Review of
'SAR deep learning sea ice retrieval trained with airborne laser scanner measurements from the MOSAiC Expedition'

This paper looks at the retrievals of sea ice classes, corresponding to new ice/open water, level/first year ice and deformed/multi-year ice using a set of X-band SAR images and airborne laser scanner data from the MOSAiC expedition. This is a good goal, and relevant at this point in time. The dataset is unique and the study is interesting, although in some places the details are not sufficient and the message is not clear. For example, the introduction indicates the paper will address the longstanding issue of sea ice classification, which has been hindered due to the lack of ground truth data and the fact that the sea ice is drifting and deforming. The authors point to the deficiencies of the ice charts as an example of inadequate labels. While there are many issues with ice charts, this feels a little disconnected because the chart labels are different than the ones used (i.e. the classes are different), and the dataset used for the paper here is also quite specialized in comparison to what is available to make an ice chart (which are typically C-band ScanSAR wide data, and other data sources). Regarding detail, much more detail is needed about the networks chosen and how they were trained so we can assess how systematic the comparison is, and if anything can be added to the discussion regarding unique aspects of the chosen architectures than can be better used, given that the test accuracies are not that high.

Major comments

- The SAR data (X-band stripmode SAR) is very niche (not everyday data). It is also high spatial resolution. Why was X-band used? What property does X-band have in comparison to C-band or L-band that make it suitable for this problem. I realize X-band was likely easier to obtain than other data, but how different would the results be if one were to use the more common C-band data at a coarser spatial resolution?

- line 92 'time between satellite measurements and helicopter measurements' how small is this? Is it really sufficient for precise for the features to overlap 'perfectly' across the entire swath for each image? For example on line 89 it says it is accurate to a couple of meters (is this +/- a couple of meters? So 4 meters?). Given that the SAR pixel spacing is 3.5 m, when classification is done pixel-wise, this may not be a wide margin (depending on what the accuracy means). If the authors think this is an issue it should be noted. Would some architectures be more sensitive to this than others?

- The assumption of a Gaussian distribution for freeboard and reflection does not reflect the PDFs shown in the paper. The PDFs shown in Figure 2 are really not Gaussian. A different type of distribution should be used or the impact of this assumption on the analysis should be assessed.

- I also don't follow how a Gaussian distribution allows one to infer uncertainties (unless the authors are referring to the use of 'soft' vs 'hard' labels, or probabilities rather than discrete classes). How does this help their overall objective (was there a comparison with using discrete labels?)

- Better justification of the choice of the architectures and implementation details are needed. While the Unet has been widely used for related sea ice studies, for ConvNext, VGG16, the CNN and the Unet variants it's not clear what special features these architectures have that make them suitable for the problem. For example, ConvNext is an improvement over ResNet, but in what way and how does this motivate using it here? In addition it's not clear if the models are trained from scratch or fine-tuned, or if the same stopping criteria are used for all of them? Could the authors show the training accuracies as well, for comparison with the test accuracies in Table 1.

- line 173 the kernel size of ConvNext is thought to be part of the problem with the poor model predictions. Could this not be changed and the hypothesis verified, or was there a reason for choosing ConvNext for which that kernel size or maintaining the overall architecture is important (for example did the authors consider it out of scope to modify established architectures)?

- lines 120-123 refer to the methods used to handle the class imbalance, but then on line 192 it says that the OW/YI classifications are the most error prone, 'expected because they are the most

sparse' in the dataset - does this imply that the methods used for imbalance were not effective, or could it be that the signatures of OW/YI were quite variable and not well represented given the sample size?

- line 154 The authors minimize the KL divergence instead of the cross entropy, which is more typical for a multi-class classification problem. I think the KL divergence should yield the same result as the cross-entropy (based on the definition of KL divergence), but if not then the could the authors explain their motivation.
- line 158 The validation set is very small (5% of data) - why is this so? Is it large enough to represent the general characteristics seen in the data (how was it chosen)?
- line 185 'generative models' - I don't think generative models were used in this study,
- the custom CNN architecture from Kortum et al. 2022 (shown in Appendix) uses a different size of input patch (smaller) than the other architectures. This leads to sharper, possibly noisier, predictions because the context over which the features are learned is smaller for the smaller patch (local or patch-size features are emphasized over global context). Similar results were found in Radhakrishnan et al. 2021 (reference below)
- line 69 'mow the lawn' pattern - if this is known terminology could a reference be given?
- Figure 1: Could something other than greyscale be used? The legend says white regions indicate no data, but it is hard to tell this apart from the very light grey. Also, there is no red circle.
- Figure 2: The PDFS shown are very smooth. Are fits over all the data points in the image shown to the left?
- Figure 4: This might be better placed as a subset of Fig 2.
- Figure 6: This is interesting in that the cyan regions are very irregular early in the season (November) and more linear later (in March) - is this typical?

K. Radhakrishnan, K. A. Scott and D. A. Clausi, "Sea Ice Concentration Estimation: Using Passive Microwave and SAR Data With a U-Net and Curriculum Learning," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 5339-5351, 2021, doi: 10.1109/JSTARS.2021.3076109.