The manuscript "Self-supervised learning reduces labelling requirements for sea ice segmentation in Sentinel-1 SAR imagery" presents an interesting approach to reduce the number of manually drawn labels for training automatic ML-based ice classification on SAR imagery. The manuscript is well written, and the description of data and methods is clear. However, there are critical errors in the labelling and interpretation of the SAR imagery used, which leads to drawing wrong conclusions in the manuscript (see the major comments below). I therefore recommend rejecting it, and provide only major/minor comments without per-line corrections.

Major comments

1.

The labels on the test images are incorrect. They do not show open water, but new and young ice freshly formed in the leads. The pattern visible on the backscatter from ice in the leads clearly shows a cyclic evolution of freeze-up with several bands of lower and higher backscatter along the lead edge.

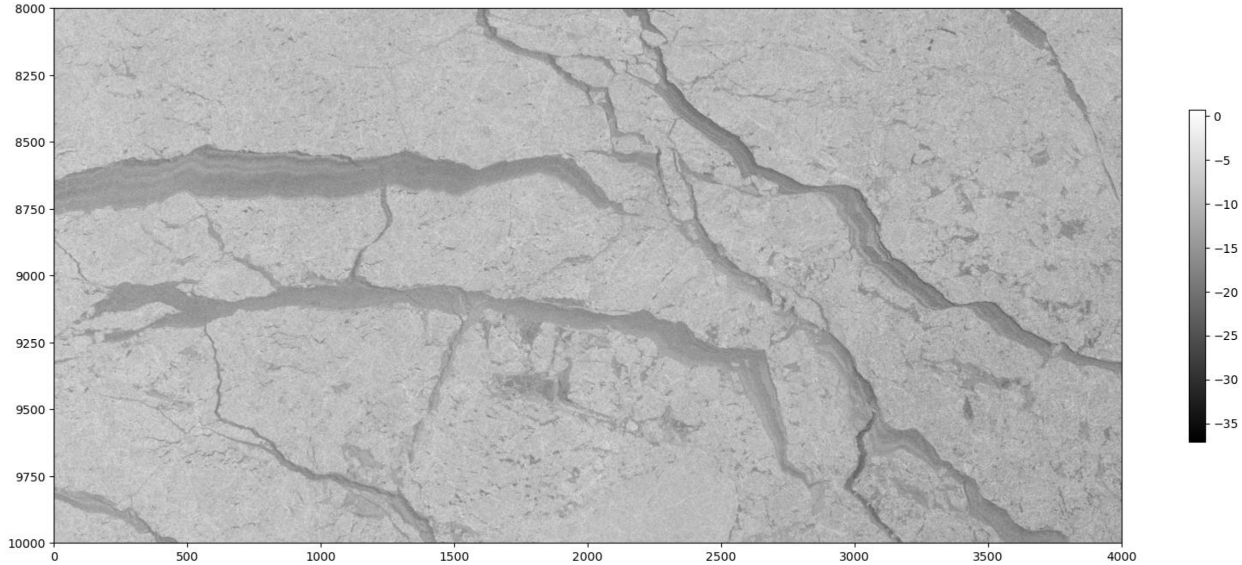

Figure 1. Cropped S1 image S1A_EW_GRDM_1SDH_20221027T161558_20221027T161702_045630_0574C3_1CD3. All dark leads are actually ice-covered. But labels on Fig. 7 in the manuscript suggest that it is open water.

