1st reply to reviewer comments for

"AI4SeaIce: Task Separation and Multistage Inference CNNs for Automatic Sea Ice Concentration Charting"

[EGUSPHERE-2023-976]

Dear Editor Ludovic Brucker,

On behalf of my co-authors and myself, I would like to thank the two reviewers for their comments on our manuscript. The reviewers have made extensive suggestions on how to improve the explainability in key areas. We have followed their suggestions to our best efforts in the vast majority of cases and where it was not possible we have given an extensive justification why it was the case. In the following, we provide a point-by-point answer to all the issues raised by the reviewers. We have gathered all reviewers' points in black, our answers are highlighted in blue, the original text in red, and edited content in green. In addition, we have uploaded an associated *difference* document highlighting the changes in the manuscript.

Best regards on behalf of the authors,
Andreas Stokholm

Reviewer 1:

General Comments:

This research intends to explore a SAR-based multistage expansion of sea ice concentration diagnoses by leveraging one UNet for bulk class prediction (open water, ice, consolidated intermediate concentrations) and another UNet to refine diagnoses of intermediate concentrations. The authors explore such an approach with a diversity of classification- and regression-focused loss functions. While this approach did not yield correlation coefficient improvements over a prior effort, an improvement was shown for intermediate concentration accuracies. As this is a quickly expanding field and the study applies to a very useful open-source dataset that will likely drive research in the near future, I believe this paper does have a place in the knowledge corpus being assembled in this domain. However, there are several key concerns that, when addressed, will make this a stronger submission. In its present state, I believe the manuscript requires major revisions.

Specific Comments:

Test set evaluation concerns

Despite the authors note, I am concerned about the level of potential cross contamination between the test set and training set imagery which could inflate metric scores. While there is brief mention of this possibility, the authors may mitigate this concern by clarifying minimal spatial/temporal overlap between test and train scenes. If the test scenes demonstrate minimal footprint overlap and/or adequate time deltas against training scenes in that location, potential strong correlation between train and test may be less impactful. Any overlapping train/test scenes where minimal changes in sea ice have between demonstrated between the two may be problematic as deep-learning algorithms are very much capable of memorizing training input. There is no mention of a specific validation dataset or early stopping based on such hold out scenes, further enhancing this concern.

We agree that spatial and temporal overlap can be an issue for leakage between training and testing data, we are confined to working within the means of the available dataset. While a lot of pixels were trainable (above a billion), relatively few high-quality scenes were available (339), thus selecting a scenario to test for difficult and interesting cases with spatial and temporal overlap is challenging to balance. We also emphasise that testing for an operational context requires significantly deeper investigation than what we were able to present here. However, for testing model strategies and exploring more fundamental challenges and strategies for further model development is less critical to perform rigorous testing as maritime safety concerns are not present in the laboratory. What is most important is that all the trained models are compared against the same baseline, which should remove any cross-model score inflation, including those scores compared with other papers. Thus a potential score inflation could simply be a bias across the reported scores.

In addition, we have carried out an analysis regarding the temporal and spatial overlap. 9 scenes were fully isolated spatially and temporally, while 11 test scenes were acquired in a swath with test scenes connecting to a training scene edge. Herein, there are 2 test scenes that have both a training scene before and after within the larger swath. While not ideal, we do not see this as a significant issue as the scenes are 400 x 400 km and thus cover a vast area. The final 3 scenes overlap from different satellites but are taken within 24 hours of each other but more than 10 hours apart. However, these scenes contain diverse microwave signatures as the satellites are either ascending or descending and thus view the area from different angles. Otherwise, typical time intervals between training and testing scenes over the same spatial location is around 5-7 days or more.

Additionally, the author mentions all scenes spanned between March 2018 and May 2019 collection dates. It sounds like scenes from the recent Auto-ice challenge dataset were leveraged for this paper. I believe that dataset reference scenes from 2018, 2019, and 2020. Using scenes from a different year altogether (2020) would go a long way towards an independent test set and would certainly strengthen the arguments in this manuscript.
We would like to clarify that the AutoICE dataset was not utilised for the experiments carried out in this manuscript, as it was unavailable. Instead, the ASID-v2 was used - a very similar dataset. While we do agree that combining the ASID-v2 and AutoICE datasets could be beneficial, all experiments and testing would have to be redone, which is a substantial undertaking.

