed lines represent the best-fit Richards's ~~s~~ Curve. (Top panel) The optimal values are extracted from the locations where the score increments per unit of each hyperparameter become lower than a threshold (e.g., 0.001). (Center panel) If the curve does not fit the grid search results well, (Bottom panel) the difference between training and test scores is used to find the locations where it does not exceed a threshold (e.g., 0.03) in order to avoid overfitting.

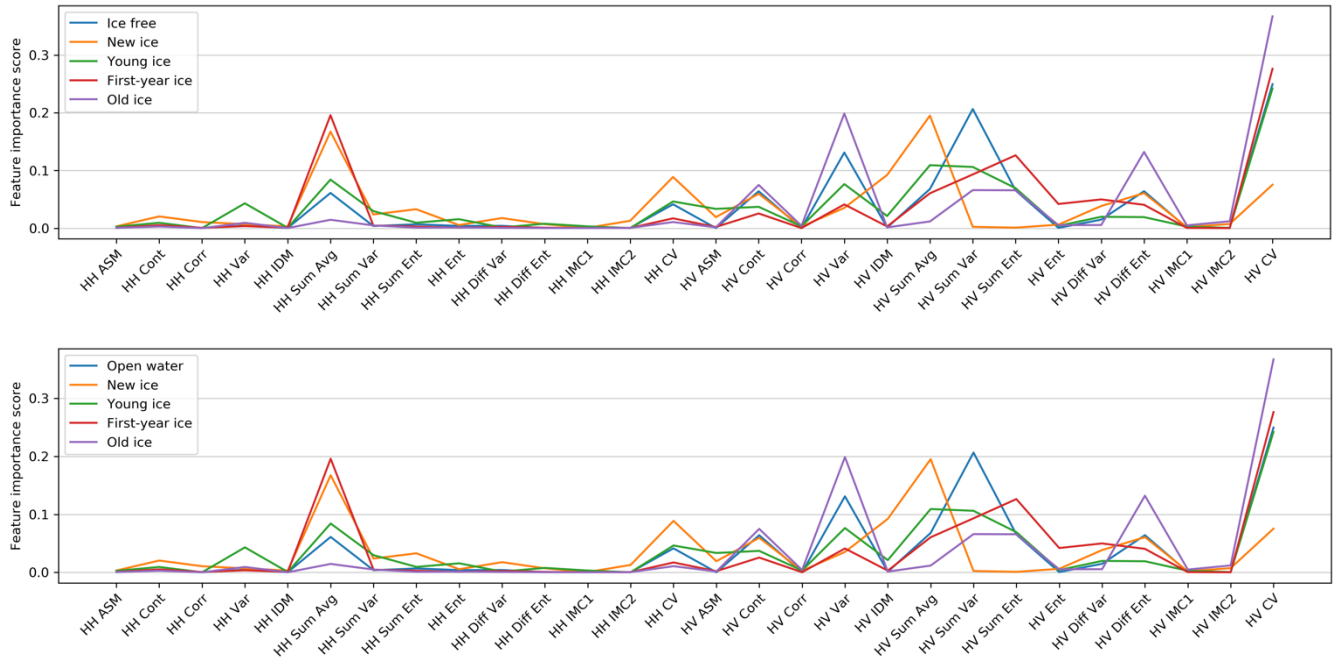
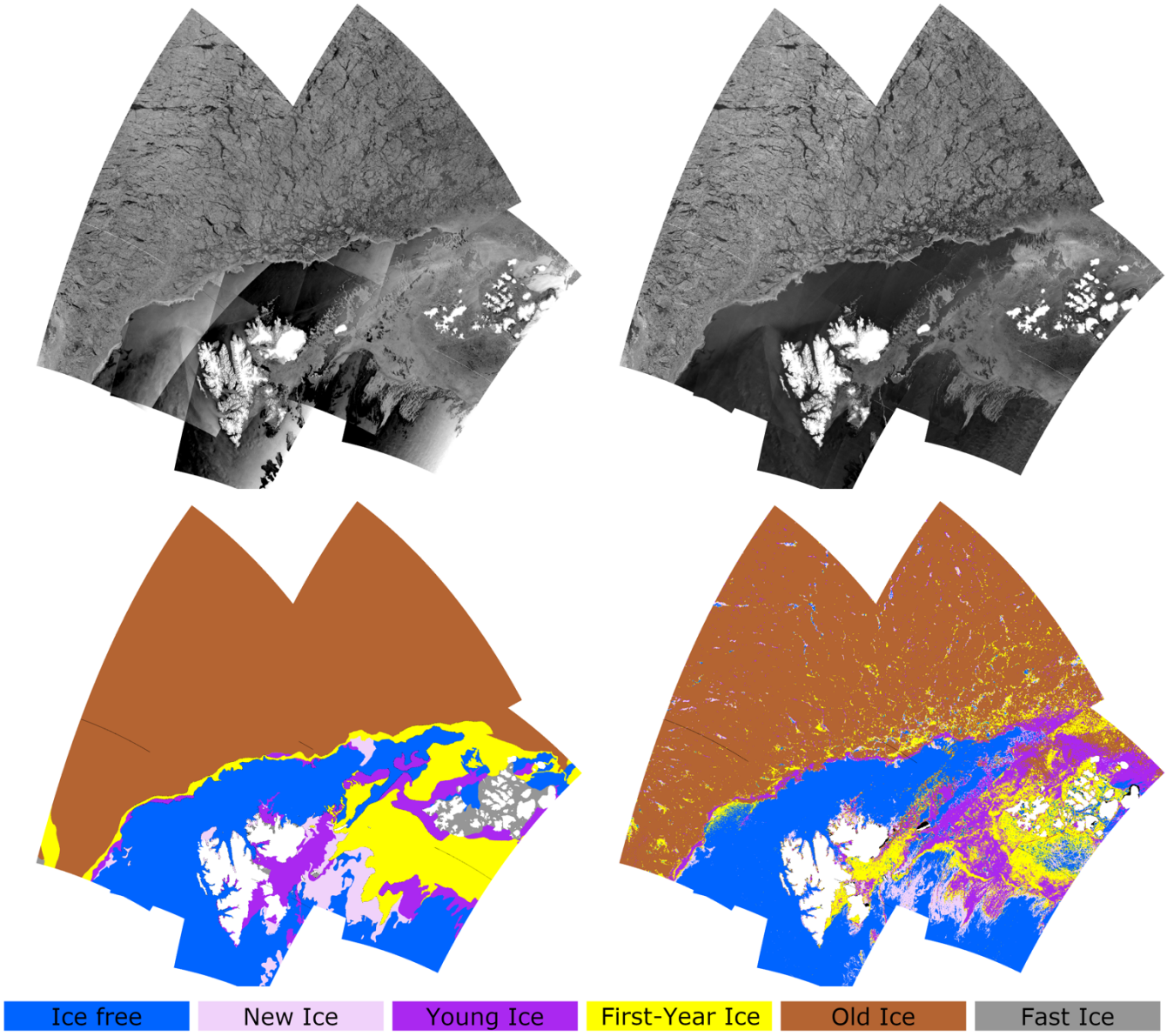


