Reviewer 1

General response

Thank you very much for your in-depth feedback and for providing us with such helpful comments. All of your comments are very much appreciated and have helped us to improve the manuscript. To summarize the most important responses: The profiles have been labeled with additional in-situ observations at hand (Micro-CT and NIR) – we added more information on the complete labeling process in the manuscript and provide a comprehensive overview in the additional complementary material. We will also include a more detailed discussion about the micro-mechanical properties of the different snow types and the relation between classification difficulty and micro-mechanical properties. In our responses you can find an explanation about the “qualitative nature” of the validation process, why we find it important to include such an evaluation and why this study cannot solve the subjectivity of snow grain classification in general. We hope that you find all your other suggestions addressed in the below responses. We found them all very helpful and will include them in our revised manuscript. Thank you for your time and helping us improve the manuscript significantly.

Point-by-point responses (following the order of the comments)

- *[2] consider rewording as, "at submillimetre intervals in snow depth."*
    - Accepted, will be reworded.
- *[12] The software "snowdragon" has not been previously explained. Incorporate a sentence within the abstract to give the reader context to this software, as to avoid confusion.*
    - Agreed, thank you so much. The following sentence will be added:
      “The findings presented will facilitate and accelerate snow type identification through SMP profiles. Anyone can access the tools and models needed to automate snow type identification via the software repository ``snowdragon''. “
- *[31] I caution against using this language, as proposing the replacement of trained scientific specialists with a "blackbox" software raises philosophical discussion which is beyond the scope of this manuscript.*
    - Thank you for pointing this out! We used to have a larger paragraph here explaining how remote sensing scientists working on just one project might prefer using a software instead of learning to categorize snow types for just one project. Now, this sentence only states the replacement of trained scientists, which is

absolutely not what we wanted to communicate, so thank you for bringing our attention to this sentence.

- We will rephrase the sentence:
  "...(3) support interdisciplinary scientists who are unfamiliar with snow type categorization…"

- *[32] Consider a segue between the previous and the next using an introductory sentence such as, "Snow type classification has previously been accomplished using supervised ML algorithms."*

  - Agreed, the paragraph starts very abruptly. New sentence:
    "Several previous works have addressed the task of automatically classifying snow grain types with multivariate statistics or machine learning algorithms."

- *[37] Consider joining the following three paragraphs into one, as they are a bit meager to stand alone.*

  - Agreed. We will join the paragraphs and add the following conclusion sentence at the end:
    "Thus, previous work showed that supervised machine learning algorithms are a promising pathway to automatic snow grain categorization."

- *[52] I think the novelty of your approach is quite apparent and this disclaimer does not add value for the reader.*

  - Accepted. The disclaimer will be removed.

- *[general] This term "ground truth" is misleading, as the interpreted profiles were not validated by an exhumed pit. Consider referring to these as 'Labeled' throughout.*

  - We completely agree that ground truth is misleading, as there is no such thing as ground truth. We will adapt this throughout the manuscript. We will also update Figure 3 accordingly.

- *[75 – 76] Without validating the SMP force profile labeling, how can you be confident in the interpretation? Some discussion is given on this point later, but I think an additional sentence or two, which clearly states that no in-situ comparisons of grain-card type cryptography, micro-CT scanning, NIR photography, or SSA measurements were collected with the SMP observations. You must convince the reader here that your entire ML methodology which relies on these labels is still valid. A statement on the confidence and expected uncertainty in these interpretations is needed in the methodology section. And a justification as to why no corresponding validation measurements were collected, even if the explanation is that it is too cold or windy on the sea ice to bother with these*

observations, needs to be provided.

