We thank the Editors and Reviewers for their detailed and constructive comments. We have carefully considered all feedback and have prepared an initial set of responses addressing the key points raised.

These responses outline the revisions made to our methodology, data processing, and interpretation, as well as clarifications that will be incorporated into the manuscript. We are currently in the process of implementing these changes and will provide a fully revised manuscript, together with a final, consolidated response to all comments.

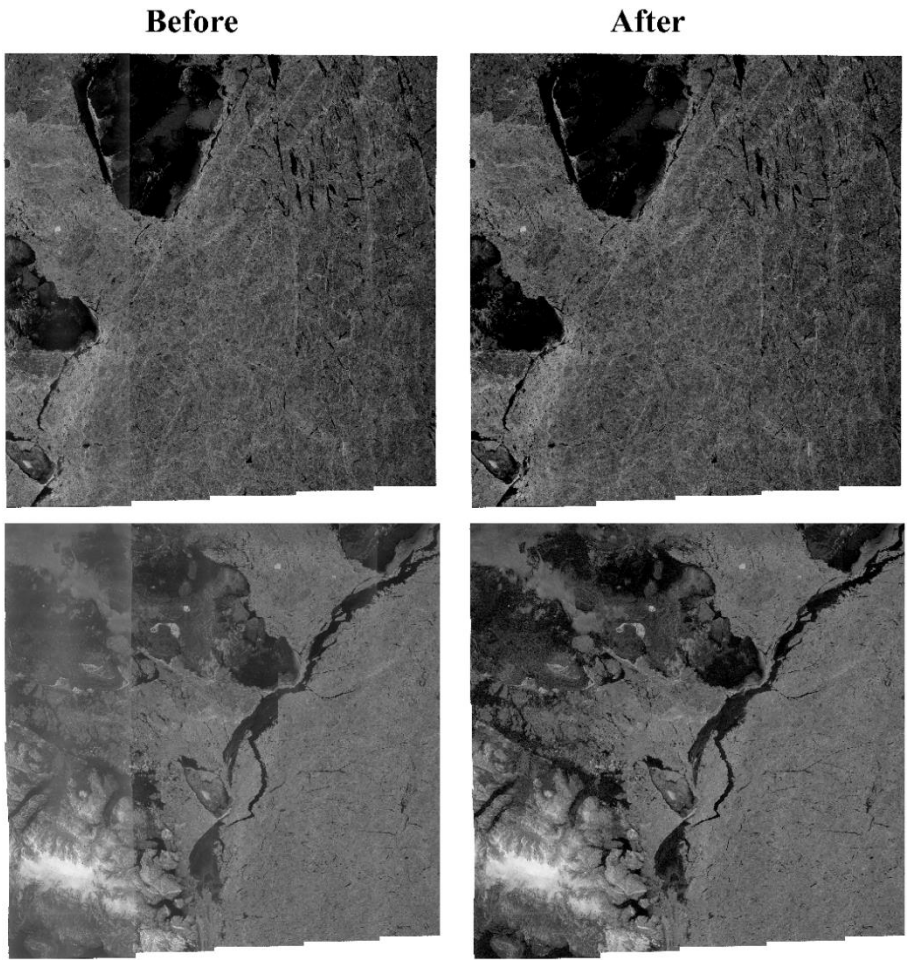
Referee 1

(1) Lack of thermal noise discussion:

Figures 7 and 8 show the classification results for both test scenes, obtained from the UNet SSL and the four baseline models. The visually most striking feature in almost all results is the S1 thermal pattern. Yet, thermal noise is not explicitly mentioned at any point in the paper. Especially the results for test scene 2 (Figure 8) are dominated by thermal noise patterns (scalloping and sub-swath boundaries), and it appears that the classification results must be highly dependent on how well each model learns these noise patterns. Although this is a well-known issue for S1 sea ice mapping, the authors do not include thermal noise correction in their pre-processing chain. At least, these patterns and their influence on the scores should be discussed.

We thank the reviewer for highlighting the influence of thermal noise patterns in Sentinel-1 EW imagery.

In response to this comment, we have reprocessed all Sentinel-1 scenes using an updated SNAP workflow that explicitly includes thermal noise removal and border noise correction prior to radiometric calibration. This substantially reduces the visible scalloping and sub-swath boundary artefacts present in the original figures.



