Response of Reviewer A

The current manuscript presents timely deep-learning approaches to detect blasts from construction sides and quarries recorded by low-cost seismic sensors. The results and conclusion appear sound but certain parts (e.g. the construction of the dataset and the evaluation of the models) need to be clarified.

We would like to thank René Steinmann for his thorough and very helpful review. In particular, we appreciate the comments addressing the evaluation of the deep learning models. We agree that certain information was missing, and we tried to the best of our ability to clarify all issues.

I also miss a comparison/discussion of simpler and non-deep-learning approaches. Is there any baseline to compare to? Do we need to train and hand-design auto-encoders for each station to detect outliers or are the simpler ways (e.g. spectral coefficients+robust covariance)?

We absolutely agree that there are other methods which can potentially be used and compared to detect events of interest in an unsupervised and non-deep-learning manner. However, we would like to emphasize that we do compare our new outlier detector to a baseline, which is in our case the simple STA/LTA trigger (although we acknowledge that our method takes the baseline detection as input). We already discussed in the paper that an STA/LTA can for certain events in a low-noise environments have the function of an outlier detector. There are certainly more simple methods based on characteristic functions of the waveform that can be used to trigger events (kurtosis, spectral amplitudes in different bands, etc.). Here, we consider STA/LTA as the best-established conventional trigger method in this context, and we believe it is therefore appropriate to compare our method with this one. As already mentioned in the paper, the STA/LTA trigger has the disadvantage of triggering also frequent noise burst that are not of interest in our study. To overcome this issue, we decided to use a more sophisticated method from the field of deep learning to select outliers among all STA/LTA-detected events. Outlier detection for seismic data was not explored so far with deep learning to our knowledge.

Another approach in this context, which we also mentioned, is to cluster the entire seismic record using deep or non-deep unsupervised learning methods, or alternatively only cluster the pre-detected events (an approach we missed to mention). The second step would then be to identify the cluster of interest, in our case rare events, i.e., outliers. Although we considered this in the start of our study, we decided to not pursue this approach further since we wanted to avoid the additional cluster identification step. Furthermore, rare events may not necessarily be caught up in a distinct cluster. Hence, our focus of this study was to test if a deep learning model can find these events directly without grouping all signals first. It would be very interesting and a worthwhile study to apply (deep) clustering to our data set and compare the results. However, we feel that this would go beyond the scope of the current paper.

As to the comment about needing a hand-designed auto-encoder, we want to emphasize that auto-encoders are relatively easily implemented and trained (compared to more sophisticated deep learning models) and are not difficult to tune for each station. We might have given this impression when describing the training data selection, but in fact we found the auto-encoder method to perform well and effective without a long tuning and “trial-and-error” phase. More about this below.

We now modified the introduction (new paragraph added), beginning of the method and the discussion section to emphasize the issues raised by the reviewer, and hope we sufficiently addressed these concerns.

In the same regard, how sensitive is the proposed approach to the set of hyperparameters? Do we need to tweak these parameters for each station?

This is a good point which we already touched in the discussion section. However, this needed indeed further elaboration. We did some tests using different number of layers and filter for the auto-encoder before training the final model. However, this did not require massive tuning and tweaking of all parameters for each individual station. The only difference in the outlier detector models is the training data selection, number of samples in the input window, and latent dimension. We acknowledge that the choice of training data and time window steps may seem a bit ad-hoc. We arrived at the final choice of these setting by testing the model performance at the two stations we selected, i.e., the model's ability to reconstruct the background noise. In this process, we also identified the number of samples in each time window and latent dimension to be the most important parameters, while the rest of the model architecture, hyperparameters, and the choice of training data (except that they should cover different noise condition at different times) does not need to be changed for each station. Having said this, application to another study area will show if certain model parameters need to be adapted after all. We updated the text in the discussion.

In general, I would recommend renaming the occurrences of "machine learning" to "deep learning", since all methods (auto-encoder for outlier detection and CNN for classification) belong to the area of deep learning, which is a part of machine learning.

Thank you for this suggestion. We agree and changed to "deep learning".