- Thank you for that valuable comment. We will add a paragraph explaining what is going on here and pointing out that Micro-CT data has been used to fine-tune the labeling where possible. We will also add an appendix that explains the complete labeling process in more detail and makes it hopefully more transparent to the reader.
- The paragraph that will be added:
  "The labeling was conducted by a snow expert and is based on the properties of the force signal (magnitude, frequency, and gradient) and the signature of the SMP-signal \citep{schneebeli1999measuring}. Micro-CT samples and NIR photography were used to fine-tune this process by validating the grain types identified from the force signal. However, these additional measurements are not available for each SMP profile for the following reasons: 1) Time constraints, i.e., only a few hours were available to perform all measurements within one snow pit; 2) harsh conditions on Arctic sea ice make snow pit measurements challenging."

  "Throughout the expedition, there were different operators conducting the snowpit measurements. As a result, stratigraphy analysis and in situ snow grain classification from snow pits would not be continuous since they vary from person to person. We reduce the subjectivity of in situ snow grain classification, which would introduce variability in the dataset. Instead, we use one person to create the training dataset. This reduces operator biases. The SMP is able to provide profiles fast, without physical labor, and independently from the person who measures them. The labeling procedure that was conducted on the collected SMP profiles is described in more detail in Appendix \ref{app:labeling}."

- *[77] What is the quantifiable difference between two expert interpretations of the snow types for the SMP signal? How can one deduce a particular snow type, and distinguish it accurately, from a qualitative look at the SMP signal?*
  - We will address those issues in the additional supplementary material.

- *[82] For reproducibility, explicitly state which features were included. I can only assume that the mean, variance, max, and min force values of sliding windows and some unnamed mechanical properties derived from the shot noise method were included for the analysis.*

  How important are the micromechanical features in classification compared to the force penetration profiles?
  - Agreed, we cut too much information here – thank you for pointing us to this. We will adapt the relevant paragraph and add a Table in the Appendix that lists all features included in the data. Regarding the importance of the micromechanical

features: It depends a bit – we did ANOVA and decision tree feature extraction, which ordered the importance of each feature. Taking ANOVA results, we see that the micromechanical features are not as important, however, the decision tree importance does actually estimate L (4 mm window) as the 4th most important feature. We added a table at the end of this document on that matter. We found it more helpful to look at the feature importance for each grain type separately because different features are more or less important for each grain type. We will reorder our Appendix, add a few lines, and also refer in the main text to the feature correlation heatmap on that matter now. In the heatmap, you can see that e.g., for rounded grains wind packed, the micromechanical features are very important, whereas for melted form of depth hoar, the force values are much more critical.
    - Adapted paragraph:
      "For each SMP profile, we replaced negative force values with 0, summarized the signal into bins (1 mm), and added mean, variance, maximum, and minimum force values for those bins. Those values were also determined for a 4 mm and 12 mm moving window. Moreover, \citet{lowe2012poisson}' Poisson shot noise was used to extract δ, f, L, and the median force value for a 4 and 12 mm window. We added further depth-dependent information by including for each data point the distance from the ground and position within the snowpack. Refer to Table \ref{tab:features} in Appendix \ref{app:features} for an overview of all features used for each SMP profile, and to Table \ref{tab:feature_corr} to see the feature importance for each grain type."

- *[84, 90 etc.] Because the SMP is measuring force as a function of depth, depth-dependent seems like a more clear explanation. It took me a while, but I understand time-dependent as the snowpack history, which accumulates in time. [and all related comments]*

    - Thank you for this feedback – the main author is used to thinking in machine learning terms and framing this as "time-dependent" information (because each data point has been measured sequentially after each other == "time-dependent"). Depth-dependent does capture this really much better, and we will adapt this everywhere.

- *[89] Although for this classification purpose, it may be convenient to compile these snow types into the rare category, ice formations and surface hoar have widely different mechanical properties. Therefore we can almost expect the classification to perform poorly on the rare class. I appreciate the brief discussion on the value of separating these snow types for avalanche hazard assessment, but if you were to take an SMP profile in the rocky mountains of the western US this year, buried surface hoar would be a widely extensive class, not rare.*

  Drafting some discussion regarding the snow mechanics as a classification device,

rather than the rare appearance of this snow type in a dataset is lacking from this work.