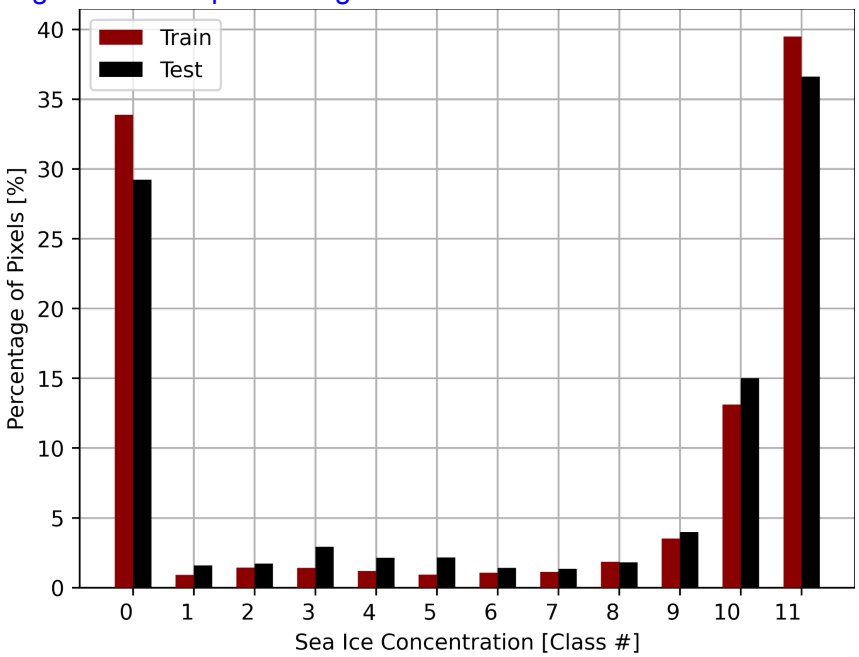
Class Imbalance
Aside from mentioning weighted losses based on class frequency, the authors make minimal effort to combat potential class imbalance issues. For example, there is no mention of attempting over/under sampling or including loss variations specifically designed for such issues (e.g., focal loss – an easy swap with CE). If class imbalance is minimal in this dataset, the inclusion of a label count table would be helpful in assuring the reader. The authors express disappointment towards not achieving greater than a 93% R^2 throughout the updated experiments. One wonders if this is simply related to class imbalance, where improvements in intermediate class results are still not enough to affect change in R^2 because it is dominated by 0% and 100%. This should be something that could be examined via confusion matrices and perturbation of such matrices to look at impact of imbalance. The matrices can be converted in this case of clean concentration intervals to match values given by the Sklearn function for correlation coefficient (I am speculating this is being used as it was noted in the Auto-ice challenge documentation). This may provide some insight.
We agree that more effort in addressing class imbalance could and should be considered, such as over/under sampling as suggested but could warrant a study by itself. We would like to add that this paper (https://www.nature.com/articles/s41598-023-32467-x) looked more into utilising class frequency weights for model optimisation. We also agree that more information regarding the class distribution could be included. We suggest adding a reference to a previous paper, which highlights the class imbalance present in the dataset.

This distribution is skewed towards class 0 (open-water), 10 (100\% sea ice) and 11 (masked pixels) being represented the most. The remaining intermediate classes are relatively equally distributed.

This distribution is skewed towards class 0 (open-water), 10 (100\% sea ice) and 11 (masked pixels) being represented the most as highlighted in Stokholm et al. 2022. The remaining intermediate classes are relatively equally distributed.

For reviewer convenience, the figure is included below, which shows that open water is by far the largest (predicted, i.e. exempting class 11, which are masked pixels) class, followed by the 100% class. It should also be noted that the combined number of intermediate classes exceeds the number of 100% SIC pixels. In addition, there are more combined ice classes than water in the dataset, and the intermediate classes are more present in the test set compared to the training set.

The mentioned disappointment may be somewhat interpreted, as achieving 93% R^2 using a classification-oriented approach was not achieved with reference to https://www.nature.com/articles/s41598-023-32467-x (notably with the same training and testing dataset). Instead, this was only achieved using regression-based optimisation losses, which had the disadvantage of scoring poorly on open water and 100% SIC. This was incidentally also the occasion for the idea for the multistage inference hoping to retain both the superior open water and 100% SIC from the classification-optimised models and the high-resolution intermediate SICs from the regression-based models. Therefore, the increased performance, from the perspective of utilising a classification-oriented loss function, is mainly due to the superior intermediate SICs segmentation while retaining the open water and 100% SIC accuracy. However, we do agree that future testing with additional scenes reflecting a more balanced class distribution would be very beneficial. Confusion matrices were also considered but impractical for this number of model experiments, though it could be possible to merely investigate the best-performing model.



