Dear editors,
Dear reviewers,

We would like to express our deep gratitude for the outstanding effort made by the reviewers in providing feedback, which has greatly improved our manuscript. With the submission of the revised manuscript, we hope to meet the high standards requested.

Kind regards,
Andri Simeon

Table of Contents

Referee 3#:

General Comments

This study develops deep-learning models to identify seismic signals of avalanches.
I am a seismologist working on non-earthquake signals (avalanches, landslides, glaciers…).
But I have no expertise on deep learning methods.
The introduction and the seismological methods are well described. But I had trouble to understand the deep-learning methods.
I think the authors did not make much effort to make their work understandable and interesting to readers that are not familiar with deep learning methods.
I did not know many terms (f1 score, recall …), but I learned from other sources. But readers of GMD are probably more familiar with deep learning methods than me.
I am thus not able to criticize the part on deep leaning methods.
The results of the model seem correct, but not as good as previous similar studies (eg, Provost et al), and maybe not good enough for real time warning.

We thank the reviewer for providing valuable feedback on our work. Introducing and explaining deep learning building blocks in detail is beyond the scope of this work. The neural network layers and metrics are state-of-the-art and well-documented. Excellent online resources exist for readers who wish to delve deeper into the underlying mechanisms. Nevertheless, we have revised the manuscript in accordance with all the reviewers' comments and enhanced the description of the machine learning models for improved clarity.
Considering the previous work of Provost et al., 2017, we would like to emphasise that a direct comparison is challenging since their objective is entirely different. They developed a model to detect entire events rather than subsequences. In this study, we adopted their feature extraction method in our baseline to make a comparison possible. However, the signals generated by landslides and avalanches are different. Landslides usually generate higher amplitudes and frequency content, which depend on the sensor-source distance. In addition, the signals generated by landslides are usually longer than avalanche signals. For instance, Provost et al. (2017) used different frequency bands to compute the seismic features and used a cross-validation strategy to evaluate the model's performance instead of using an independent test set.

Reference:
Provost, F., Hibert, C., and Malet, J. P.: Automatic classification of endogenous landslide seismicity using the Random Forest supervised classifier, Geophysical Research Letters, 44, 113–120, https://doi.org/10.1002/2016GL070709, 2017.

Specific Comments

1) Objectives
It is not clear to me what are the long-term objective of this work, beyond showing that machine learning can detect seismic signals of avalanches, more or less as well as experts?
If the final goal is to monitor avalanches and investigate their flow properties, experts could do that better and faster than machines, given the relatively small rate of avalanches.
If the goal is real-time warning, do you think that there would be enough warning time (from detection until propagation to roads, villages …) to allow mitigation actions?
We have tried to make our objectives more transparent in the introduction (Lines 72 -76). Monitoring avalanches on a large scale in (near) real-time is a long-term objective of this study, which is not feasible to carry out manually. Avalanches typically occur during winter storms, making it impossible for humans to monitor releases, especially in remote areas. Also, automatic cameras cannot help in these cases. Therefore, expensive Doppler radars are the most reliable solution to monitor a single avalanche slope. In contrast, seismic detection systems provide a cost-effective solution that is scalable to cover wider areas and, therefore, offers a higher spatial resolution of avalanche activity. But then again, having experts interpret all this data in real-time is not practical. Thus, developing reliable algorithms to automatically classify signals in seismic data is a necessary first step before setting up an early warning system. Hence, we studied and developed models that could provide a solution in the future. A question that naturally appears when trying to monitor avalanche activity in (near) real-time is that of real-time warning. However, since developing a real-time warning is a desirable future application of this study, we presented and discussed this as a broader outlook in the Section «Applicability to early warning and monitoring systems» (Section 6.4).

2) Time window
Seismic signal is divided in windows of 10 s, much smaller than the typical duration of avalanche seismic signals.
I understand that a short window is preferable for early waring (but is this your goal and is it feasible?).
But for classification, I think a longer time window (about 60 s) would improve the results by allowing to better detect the characteristic spindle space of avalanche signals and the consecutive P, S and coda waves of earthquakes
A shorter time window is preferable for early warning and real-time monitoring of avalanche releases. Most importantly, using small window sizes enlarged the avalanche data used to train the models. Additionally, working with smaller time windows would, in principle, allow for the segmentation of different events, hence capturing the duration, release and setting of avalanches and offering additional data to study events. Developing a neural network from scratch using only 84 avalanche signals, i.e., the number of recorded avalanches, is unrealistic. However, we agree that longer windows would better capture the characteristic shapes of different signals, and we will consider this for future work.

3) Network based attribute
Why did you choose not to use network attributes, although you have a network of 5 sensors and network attributes were found to improve the results in the study of Provost et al. (2017)?
For instance, computing cross-correlation between sensors provides the peak correlation and time delays, which could be useful to distinguish avalanches from earthquakes or noise sources.

The correlation between sensors is much larger for distant sources (earthquakes, quarry blasts) than for nearby signals (avalanches and noise).
The time delay depends on apparent velocity, which is larger for deep and distant sources (earthquakes) than shallow and nearby source (avalanches, acoustic waves).
We avoided using network attributes since we wanted to explore whether a single sensor suffices to detect avalanches. Our reasoning was based on anticipating future scenarios where having multiple sensors cannot be taken for granted when moving to large-scale avalanche monitoring. Again, like the windowing, using the five sensors separately provided us with more data to develop the models. In practice, when multiple sensors are available, their integration, as well as derived data, can be seamlessly incorporated into the system.

4) Infrasound
Each seismic sensors was colocated with an infrasound sensor. Why didn't you use infrasound data?
I guess that the amplitude of the infrasound signal (relative to the seismic signal) should be much higher for avalanches than for earthquakes?
Incorporating infrasound data was not within the scope of this study. Furthermore, the acoustic system did not function correctly during the first winter season, that is, 2020-2021. Consequently, the database available for developing machine learning models was too small. Nevertheless, we agree that this system could provide valuable additional insights and will be considered for future work.