(2) Figure quality:

The quality of some figures in their present form is not sufficient and should be improved (in particular Figures 2, 7, and 8). Suggestions for improvement are given in the detailed comments below. Generally, I suggest including all figures as vector graphics instead of pixel graphics, as this will highly improve the quality especially when zooming in on details.

We thank the reviewer for highlighting the limitations in figure quality. In response, we will update all figures to improve clarity and visual interpretability. Specifically, figures will be regenerated at higher resolution and, where appropriate, exported as vector graphics to ensure improved quality when zooming and viewing fine-scale features.

(3) Limited training and evaluation:

While the authors explain that "the primary contribution of the study lies in demonstrating relative performance gains under reduced annotation budgets, rather than maximising accuracy on a specific benchmark", I am sceptical that the results ("MCC improvement of 24% and 44%") are transferable to "more realistic scenarios" with significantly more training data. Generally, all performances presented in Figures 7 and 8

are still quite poor and heavily affected by thermal noise (see comment (1)), and any algorithm that is moving closer to operations needs to produce much more accurate and reliable ice-water maps, requiring additional training for the UNet SSL as well as the baseline methods. While it may be worthwhile to demonstrate the potential improvements by the SSL approach (as done in this study), the authors should probably discuss the transferability/scalability of the improvements when a lot more training data is used to achieve overall good enough results (with any method).

The reviewer raises an important point regarding the transferability of the reported improvements to scenarios with substantially larger labelled datasets.

In response, we have expanded and refined the labelled dataset to improve its robustness and relevance to the classification task. Specifically, six additional scenes have been added to the labelled dataset. At the same time, two scenes have been removed from the labelled set, as under the revised labelling protocol they do not contain any open water and therefore do not contribute meaningfully to a binary ice–water classification. Retaining such scenes would introduce a bias towards a single class and reduce the effectiveness of both training and evaluation. However, these two scenes are retained for use in the self-supervised (BYOL) pretraining stage, where class labels are not required and the scenes still provide useful SAR texture and structural information. These changes ensure that the dataset more accurately reflects the intended classification problem and includes a broader range of ice–water conditions, reducing the likelihood that results are driven by a small number of specific scenes.

We emphasise that the primary objective of this study is not to maximise absolute segmentation accuracy, but to evaluate relative performance under constrained annotation conditions. Specifically, the study is designed to assess whether self-supervised pretraining enables improved segmentation performance when only a very limited number of labelled SAR scenes are available. As such, the results should be interpreted in terms of label efficiency rather than as a benchmark of operational performance.

We agree that, in more realistic operational scenarios with substantially larger labelled datasets, all models would be expected to achieve higher absolute performance. As the amount of labelled training data increases, the relative performance gap between self-supervised and fully supervised models may decrease, as all models benefit from additional supervision. However, the advantage of self-supervised pretraining is not expected to disappear entirely, as it provides a more informative initial representation that can improve convergence, robustness, and generalisation, particularly in complex or heterogeneous ice conditions.

The key implication of our findings is that self-supervised pretraining reduces the amount of labelled data required to reach a given level of performance. In this study, the BYOL-pretrained UNet trained on three labelled scenes matches or exceeds the performance

of a fully supervised UNet trained on more than twice as many labelled examples, demonstrating a clear gain in label efficiency. This behaviour is particularly evident in more challenging regimes, such as the fragmented ice conditions in Scene 2, where the BYOL-pretrained model shows improved robustness and reduced false negatives. These results suggest that the representations learned during self-supervised pretraining capture structural and textural features of SAR imagery that generalise beyond the specific labelled examples used for fine-tuning.

We note that the Sentinel-1 scenes have been reprocessed to reduce thermal noise effects (see response above), improving the reliability of the evaluation.

We therefore view self-supervised learning not as a replacement for fully supervised approaches in large-data regimes, but as a complementary strategy that enables more efficient use of labelled data. This is particularly relevant for Arctic SAR applications, where manual annotation is costly, time-consuming, and geographically limited. In this context, the demonstrated reduction in annotation requirements represents a meaningful step towards scalable, data-efficient sea ice mapping workflows.

---

(4) Random Forest performance:

Although it is only a "baseline" algorithm for comparison, the performance of the RF algorithm is noticeably bad and concerning. The presented RF basically maps everything as sea ice without even identifying the dark leads in the test scene 1 (Figure 7). Although the deep-learning approaches are probably expected to perform better than a "traditional RF", this exceptionally bad performance is still quite surprising, since RF have been used quite successfully for sea ice mapping in previous studies (e.g. Park 2020, Lu 2023). Without any additional explanation, this raises concerns that (a) the RF and the texture features for it are not well designed and/or (b) that the training data is simply not sufficient for any reasonable assessment. I think at some discussion/explanation of this poor performance compared to RF in previous studies is needed.

