Reply to reviewer #1

Dear Editors, dear Reviewer,

We thank the reviewer for reviewing our manuscript. We went through all comments carefully and replied to each of them in the following. The reviewer's comments are listed and repeated in italic. Clarifications and improvements suggested by the reviewer will be taken into account in the revised manuscript.

"This paper is very interesting because it proposes for the first time to use Convolutional Neural Network (CNN) on seismic signals recorded on unstable slopes. The analysis of the catalog produced is very exhaustive, relevant and brings very interesting elements to better understand the link between the micro-seismicity endogenous to a landslide and the external forcings, and the associated mechanisms. The article is very well written, easy to follow and understand, and the figures are of excellent quality."

"Overall, I find this article almost publishable as is. I nevertheless have some minor comments and questions, especially on the Machine Learning part. I detail my comments below:"

- *"L202-205: I think the training set should be better described. In any implementation of supervised classification algorithm it is very important to know the exact number of events used to train and to test the algorithm, as those can greatly influence the performance of the algorithm. Then the description here is confusing: first you say that "most classes constitute around 12% of the training set", but then that this is not the case for the HF SQ and RE classes. I would suggest adding two columns in table 1 with the number of events in the training and in the testing sets for each class."* → we agree that this may be confusing. The composition of the training and test sets is shown in the form of pie diagrams in the appendix only (Figure A11) and numbers are provided there, but we will make changes to ensure they are better described in the main body of the manuscript by adding columns in Table 1 as suggested.

- *"L216-219: Some previous studies also proposed features computed on the spectrograms, and they usually are amongst the best features for the classification (e.g. Provost et al., 2017; Hibert et al., 2017; 2019 ; Maggi et al., 2017; Wenner et al., 2021). This should be mentionned here. You propose a new approach based on the spectrogram but you are not the first to propose to use this data transformation and this should be acknowledged in your description here."* → you are right, these papers are cited earlier in the manuscript but not specifically linked to the features definition/extraction part. In the revised manuscript, we will make sure this is the case.

- *"L222-227: I don't understand the drawbacks described here. Features computation is not more complicated than the computation you do to transform your data into spectrograms. I understand that you need some arguments to tell the readers why you choose to work with spectrograms, but I don't think the arguments you propose here are valid."* →

we agree that computing features is not substantially more complicated. But they need to be defined beforehand, resulting in a large number of different features. Moreover, it certainly depends on the way features are extracted. In a previous approach that we tried, we found that the event duration was a major feature for discrimination – but it was also not the easiest feature to extract (in the sense that it is rather mistake-prone). Since we also extracted most features within the event duration window, if this was not well estimated, the other features usually contained more errors and could not be well estimated (e.g. the energy was over- (or under-) estimated,...). But we agree that this was the case for only a few data and that this cannot be generalised or used as a good argument "against" feature-based algorithms.

- *"One could argue that your approach based on spectrograms is more susceptible to computation parameters, as the resulting spectrogram is completely controlled by the way you compute it (e.g., what is the influence of the spectral transformation you use? Between DFT, wavelets, Z-Transform, etc? What about the window shape? the window length? the overlap?)."* → We have not tested different transforms, although they might of course slightly influence the final results. Extracted features are dependent on computation parameters in a comparable way: for example, features extracted from spectrograms will still depend on the way spectrograms are computed and spectral features may also vary if the FFT or the PSD is used.

- *"What would be the best option is the possibility to input directly the raw signal into the machine learning model, but I'm not sure you can do this with CNN."* → it is possible to use 3C data as input as done by e.g. Köhler et al., GJI, 2022 (Classification of seismic calving events in Svalbard). In the Åknes case, it has been briefly tested, but spectrograms seem to yield better results.

- *"I think that there is a simple argument in favor of CNN that you should make, which is that supervised machine learning algorithms based on curated features might miss critical information within the signals that CNN will find because it does not need any manual hence subjective definition of the features. CNN use images (most of those models at least), so in order to use them on seismic data you need to transform the signal into something that has the same properties as an image, which are spectrograms. This is largely sufficient to motivate the use of CNN and your study I think."* → Thank you - indeed, this was also our point, but it was maybe not as clearly expressed. Based on this comment and the previous comments listed above, we will reformulate the section in the revised version of the manuscript.

- *"L249: You choose to stack the spectrograms, but how is this influencing the global noise of your final spectrograms? Would it be better to calculate the product of the different spectrograms? I might be wrong, but by multiplying the spectrograms I think that you will reduce the noise and the influence of propagation effects while bringing out the part of the seismic energy generated by the source? It might be worth testing in a future work."* → the stacking approach indeed enables to enhance the signal-to-noise ratio. We did not think about multiplying the spectrograms instead of stacking them, but we thank you for this good idea as it should indeed enhance the SNR even more.