Minor comments;
Figure 8: what are the main features shown in x axes? ("ES[3]" "DISTQ3Q1" "F21"...?).
We have added a brief description of the three engineered features in the caption. ES[3] is the energy in the frequency band [6, 9] Hz, DISTQ3Q1 is the mean distance between the 3rd and the 1st quartile of all FFTs as a function of time and DISTQ3Q2 is the mean distance between the 3rd and the 2nd quartile (Feature 36, 57 and 56 in Table B2 and B3). The features starting with 'F' are the features from the autoencoders. They do not have any direct physical meaning.

Figures 2 and 12: Could you also show the spectrogram of the signal? Avalanches are often easier to identify by looking at the spectrograms than seismograms.
We have added all spectrograms alongside the waveforms in the revised version.

l349 "two-thirds of the avalanches were not verified". You mean, by cameras or radar? But they were verified by experts?
One-third has been verified by radar and/or cameras, so it can be considered ground truth. The remainder was labelled by experts and, thus, are not verified ground truth.
We clarified this by adapting the text to (Line 404):
... (two-thirds of the avalanches were neither verified by the radar nor the cameras).

l424-425 What do you mean by " Thus, the models could, in turn, be considered to annotate large datasets, which in turn can be used to detect fine precursor signals."?
We realised this sentence was not clear and informative. Therefore, we have removed it from the manuscript. Initially, we wanted to say that the models could label new or unlabelled datasets. These datasets could then be used to develop new models to detect the initial signals of an avalanche for early warning.

l462 "Also, these event types typically generate signals with a higher signal-to-noise ratio than avalanches".
I don't agree. All gravitational and tectonic processes generate a majority of small events (eg, Gutenberg Richter law for earthquakes). The minimum size is always the detection threshold.
We have removed it from the manuscript.

l469 "In contrast to our approach, they only considered avalanches verified on camera images or manually 470 picked events."
I don't understand. I thought you also manually picked the avalanche seismic signals and verified part of them with radar and cameras?
We have removed it from the manuscript.

Referee 4#:

General Comments

This manuscript develops unsupervised machine learning techniques of seismic data recordings to detect snow avalanches, showing a success level greater than expert interpretations and providing insights and future improvements in the study. The work is presented with respect to the previous studies employing machine learning for seismic event detection while highlighting their significant and novel contribution of unsupervised feature extraction. I found the paper to be informative and complete in analysis yet lacking full clarification and description of key methodological steps. I recommend the paper for publication after minor, clarifying revisions, which will improve the readability and methodological completeness of the manuscript.
We appreciate the time and effort that the reviewer has invested in providing valuable feedback on our work. We have thoughtfully considered the points raised and have endeavoured to improve accordingly. Our responses to the reviewer's suggestions are detailed below.

Specific Comments

Section 4 is difficult to wade through with lengthy explanations which leave out important methodological developments.
Reviewer #6 had similar concerns. Therefore, we put a lot of effort into rewriting and restructuring the method section. For details on this section, we refer to the marked-up manuscript, which contains all the changes.
Section 4 is now structured as follows:

(4) Model development
- (4.1) Baseline features
- (4.2) Autoencoder features
  - (4.2.1) Architecture
  - (4.2.2) Training Regime
  - (4.2.3) Validation
  - (4.2.4) Model selection
- (4.3) Feature classification
  - (4.3.1) Random forest model
  - (4.3.2) Cross-validation
  - (4.3.3) Inference and post-processing

Section 4.3 contains a significant amount of information which reads as discussion material and should be placed into Section 6 to improve the readability of the manuscript. Paragraphs have been highlighted in the marked-up document to
convey this.

We have combined and relocated both of these paragraphs to the discussion section. This should enhance the reading flow and minimise duplicated text.

Line 281 “During inference, each tree prediction is aggregated to form a final majority vote, from which it is possible to retrieve class proportions, often interpreted as probabilities.”
The statement on Line 281 contains the entirety of the description of the model output. More care and consideration should be taken to clarify exactly how random forest classification is interpretable as a probability. It should also be stated why 0.5 probability was chosen as the threshold for event detection.
We acknowledge that the actual model output was not described in full detail. Consequently, we have taken this comment into account and added a subsection titled «Inference and post-processing» to the section «Feature classification». This subsection explains how the random forest model arrives at its predictions and what steps we have taken to post-process them.

The authors should refine the work to place methodologies in the appropriate sections.

The process of multi-sensor data aggregation is in introduced within the results on line 306, and vaguely at that. This idea can be introduced in Section 4. Additional examples of this have been identified in the marked-up document.
To clarify which tasks we evaluated our models on, we have added the aforementioned «Inference and post-processing» section, which introduces all the aggregations we utilised.

Figures in Section 4 are lacking the necessary detail to be supportive/educational information. The authors should at a minimum improve the captions to better convey the depictions.
Thank you for this suggestion, the captions were indeed relatively sparse. Therefore, we have added more descriptive captions to Figure 5 (temporal autoencoder) and Figure 6 (spectral autoencoder). Additionally, taking into account another reviewer’s feedback, we redesigned the entire Figure 4.

Additional Comments:
See the marked-up document included.

With Regards,
Tate Meehan

Referee 5#:

General Comments

The manuscript by Simeon et al. presents a novel method to automatically detect snow avalanches in seismic data collected at a test site above Davos, Switzerland. Specifically, the performance of three different algorithms is assessed: seismic attributes, temporal autoencoder, and spectral autoencoder. Based on this, they find that the inclusion of features from the frequency domain improves model performance and unsupervised autoencoders show potential as an alternative to the standard expert-based seismic attributes classification.
I believe the science behind this study is sound and aligns with the focus of Geoscientific Model Development. However, I suggest some restructuring and clarification prior to publication. In particular, the Discussion section is quite long and, as pointed out by another referee, the manuscript still contains some repetitive information. Considering Provost et al. (2017) found higher true positive rates for non-windowed signals, I believe determining the effect of different window lengths would add value to the manuscript. At least the potential effects of choosing a different window length should be discussed. I recommend the authors also take my specific comments listed below into account.
Thank you very much for reviewing our work and providing insightful feedback. Since other reviewers also suggested some restructuring, we have improved the readability of the paper by reorganising the methods and discussion sections. In parallel, we have removed duplicative information where we thought it was not needed or explicitly highlighted by a reviewer. We believe that some level of repetition is essential to provide context for related information. For the responses to the specific comments, we refer to the list below.