The reviewer raises a valid concern regarding the relatively poor performance of the Random Forest (RF) baseline compared to previous studies.

We note that successful RF-based sea ice classification approaches typically rely on extensive feature engineering and large training datasets. For example, Park et al. (2020) use GLCM and Haralick texture features with explicit preprocessing and train on hundreds of Sentinel-1 scenes, achieving good overall accuracy but still struggling to distinguish similar ice types, especially in summer conditions.

In contrast, the RF model in the present study is deliberately implemented as a simple baseline operating on limited labelled data and without extensive feature engineering. The study focuses on evaluating model performance under constrained annotation conditions, and therefore does not fully optimise the RF model to the level of more extensively engineered operational systems.

The observed tendency of the RF model to classify most pixels as sea ice is therefore likely a consequence of both the limited training dataset and the absence of spatial context and texture-based features. In SAR imagery, discrimination between ice and open water often depends on spatial structure and contextual information, which are not explicitly captured by pixel-wise RF approaches.

In comparison, the deep learning models used in this study are able to learn hierarchical spatial and textural representations directly from the data. The improved performance of the self-supervised model further suggests that robust feature learning is particularly important under limited-label conditions.

We will clarify these differences in the manuscript and emphasise that the RF results should be interpreted as a baseline under constrained conditions, rather than as a fully optimised implementation of RF-based sea ice classification.

---

(5) “Applicational” motivation of the study:

It is not entirely clear whether the main applicational focus (i.e., why do we need improved automated sea ice mapping) is on navigation or environmental studies (or both). The initial mentioning of “maritime safety” and explanation of operational ice charting suggests an operational focus on tactical navigation, but later this is repeatedly mixed with the importance of leads for ocean-atmosphere interactions, which suggest an environmental focus. Both applications are valid and important, but they should be clearly distinguished in their explanations.

The reviewer raises an important point regarding the distinction between operational and environmental motivations in the study.

The primary focus of this work is on improving segmentation performance under limited-label conditions, with the goal of enabling more efficient and scalable use of SAR data for operational sea ice mapping. This is directly relevant to applications such as maritime navigation and safety, where timely and reliable ice–water discrimination is essential.

While accurate identification of features such as leads may also support broader scientific studies of sea ice dynamics, these are not the primary focus of this work. Rather, they are a secondary benefit arising from improved segmentation performance.

To address the reviewer’s comment, we will revise the manuscript to clearly distinguish between these application domains. In particular, we will present operational ice mapping and navigation as the primary application context, with other uses framed as extensions of the same methodology.

We therefore consider that this clarification will improve the overall structure and motivation of the study.

---

(6) Roughness scales:

Roughness is always a relative term, depending on scale. Whenever mentioning roughness throughout the manuscript, the authors should specify the roughness scales they are thinking of: small-scale (on the order of the radar wavelength) or large-scale (on the order of meters). Large-scale roughness could also be referred to as large-scale deformation and is directly related to ice type (deformed FYI or deformed MYI), whereas small-scale roughness can in fact vary significantly within one ice type, especially young ice with/without frost flowers or finger rafting. It might be worth to explicitly mention somewhere that the interplay of both small- and large-scale roughness effects contributes to the challenging interpretation of sea ice in SAR imagery.

The reviewer raises an important point regarding the interpretation of roughness in SAR imagery and its dependence on spatial scale.

We agree that the term "roughness" can be ambiguous if the relevant scale is not specified. In the revised manuscript, we will clarify that roughness in this context refers to both small-scale (on the order of the radar wavelength) and large-scale (on the order of metres) surface variations. Small-scale roughness influences the scattering behaviour at the wavelength level, while large-scale roughness is associated with ice deformation and floe structure, and is more directly related to ice type.

We will update the manuscript to reflect this distinction and improve clarity in the discussion of SAR backscatter mechanisms.

---

(7) Description of the 2 test scenes:

The two selected test scenes are throughout the manuscript described as "consolidated ice pack" (scene 1, Figure 7) and "MIZ" (scene 2, Figure 8). The MIZ is "traditionally" defined as the area with 20-80% SIC, and more recently physics-based approaches as the "area affected by waves" are more common. Based on visual inspection of Figure 8, neither of these definitions makes me think that this image is in the MIZ. The authors should explain the reasoning behind this description of the test scenes and maybe consider to change them.