- We are completely aware that the classification performance on the class “Rare” will be bad due to the different mechanical properties. We did this to simplify the evaluation of the models and in order to not skew the balanced metrics too much: If we have many rare classes where it is almost impossible to achieve good performance (not enough data for ML), this will lower the overall accuracy heavily since each class is weighted not according to occurrences but according to the overall number of classes. Hence, we had the feeling that the overall performance of the models is easier to evaluate for everyone if we summarize those very rare grain types into one class. We also could have dropped them completely as commonly done in previous work, however, we still wanted to include those occurrences because it was important to us to show how models perform on a real-world dataset – and each dataset will usually have some rare classes, and in our opinion, it would be a loss to force practitioners to drop those profiles.

- Regarding the notion of which grain types are “rare” – we completely agree that this is heavily dependent on the dataset at hand, and we invite everyone to retrain snowdragon for their specific datasets. And we highly encourage summing up other grain types to “rare” in those contexts. We will include a more detailed user guideline to make this more transparent. We will also adapt the text to make clear that the notion of “rareness” only applies to the MOSAiC dataset and is not meant as a general categorization.

- Adapted text:
  “The few occurring ``Ice Formations'' and ``Surface Hoar'' instances in the MOSAiC dataset are summarized in the class ``Rare''. While a high classification performance cannot be expected for the rare classes, we still include them to show how the models perform on a ``real-world dataset'' that in most cases will also include classes with few occurrences.”

- Regarding your comment on snow mechanics as a classification device, we will answer this later in our response.

- *[93] Balance and imbalance are ML jargon terms that could be more clearly defined to improve the readability and interpretation of the results.*

  - Yes, that is true, thank you. This will be adapted in the following way:
    “The resulting dataset has the following properties: (1) There are multiple, noisy, and overlapping classes. (2) There is a between-class imbalance, i.e. some grain types occur much more frequently than others. (3) There is a within-class imbalance, i.e., some grain classes contain different sub-grain classes, but some

of them are more frequent than others."

- *Minor corrections regarding the abbreviations of AUROC etc.*
    - All accepted, will be changed accordingly.
- *Minor corrections regarding "generalized data"*
    - We prefer naming this specifically "out-of-distribution" data.
      Generalized data transports the message that if we train a model on this "generalized data", it can actually generalize to anything. If we use "out-of-distribution" data, we can test the generalization capabilities of a model, but it does not entail that the model can generalize to anything. A model trained on "generalized" data sounds like it could generalize to anything.
    - We still will make the following adaptations:
      "(3) The generalization capability is tested by running the best-performing model on 100 random profiles from different parts of MOSAiC winter data. These profiles are outside the distribution of the training, validation, and testing data, and we refer to them as ``out-of-distribution profiles``. Here, the ``out-of-distribution'' profiles contain the same classes as the training data, so the model still has a chance to predict the correct labels."
- *Other linguistic suggestions on Page 6*
    - All accepted, will be changed accordingly.
- *[167] This is an assumption, as the generalized data are not validated. It is better to write, "however, if the general data contains the same classes as the training data…"*
    - Yes, we made sure to choose an out-of-distribution dataset that contains only labels known to the ML models. We will adapt the text to make clear that specifically in our case, we made sure that the out-of-distribution dataset from which we draw the profiles contains only labels known to our models. (See paragraph mentioned above).
- *[162] Expand this section to include more detailed description of how accuracy is calculated, how balanced accuracy differs/what information is conveyed by this metric, how weighted precision is used as an uncertainty metric. This is explained for AUROC, and should be explained for the other metrics to give a general reader context to the evaluation metrics. F1 score is undefined*
    - We agree that a more detailed description of the metrics is important for all the readers who do not work with those metrics on a daily basis. We found it most

helpful to add another appendix to this end so we are able to provide both formulas, definitions, and intuitive explanations for all metrics. We are also defining the F1 score now.