Figure 6: Feature importances of the binary sub-classifiers. ASM: angular second moment; Cont: contrast; Corr: correlation; Var: variance; IDM: inverse difference moment; Sum Avg: sum average; Sum Var: sum variance; Sum Ent: sum entropy; Ent: entropy; Diff Var: difference variance; Diff Ent: difference entropy; IMC: information measures of correlation; CV: coefficient of variation. For definitions of each parameters, please refer to Haralick et al., 1973.


Ice free
New Ice
Young Ice
First-Year Ice
Old Ice
Fast Ice

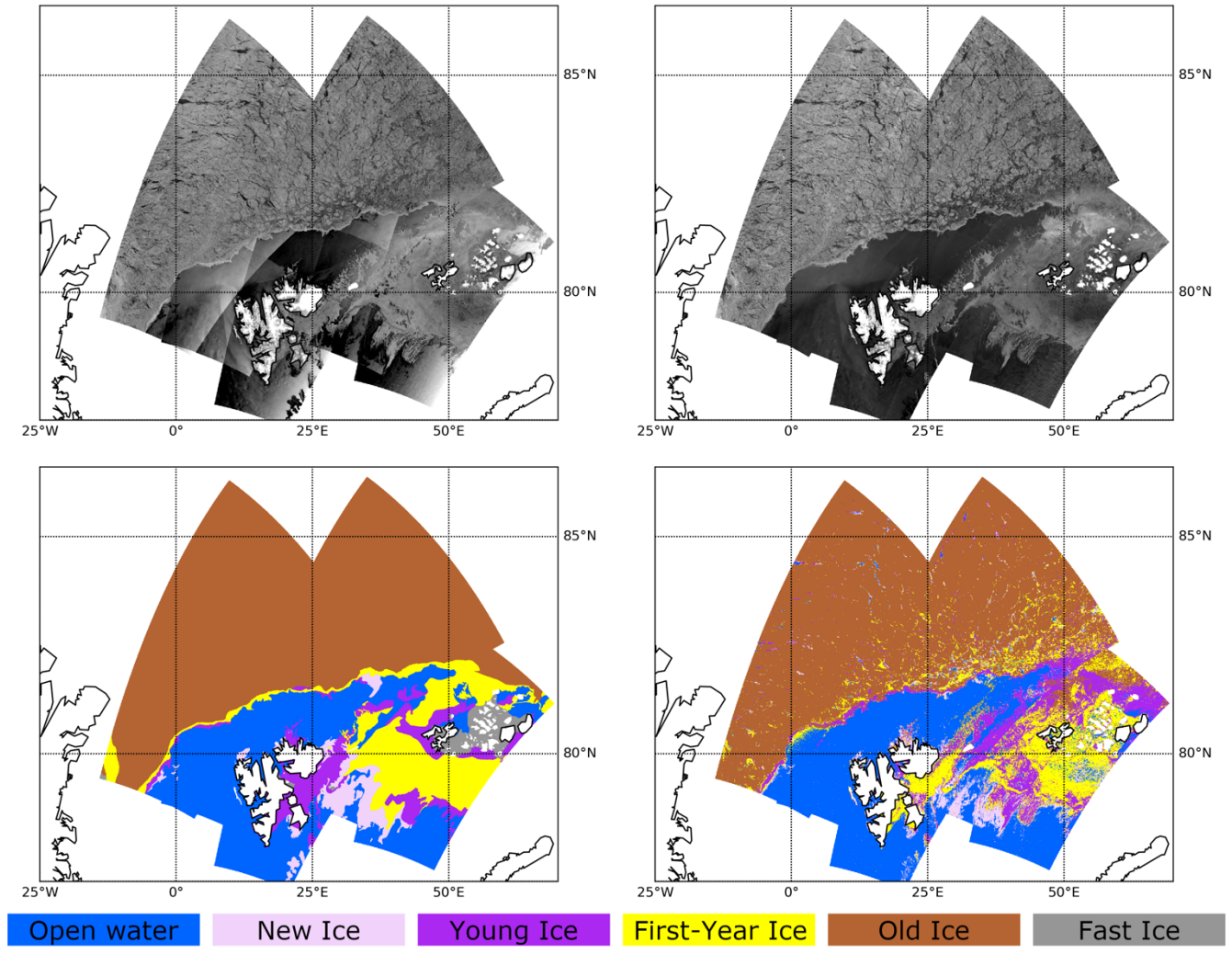
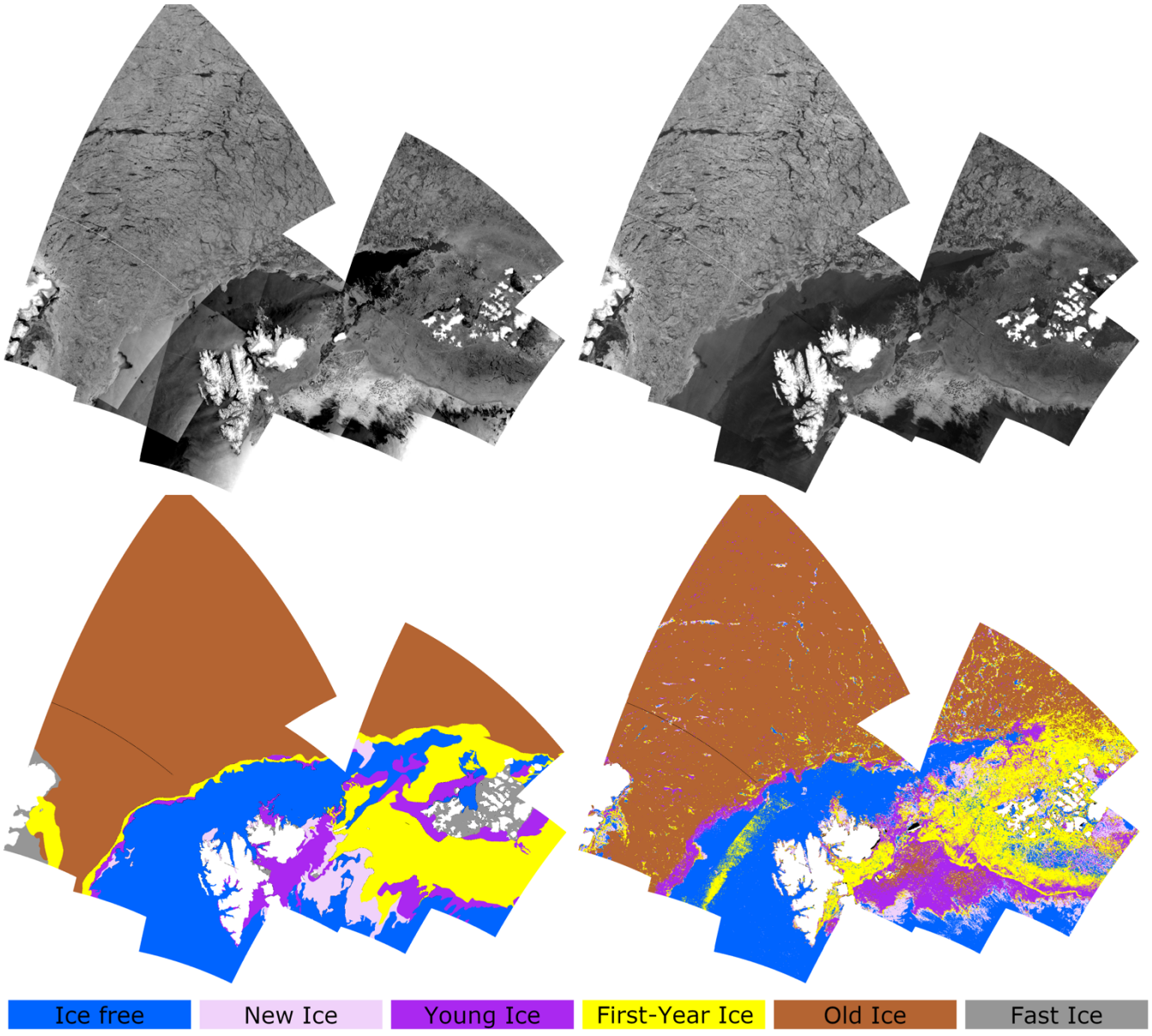


