EGUsphere, referee comment RC2
https://doi.org/10.5194/egusphere-2022-938-RC2, 2023

Comment on egusphere-2022-938

Pascal Hagenmuller (Referee)

Referee comment on "Automatic snow type classification of snow micropenetrometer profiles with machine learning algorithms" by Julia Kaltenborn et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-938-RC2, 2023

Review of Automatic classification and segmentation of Snow Micro Penetrometer profiles with machine learning algorithms by Julia Kaltenborn et al.

Summary :

The authors use machine learning approaches to learn the relation between the penetration profile of the snowpack and grain shape class. The final goal is to have a quick and easy way (measurement of penetration profile with SMP) to capture one key descriptor of snow stratigraphy (grain type) which is usually time-consuming to measure. The predictors comprise different quantities derived from SnowMicroPen (SMP) measurements, including mean force, standard deviation, position in the snowpack (height), and probably noise features obtained with a statistical model (shot noise model). The goal variable is the grain shape class estimated by an expert solely from the SMP profile signature. The data comprises 164 SMP profiles measured on Arctic sea ice during winter 2019-2020 and manually labeled by a single expert. This classification leads to a segmentation of the snowpack profile into distinct layers by assuming that contiguous points of the same grain shape class belong to the same layer. The authors show that the different algorithms are able to reproduce the choice of the expert (max ROC AUC of 0.94).

Main comments:

- **I am not fully getting the final objective of the paper. What is the scientific question we want to address by automatically reproducing the grain shape class inferred from the penetration profile based on undescribed expert analysis?**

- By definition, the grain shape class or snow types (Fierz et al., 2009) is related to the shape of the grains and is traditionally derived from the observation of single grains on a crystal card with a magnification lens. This measurement remains manual, is very time-consuming, inevitably contains some subjectivity, and the use of classes is limited to capture the continuous nature of snow types. Trying to overcome some of the two first limitations by automatic classification is of great interest. Different attempts exist to relate the SMP signal to scalar microstructural features of snow based on the physical interpretation of the penetration process (e.g., Löwe & van Herwijnen (2012), Lin et al. (2022)) or with direct statistical / machine learning approaches (e.g., Proksch et al. (2015)). In particular, King et al. (2020) and Satyawali et al. (2009) used the latter approach to relate MEASURED grain shape class to SMP profiles. Here the ground truth is not the measured grain shape on independent data but corresponds to the interpretation of solely the SMP signal signature.

- This direct identification has never been documented so far. The description in the text is elusive, with a reference (l.76, Schneebeli et al. 1999) that does not describe the procedure. Besides, the data presented here relies on the interpretation of a single expert (l. 75-77). One cannot evaluate any reproducibility of the procedure or agreement with ground truth based on manual observation in snow pit data. Moreover, it is highly likely that the estimation is subjective. For instance, in Fig. 1, one may wonder why only the specific layer at a depth between 98 and 102 mm is labeled as « Depth hoar wind packed » and not other layers below that show similar features. In addition, there are obviously « inconsistencies in their ground truth labeling » (l. 324) and the results are linked to « different classification styles of experts » (l.332) and the evaluation is qualitative (« classification patterns […] were satisfying to domain experts » l.368). The discussion is not convincing based only on the feeling that «in the view of the authors, a temporally consistent classification is more relevant to the interpretation of the development of the snowpack, even if there is a certain, but unknown, bias to an expert interpretation » (l. 255-257). To me, it appears, in the end, that the presented algorithms are able to reproduce one analysis of one single expert on specific snowpack types. In my opinion, this limits a lot of the interest and generalization to the snow community.

- The authors refer to Fierz et al. (2009) for describing the different snow types referenced in the paper (l. 74). However, it is not very clear how the different classes presented in the paper (see legend of Fig. 1) are defined as they are not present in the international classification described in Fierz et al. (2009).

- Grain shape class has been used since the beginning of snow science and was first motivated by avalanche forecasting. It remains the most common descriptor in snowpit observations but has many known limitations. It is a discrete class whose evolution cannot be described by differential equations in models. It cannot be quickly and objectively described. Currently, the international classification is not necessarily adapted to any snow on Earth (e.g., here, the authors added classes that are not in the classification). Therefore, one may wonder why, in general, we want to stick to this description of snow.

- The interest of the algorithm is described in grandiose terms: they make « training of interdisciplinary scientists in snow type categorization obsolete » (l. 31), «can be directly employed by practitioners for their own SMP datasets in the field » (l. 250), «

These findings will enable SMP practitioners to automatically analyze their SMP measurements. To that end, an SMP user must simply decide on one of the fourteen models provided » (l. 369-370). However, I do not understand these sentences. I understood that everything relies on a single expert analysis, that the model must be retrained on other data (e.g., snow data in other places around the world) and that without this expert, no model can be retrained. In contrast, the limits of previous studies are somehow presented unfairly. For instance, it is indicated that « This [generalization] would not have been possible in previous works such as Satyawali et al. (2009) since knowledge rules for one snow region and season do not transfer to other regions or seasons » (l. 335), but the exact same applies to their work as the model must be retrained in any case to be used on other snowpack climate or expert analysis in the end (the model of Satyawali et al. (2009) could be retrained too).

- The authors positively present the work as both « automatic classification and segmentation » (title and in the text) of snow profiles. It appears that no segmentation procedure is present in the paper. Indeed, the segmentation consists of saying that connected (i.e., neighboring) points with the same label belong to the same segment.

2. On the form, the description of the work is sometimes vague and incomplete.

- The objective of the paper described in l23-31 seems rather unclear to me. It took me several reads to understand that the goal is to reproduce the classification of one expert on SMP data.
- There is a welcome short bibliography on previous attempts to classify SMP profiles automatically. The description of the selected articles (Satyawali et al.,(2009), Havens et al., (2012), King et al., (2020)) would benefit from more detailed statements to capture what was really done in these papers. For instance, what is « too small to be representative » (of what?) (l. 34), « including knowledge-based rules » (l. 35), « good accuracy » (l. 42), and « additional snowpit information » (l. 42)?
- Fig. 1, the international classification (Fierz et al.,2009) provides a color code. Is there a specific reason for not using it?
- One key piece of information about the procedure is the list of predictors used as input for the ML model. They are very shortly described l. 79-86. But the description is too elusive to understand which variables are used. What are « added additional features», « time-dependent information » (where is time here ???), and « including variables of the shot noise model » (which variables?)?

Overall, I did not understand the overarching objective of the work and its applicability to different data. In addition, the presentation of previous and present work is too vague to capture the key ingredients of the methodology. I, however, acknowledge an important work to test numerous machine learning models (14) and the effort to provide the code source directly on a git repository.

I do not feel that the raised issues can be solved before publication.

Pascal Hagenmuller