- *[171] Consider joining this section with the previous Section on Evaluation, as the section is quite short to stand alone.*

    - In machine learning papers, it is common to separate the evaluation and the experimental setup, so the reader can look up the experimental setup immediately to check if they have the means and resources to reproduce the experiments. We understand, though, that the sections were quite short to stand alone. We are going to add some more details to the experimental setup, and with the planned changes of the evaluation section, we hope the paragraphs will be able to adequately stand alone.

- *[175] Define/Explain hyperparameter tuning*

    - Adapted paragraph:
      "Hyperparameter tuning is the process of searching for the optimal internal learning settings of an ML model. Hyperparameters control the learning process of the models, whereas parameters are learned by the model. The tuning is performed on the validation data and the hyperparameters that achieve the highest performance for their model chosen for subsequent model evaluation. Here, hyperparameter tuning was applied moderately and with a simple grid search. All tuning results can be found in the GitHub repository. Specifications of the machine on which the experiments were run can be found in Appendix \ref{app:machine_specs}, and descriptions of the model setup can be found in Appendix \ref{app:model_setup}."

- *[Figure 3:] This example between 150 - 200 mm on the Medium depth profile gives me a bit of pause. I would agree with the LSTM model here, which defines this fairly obvious series of layers. If I were to interpret this data, I would have a very difficult time discerning the snow type of this layer. I am not calling into question here, the interpreter's decision, but without validation I struggle with confidently recognizing such layering as a homogeneous snow type.*

    - We agree that without validation, it is arguable if that layering is a homogeneous snow type. The two peaks in the "medium" profile could (probably) be more traditionally classified as "wind crust", however, we did not include this class. It is interesting though that all three ML classifiers classify the peaks differently, hinting to a larger uncertainty in these predictions.
    - Your observation is an example of how ML models can support practitioners in their analysis: Essentially, the LSTM model tells the user: "Inferring from how you labeled your other profiles (training data), I would suggest the following series of

layers”. Throughout the process of this study, we observed that the LSTM model can actually help to discover inconsistencies in the labeled data or human mistakes.

- *[217] This result leans into the hypothesis that it is mechanical properties that are more differentiable between classification than the count of appearances. Precip particles are a unique class distinguished by the relative lack of bonding among fresh snow.*

    - Thank you for pointing this out – we will now discuss this in more detail in the discussion section than we did before.
    - Mechanical properties of the snow influence the penetration force signal both in magnitude and characteristics of the signal. By evaluating the signal we have taken into account the mechanical properties of the snow layer. This will be explained further in the additional supplementary material.

- *[219] Explaining exactly which characteristics would significantly increase the significance of this work.*

    - The characteristics of each snow grain classification will be outlined in the new supplementary material.

- *[224 – 225] Possibly because "Rare" is comprised of mechanically different snow types, while "Precip Parts" is comprised of mechanically similar snow. Analysis of the micromechanical properties that are inverted for via the shot noise approach would increase the value of this and subsequent discussion.*

    - See response for [217].

- *[general] A general response to all comments regarding the micromechanical properties of the snow types and how they can be used as a classification device.*

    - We will give a physical reasoning why these different snow types differ in their micromechanical properties and how these properties develop through metamorphic processes.
    - That they can be used as a classification device is already entailed since the micromechanical properties create different types of SMP signals, and we classify exactly those signals with our models.
    - For us, it was a general problem that the International Classification does not represent very well snow types occurring on arctic sea ice.
    - We will add parts of this - where appropriate - in the discussion and provide more detailed descriptions in the supplementary material.

- *[254] Analysis of the spatial variability of snow class composition derived from ML prediction would be a valuable contribution which could justify this claim. In lieu of such*

analysis, draw from the literature to better quantify the length scales of variability for snow on sea ice.

    - We believe spatial variability analysis of this dataset is beyond the scope of this paper, but we wish to conduct this analysis in the near future.

- *"grouping" instead of "island"*

    - Accepted, will be changed accordingly.