On a related note, the paper shows relatively minimal sensitivity to the number of intermediate classes used. Is this again connected to class imbalance, or one or two intermediate classes are far more dominant and always selected, leading to similar results?
It does indeed seem like the number of intermediate classes utilised in model training does not have a large impact on the model's ability to differentiate between the macro classes (open water, any intermediate or 100% SIC). However, the CE model performing the worst at this distinction is the typical 11-class model, so there may be some impact, though small. In addition, as the macro bins are the average of each macro class, the intermediate classes are weighted equally to the open water and 100% SIC performance.
In a related paper, https://www.nature.com/articles/s41598-023-32467-x, the per-class accuracy was investigated, which revealed that CE-optimised models utilise all the intermediate class, however, models optimised with the EMD2 had very low performance in a select number of classes, which may indicate that these are less prioritised. We also carried out an experiment (unpublished) using a loss function, which rewarded the model equally for guessing neighbouring intermediate classes of the correct class. However, this quickly leads to the model developing a strategy for only utilising every second class, effectively reducing the number of intermediate classes by half. This is somewhat similar to reducing the total number of intermediate classes but was not the desired effect we had hoped for.

Multiple Models
While my guess is inference is relatively fast with standard CPU+GPU resources, even with two models, a multi-model approach may not be desirable in a limited resource scenario. Additionally, it is very likely weights from both models are performing similar feature extraction functions, especially along the encoder path. Have the authors considered a single model approach, perhaps a UNet with two heads, where one head predicts macro classes, and the other intermediate pixels, masking the loss from 0% and 100% pixels? Combining the results of losses from the two heads (e.g., Total Loss = w_1*CE+w_2*MSE) is not uncommon. Have the authors explored such an approach? If so, and the multistage approach outperforms, that would make an even stronger case for two models.
These are multiple very good points both on resources and feature extraction. The idea of creating a model with two heads to perform either intermediate or macro pixels is sound, though practically, needing to use the outputs from one head to select the desired pixels to use from the second head before the backpropagation. This could easily be an avenue for further developments or another entry into the model development series. However, for this manuscript, which highlights the idea of combining multiple losses to target specific tasks within the SIC assignment, we argue that it is out of scope, particularly considering the lengthy extent of the current experimentation.

Composition Quality
There are a number of instances across the manuscript where the writing style confuses the intended message. I will note a number of such examples in "Technical Corrections", but the

authors should make a concerted effort to go line by line through the manuscript and rewrite/reword when necessary. Please note some technical corrections below will contain technical and grammatical suggestions.
Thank you for these technical and grammatical suggestions. The manuscript was complicated to compose with so many experiments to cover. Choosing certain wordings was not easy, thus, we welcome these suggestions. We have made an effort to illustrate the suggestions that we have added to the manuscript.

Technical Corrections:

Line 4: “apply” to ”applies”
Incorporated.

Line 10: A) Refrain from using the world “detections”. “Detection” is a specific task in the computer vision community and use of this is confusing when doing segmentation. “segmentations” may be a better term. B) “multistage synergises” to “the multistage approach synergises”
We have incorporated both suggestions.

Line 12/13: “compressed” may also not be the best term as data compression is again an entire field of its own. Perhaps “consolidated” or “merged” is a better choice throughout manuscript.
We have changed the wording from “compressed” to “merged”.

Line 16-19: Rework this. It’s cumbersome in its current form.
We have tried to rewrite the section.
Charting the ever-changing sea ice is important for navigation in the remote and cold Arctic to both circumnavigate and traverse sea ice safely and quickly. Effective navigation thus makes high-resolution sea ice charts detailing the local sea ice conditions indispensable. This is particularly relevant for local and indigenous populations, e.g. around the Greenland coast. Livelihoods in these regions depend on fishing and infrastructure is sparse, where goods and people are primarily transported by boat.
Charting the ever-changing sea ice is important for navigation in the remote and cold Arctic for traversing safely and quickly. High-resolution sea ice charts detailing the local sea ice conditions are indispensable for effective navigation. Due to sparse infrastructure, local and indigenous populations along the Greenland coast rely on goods transported by ship while many livelihoods depend on fishing, which makes ice charts important economic enablers.

Line 17: Remove second “sea ice” mention.
See above.

Line 23: “continue to cause” to exacerbate
Incorporated.

Line 24: “could” to “should”; remove “furthermore” as it was just used above.
Both suggestions were incorporated.

Line 31: “distant” location to remote? Not a great sentence in general. A rewording may flow better.
We suggest the following rewording:
The Arctic is particularly challenging to monitor because of its distant location, sparse infrastructure and vastness of the area.
The Arctic is particularly challenging to monitor because of its vastness, remoteness and lack of infrastructure.