The reviewer raises a valid point regarding the absence of a true Marginal Ice Zone (MIZ) in the second test scene.

We acknowledge that the classical oceanographic definition of the MIZ refers to the transition between consolidated pack ice and the open ocean. In this study, we recognise that the selected scene does not strictly meet this definition. Instead, it represents a more complex ice regime than the other test scene, characterised by a mixture of ice types and ages, fragmented floes, and heterogeneous backscatter signatures, which present similar challenges for segmentation.

Such conditions introduce increased structural and radiometric ambiguity, making the scene a suitable test case for evaluating model performance under difficult segmentation conditions.

We therefore consider that the use of this scene remains appropriate for the comparative analysis presented in this study.

Referee 2

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.

We thank the reviewer for highlighting the ambiguity between open water and newly formed ice in low-backscatter regions. In response, we have revised the labelling approach to explicitly address this concern.

Following the reviewer's comments, we have refined and standardised the labelling protocol to improve both consistency and physical interpretability across the dataset.

In the updated workflow, low-backscatter features such as dark leads are no longer automatically classified as open water. Instead, thresholding is used as an initial, physically informed guide that incorporates both HH backscatter and incidence angle, alongside an HV constraint to account for noise characteristics. Specifically, pixels are initially identified as open water where HV ≤ −25 dB and HH follows an incidence-angle-dependent relationship.

This formulation is informed by the relationship between σ^0 and incidence angle reported in Lohse et al. (2020), which shows a systematic decrease in HH backscatter with increasing incidence angle. For open water, Lohse et al. (2020) report an HH–incidence-angle slope of approximately −0.72 dB per degree. In our implementation, we use a simplified threshold of HH = 3.7 − 0.72 × IA. This preserves the slope reported in the literature, while the intercept is dataset-dependent and adjusted to reflect the empirical distribution of backscatter in our scenes.

We will include a figure illustrating the relationship between HH and incidence angle in our labelled data. This exhibits a similar negative trend, but with substantial overlap between open water and ice classes. It demonstrates that a single linear threshold cannot fully separate the two classes across all incidence angles, particularly in the mid-range (~30–40°), where scattering signatures are ambiguous.

All labels were manually reviewed and corrected across the entire dataset, with manual interpretation integrated throughout the labelling process. Manual interpretation always

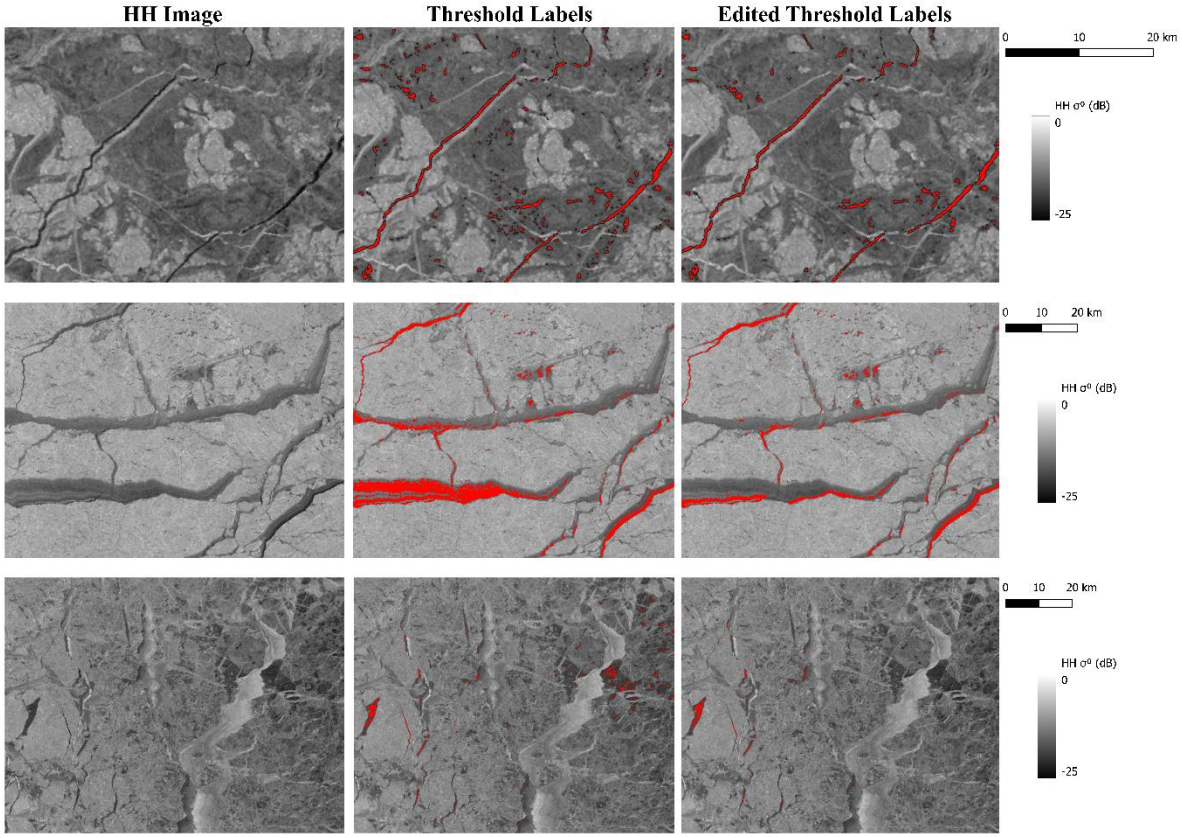
overrides threshold-based classification in cases of disagreement. In practice, this involves identifying and excluding features that exhibit spatial structure inconsistent with open water.