Specific Comments

L41 – radius of several kilometres: This becomes clearer later on in the manuscript, but I would suggest being precise from the very beginning.
This definition should provide the reader with a vague idea of the detection radius of such systems. Seismic waves are rapidly attenuated with distance, giving rise to a natural limit of detection of signals produced by avalanches. The detection range of a seismic network varies depending on the type and size of the flow, the characteristics of the terrain, and the background noise level. Hammer et al. (2017) used a seismic station of the Swiss seismological network to detect large wet-snow avalanches up to 30 km. Pérez-Guillén et al. (2019) detected large avalanches released on Mt. Fuji (Japan) at a maximum distance of 15 km. Therefore, making a quantitative statement on the detection radius that holds in general is difficult.

References:

Hammer, C., Fäh, D., and Ohrnberger, M.: Automatic detection of wet-snow avalanche seismic signals, Natural Hazards, 86, 601–618, https://doi.org/10.1007/s11069-016-2707-0, 2017

Pérez-Guillén, C., Tsunematsu, K., Nishimura, K., and Issler, D.: Seismic location and tracking of snow avalanches and slush flows on Mt. Fuji, Japan, Earth Surface Dynamics, 7, 989–1007, https://doi.org/10.5194/esurf-7-989-2019, 2019.

L46 – other types of mass movements: Consider adding a few examples.

We have added:

... such as landslides, debris flows, and lahars.

L47 – other seismic sources "such as earthquakes"?

We have changed this sentence to:

These patterns have frequently been used to detect and identify avalanche signals.

L71: You describe what an encoder is, but not a decoder.

Following another reviewer's suggestion and in order to avoid duplication, we have removed the detailed description of the autoencoder model from the introduction. Nevertheless, thank you for noting. A thorough introduction to the autoencoder architecture can be found in the methods section.

L80: Why did you use 10 s windows and not, e.g., 5 s, 20 s, or maximum length of picked avalanche events?

Referee #3 expressed a similar concern. In summary, by using smaller windows, we could generate more avalanche samples, which were needed to develop the autoencoders. Moreover, a short time window is preferable for potential real-time applications. Thus, 10 seconds is a compromise between compiling a larger number of avalanche samples for training and avoiding a reduction in length to a point where the avalanche is not detectable anymore. Given that the shortest avalanche recorded at our site was 13 seconds, we thought empirically that this was a good choice. Nonetheless, we also intend to further analyse this in future work, given the number of avalanche samples increases.

L91: What are the advantages of a star-like pattern compared to others?

This is used to localise avalanches through array processing methods. To clarify this we have added the following sentence:

This spatial configuration allows for the localization of avalanches (Heck et al., 2018a).

Reference:

Heck, M., Hobiger, M., van Herwijnen, A., Schweizer, J., and Fäh, D.: Localization of seismic events produced by avalanches using multiple signal classification, Geophysical Journal International, 216, 201–217, https://doi.org/10.1093/gji/ggy394, 2018b

L127 – non-background noise signals: having 912 non-background noise signals that then split into avalanche and noise events is a confusing terminology. Consider using a different term for non-background noise signals.
We have simplified this formulation by just naming it «events»:
In total, we picked 912 events…

L139 – Fig. 2b: The labels of the panels are missing in Fig. 2.
We have improved this figure according to another reviewer's comment and therefore adapted the phrasing to:
… middle column in Fig. 2).

L148 – scores exceeded 1.5 : Consider changing to was at least 2.0. Readers just quickly skimming the manuscript might misinterpret it as ≥ 1.5
We have changed the sentence accordingly.

L167 – even: change to roughly even.
We have changed the sentence to: …were approximately balanced.

L167 to 169: You differentiate between dry and wet avalanches here but never address it in the results.
With this information, we want to express that we want to develop a model capable of detecting all types of avalanches. Therefore, we must ensure that the training and testing sets contain both types of avalanches.

L173: Sect. 4.3 referenced before 4.2.
These references have been changed through the restructuring of section 4. However, we have made sure that the subsections are referenced sequentially.

Fig 3: Could add brackets to the y-label clearly indicating which are training and test folds.
We have improved this figure to be more informative. In this process, we have considered this feedback and labelled the folds explicitly as train or test folds.

Fig 4: Only the encoder is shown. Is the decoder part of the autoencoder not used?
Yes, the decoder is discarded during inference. To make this clear, we have added this information to the figure's caption. It is also provided in the main text under the section «Autoencoder features».

L194: Why is the decoder discarded?
The decoder is only used during training. Its main purpose is to decompress the features and reconstruct the given input signal. The network is optimized by comparing the input and output with the mean squared error loss. However, during inference, meaning when predicting, the random forests only need the learned features for classification.
We have made this clearer (Lines 214 - 216).

L264: Could include these results as supplementary material.
We have removed this sentence from the manuscript, as incorporating this early stage of development is outside its scope. Since we used the method from Provost et al., 2017 as

our baseline, we implemented the same classifier on top of the autoencoders for better comparison between the methods.

Reference:
Provost, F., Hibert, C., and Malet, J. P.: Automatic classification of endogenous landslide seismicity using the Random Forest supervised classifier, Geophysical Research Letters, 44, 113–120, https://doi.org/10.1002/2016GL070709, 2017.

L280 – Gini information criterion: Missing citation.
We have added a reference to the work of Breiman, 2017.