- [310-311] This hypothesis is a bit unrefined. All snow types are transformative and related to one another through metamorphic processes. Please clarify this statement with more physically-based reasoning. As it is written, precipitation particles should be equally non-separable because all snow has metamorphosed from fresh precip. Include indurated hoar in this discussion. What about the mechanical properties of these snow types is similar (or different) which may cause difficulties in their classification?

    - Thank you for your comment, we will adapt this paragraph and discuss the metamorphic processes and the mechanical properties of these snow types in more detail as mentioned in our previous responses. Once again, the mechanical properties are the underlying driver for the classification, even when we speak about metamorphism. If two data points are very close to each other in terms of metamorphism, it entails that their micromechanical properties are close to each other. The transformation between the snow types means a transformation of their micromechanical properties. And similarly, as you suggest that some snow types have mechanical properties that are more similar to each other, we suggest here that there are metamorphism states that are more similar to each other than others.

- [314] Despite the evidence supporting that 80% accuracy appears to be a contemporary maximum for classification accuracy, the claim that 100% accuracy is virtually impossible is not justified. This language presents the notion that further advancement in this field of science cannot be achieved, and I caution against communicating in a way that paints you into a corner based on this opinion.

    - In the ML community, it is quite normal to assume that 100% accuracy cannot be achieved – such a model would just be overfitting. Of course, there is a lot of space left between 80% and 100%, and we do absolutely think that further improvement is possible. But we also want to communicate that one cannot aim for such high accuracy as the ML community is usually aiming for, e.g., on the MNIST datasets. We found it important to communicate that a relatively low accuracy of 80% (for the ML community low, given their "perfect" datasets) is still a lot in a setting where classes are not clearly separable from each other.

    - Nevertheless, we will adapt the wording since it was apparently a bit too extreme:
      "it is currently impossible to reach 100% [...]" instead of "it is virtually impossible to reach 100% classification accuracy on every snow type since some snow types will always lie between two categories".

- *[319] This type of uncertainty can be reduced with in-situ observations of snow type through methods of crystallography etc. The labeling process is not intrinsic to the SMP analysis. The design of the experiment presented relies solely on interpretation of SMP profiles, and this choice should be discussed here. Any quantifiable uncertainty that was learned through repeated expert labeling should also be included to shed light on the value of ~80% accurate snow-type classification.*

    - Yes, we agree with you, thank you for bringing this up. We will adapt this part in the following way:
      "The uncertainty during labeling is an inherent problem of SMP analysis: The annotation of SMP profiles is subjective, meaning that two different snow experts may produce two different labeled and segmented profiles for the exact same measurements \citep{herla2021snow}. This intrinsic uncertainty can be partially mitigated by supplying additional in-situ observations of the snowpack, e.g., through methods of crystallography or Micro-CT measurements. But even with additional observational data, experts might provide different annotations of the same profile. Both experts might agree that both labeled profiles are valid analyses of the same profile though. In conclusion, the model's performances cannot only be measured in terms of accuracy because models with low accuracy might still produce sensible, directly usable predictions."

    - We are bringing up the topic of quantifiable uncertainty as well now in the discussion – thank you for coming up with this idea:
      "As previously discussed, the uncertainty of the expert labeling is a general limitation of this particular study. While this uncertainty might be partially mitigated further by using a dataset for which many additional in-situ observations exist, it would still remain an issue. One approach for future work would be to quantify the uncertainty that is inflicted upon the labeled profiles. Subsequently, a machine learning model could be trained to classify not only grain types but provide a \textit{probabilistic} classification."

- *[323] fix the spacing surrounding the clause*

    - Accepted, will be changed accordingly.

- *[351] This sentence adds little to the discussion. It seems as though this result is "completely unclear" because such analysis has not been completed. One possibility to explore, in lieu of this large and complete data set, would be to train an LSTM model on*

SMP data collected from a different time and place with similar snow types, predict the classification on the winter mosaic data, and validate the results on the labeled profiles. While this analysis would not clarify if one large dataset driven model would be beneficial, it would give clarity on the spatio-temporal transferability of this technique.