Line 33: Move “almost indistinguishable” to after “reflection”
Incorporated.

Line 40: Use of “spatial resolution” but referencing pixel spacing – spatial resolution and pixel spacing are different concepts in this case. Maybe “...versatile measurement with pixel spacing typically between X and Y”. Also note the 10-40m pixel spacing is not inclusive enough for SAR in general. S1 EW GRD is 40m, but RADARSAT-2 and RCM (similarly useful C band collectors) have different values depending on product.
We agree that the distinction between spatial resolution and pixel spacing is confusing. We also acknowledge the usefulness of the RCM, which would be a natural addition to the training data. We suggest changing the sentence as follows:
Consequently, Synthetic Aperture Radar (SAR) images are the backbone for sea ice charting, as they offer versatile measurements in a high spatial resolution (10 - 40 m pixel spacing) independent of sun illumination and clouds.
Consequently, Synthetic Aperture Radar (SAR) images are the backbone for sea ice charting as they offer versatile measurements with pixel spacing typically between 10-40 m independent of sun illumination and clouds.

Line 46: Remove “a” before “resource”.
Incorporated.

Line 78: “in relation” to “relative”?
We agree this has been incorporated.

Line 79/80: Is there a reference for rolling up to 11 concentrations? It sounds like total concentration was the label. Does your source contain values like “81”, if so, how is this treated, or is this different than typical products available through sources like the USNIC?

We utilise the SIGRID-3 codes, which do also include sea ice concentration intervals. However, to our knowledge, the DMI charts do not utilise these intervals in the ice charts from this time period.

Line 79: Define SIGRID-3 here.
We suggest the following change at line 79:
SIC is defined by the World Meteorological Organisation as part of the SIGRID-3 code
SIC is defined by the World Meteorological Organisation (WMO) as part of the Sea Ice GeoReferenced Information and Data code, also known as SIGRID-3.
In addition, we suggest this change in line 147:
The ice information is conveyed using the 'Egg Code', which follows the World Meteorological Organization (WMO) code; Sea Ice GeoReferenced Information and Data (SIGRID3), and…

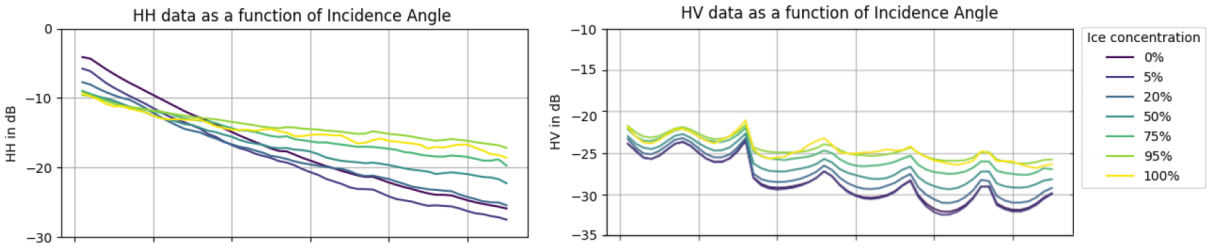
The ice information is conveyed using the 'Egg Code', which follows the WMO code SIGRID-3, and …

Line 94: How is it possible a model trained on hand drawn labels could exceed the original spatial resolution? Is it more related to a mix of labeling styles, where the more detailed analyst reports are showing through?
We believe this can be the case when the labels are associated with individual pixels, which the model is trained on at random while not necessarily seeing the entire hand-labelled ice polygon. The model produces some representation of the variety of polygons and ice conditions, which, as suggested, contain both different levels of labelling resolution (or individual personal styles) but may also be from areas that have received more or less attention. In addition, the hand-labelled ice polygon is the average ice concentration but it may vary significantly within the polygon. I.e. it can be argued that it is somewhat of an average representation. Though this is mostly speculation it does appear in the qualitative inspection that the frequency of change between classes is higher for the model outputs in comparison to the hand-labelled ice charts.

Line 103: "compress" again – may be a better word choice available.
The wording has been changed to "merge" and "merging" in the proceeding appearance on line 104.

Line 125: "contributed" to "attributed" – sentence is cumbersome.
Incorporated.

Line 129: Comment on narrow dynamic range dB – I think this could be written more clearly.
Here, we reflect on the figure below cited from another paper, which examines the HV dB intensities.



The narrow dB range is simply that the dB in the HV channel appears to be within ~-22 to -33 dB compared to the HH channel, which is closer to -3 to -28 dB. We suggest a small rephrasing (removing dynamic):

The disparity between sea ice and open water is largest in the HV image with a narrower dB dynamic range “Author2021”, though the strongest backscatter signal is in the HH image.