Figure 7: One~~1~~-day mosaics of Sentinel-1A/1B images (Top left: HH, Top right: HV) and the ice classification result (Bottom right) ~~for~~ on 5 February 2019. The ~~date of~~ publication date of the reference weekly ice chart ~~(Bottom left)~~ is 8 February 2019 (Bottom left).


Ice free
New Ice
Young Ice
First-Year Ice
Old Ice
Fast Ice

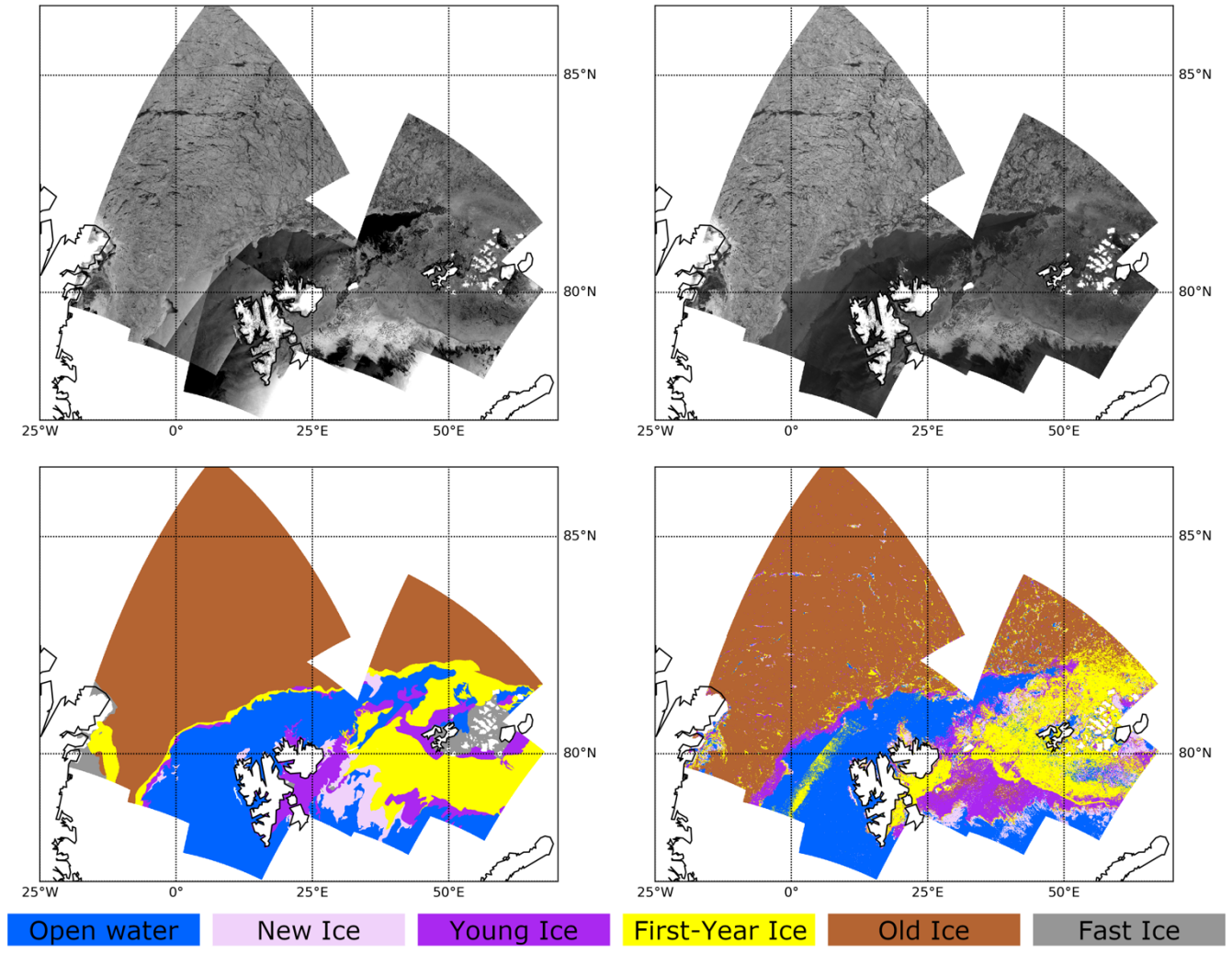
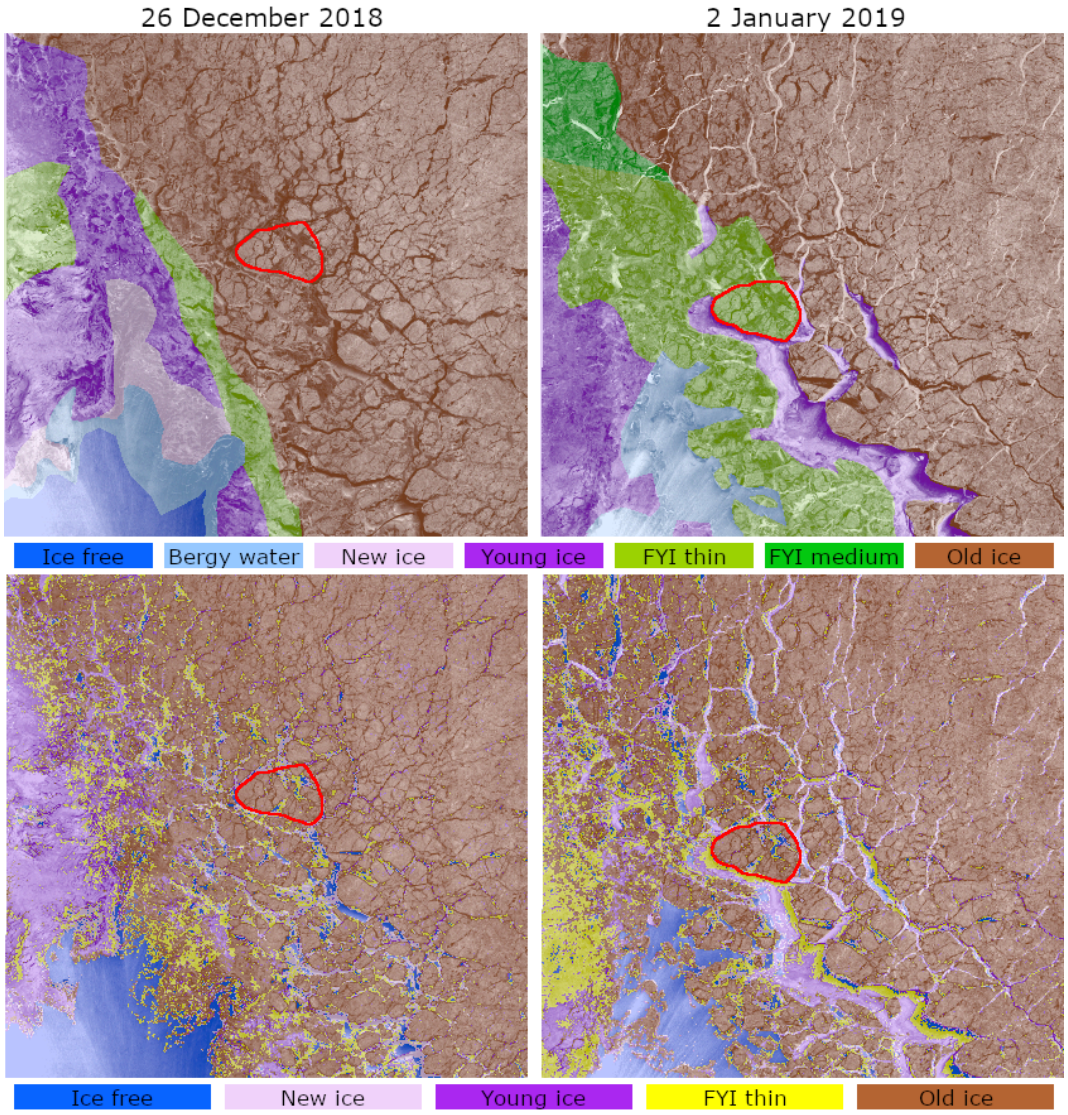