    - Thank you. We will adapt our paragraph following your suggestion:
      “In theory, a large enough model trained on a large enough dataset could be able to produce direct predictions for any SMP users. Thus, it would be interesting to train an ML model on a generalized dataset and validate its' performance on the specialized MOSAiC SMP dataset. This would shed new light on the spatio-temporal transferability of the ML models presented here.”

- *[367] The qualitative sell of this method is my largest grievance with this work.*

    - We would like to point out that the work provided here is really a methodological paper that compares different machine learning algorithms for classify SMP profiles. The qualitative aspect of snow grain classification is a general issue and well known in the snow community. We are here introducing a new method to classify snow grain types but we are not suggesting that we are removing the subjectability of this research field. We are providing tools that could be used as an alternative and in the future, when the conditions in the field allow or need such a tool.
    - Your comment seems to target the fact that we have not only used numerical metrics to analyze the performance of the models but also qualitative measures, namely the feedback of snow experts. (Snow expert provides labeled profiles and checks if the models create profiles that are consistent with his/her/their labeling). We do this, because this is considered an important measurement within the field of applied machine learning. While numerical measurements such as accuracy and precision might give the impression that a task is well tackled, it often can happen that domain experts criticize the predictions of the model based on features that have not been captured by the numerical metrics. Thus, it is an important addition for us, to show that domain experts have looked at the predictions of the models and deemed them as “usuable” in the field, because this is a qualitative measure often missing in ML studies. We want to make sure that the impact of this study is not purely theoretically and this is what we want to express with this sentence in our conclusion.
    - An alternative qualitative measurement would be to compare NIR profiles with the ML-classified profiles. One could argue that such a comparison has a higher confidence. However, it still remains a qualitative measurement and will not become more objective. If you want to create labels that are very much aligned to NIR profiles, you can label the training data while using NIR photographs for each profile. The models, e.g. the LSTM, will then be able to create predictions that are particularly near to NIR photography because they were trained to do so. How well they are sticking to the desired prediction can only be evaluated

qualitatively, since we have no ultimate objective ground truth data.

- *[373] "Snowdragon" was mentioned in the abstract without definition, and was largely left out of the manuscript. I think an introduction to snowdragon is needed in the abstract, and the recap of the snow dragon repository needs to be explained in a more straightforward manner in the conclusion.*
    - Yes, thank you – somehow, this slipped our attention, we try introducing snowdragon properly now:
        - Abstract:

          "The findings presented will facilitate and accelerate snow type identification through SMP profiles. Anyone can access the tools and models needed to automate snow type identification via the software repository ``snowdragon''. Overall, snowdragon creates a link between traditional snow classification and high-resolution force-depth profiles. With such a tool, traditional snow profile observations can be compared to SMP profiles."

        - Contributions:

          "[…] The snowdragon repository that provides the tools to automate SMP labeling"

        - Conclusion:
          We will remove it here at the point where you left the comment. This paragraph was more about pointing out the general implications for the field and was not supposed to be so much about snowdragon itself. It will still be mentioned in the conclusion as a "repository", but we will separate between study and repository more strongly in the conclusion.

        - Appendix:
          Proper user guidelines on how to use snowdragon

- *[378] delete "already today"*
    - Accepted, will be changed accordingly.

- *[380] ~ Glittering Generalities ~ Describe how automated snow-type classification is essential for understanding patterns of climate change. The context as to how this work would mitigate climate change impacts is not described. Consider revising this language to more clearly state what is accomplished by this work and how it is beneficial.*

- Yes, agreed, the last sentence is indeed a bit too broad.

- New suggestion:
  “Snowdragon enables the analysis of the SMP MOSAiC dataset, a dataset containing detailed information about snow on Arctic’s sea ice. In times of climate change, this information is crucial: We need to understand the state of the sea ice in order to understand in which state the Arctic system is. For the first time, MOSAiC enables the scientific community to have access to such a detailed and large dataset. And snowdragon is one example of how ML can help us to access the knowledge behind all the data.”