Reference:
Breiman, L.: Classification and regression trees, Routledge, https://doi.org/https://doi.org/10.1201/9781315139470, 2017

Fig. 7: Consider adding a colour bar.
We believe a colour bar would not add much information since the percentages and number of samples are already shown within each confusion matrix. The colours encode the percentages calculated row-wise and should only provide a fast visual representation.

Sect. 5.1: I recommend integrating Sect. C1 into the main text. Otherwise, it is difficult to interpret the results.
Thank you for the suggestion. However, these classification metrics are widely used and well-known in classification and detection. We assumed that readers of the Geoscientific Model Development journal are familiar with them. For this reason, and to avoid elongating the entire manuscript, we kept them in the appendix. Likewise, we did not introduce convolutional layers or long short-term memory (LSTM) cells used in the temporal autoencoder and referred the reader to the original papers.

A nice overview of frequently used classification metrics is presented in:

Hossin, Mohammad & M.N, Sulaiman. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 5. 01-11. 10.5121/ijdkp.2015.5201.

Fig. 8: It is not clear what the x- and y-labels are.
We have clarified this in the figure's caption.

Fig. 10 – up/down: Change to top/bottom.
We have changed it in the revised version.

L331 and 385: Sect. 6.4 referenced before 6.3. I recommend to first discuss the missed avalanche windows and false alarms, and then the applicability to early warning systems.
Thank you for this suggestion. We agree and hence have swapped the sections.

L367: Same as the first sentence of the paragraph.
We have removed the duplicate. Thank you for pointing out.

L378: dotted blue line in Fig. 11.
We have added «…in Fig.11».

L409 to 411: It would be interesting to see how the issue of identifying avalanche onset evolves under different signal window lengths (e.g., trading slightly later detection for overall better performance).
Thank you for this great suggestion. We will consider it for future work.

L441: What could these unknown sources be?
Speaking from experience with these models, a source that was often mistaken for avalanches is aeroplanes or helicopters. These air vehicles display the characteristic spindle-shaped waveform of avalanches, but they show higher frequencies in the spectrograms. The other false detections were impossible to identify.

L445: Compared to previous studies, why can the models discussed here better differentiate between avalanches and earthquakes?
The previous studies do not explicitly state that the models had difficulties differentiating between avalanches and earthquakes. Looking at the study of Provost et al., 2017, they did not have problems distinguishing earthquakes from landslides either.

L466: How do the approaches of Bessason et al. (2007) and Hammer et al. (2017) compare the this study?
We did our best to compare our results with previous studies. However, both of these studies do not present metrics to compare, which makes it difficult. In general, in earlier studies, there seems to be no consensus on which metrics to present. This is part of why we reported all metrics, potentially making this work a basis for comparing future implementations and similar studies.

L504: I recommend ending with your key takeaway.
Thank you for the suggestion, we have modified the last paragraph in the conclusions, trying to present the key takeaway (Lines 555 – 565).

I hope the authors find my comments helpful.
Sincerely, Kevin Hank

Referee 6#:

General Comments

The study at hand investigates the possibility of detecting avalanches in seismic sensor signals with the help of representation learning. It demonstrates that an auto-encoder can learn useful features from seismic sensor signals in an unsupervised fashion. The usefulness of those features becomes apparent when training RF classifiers on top of the autoencoder to address binary avalanche detection tasks. Those validation tasks have a high scientific significance and are relevant for operational services in the avalanche warning community.

Scientific significance: The study could contribute significantly to the current field of avalanche detection from seismic sensors. Previous work is limited to random forest approaches and manually engineered features. The authors' approach to learning a representation of the avalanche features and training a classifier on top is a well-respected method in machine learning. While the benefits of this method have not been very well presented in the paper, I acknowledge and recognize the high scientific significance of this work, as well as its novelty.

Scientific quality: The scientific approach is sound and convincing. However, the usefulness of the novel methodological approach is not evident, given that related work outperforms this approach. Potential advantages of the novel method are claimed, but they are neither backed by literature nor experiments. Moreover, the experimental setup must be revised to address relevant concerns regarding the data processing (concerning data split and data normalization).

Scientific reproducibility: The ML part of the work is fully reproducible. The data and the code are accessible and executable. The experimental setup to collect the data is made available and could be reproduced for different site locations. It is unclear how to reproduce the processed data from the raw data.

Presentation quality: The presentation quality could be improved significantly. All requests on that end are "minor," i.e., they involve mainly restructuring bits and pieces, including literature to back up claims, building a more substantial narrative throughout the paper, and editing the paper to respect "context-content-conclusion." Even though my vote is low in that category, I am convinced that the authors can address all those concerns.

Detailed feedback can be found in the referee report.

I am listing here the major revisions requested that I consider the most relevant (emboldened in the feedback document):

Presentation:

- Research questions and contributions should be described more explicitly.
- Improve the narrative/storyline of the paper.
- Structure the paragraphs with "context – content – conclusion" format.
- Use names/identifiers (for models) and technical terms consistently throughout the paper.
- Use booktabs to format the tables.
- Improve figures by adding more descriptive subcaptions. Each figure should bring one main point across. Make the content in the figures more accessible (see comments).
- Discussion: Back up some of the claims on the ML side with literature.

Methods:

- Data split: How is data leakage via correlated samples prevented? Are you employing a stratified split? Are you separating sensors/locations?
- Normalization of the data: This is a major revision that is necessary. See the referee report for reasoning.
- Balancing of data: Address this in the paper (explain what you do to address the imbalance).
- Learning curves: Include them in the appendix so the reader can see whether the models are learning well and whether they underfit/overfit.
- More runs: Run the models at least 3 times with different random seeds. If you have already done that, indicate this in the paper and add error bars where applicable. This is necessary to compare the different models. The results are so close that the differences might not be significant if re-run.

Code and Data:

- Make sure to store the data on zenodo separately from the code.
- Minor revisions on the code; see referee report.

