Response to Reviewers

The authors of the manuscript entitled "A study on the effect of input data length on deep learning-based magnitude classifier" present a very interesting contribution into addressing some crucial aspects of the automated techniques (powered by deep-learning) to estimate the earthquake magnitude. Despite the large perspectives that deep-learning techniques provide to the EEW framework, the choice of the right neural architecture, the dataset preparation and interpretation of results is crucial into determine their success. The authors have addresses many of the concerns that this sensitive deep-learning application has, showcasting a powerful neural classifier that can discriminate between noise, low-magnitude and high-magnitude earthquakes. Still, some open questions remain, which is why this reviewer suggested a major revision, so to address them more into detail and so to promote the scientific debate on it.

As metnioned, the manuscript is very interesting and it raises many questions about the decisions the authors made to perform their classification benchmark. The list of comments below is non-exhaustive: please, refer to the attached reviewed manuscript for further comments.

- The decision boundary between low-magnitude and high-magnitude is rather arbitrary, as stated by the reviewers and as shown in Fig.7a. Plus, It highly depends on the seismic context of interest and the risk assessment and vulnerability policies of each country/region. It would be interesting to test at least one another decision boundary or at least test somehow the sensitivity of the classifier to this choice.

  - The authors thank the reviewer Dr. Filippo Gatti for this comment. We have experimented with decision boundaries of magnitude 3 and 4. The accuracy, precision and recall values were found to be similar to what has been presented in the manuscript and did not show any clear dependence on the length of input data. A comment on this has been added to the revised manuscript (lines 53-55).

- The effect of the source-to-site distance seems to have been disregarded. Maybe, separating the waveforms in different bins, based on source-to-site distance, could unveil some interesting aspects of the classification performance. Some comments on it would be beneficial to the manuscript overall clarity.

  - The authors agree with the point raised by the reviewer and thank him for addressing it. We have analyzed the model performance for different source-to-site distances (please refer to figure 8(a) in page 8 of the revised manuscript) and observed that the model is indeed capable of performing reliably over a wide range of hypocentral distances. In other words, no clear dependence between the model performance and hypocentral distance can be observed. This discussion can be found in the revised manuscript (lines 213-219).

- Sometimes, it's rather useful to analyze the waveforms at stake in the Fourier's spectrum domain. The corner frequency is strictly related to the source spectrum, which mostly determines the magnitude (along with the distance) In this case, this reviewer suggests to check the spectrogram of the

classified waveforms, so to verify that the duration is compatible with the associated frequency corner value for the correspondent moment magnitude (see the statistical relationship between corner frequency and moment magnitude presented by *Courboulex F, Vallée M, Causse M, Chounet A (2016) Stress-drop variability of shallow earthquakes extracted from a global database of source time functions. Seismol Res Lett 87(4):912–918*

- We noticed that the model is capable to perform correct classifications over a wide range of hypocentral distances and magnitude ranges suggesting that it is capable of learning the frequency characteristics of the waveforms.The use of Fourier spectrum in addition to waveform data was tested during our initial experiments, and it achieved results comparable to the model which used only waveform data as input waveform.

- Have the authors considered the earthquake type when preparing the dataset? A comment on this aspect would be very interesting.

- We have analyzed the effect of hypocentral distance and SNR on the model performance. While we do not see any clear dependence on hypocentral distance, the SNR of the data seems to play a role in the classification of waveforms. (Please refer to figure 8 and lines 213-219 in revised manuscript). On the other hand, due to unavailability of the Information on focal mechanism in the metadata we were not able to experiment with this. However, the role of the earthquake source type could be considered further in a separate study.

"specific comments"

Comments for the last paragraph in 'introduction':

The authors said boundary of low and high magnitudes are arbitrary chosen and does not influence the model performance. However, the reviewer think boundary selection could affect the performance, because the faulting process become more complex for larger earthquakes so that initial P-wave does not necessarily has large amplitude during the P-wave trains of the larger earthquake. In the paper, analysis durations does not affect the results, but this results are only examined for the magnitude boundary of 5.0. If the boundary shifts larger (like 7.0), analysis duration could affect the performance, although such analysis is difficult for STEAD.

- It is difficult to experiment with decision boundaries above 5 because the number of waveforms for such high magnitudes present in the dataset is severely limited. Although we experimented with decision boundaries of magnitude 3 and 4 and got similar results. (Comment added to the revised manuscript, lines 53-55)

Comments for the description of data used:

Are there any selection criteria in source-to-site distance and station?

- Currently no selection criteria are applied to the source-to-site distance (see next comment).

STEAD includes from small to large distance data. In the scheme of the paper, the station(s) nearest to the epicenter seems appropriate for the analysis, because the rapid warning is the purpose. Please add description of selection criteria for distance/station if exists. Also, please add distance distribution like Figure 1 irrespective of existence of the criteria.

- Thank you for the suggestion. We have added Figure 1(b) showing the distribution of source-to-site distances to the revised manuscript.

The reviewer is wondering that use of large-distant records increase the difficulty of classification, because such record become very complicated waveform due to the propagation of long distance in complex media.

- The analysis of the model performance for different source-to-site distances is shown in Figure 8(a) of the revised manuscript. No clear dependence between the model performance and hypocentral distance can be observed. As one can observe from Figure 8(a), the model can perform correct classifications over a wide range of hypocentral distances. This discussion can be found in the revised manuscript (lines 213-219).

Comments for Model Architecture:

Please describe why the authors choose the model architecture in Figure 3. (Please explain how each part contributes.)

- Convolutional Neural Networks have often been found to be useful for seismological data analysis as they are capable of extracting patterns in the data (features) without any temporal dependence. When combined with LSTMs the temporal relations between these features can be obtained. In applications such as magnitude-based classification of earthquakes, this aids in the effective analysis of signal features as compared to the pre-signal background noise. The dropout layers are used to prevent the model from overfitting and the maxpooling layer is a method to reduce the data dimensionality so that only relevant features can be retained. The final layer is a softmax layer which outputs the probabilities corresponding to each of the three classes that the data is classified into. This description has been added to the revised manuscript (lines 123-128).