The disparity between sea ice and open water is largest in the HV image with a narrower dB range “Author 2021”, though the strongest backscatter signal is in the HH image.

Line 132: Is wind also a factor in the brightening here?

It is possible that there is also some wind here though difficult to assess with so little water texture. At least we do not believe this to be the largest contributing factor to the brightening,

Line 148: SIGRID definition should have been earlier.

We have addressed this comment in response to the comment on line 79.

Line 190/191: Be cautious about how ignore index for the CE function in PyTorch may operate differently than multiplying by 0 in the other case. Depending on options, the “reduction” could be impacted.

We did carry out experiments to investigate this, and it appears that the loss computation is identical to multiplying with 0. We suggest removing the sentence to remove any ambiguity:

In the case of CE, the \textit{ignore_index} argument is used instead, having the same effect.

Line 215-220: I believe this text could be reduced upon rewrite. I would also avoid calling out a “new” metric as this isn’t really novel. Simply calling it an average accuracy as you note is sufficient.

We would argue that the metric is sea ice concentration-specific and thus a new metric in this specific regard. Furthermore, it helps distinguish this metric from many of the other reported accuracies, which could otherwise further complicate the reading of the manuscript. We have attempted to shorten the paragraph:

To evaluate the IceDisc models, we define a new metric that we denote \textit{"MacroBins"}, which is an altered version of the Mean Producer Accuracy or the Average Class Accuracy. The

objective of this metric is to measure how well the model is capable of separating open water, intermediate classes and 100\% sea ice. We define this metric as: {equation}

where TP is the number of True Positives, NP is the Number of Pixels belonging to the class and \textit{int} are intermediate class pixels. In essence, this is the average accuracy of water, 100\% sea ice and whether true intermediate class pixels are predicted as any intermediate class.
To evaluate the IceDisc models, we define a new SIC-specific metric, \textit{"MacroBins"}, an altered version of the Mean Producer Accuracy or the Average Class Accuracy. The metric measures the models' capability of separating open water, intermediate classes and 100\% sea ice. We define this metric as: {equation}

where TP is the number of True Positives, NP is the Number of Pixels belonging to the class and \textit{int} are intermediate class pixels. Effectively, it is the average accuracy of water, 100\% sea ice and whether true intermediate class pixels are predicted as any intermediate class.

Line 262: Please be as clear as possible when using the Kucik and Stockholm (2023) models. I think these are your reference models, but you never actually call them the reference models explicitly when talking about the prior study. This applies here and later on.
We suggest clarifying:
In addition, two 11-class IceDisc models for each loss function are presented as references.
In addition, two 11-class IceDisc models for each loss function are presented as references, similar to those trained in Kucik and Stokholm 2023.

Line 272: "reference 11-class models" to clarify as unweighted?
We suggest clarifying:
Inspecting the 2-class models, the performance for 0\% is on par with most of the IceDisc models with additional classes but lower than the reference 11-class models.
Inspecting the 2-class models, the performance for 0\% is on par with most of the IceDisc models with additional classes but lower than the reference unweighted 11-class models.

Line 274-275: Rewrite – placement of "outperforms the models" makes sentence cumbersome.
We suggest the following modification:
For the CE IceDisc models, the 7-class model outperforms the models trained on fewer classes, particularly in terms of \textit{int in int} and 100\% sea ice while performing the worst at open water.
For the CE IceDisc models, the 7-class model exceeds those trained on fewer classes, particularly in terms of \textit{int in int} and 100\% sea ice while performing the worst at open water.

Line 277: “much better” – please just quantify based on tables – 4-7% improvement, etc.
We agree and suggest the following change:
We also see that IceDisc models with 3-7 classes score much better in terms of \textit{int in int} compared to the 11-class IceDisc, which scores highest in 0\% and 100\% SIC.
We also see that IceDisc models with 3-7 classes score 4-7\% better in terms of \textit{int in int} compared to the 11-class IceDisc, which scores highest in 0\% and 100\% SIC.

Line 280: IceDiscstoIceDisc models
This comment is a bit ambiguous, we interpret it as a difficulty in the sentence structure. We suggest a slight modification:
With respect to the $\emd^2$ IceDisc models, we see fluctuating performances with the highest macro classes distributed between different IceDiscs.
With respect to the $\emd^2$ IceDisc models, we see fluctuating performances with the highest open water, \textit{int in int} and 100\% accuracies distributed among models with different numbers of classes.