I recommend the paper for publication after major revisions. I consider the work done here an essential contribution to the field - a step towards automating avalanche detection and warning. Using representation learning in avalanche detection is a major novelty, and the work presented here has the potential (after revisions) to be an outstanding example of using representation learning for a relevant task.

We want to express our gratitude for the thorough and careful review. The feedback was immensely helpful in enhancing the manuscript and strengthening the storyline. We have tried to follow the extensive suggestions while also considering the comments from earlier reviewers. Below, we provide comments and responses to the reviewer's list of suggestions. For detailed implementations and modifications, we refer to the marked-up manuscript.

Specific Comments

Comments in Manuscript

Major revisions are emboldened.
Minor revisions are not emboldened.

1. General
    1. **Research questions and contributions should be described more explicitly.**
    2. Add task description (and why those tasks address your research questions)
    3. **Improve the narrative/storyline** of your paper. Sometimes, explanations or context is missing, and "why" you have made certain decisions is often unclear. You also have very strong arguments at hand from time to time that you are not mentioning or considering.

    Thank you for the general suggestions. We have carefully reorganised and improved the introduction section to make the research question clearer (e.g. Lines 72 - 76). In addition, we tried to explicitly describe the different tasks at the end of the introduction (Lines 85 - 87).
    Considering the storyline, we have made a considerable effort to improve several sections, as suggested by the reviewer. In particular, we restructured the introduction, the methods and the discussion section. For instance, in the introduction, we have concatenated all traditional approaches to detect mass movements into one paragraph (Lines 46 - 60), highlighted the field of representation learning (Lines 61 - 71) with references to high-impact works and justified the choice of this approach. The «Model development» section was entirely restructured following the reviewer's suggestion to clarify the single steps and decisions (see point 7. below). In the discussion, we have moved the « Applicability to early warning and monitoring systems» section behind the «Missed avalanche windows» and «False Alarms» sections. Moreover, as suggested by Referee #4 we have moved parts of the methods section to the discussion.
    Finally, we have aimed to motivate our decisions explicitly by referring to the model tuning process we applied to find all the parameters associated with the models. We also improved the description of the data preprocessing and motivated the chosen normalisation in section 3.3 (Lines 175 - 179).
    Additionally, the reasoning behind why we have chosen, e.g., to separate the feature extraction and classification, can be found in the discussion.

2. Structuring
    1. Sections sometimes need to be structured better.
    We have restructured most of the introduction and method sections. For details, see point 1. or the marked-up latexdif file at the end of this PDF.

    2. **Context – content – conclusion:** For most parts, you adhere to that, but several subsections do not follow that structure. Using CCC

consistently would significantly improve the flow and accessibility of your manuscript
We understand the advantages of this structuring and therefore have tried to adhere to it more closely. Nevertheless, where we thought conclusions were not needed or would add duplicate information, we have not. We generally followed the reviewer's marked comments in the manuscript on this concern.
Specifically, we have applied the context-content-conclusion structure to the last paragraph of the introduction, the «Study site and Instrumentation» section, the «Signal windowing, normalisation and dataset splitting» subsection, the «Feature classification» section and all newly created sections.

3. Language
    1. **Consistency**: The paper's language is sometimes inconsistent. You often switch between descriptors (sensor array vs multiple-sensors). You do not refer to your different models consistently with the same name. For you, the synonyms are clear, but for the reader, this can be very confusing: It is unclear if two terms are referring to the same concept or (slightly) different concepts and whether that difference is relevant.
    Thank you for pointing out this potential confusion. We have carefully reconsidered the naming of different concepts in the manuscript and named them consistently throughout the text. We refer to the methods as a) the baseline for the engineered feature extraction, b) spectral autoencoder or SAE and c) temporal autoencoder or TAE for the learned feature extractors. To remind the reader what we mean by baseline, we specify it now and then, e.g. «…the baseline, i.e. feature engineering…» Moreover, we ensured to always refer to the seismic sensor array as «sensor array» or «array of sensors», as we understand this might be confused with the array as a data structure.

    2. Explicit is better than implicit (explain things as explicitly as possible).
    We have followed this suggestion whenever it was appropriate. For instance, introductions of general concepts are sometimes still implicit, but we made all explanations of the implementations explicit.

    3. Consider running the manuscript through a language correction tool – I am not a native speaker, but some formulations sound "off" to me.
    We have already used language correction software for the last submission and plan to do the same for the revised submission.

4. **Formatting**
    1. Make sure to add links for figure and section labels.
    All links to sections and figures were tested to work in the last submission with a standard PDF reader. We will test all of them again before submitting the revised manuscript. For clarification, LaTeX links the

numbers following the section or figure label, not the label itself (e.g. Sect. **1** or Fig. **1**).

2. Make sure all figures are highly resolved.
   We have changed the format of most figures from PNG to PDF, particularly all figures highlighted by the reviewer.

3. **Change tables to booktabs.**
   From our point of view, the tables present a complete view of the model performance, which is intended to facilitate comparability for future implementations in this field. Regarding the format, we have already used the LaTeX booktabs package and followed the Copernicus template as required. As we understand the submission and publication process, the manuscript, if accepted, undergoes another stage of formatting and spelling corrections to adhere to the GMD journal standards. We will follow the editor's instructions.

4. Fix appendix issues.
   Again, we have followed the Copernicus template and trust these issues will be resolved during the potential final production. However, we have tried to reorganise the appendix accordingly.

5. Data
   1. Label incompleteness: What if experts have missed avalanches (unknown avalanche activity?)
      By consulting the imagery from the automatic cameras and the radar detections, we were able to identify all days with avalanche activity. We then analysed them by visually scanning the 24-hour data streams and picking all signals exhibiting a high signal-to-noise ratio. The three experts then labelled the picked signals.
      To clarify this, we have reformulated parts of the «Event picking and signal processing» section.