In particular, regions corresponding to newly formed or thin ice are no longer labelled as open water. Features showing cyclic freeze-up signatures—characterised by alternating bands of lower and higher backscatter along lead edges—are consistently interpreted as young ice based on their spatial structure and contextual relationship to surrounding ice.

Manual correction also acts as a denoising step, removing isolated low-backscatter pixels and artefacts arising from sensor noise or thresholding errors. This ensures that open water labels are restricted to spatially coherent regions with consistently low backscatter and minimal internal structure.

Overall, this ensures that the classification task remains a binary ice–water distinction, rather than a proxy for ice age or surface roughness.

We therefore consider that the revised labelling approach resolves the issue raised by the reviewer and provides a more physically consistent distinction between open water and sea ice.



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. 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.

We acknowledge that per-image normalisation removes absolute backscatter scaling, meaning that the networks do not observe consistent σ^0 values across scenes. This is an inherent trade-off of such normalisation strategies. However, this approach was adopted to account for the substantial variability in SAR backscatter across Sentinel-1 EW scenes, which arises from differences in incidence angle, environmental conditions, and scene composition. Normalising each image independently reduces the risk that models overfit to scene-specific intensity distributions.

We agree that this normalisation reduces the direct interpretability of absolute backscatter magnitude. However, the segmentation task considered in this study relies not only on intensity but also on spatial structure, texture, and contextual information. In this context, per-image normalisation allows the models to focus on relative contrast and structural features, which are key discriminants between ice and open water in SAR imagery.

The reviewer suggests that the observed differences between models may be influenced by their sensitivity to normalised intensity. While this is a valid consideration, the improved performance of the self-supervised model, particularly in structurally complex regions, suggests that it is learning more robust feature representations that extend beyond simple magnitude-based discrimination.

We agree that alternative normalisation strategies, such as fixed global scaling, may influence model behaviour and provide additional insight into the role of absolute backscatter. In response, we will adopt the suggested approach of computing normalisation statistics across multiple images and applying a fixed scaling during training and evaluation.

We note that thermal noise removal has now been incorporated into the preprocessing pipeline (see response above), reducing one potential source of variability in backscatter magnitude.

We therefore consider that while normalisation influences the representation of input data, the relative performance improvements observed for the self-supervised model remain indicative of its ability to learn transferable SAR feature representations under limited-label conditions.

---

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.

See response above.

---

4. 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.

The reviewer notes the use of Conditional Random Field (CRF) post-processing.

In this study, the CRF is applied as a post-processing step to the outputs of the deep learning segmentation models. Its role is to refine spatial coherence and reduce noise in the predicted segmentation masks, particularly along boundaries and in heterogeneous regions.

Because the CRF is applied uniformly across all models, it does not introduce bias in the comparative evaluation. Instead, it serves as a common refinement stage, ensuring that differences in performance reflect the underlying model outputs rather than differences in post-processing.

We acknowledge that CRF post-processing can influence the final segmentation results. However, its effect is limited to local boundary refinement and does not alter the overall structure of the predictions. As such, the relative performance differences between models remain driven by the models themselves rather than the post-processing step.