Figure 8: One~~1~~-day mosaics of Sentinel-1A/1B images (Top left: HH, Top right: HV) and the ice classification result (Bottom right) ~~for~~ on 8 February 2019. The ~~date of~~ publication date of the reference weekly ice chart ~~(Bottom left)~~ is 8 February 2019 (Bottom left).

5


26 December 2018
2 January 2019
Ice free
Bergy water
New ice
Young ice
FYI thin
FYI medium
Old ice
Ice free
New ice
Young ice
FYI thin
Old ice

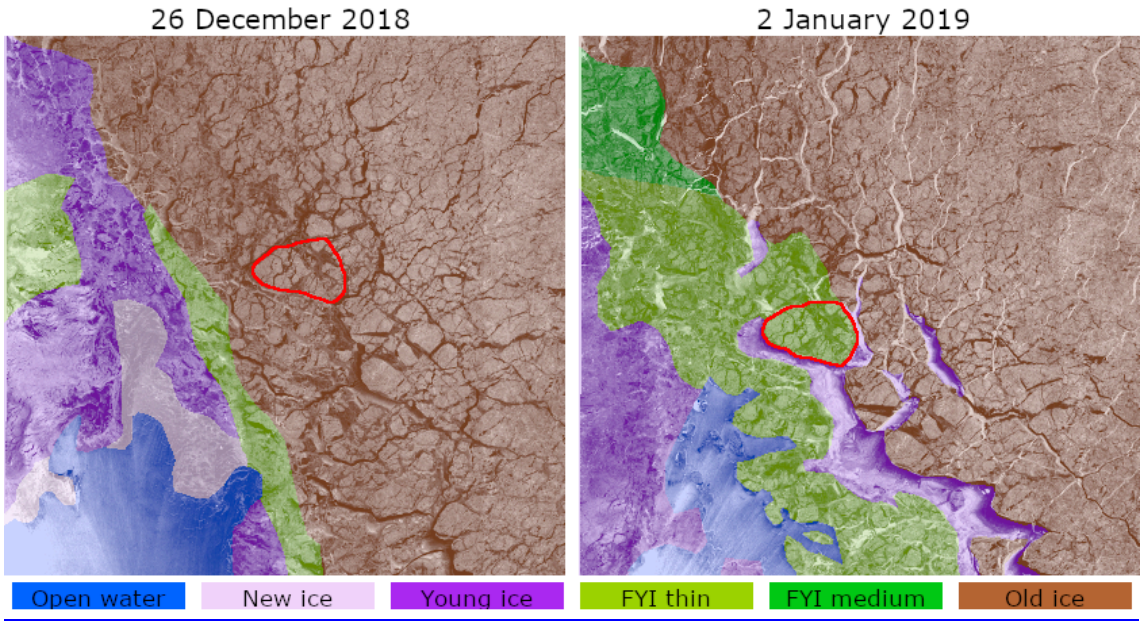


Figure 9: An example of the inconsistency ~~in ice types in~~of the ice charts. The SoDs from the NIC ice charts on different dates (26 December 2018 and 2 January 2019) are superimposed on the Sentinel-1 backscattering ~~images~~image of the corresponding dates. The ~~type of the~~ same ice floe (red outline) is ~~annotated~~ classified differently in each ice chart ~~the two ice charts~~ (old ice on the left panel and first-year ice on the right panel) ~~while it looks almost the same~~despite of the similarity in the SAR backscattering images.