      Nevertheless, through the later expert labelling process, some minor avalanches or those flowing at the sensors' detection limit may have been mistakenly assigned to the noise class, i.e. false negatives. These are the potential avalanche signals that received an expert score of 1.5 or lower. However, missing some of these avalanches is inevitable and inherent to the challenging task of detecting avalanches.

   2. Label uncertainty: Should the task be considered a regression task instead of a binary classification task?
      Thank you for this suggestion. In our setup, we aim to classify a window into clear and distinct classes. We can imagine a setup in which it could be viewed as a regression task using the expert scores. However, in this case, the correctness and completeness of

the label would be even more critical, which is a step our initial implementation could address. We could relate a probability of classification to the expert scores and verify that the classification is, so to say, calibrated. But we leave this for future work.

3. Label noise: The agreement score of expert labels is relatively low. Evaluating the performance of the DL model is difficult with that.
4. Question: How would you evaluate the success rate of expert labeling? Can you evaluate the ML model independently of expert labels by any chance? Is there a theory on what avalanche signals look like in the Fourier domain, and could you calculate some score based on that?
   To our best knowledge, the set of data we used to validate and compute generalisation scores is as correct as possible. To this end, we consider this labelled set as accurate as possible, and the ML models are now compared to this gold standard. In a practical implementation, labels won't be available until after an expert assesses a given prediction, which, in this case, we would evaluate the precision of the model (i.e., out of the model's predictions, how many are actual correct events). This would be a straightforward evaluation without requiring experts, but it is surely weaker. The decomposition of signals into the Fourier domain largely depends on the size of avalanches, density/compactness of snow, soil/rock types, water content, and the distance from the avalanche to the recorded sensor. Although there is some theoretical understanding of this aspect, using this knowledge to assess the model's output would still require the latter to be accurate, so we did not venture into this evaluation setup.

6. Data processing
   1. How do you get the data provided in your repository? I assume this is not the raw data from the sensors? Making that bit accessible / describing it would be important to improve reproducibility.
      The data in the repository are the 10-second windows bandpass-filtered from 1 to 10Hz. To clarify this to users, we added a more descriptive readme to the data directory. The provided data and the code repository let the user reproduce all results. We did not provide the raw data and the accompanying processing code since the main objective of this study is model development, not data acquisition. However, we have decided to upload the raw seismic events with the corresponding labels to a separate Zenodo repository (DOI: 10.5281/zenodo.14892926).

   2. Extending the dataset: Have you considered extending your dataset with samples from previous publications? (Provost et al. (2017) + Rubin et al. (2012))
      No, we have not considered it for this study. This is a good point and will surely be a topic to address in the future. To have accurate systems at different geographical locations, we must guarantee that the ML system can generalise to other datasets, even potentially acquired by

different sensors and settings. Once verified, we can assess performance by including several datasets in the training set and studying how the model generalises. This would be a crucial step before applying such a system in practice.

3. **Data split**: you should also separate locations / sensors, so you do not have leakage via correlated samples. Are you doing a stratified split? I.e., are the same amount of pos/neg samples available in each split? Or is it a random split?
   We carefully picked three dates to split the dataset to a) prevent leakage between the four folds entirely and b) have roughly even class distributions (see Figure 3). In that sense, we manually defined stratified splits. The correlated samples from the five sensors were always assigned to the same fold, hence not leaking any information. We purposely avoided any look-ahead implementation through appropriate splitting and data normalisation.

4. **Normalization of your data.** Look at feature transformation and data normalization techniques, make your choices, and describe them in the paper.
   We developed the presented methods based on 10s seismic signal sub-sequences generated with a windowing algorithm applied to the entire event signals. In general, input normalisation is crucial when training neural networks (Sola and Sevilla, 1997) and even more so when using subsequences of entire time series (Rakthanmanon et al., 2012; Lima and Souza, 2023). In our case, we must avoid implementing a look-ahead normalisation since, at inference, the characteristics, e.g. maximum absolute amplitude, of an event signal are not known in advance. Thus, we normalised each subsequence independently.
   Alternatively, normalising all data by any high-level value of the training dataset was not intended. Due to the unsupervised nature of the autoencoders, we were not allowed to differentiate between classes when retrieving such high-level dataset values. Therefore, normalising the entire dataset by the maximum absolute amplitude in the train set, for instance, would have meant normalising avalanches by the maximum amplitude within the noise class. This would have led to input samples differing by order of magnitudes and destabilised training. We made the choice of normalisation explicit by adding it to the «Signal windowing, normalisation and dataset splitting» section.

   References:
   Sola, J. and Sevilla, J.: Importance of input data normalization for the application of neural networks to complex industrial problems, IEEE Transactions on Nuclear Science, 44, 1464–1468, https://doi.org/10.1109/23.589532, 1997.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., and Keogh, E.: Searching and mining trillions of time series subsequences under dynamic time warping, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, p. 262–270, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/2339530.2339576, 2012.

Lima, F. T. and Souza, V. M.: A Large Comparison of Normalization Methods on Time Series, Big Data Research, 34, 100 407, https://doi.org/https://doi.org/10.1016/j.bdr.2023.100407, 2023.

5. My thoughts are to do batch normalization or normalize with sensible physical values (max value recordable via sensor).
    1. 1) Generally, data normalization is necessary for your model activation functions and to achieve good and stable learning.
    We strongly agree, and therefore, we used input normalisation and implemented batch or layer normalisation.

    2. 2) You want to focus on the data's pattern, not the scale. I hypothesize that this will improve your model's ability to detect small avalanches or avalanches that are farther away.
    We want to detect all avalanche sizes, types, and early stages of the avalanche motion characterised by lower amplitudes due to a lower generation of seismic energy and longer source-receiver distance. This is precisely why a normalisation should be applied to the windows independently, as we did. With that, we solely focus on patterns, not the scale.

    3. 3) You have different sensors at different geo-locations.
    Yes, but our objective was to implement a method based on a single sensor. Considering possible future scenarios at various sites, the deployment of sensor arrays is not necessarily guaranteed.

