First of all, I would like to thank the reviewer for taking the time and effort to read and review the manuscript and lending their expertise. Their comments are much appreciated.

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.

We have rewritten part of the introduction to clarify the aims of the paper. It does by no means solve the issues of sea ice classification: If anything, it offers a measurable disconnect between the commonly reported accuracies on manually or ice chart data and labels derive from measurement, which are not affected by human labelling biases. The connection to ice charts should be seen as a contrast as they are the only widely available source of label information. We take your comment with the amount of detail of the architectures very seriously. However, optimizing these architectures would offer little benefit in the long run, as the classifiers trained here will never be performant on a larger scope of ice conditions due to the constraint to a region near the MOSAiC floe and further optimization is very much dependent on the underlying data. Thus, statements about the usefulness of optimization techniques cannot be assumed to generalise.

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?

As the you rightly suggested, other available data would not have been available at the resolution or frequency to enable high spatial overlap (in terms of pixels) at small enough time differences. At higher wavelengths, especially L-Band we might expect at least a higher correlation between radar backscatter and surface roughness, as this was measured at spatial intervals of 0.5 metres, this would probably translate to higher classification accuracy for deformed ice. The complexity of the spatial distribution of classes we would not expect to drastically change at these coarser resolutions. This can be argued with the idea of fractality of structures of the Arctic sea ice across these length scales (although at resolutions as small as 3.5 m this might start to break down). Thus, the core advantage of being able to use intra-label relationships, that semantic segmentation models have over centre-pixel classifiers, is probably similarly useful even at larger length scales. There is also little evidence to suggest that the classification problem becomes easier at larger scales.

• 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?

From a visual perspective, the features really do line up perfectly between the two measurements over the entire helicopter data. Any residual inaccuracies are on the scales of a single pixel and as most features are significantly larger than that the influence of those inaccuracies should be small. Great comment on relating this to the architectures. The central pixel classifiers are probably less affected, as they have no intrinsic knowledge of classifying exactly the centre. Realistically they have a smaller ‘effective resolution’ – not necessarily classifying what is in the central pixels but rather what is in a wider notion of the centre of the image, therefore small shifts in the matching would have less of an effect on training.

• 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.

Note that the gaussian assumption is only used for local distributions (at a single TS-X Pixel), which are usually quite far from the bounds of the global pdfs. Some analysis has been added to the manuscript, that showed the mean and medians of these distributions to be different by less than a percent of the span of the distribution on average. This leads us to believe the impact of simplifying to gaussian distributions is minimal.

• 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?)

You are correct in the assumption of using ‘soft’ labels. Originally, we based this on past experiments. To follow up on your comment, we conducted a comparison with discrete labels which showed reduced accuracies by about eight percent for the best performing U-net architecture, which is quite significant. This confirmed our past experiences and will be added to the paper.

• 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.

VGG16 style architectures have also been used for ice charting in the past. The ConvNext model is a natural progression for such architectures, the same way the Unet++ is for the Unet, thus including them was found to be sensible. Although they are quite varied it strengthens the case of different results from centre-pixel vs semantic segmentation models that is highlighted in this research. The models were all trained from scratch and stopped under the same conditions. Due to premature stopping of the training run based on validation set results, training set accuracies were only a little better than on the test set (0-5%). Thank you for the comment, we will make sure to have more detail about the training procedure and the implementations in the next iteration of the manuscript.

• 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)?

In general, the aim was not to modify the existing architectures, as a lot of time could be spent on optimisation, that would in the end not be meaningful for further research, as they would depend quite strongly on this unique dataset. It would also open up the question how much the optimisation contributed to the result and how much effort would be spent on optimising etc. However, out of interest, we have redone the calculations with smaller kernel sizes: ConvNext with 3x3 kernels performed better by around 5% accuracy, whilst the population variance remained similar.

• 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?

Although the balancing strategies help mitigate the problem, they cannot fully overcome such a heavy imbalance. The variable signatures are definitely also a problem, that should be mentioned, thank you!

• 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.

You are right the optimization will be the same. The parameter itself gives more accurate representation of the amount of information that is lost, rather than just the accuracy of the result. We will rephrase some sentences to better reflect this.

• 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)?

In general, the amount of data we have available is not large and we wanted to make use of it as efficiently as possible. The sole purpose of the validation set is to set a ‘fair’ stopping point for all models, which is why we kept it the smallest of all the dataset. It is a random slice of the data, so it was not chosen in a deliberate manner. This also ensures less human influence on the results.

• line 185 ‘generative models’ - I don’t think generative models were used in this study,

Apologies for the wording, you are right it is misused here.

• 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)

Thank you for the reference, we will follow up and incorporate this into the text. The context (in terms of distance) is not smaller because the inputs are down sampled at different rates. But the amount of contextual pixels that information could be drawn from certainly are. This architecture was quite heavily optimised for different ground truth data, which did not translate to the data used here.

• line 69 ‘mow the lawn’ pattern - if this is known terminology could a reference be given?

It is the best descriptor we could come up with, the data product itself could be referenced but this would not help didactically. We can change to a more descriptive approach detailing how the data’s narrow parallel swaths were acquired in such a manner that the individual subswaths had small overlap along the long side, until the desired area was completely covered.

• 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.

Thank you, we will improve this for better visual readability.

- Figure 2: The PDFS shown are very smooth. Are fits over all the data points in the image shown to the left?

Some Gaussian kernel smoothing has been applied here to smooth over the PDFs. We will update the text to reflect this

- Figure 4: This might be better placed as a subset of Fig 2.

Good suggestion, we will change this.

- 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?

Interesting observation. The shape of leads is in some way's indicative of the breaking processes and rheology. We have thought about creating a lead-shape dataset to assess this quantitatively with some colleagues, but need to work on an efficient outlier detection scheme to automate this process.

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.

Thank you very much for your helpful comments and interesting ideas. These are sure to improve the quality of the work and give some valuable ideas for future work. Your time and effort is very much appreciated!