Tables

Table 1: ~~Values of hyper parameters~~Hyperparameters used for grid search

Parameters	Values						
N_T	1	2	4	8	16	32	64
D	1	2	4	8	16	32	64
N_F	1	2	4	8	16	28	

Table 2: Confusion matrix ~~for~~ of the five~~5~~-class RF classifier which was trained ~~for 2018 data~~with and applied to the 2018 dataset

		~~Predicted~~														
		~~IF (ice free)~~			~~NI (new ice)~~			~~YI (young ice)~~			~~FYI (first-year ice)~~			~~OI (old ice)~~		
	~~case~~	~~FC1~~	~~FC2~~	~~FC3~~	~~FC1~~	~~FC2~~	~~FC3~~	~~FC1~~	~~FC2~~	~~FC3~~	~~FC1~~	~~FC2~~	~~FC3~~	~~FC1~~	~~FC2~~	~~FC3~~
~~Actual~~	~~IF~~	~~**93.7**~~	~~**94.6**~~	~~**95.6**~~	~~1.8~~	~~1.4~~	~~0.9~~	~~0.4~~	~~0.4~~	~~0.4~~	~~4.1~~	~~3.7~~	~~3.1~~	~~0.0~~	~~0.0~~	~~0.0~~
	~~NI~~	~~20.4~~	~~19.1~~	~~18.7~~	~~**32.5**~~	~~**33.8**~~	~~**58.3**~~	~~31.4~~	~~31.4~~	~~14.6~~	~~13.3~~	~~12.8~~	~~5.9~~	~~2.5~~	~~2.9~~	~~2.6~~
	~~YI~~	~~2.0~~	~~2.0~~	~~1.9~~	~~4.5~~	~~3.9~~	~~6.9~~	~~**60.5**~~	~~**59.1**~~	~~**61.3**~~	~~26.5~~	~~29.4~~	~~25.2~~	~~6.5~~	~~5.7~~	~~4.6~~
	~~FYI~~	~~4.4~~	~~4.2~~	~~3.4~~	~~3.1~~	~~2.8~~	~~2.9~~	~~22.3~~	~~19.8~~	~~17.8~~	~~**56.8**~~	~~**60.7**~~	~~**64.5**~~	~~13.3~~	~~12.5~~	~~11.5~~
	~~OI~~	~~0.3~~	~~0.3~~	~~0.4~~	~~0.9~~	~~0.9~~	~~1.7~~	~~5.8~~	~~5.3~~	~~3.6~~	~~7.9~~	~~7.6~~	~~6.2~~	~~**85.1**~~	~~**85.9**~~	~~**88.1**~~
		Predicted														
		OW (open water)			NI (new ice)			YI (young ice)			FYI (first-year ice)			OI (old ice)		
	case	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3
Actual	OW	**94.5**	**95.2**	**96.7**	1.4	1.1	0.6	0.3	0.3	0.3	3.7	3.4	2.4	0.0	0.0	0.0
	NI	19.3	17.3	14.8	**33.1**	**38.9**	**68.7**	33.3	31.1	9.3	12.1	10.6	5.5	2.2	2.1	1.7
	YI	1.9	1.8	1.6	3.8	3.6	6.7	**62.3**	**62.8**	**64.5**	26.1	27.5	23.8	5.9	4.3	3.4
	FYI	4.2	3.6	2.4	2.6	2.6	2.1	21.7	20.6	15.6	**58.1**	**61.1**	**69.8**	13.4	12.1	10.1
	OI	0.3	0.3	0.4	0.6	0.8	1.4	5.8	5.0	3.0	7.3	7.3	4.1	**86.0**	**86.7**	**91.2**

Table 3: Confusion matrix ~~for~~ of the five~~5~~-class RF classifier which was trained ~~for~~ with 2018 dataset and applied to the 2019 dataset