Line 281: Reorder wording
We are in doubt of the exact words in mind. However, we suggest:
The 6-class model achieves the highest 100\% accuracy but at the expense of the lowest 0\% and \textit{int int int} scores across all the CE and $\emd^2$ IceDisc models.
The 6-class model achieves the highest 100\% accuracy but at the expense of the lowest \textit{int int int} and 0\% scores across all the CE and $\emd^2$ IceDisc models.

Line 289: Is the statement about the 7 class CE “approaches the class weighted reference model” correct? I am seeing 86.16% vs. 97.18%. Am I reading this incorrectly? Again, in this section, please be clearer on the exact origin of the reference models – they are from the prior paper?
This is an exaggeration. We imply that the 7 class CE model is performing the closest to the weighted class optimised model among the unweighted models.
In contrast, the 7-class CE model approaches the class weighted reference model on \textit{int in int}.
In contrast, the 7-class CE model exceeds both 11-class CE models by 6-8\% on the \textit{int in int} accuracy.
To clarify, the reference models are identical to those trained in Kucik and Stokholm 2023 but trained from scratch. The reference model parameters were selected based on achieving the highest R^2 score, while the other 11 class-trained models are selected based on a high MacroBins score. Thus the reference models exhibit the performance that the models in Kucik and Stokholm 2023 would have.

Line 299: Clarify this statement – what difference is being referenced?
Here, we refer to the internal metric performance between the classification and regression loss

function optimised models, e.g. the performance between CE and EMD^2. We suggest the following modification to make this more obvious:
The difference between the loss functions within regression and classification is, however, performing very similarly.
However, the performance difference between the classification loss functions across the 3 metrics is small and similar for the MSE and BCE-optimised models.

Line 311: The 3-class gives 93.01%, not the 6-class according to the table. Please clarify or fix.
This is a writing error and changed:
The highest R^2 and int R^2 scores are achieved by the $\emd^2$ 6-class IceDisc and MSE IntDisc combination with 93.01\%.
The highest R^2 and int R^2 scores are achieved by the $\emd^2$ 3-class IceDisc and MSE IntDisc combination with 93.01\%.

Line 315: “clear tendency” statement is confusing. You mention 6 classes performs best, but the statement doesn’t make sense to me since 7 and 11 are both worse and are at times outperformed by <5 classes.
This is a general statement referring to the top performance among the CE IceDisc models on the R^2 metric, FAcc, IntFAcc and MacroBins are achieved by the 6, 7 and 11 class models. Though we agree, that this may be more confusing than beneficial for the reader. We suggest removing the phrase.
For the CE IceDisc-based models, we see a clear tendency for top-scoring metric performances for models trained with more classes.

Line 440: You quote a “performance” – please be clear – is that R^2 or accuracy?
We agree that this is unclear and suggest the following edit:
The main downside of these models was the inadequate 0\% and 100\% performances with scores of 83.6\% and 80.41\%, respectively. For the MSE and 90.3\% and 83.76\% for the BCE optimised model.
The main downside of these models was the inadequate 0\% and 100\% accuracies with scores of 83.6\% and 80.41\%, respectively, for the MSE and 90.3\% and 83.76\% for the BCE optimised model.

Line 441: “labelling” – using “labelling” implies the ground truth is incorrect. Call it “segmentation” if you are referring to the prediction.
We agree that “labelling” is an unfortunate choice of wording, and as suggested we replace it with “segmentation”:
These low scores often occurred due to incorrect labelling of areas with high SAR noise, bright near-range fields or wind-roughened open water areas.
These low scores often occurred due to incorrect segmentation of areas with high SAR noise, bright near-range fields or wind-roughened open water areas.

Line 446: See specific comment on “Class Imbalance”.
Noted.

Line 456: This is an incomplete thought on my copy. I would be very careful in bringing up super resolution without a citation and clear thought. Super resolution is a huge topic in its own right. Dropping this in here without clarification or some evidence to do so is confusing. This is also probably better off in future work.
We agree and have removed the sentence.

Reviewer 2:

The key issues that this paper tries to address are to explore the benefits of combining classification with regression, and the performance differences of different losses. From a machine learning methodology perspective, the contribution is marginal. From the application perspective, it is also vague to know what are the implications and relevance for operational ice charting in terms of improving the current state-of-the-art approaches, e.g., the AutoICE competition models and results. The data used in the research is also partial. Please find below my detailed comments.

Thank you for the insightful suggestions and comments. While we do agree that the manuscript contribution from a machine learning perspective is minimal, however, from the application perspective, we are exploring new approaches to use domain knowledge to better model the problem at hand, introduce new ways of looking at sea ice concentration charting and new metrics to analyse the complicated model output. We are not trying to create the best possible model but rather to investigate strategies to create better models using, among others, as few data sources as possible, which also has advantages from an operational perspective with a simpler operational pipeline. We compare our models with previously published models from the literature. We argue that our findings and discussions are still useful to the community despite not comparing them with the recently closed AutoICE challenge.