6. **Imbalanced data:** Explain how you address this (via sampling strategy?)
    As already mentioned in the training procedure and described in Appendix D, we used the weighted random sampler method to address the class imbalance during training. We agree that this was an essential part of the model development. To emphasise this step, we moved and adapted the explanation from the appendix to the main text (Lines 249 - 257).

7. Data windowing: Are you adding additional features from larger window sizes? Such as mean, std, and frequency spectra – to provide long-term context?
    No, we did not. The methods were derived from 10s seismic signals without considering any longer-term context. We understand and

agree that by reducing the window size, the characteristic spindle shape of avalanche events is lost at some point. However, we found ourselves in a situation where we had to balance the number of training samples with the window size. In earlier studies, similar window sizes were used, so we favoured a larger dataset. Given more avalanche events, we suggest optimising this length in future work.

7. Model section: Please restructure the section. Make clear separations between the feature extractor and the evaluation task. Make it clear that you validate the autoencoders on a separate validation task. Separate our hyperparameter tuning. Etc. I recommend reading a couple of ML papers to get an intuition of how those sections are usually separated and written in a paper.
   We have considered this suggestion and made significant revisions to the structure of the model's section:
   (4) Model development
      (4.1) Baseline features
      (4.2) Autoencoder features
         (4.2.1) Architecture
         (4.2.2) Training Regime
         (4.2.3) Validation
         (4.2.4) Model selection
      (4.3) Feature classification
         (4.3.1) Random forest model
         (4.3.2) Cross-validation
         (4.3.3) Inference and post-processing

   Moreover, we have explained the random forest's inference process in more detail (Lines 311 - 319) and explicitly introduced the prediction post-processing (evaluation tasks) (Lines 320 - 330). Finally, we have described the autoencoder validation in a separate section (Section 4.2.3).

8. Model architecture: Please assign speaking names/identifiers to your three models that can be tracked throughout the complete paper (figures, text, abstract, appendix, tables) early on
   1. TAE: Why 1d convolutional layers?
   2. SAE: Could one use Fourier Neural Operators for this?
   3. Have you considered looking into methods to learn from imbalanced data? There are several models out there. Most commonly, people change their sampling strategy for their model to make sure the model sees positive and negative examples equally frequently.

   We have followed this suggestion and assigned the model identifiers: baseline, temporal autoencoder (TAE) and spectral autoencoder (SAE).
   1. Since we only use vertical components of the sensors and treat each sensor independently, the waveform time series is one-dimensional, i.e. 2000 samples in one channel. Accordingly, we used the frequently used implementation of one-dimensional convolutions for time series (Kiranyaz et al., 2021). To clarify, we have included the survey of

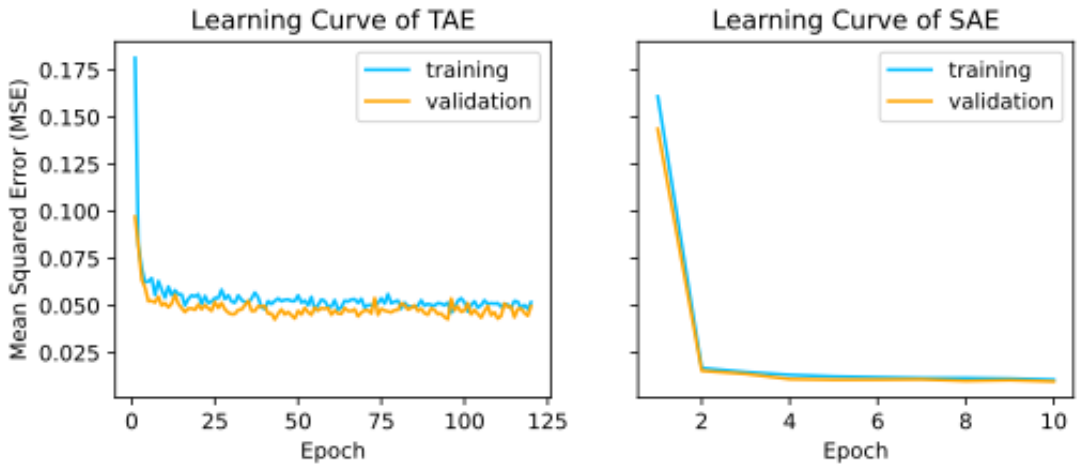
Kiranyaz et al. (2021) in the manuscript.

2. To our understanding, neural Fourier operators are useful ways to learn the embedding of functional spaces, where the learned representation is invariant of the data discretisation and would allow the model to learn complex dynamical systems with long- and short-range dependencies efficiently. Although not impossible, and some properties are indeed appealing, it is not apparent why a model relying on such layers would be performing, particularly because our output is a label space. Hence, the aim is not to learn mapping from dynamics to dynamics. We take this as a suggestion for future work and will spend some more thinking about this. Thank you very much for the suggestion.
3. Please see point 6.6. above. We used a weighted random sampler to compensate for the class imbalance by oversampling samples from the avalanche class. To make this critical point in training clearer, we have adapted and moved the section «Appendix D: Weighted random sampler» to the main body of the text.

References:

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J.: 1D convolutional neural networks and applications: A survey, Mechanical Systems and Signal Processing, 151, 107 398, https://doi.org/https://doi.org/10.1016/j.ymssp.2020.107398, 2021

9. Model learning: **Include learning curves** and show that the models are actually learning well. The reader wants to know if you are overfitting/underfitting. These figures can be included in the appendix, but they are essential to strengthening your paper. If your models are underfitting or not learning incredibly well, this would be a strong indicator that you can achieve even better results.
   We have included the learning curves of both autoencoders in the appendix «E1 Learning curves».