		~~Predicted~~

		IF (ice free)			NI (new ice)			YI (young ice)			FYI (first-year ice)			OI (old ice)		
	case	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3
Actual	IF	**89.6**	**90.0**	**87.3**	3.5	3.5	5.1	1.1	1.1	1.7	5.7	5.4	5.9	0.0	0.0	0.0
	NI	19.9	22.9	31.2	**29.1**	**25.3**	**20.2**	40.3	40.7	41.1	8.0	8.5	5.7	2.8	2.7	1.8
	YI	7.3	7.2	6.9	4.1	3.6	2.3	**42.7**	**41.4**	**50.4**	36.2	38.9	33.6	9.7	8.8	6.8
	FYI	6.2	5.8	5.9	4.3	3.9	2.0	24.4	23.3	25.1	**39.2**	**41.9**	**47.6**	25.9	25.0	19.4
	OI	0.6	0.6	0.6	1.4	1.3	0.5	2.4	2.7	7.1	3.1	3.5	21.1	**92.5**	**92.0**	**70.8**
		Predicted														
		OW (open water)			NI (new ice)			YI (young ice)			FYI (first-year ice)			OI (old ice)		
	case	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3
Actual	OW	**90.1**	**90.6**	**85.4**	3.1	2.7	5.7	1.0	1.1	2.0	5.8	5.7	6.9	0.0	0.0	0.0
	NI	20.1	24.5	28.3	**28.0**	**23.0**	**23.9**	42.0	42.4	40.9	7.6	7.9	5.5	2.4	2.1	1.4
	YI	6.7	6.1	6.3	3.3	3.4	3.1	**44.7**	**44.6**	**51.5**	36.0	38.2	33.5	9.3	7.7	5.7
	FYI	5.4	4.4	4.9	3.6	3.8	2.7	25.8	25.3	27.5	**38.9**	**42.0**	**46.0**	26.3	24.5	18.9
	OI	0.5	0.5	0.5	1.3	1.2	0.7	2.7	3.0	7.7	2.8	3.6	24.9	**92.7**	**91.7**	**66.3**

Table 4: Classification accuracies before and after applying textural denoising

class	case								
	FC1			FC2			FC3		
	Thermal denoising only	Textural denoising applied	difference	Thermal denoising only	Textural denoising applied	difference	Thermal denoising only	Textural denoising applied	difference
OW	88.4	90.1	+1.7	88.9	90.6	+1.7	88.0	85.4	-2.6
NI	30.2	28.0	-2.8	27.7	23.0	-4.7	31.8	23.9	-7.9
YI	34.9	44.7	+9.8	36.2	44.6	+8.2	43.4	51.5	+8.1
FYI	29.3	38.9	+9.6	30.4	42.0	+11.6	38.0	47.0	+9.0
OI	91.5	92.7	+1.2	90.3	91.7	+1.4	75.2	66.3	-8.9
kappa	0.62	0.67	+0.05	0.62	0.67	+0.05	0.54	0.49	-0.05

Table 54: Confusion matrix for of the three3-class RF classifier which were trained for 2018 data and applied to the 2018 dataset

		Predicted								
		IF (ice free)			FYI (first-year ice)			OI (old ice)		
	Case	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3
Actual	IF	**96.5**	**96.7**	**96.9**	3.5	3.3	3.1	0.0	0.0	0.0
	MFYI	5.8	5.7	4.8	**84.5**	**85.8**	**87.2**	9.8	8.6	7.9
	OI	0.5	0.5	0.5	14.4	13.9	12.4	**85.1**	**85.6**	**87.1**
		Predicted								
		OW (open water)			mFYI (mixed FYI)			OI (old ice)		
	Case	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3
Actual	OW	**96.7**	**97.3**	**99.1**	3.3	2.6	0.9	0.0	0.0	0.0
	mFYI	5.2	4.7	2.5	**85.8**	**87.6**	**92.3**	9.0	7.7	5.2
	OI	0.4	0.4	0.2	13.2	12.4	6.0	**86.4**	**87.2**	**93.8**

Table 56: Confusion matrix for of the three3-class RF classifier which was trained for with the 2018 dataset and applied to the 2019 dataset

		Predicted								
		IF (ice free)			FYI (first-year ice)			OI (old ice)		
	Case	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3
Actual	IF	**93.4**	**93.4**	**91.9**	6.5	6.6	8.1	0.0	0.0	0.0
	MFYI	9.8	9.2	8.9	**71.0**	**72.9**	**75.3**	19.2	17.9	15.8
	OI	0.7	0.7	0.6	6.8	7.7	18.5	**92.5**	**91.6**	**81.0**
		Predicted								
		OW (open water)			mFYI (mixed FYI)			OI (old ice)		
	Case	FC1	FC2	FC3	FC1	FC2	FC3	FC1	FC2	FC3
Actual	OW	**93.6**	**93.6**	**86.3**	6.4	6.4	13.6	0.0	0.0	0.0
	mFYI	8.8	7.5	7.1	**72.4**	**75.3**	**81.6**	18.8	17.2	11.3
	OI	0.6	0.6	0.4	6.7	8.1	39.8	**92.7**	**91.4**	**59.7**