(1) By “In the same inference combination the 6-class version scores best on the multi-accuracy metric”, do you mean “In the same inference combination the 3-class version scores best on the multi-accuracy metric”. Please carefully review the manuscript to address similar problems.

We have changed the 6-class version to the 11-class version.

(2) The analysis of the comparison results of these losses are very “shallow” in the sense that it only answers “what are the performance differences” but does not answer “why are they different”. Considering that a significant proportion and contribution of this paper is the comparison between different losses, e.g., CE, BCE, MSE, EMD, I suggest the authors introduce the differences of these losses in the methodology section of this paper, and use these theoretical differences to explain the patterns in the experimental results in sections 4.1, 4.2 and 4.3 of this paper.

We opted not to include an in-depth analysis of the available loss function approaches, as this has already been carried out in our earlier paper, https://www.nature.com/articles/s41598-023-32467-x, which was solely devoted to this problem. Instead, we suggest rephrasing the sentence on line 194:

More details are available, including a definition of the $\emd^2$ loss function in \cite{Kucik2023AI4SeaIce}.

In Kucik and Stokholm 2023, a detailed analysis of various loss functions for the sea ice concentration classification was carried out. In this paper we perform further experiments to

explore further the potential of the loss functions: CE, BCE, MSE, and EMD2 discussed therein."
We agree that more discussion regarding the effect of the different loss functions on the output patterns could be made. We have added these points to this:
Line 392:
$\emd^2$ has some degree of inter-class understanding as it calculates the difference between the assigned individual class probability distribution and the reference distribution. However, It is evident that this additional inter-class relationship capability of the $\emd^2$ over the CE loss function does not have a significant impact on the model's ability to separate the macro classes.
Line 394:
Given the strong inter-class relationships of the intermediate classes, it is natural that regression-based loss functions excel at reconstructing these classes, as MSE and BCE model the loss based on a geometrical distance and a general ice probability, respectively. The CE and $\emd^2$ loss functions on the other hand have a lower capability of modelling the inter-class relationships with the CE assigning individual class probabilities. Here, however, the additional inter-class capability of the $\emd^2$ loss does appear to have a superior influence over the CE loss. With the degree of inter-class relationships ranging from none to most, CE, $\emd^2$, BCE and MSE, are in the same order of the IntDisc R^2 performance.

(3) The “best” accuracies are highlighted in the tables. But, what does these accuracies mean from an application perspective? For example, in table 5, the “best” Int R2 value is 58.49%. Is this accuracy good enough for operational mapping, and why is the proposed approach significant for operational SIC charting? I suggest the authors illustrate all “best” accuracies from this perspective to better justify the proposed approach.
This is a good point and it is an open discussion in the automatic sea ice charting community. What constitutes “good enough” is difficult to say, depending on whether we want to strictly mimic the human-labelled charts or create something new and potentially better. A higher int R^2 value implies that the intermediate class predictions in the model-produced ice concentration map are more similar to the human-labelled charts, which at least from training perspectives is a good measure of improvement since this is what we currently compare the outputs against.

(4) What are the 11-class reference models? Why are comparisons with these models important? Are they published state-of-the-art models using the AutoICE dataset? It is more relevant to compare with the results from the winning solutions from the AutoICE competition.
The 11-class reference models are used for comparison to previously published models. Like the models trained and presented here, they utilise the previous version of the dataset, the ASID-v2 data and utilise the same test set and thus a more fair comparison can be drawn. While we agree that comparing the models with the winning solutions from the AutoICE competition would be more appropriate, these were not available at the time of manuscript submission (11th of May) and were still not published at the time of writing.

(5) Scene acquired September 3, 2018 -> Scene acquired on September 3, 2018
This has been adjusted as well as the other instances of similar occurrences.

(6) and the third row with the Earth Mover's Distance squared loss function. -> and the third row with the EMD squared loss function. Please address similar issues.
The figure caption is revised as suggested.