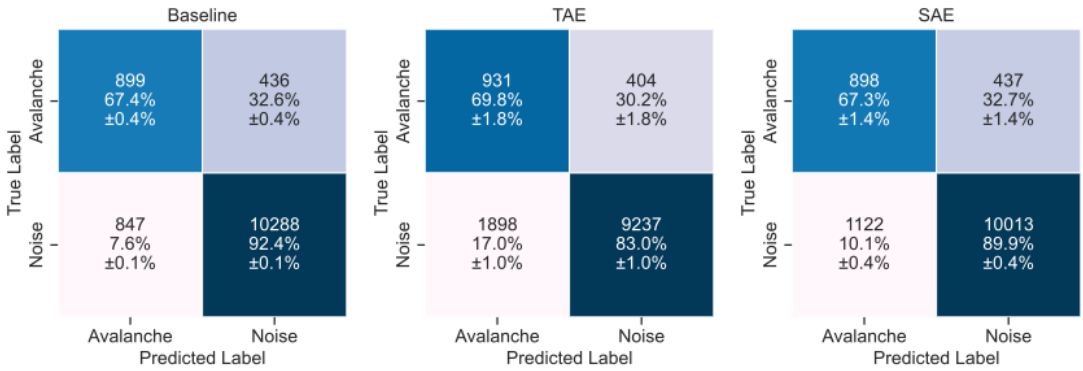
10. Model comparison: **Have more runs** (change random seed) to show whether differences are relevant/significant. Include error bars in your results. If you

already run models across multiple seeds, report the number of runs and error bars.

Thank you very much for your suggestion. We carefully developed the models so as not to be biased towards a specific type of avalanche (wet/dry) or size. Particularly, we made sure the independent test set contained all types and sizes of avalanches and, therefore, was representative of our test site. Additionally, we prevented any data leakage between the data folds by selecting and splitting by specific dates. Considering the reviewer's concern, we understand that having multiple runs is beneficial. Therefore, we have retrained them 20 times with varying seeds (powers of two starting at 2^0). Similarly, we have retrained the random forest model of the fixed baseline features with these seeds and the random forest models of the autoencoder features with these seeds on the correspondingly extracted seeded features. The following figure and table compare the models across these runs. We observed consistent performance and stability with avalanche f1-score variability being at most 1% in all three methods.

As we have considered this suggestion particularly valuable, we have adapted all figures and metrics in the tables as well as the text showing the mean value and the standard deviation.

	Baseline	TAE	SAE
Avalanche f1-score	0.58 ± 0.00	$0.45 \pm 0.01\%$	$0.54 \pm 0.01\%$



11. **Figures**
    1. Each figure should bring one main point / finding across. The current figures show results, but they are not all structured or designed in a way that makes results easily accessible. A reader should be able to read a paper and access all main findings by reading the figures + captions.
    2. Some figures need more descriptive captions.

Thank you for this mindful suggestion. We have changed the format of the figures to PDF to ensure they are highly resolved when zoomed in. Moreover, we have considered the reviewer's comments in the manuscript and adapted the figures accordingly. In particular, we have improved the figures related to the

methods (Figure 4). Additionally, we have considered all comments in the script regarding the captions and written them more descriptively.

12. Discussion
    1. **You need to back some of your claims (scalability, ability to generalise, etc.) with literature.**
       We agree that these claims need to be proven first. Therefore, we have removed or modified them and referred to future work investigating these points.

13. Conclusion
    1. Adapt the sentence "We have shown that it bears strong potential..." (see comments). You have shown it can keep up with current state-of-the-art avalanche detection methods. Still, you have not demonstrated its potential (scalability, generalisation ability, etc) in your experiments.
    2. A more memorable last sentence - talk about downstream applications (operational! Avalanche warning! - You have such a strong and relevant use case, and your audience wants to hear about that.) - would strengthen your conclusion.

    Thank you for these suggestions. We have modified the last paragraph of the conclusions accordingly (Lines 555-565).

14. Code
    1. Looks good! Clean repo, nicely coded, well done.
    2. You could add more documentation.
    3. You have the code two times in the repo – one seems to be for MacOS? Is this on purpose?
       No, this was not on purpose. We have removed these files and all «DS_Store» files.

    4. Mention the Python version in the Readme.
       We have used the Python version 3.9.7 and have added it to the Readme.

    5. Add versions of your packages in requirements. Your code will not be reproducible later on if you do not report that (and it is not maintained).
       We have included all package versions in the requirements.

    6. Dir "models" and "lib" should be in "code" (semantically). I also get import errors if I do not move them over there.
       We have tested the code again, and indeed, it showed import errors. Thank you for noting. One workaround was to change the PYTHONPATH variable, i.e. «export PYTHONPATH=${PYTHONPATH}:${HOME}».
       To avoid this, we have moved all files from inside the «code/» directory to the parent directory.

7. Using your panda version broke the code for me (numpy – pandas incompatability). I upgraded pandas to fix it.
   This has been resolved by specifying the package versions (see Point 5).

8. Consider using Pathlib to handle path os-independently.
   Thank you for the hint. We have modified all path-related code lines to use the package Pathlib and successfully tested all functionalities again.

9. Add in your readme where people must change paths to get your code running on their machine.
   This information has already been part of the Readme but might have been misleading. We hope to have clarified it by using Pathlib and adding the following to the Readme:
   Upon successful installation the root directory needs to be changed in lib/utils/variables.py!
   ROOT_DIRECTORY = Path.home() / 'path' / 'to' / 'project'

15. Data
    1. **Make sure to store the data on zenodo separately from the code.** You will want to update the data without updating the code in the future. If you want the data to be accessible (for other researchers and their projects), it needs to be maintained separately and with its own version control.
       We understand this point and also believe research should be open to everyone. Therefore, we have created a new data repository where the raw seismic data of all events used in this study and their corresponding labels are available (DOI: 10.5281/zenodo.14892926). In the future, we can update this repository with the latest events recorded at our test site. Nonetheless, we have retained the processed 10-second windows in the code repository to maintain conciseness and facilitate the models' downloading and testing. We hope to serve the research community with this.