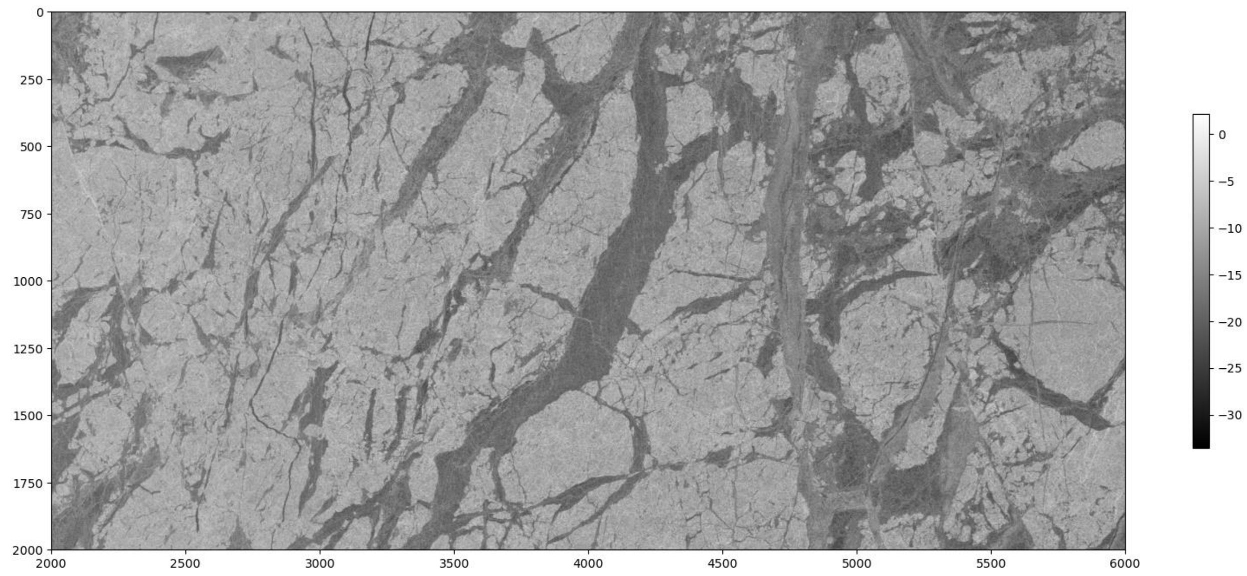

Figure 2. Same for S1 image S1A_EW_GRDM_1SDH_20230305T155314_20230305T155414_047511_05B46C_E347.

It is highly likely that the U-nets were trained on incorrect labels and are predicting not the ice/water classification, but the "older ice" / "younger ice" classification. In that case, both the Control and SL U-nets are performing better on the second image, where a clear distinction between older, rougher, brighter and younger, smoother and darker ice is visible. And the SSL U-net derives 'ice' which should be interpreted as 'old ice' in the pixels with obvious first-year ice. In this case, the main conclusion that unsupervised pretraining increases accuracy is actually compromised.

To prove that unsupervised training improves the classification accuracy, the training and the testing labels should be correct and consistent. It is highly recommended to include a sea ice expert in the team.

Moreover, two test scenes are obviously not enough to draw sound conclusions. As demonstrated by the authors, it is easy to create a solid argumentation based on false results from a few samples. Large statistics reduce chances. It is also recommended to include scenes with true MIZ (see comment 3) and share the training / testing labels online (e.g., Zenodo) for simplifying the review process. Alternatively, the existing, widely used datasets with SAR imagery and ice-chart based labels (e.g., AutoIce) can be safely used.

2.

In the preprocessing step, the samples are normalised using a z-score statistic computed on a per-image basis (L. 431). This normalisation scales the sigma0 values from dB to relative units which are dependent on a range of values on each image. As a result, the networks never see the same values of backscatter for the same type of surface (water, young ice, older ice). That may explain why the SSL network works worse (or better, from the author's perspective). As it was trained on more images than the Control and SL, it probably learned to ignore the magnitude of the normalised signal, whereas the Control and SL networks rely only on the magnitude. As the contrast increased on the second image, the magnitudes of the signal from the older ice and younger ice became very different and the Control and SL networks easily distinguished them, whereas the SSL network failed.

Figure 9, supports my hypothesis that the image-based scaling changes performance of the networks from image to image. If the images were normalized using a fixed scaling, the HH dependence of accuracy should be very similar on two images.

It is recommended to normalise the images, but the z-score statistics should be computed from many images and kept fixed. Moreover, it is better to perform incidence angle correction and proper thermal noise removal to exclude these factors from the study. Otherwise, it difficult to say what is actually playing a larger role in the undertrained network performance - is it a super sensitivity to noise or other factors.

3.

There is no MIZ on the second test image. Not because there is almost no open water, as discussed in comment 1, but because MIZ separates pack ice from the open ocean.

Good examples of scenes with MIZ:

A very wide MIZ. All ice is broken into 10 - 300 m floes:

S1C_EW_GRDM_1SDH_20260224T180149_20260224T180254_006501_00D19C_9D7E

A relatively narrow MIZ separates the pack with a mix of older and younger ice from open water:

S1A_EW_GRDM_1SDH_20260224T073006_20260224T073106_063358_07F521_E549

It is highly recommended to include a sea ice expert in the team. It is also recommended to test the networks in different ice conditions, including open water scenes.

Minor comments

1. The Control UNet was trained on only two images, whereas the SSL on seven plus two images. That makes the comparison of these two networks unfair. I would rather train only two networks (SL and SSL) on the same amount of SAR images excluding labels from some samples for the SSL. That also enables an experiment - how many labels can be dropped to achieve a similar result.

2. Although the training procedure description is quite verbose, it is difficult to understand without a flowchart or a concise algorithm description. I would recommend adding an Algorithm (https://www.overleaf.com/learn/latex/Algorithms) with notations of the online and target networks.

3. The CRF post-processing may significantly alter the results of classification depending on spatial distribution of probabilities returned by different networks. It is better to exclude it from the experiment, as it adds an unknown factor (similar to varying thermal noise, etc.) influencing the accuracy.

4. Figure 3 needs improvement. The arrows are too long, leaving too much space between too small boxes with too small text (for my eyes).

5. Figures 7 and 8 need some improvements:

* Labels on the images should be rotated to view the figure in the portrait layout,

* both HH and HV should be presented,

* the backscatter should be presented after normalisation with a colorbar,