- *"L274: For how many of those 59.608 events the class has been confirmed manually?*

The 2554 you included in the test set or more than this?" → Only the 2554 events constituting the test set have been QC'd manually.

- *"If more, what guided your choice of the 2554 events you have in your test set?"* → The test set contains all detected events since November 2020 when the classifier was first implemented in the near real-time workflow. In addition, it contains randomly selected events that were also checked while testing the classifier before its implementation.

- *"Are the events in the training set from those 2554 events?"* → No, all those events are different from the training set.

- *"A better description of your training and testing data sets is needed I think, as suggested in a previous comment."* → Columns with number of events in each class for both the training and test sets will be added in Table 1 as suggested in your previous comment.

- *"L303-304: So you did scan all the 59.608 events manually to remove the electronic spikes? See comment above. It should be very clear for the readers if the catalogue you interpret in the following sections is fully automatically made, automatic but fully manually controlled, or automatic but partially manually controlled. Provide numbers."* → Thank your for this comment. This part was indeed not well described in the manuscript. The catalogue was only adjusted for the cumulative energy curve in Figure 6, where all events not related directly to the slope movements were removed (i.e., spikes, noise, regionals). Using the automatic classes, we sometimes observed large jumps in this curve and checked which events caused them. We found that most of these jumps were due to spikes being wrongly classified, and we subsequently decided to remove them (450 events, i.e., 7% of the spikes were not well classified). However, in the rest of the paper (histograms, ...), we only used the uncorrected automatic catalogue (given the large number of events, those misclassified events count only for less than 1% of the dataset and do not affect significantly the results). We will strive to make all this clearer in the final manuscript.

- *"L400 & 414-415: Would it be possible to process data with different sizes with other CNN implementations?"* → using CNN implies that input images have the same size, but FCN (Fully Convolutional Networks) allow for different sizes. Alternatively, one can transform the image to the desired size before feeding it into the CNN. Since we work with triggered data with fixed length, we did not feel a great need to explore more any of these options.

- *"This is a huge advantage of methods based on curated features (RF, SVM), they can work with signals with different durations. However this needs a pre-detection of the event, which can be tricky. This is why we start to see implementations of those approaches on moving windows (Wenner et al., 2021 ; Chmiel et al., 2021). Would it be possible to do the same using a NN such as AlexNet? If so, it can be interesting to tell the readers in the discussion or in the perspectives how such an implementation could be done and what could be the difficulties."* → Thank you for this question and the papers which are very interesting. Although we did not have the occasion to test our workflow on continuous data, we think this should be possible to apply a similar approach (i.e. sliding window) with a CNN since the filters in the CNN can be seen as features. A

work going into this direction has been published by Takahashi et al. (Earth, Planets and Space, 2021) with the aim of detecting and classifying earthquakes, tremors and noise using CNN.
Moreover, in a CNN, the detection step could be performed through class activation maps (CAMs) which help visualizing and highlighting which part(s) of the spectrogram image the CNN focuses on. It should then be possible to define activation thresholds above which an event would be first identified and then classified.
Lastly, in cases when several stations are available, one could also imagine classifying each trace separately and decide whether an event occurred by voting. For example, if more than 50% of the automatic classes at each individual traces are classified as noise, then there is no event.

- *"L435 – 437: Indeed. This sounds like it could be easily tested. What prevented you from doing so for this study? Was it too costly in term of computation time?"* → we would like to draw your attention to a follow-up work that has been performed and is currently accessible on ArXiv: Lee et al., 2022, https://doi.org/10.48550/arXiv.2204.02697. In this work, the classifier has been improved by using ensemble prediction and self-supervised learning approaches. A direct comparison of classification performances for AlexNet when individual spectrograms are considered instead of stacked spectrograms is provided in the paper: unsurprisingly, the ensemble prediction approach turns out to be one of the major factors of improvement. As a side note, please note that this work has neither been applied to the entire Åknes dataset, nor in the near-real time workflow yet (although planned in the future).

- *"Conclusions: I found the last sentence/paragraph a bit vague and underwhelming. I think you have plenty of insights from this first implementation of a CNN that you should share with the readers. They should be highlighted in the conclusion."* → thank you for this comment, we agree that this sentence is too fuzzy. We will reformulate the sentence and expand a bit in the revised manuscript.