(7) In Figure 4, the reference model using CE loss tends to generate a map that is more visually consistent with the HH and HV channels, although it is more inconsistent with the ice chart. How do you account for the uncertainties and even errors in polygon definition and SIC values assigned by human experts? To what extent do the errors and uncertainties in the ice chart lead to misleading conclusions in your experiments?
Currently and to our understanding, human-labelled ice charts are deemed the best source of high-quality sea ice charts, though as highlighted in the manuscript, there are documented uncertainties in the charts. Much of this uncertainty boils down to how the SAR image is both interpreted (e.g. who made it) and where the polygons are drawn, which can alter the ice concentration. As the ice analysts are in a somewhat rush to publish new ice charts, some areas receive more detail and resolution can be coarse. These factors are difficult to model as we do not have many quantitative or qualitative numbers. However, one way we attempt to acknowledge that there is some uncertainty in the ice charts, is by using the flexible-class accuracy as this metric ignores small errors in the intermediate predictions, and thus arguably better captures the model performance. As we are not professional ice analysts, it is difficult to argue whether something is an actual error. However, we try to filter out scenes (or maps) that have obvious flaws (e.g. scenes with polygons that were not closed, as thus cover large chunks of open water), though far apart. It is true that model metric performance can be disproportionally penalised, such as in the case of Figure 5, which is ultimately a question of where the polygon is drawn. Therefore, qualitative evaluation is important but limited by time.

(8) For the question "It is somewhat surprising that separating pixels into fewer categories does not lead to greater separability improvements, as logically, it should be an easier task to predict fewer classes.", the authors are encouraged to do analysis from the increased inner-class variation perspective. For example, combining classes will lead to large class heterogeneity/variation, leading to vague definition/signature of a class. Combining classes leads to a "weakly" supervised problem with larger ground truth vagueness and inexactness, but using 11 classes is a "strongly" supervised problem with fine-grained ground truth that is potentially more helpful for the classifiers to learn class signatures.
This is a valuable suggestion, which we would like to explore in further detail. However, we believe that the current scope and length of the manuscript leave little room for this. So to give it the appropriate attention, we prefer to investigate it in our future research.

(9) Another avenue could involve investigating super resolution -> Another avenue could involve investigating super resolution.
Based on the other reviewer's comments, we have removed this sentence.

(10) Why the authors did not use the ancillary data in AutoICE dataset, e.g., passive microwave (especially considering that the authors said that PM can be helpful), weather data, and distance to land data? Using these data can address many problems identified in this paper, and also allow fair comparison with the AutoICE competition results to better justify the proposed approach. So the authors are strongly suggested to do this.
The dataset used here is not the AutoICE dataset, but rather the predecessor, the ASID-v2 dataset. Only the SAR data is used to better compare with the previous AI4SeaIce models, https://www.nature.com/articles/s41598-023-32467-x and https://ieeexplore.ieee.org/document/9705586, which are exploring and documenting different strategies for developing better SIC models. While we agree that passive microwave data, weather data and distance to land could address or improve on some of the obstacles mentioned in the manuscript, there is, at least to the authors' knowledge, little empirical published evidence to support this.

(11) Why was the patch size set to be 768? Why did you use 8 blocks for Unet? Have the authors tried other patch sizes? How do you ensure that the base model is optimized in terms of input data preparation and hyperparameter tuning?
We utilise the same hyperparameters for easier comparison with previously published manuscripts, i.e. https://www.nature.com/articles/s41598-023-32467-x and https://ieeexplore.ieee.org/document/9705586, where the latter experimented with multiple patch sizes and model sizes. However, we do agree that more could be done for better hyperparameter tuning. We suggest the following edit to better reflect the first point:
We train models with the U-Net CNN architecture \citep{Ronneberger2015U} containing 8 encoder-decoder blocks (16 and 32 filters in the first two and 64 filters in the remaining) for 100 epochs, each with 500 batches (training steps).
We train models with the U-Net CNN architecture Ronneberger et al. 2015 containing 8 encoder-decoder blocks (16 and 32 filters in the first two and 64 filters in the remaining) for 100 epochs, each with 500 batches (training steps), similar to those utilised in \cite{Kucik2023AI4SeaIce, Stokholm2022AI4SeaIce}.

(12) In table 3, what are the differences between 11 cls and 11 cls in the reference model?
To clarify, the reference models are identical to those trained in Kucik and Stokholm 2023 but trained from scratch. The reference model parameters were selected based on achieving the highest R2 score as mentioned in the paragraph starting on line 246. The other 11 class-trained models are selected based on a high MacroBins score. Thus the reference models exhibit the performance that the models in Kucik and Stokholm 2023 would have.

(13) there are too many measures of accuracy, some of which are not defined. For each measure, please define its equation explicitly, and explain the differences among these measures, and justify why so many measures are needed. Now, it is very confusing.

We agree that the number of accuracies can be confusing, particularly because during an internal review, we changed the “multi-accuracy” to “flexible-accuracy” and we missed some occurrences (given that we use the “multistage” wording as well, there are many instances of “multi” present in the manuscript..). We hope that this is more clear now.