In the discussion part, I also miss two points.

Firstly, I think it would be interesting to close the loop to the introduction, where you propose to include seismic sensors in smart-city applications for monitoring blast activity. How do you see these methods deployed in real-time for the Agency of Emergency Planning or the water and sewage department? What are the impacts of false negatives or positives for the local authorities?

Thank you for this comment. This was indeed missing, and we added a paragraph in the discussion section to close the loop.

Secondly, how well does your approach generalize to new stations within the network? It would be convenient if you could train one model which generalizes across all stations. Or what are the challenges in doing so?

L120: „Our final choice of best performance…" - what means best performance here? What metric did you use to make that choice?

We address both comments above together since answering the first one requires addressing the second one first.

The outlier detector and blast classifier performance are evaluated with respect to our reference data set of blasts. We want to achieve a high recall (not missing blasts) and high precision (low number of other events triggered). In contrast to a conventional event detector, the decision about what is a false and what is a true positive is a moving target for outlier detectors. Additional events not being part of the refence data, could still be events of interest. Nevertheless, we can still use the recall-precision metrics as a proxy to compare model performance relative to each other. To decide on decision thresholds for the outlier detector (correlation coefficient) and blast classifier (blast probability), we also evaluate these recall-precision curves. We do this before the Rg localization step. Showing these evaluations to the reader was indeed missing, and we therefore added additional figures to justify our choice for the decision threshold, model parameters, and training data for the outlier detector as well as for the blast classifier.

New Supplementary Figure S1 shows recall-precision curves for three outlier detectors with three different latent dimensions. The results suggest that the dimension of 2xT is more optimal than 1xT (more data compression) and 3xT (no data compression). Note that precision is not larger than about 0.85 because we do not sort out unlocatable events when computing these metrics. This provides a more objective performance evaluation. However, precision will become higher when we would apply the automatic location procedure subsequently. Those curves also allowed us to choose an optimal decision threshold for the 2xT model, i.e., the correlation coefficient producing a recall and precision point closest to the upper-left corner of the plot.

As to the generalization ability, we need to train a new model for the outlier detector for each station because each station has its specific background noise conditions. However, we now test a common model for two stations (see next comment and new Supplementary Figure S2), which shows that the new outlier detector performed worse than the single station mode. In the case of the blast classifier, our goal was indeed to train one model applicable to all stations. As we briefly mentioned in the discussion, we tested this approach, i.e., training a model with data from all stations where blasts were observed. In principle this is no issue with our implementation, however, it needs a larger and more balanced training data set if the goal is good performance for stations in the East of Oslo, where blasts are infrequent. This is the reason why we choose to train only one model for the station on which we also applied the outlier detector in the West of Oslo, where most blasts were observed. Consequently, the model learns station-specific features and does not generalize well to stations in the East, or even close stations in the West. With longer time series of blasts being available in future, we would like to generalize the blast classifier for more stations, and most important, for other source areas. For now, we believe it makes sense to focus on a single station in the West of Oslo for blast classification. However, to follow the comment of the reviewer, we now also trained a new model using data from two stations in the West of Oslo and compared the results. Supplementary Figure 3 shows that the generalization ability is as expected not very good. The classifier trained on OSLN2 and applied to OSLN3 does not perform good at all. The model trained on both stations and applied separately to both stations perform better at OSLN3, but still clearly worse than our preferred model. The model trained on both stations and applied to OSLN2 performs slightly worse than the original model we used in the paper. Hence, these tests suggest that more training data are needed to train a model able to classify blasts on all stations.

We added the new results and this discussion in the manuscript.

L136-L149: Is it necessary to train two auto-encoders with such a different setup? I guess the simplest solution would be to train one auto-encoder for both or at least keep the same setup. As I understand, the different setup results from visually inspecting the data, but I would assume that the models performance would not be worse with a similar setup or one auto-encoder for both stations. Did you try that? If yes it would be interesting to make show a performance comparison in the appendix.

As mentioned above we need to train auto-encoders for each station because of different quite background noise and signal characteristics. An autoencoder trained on all stations of the network would not be able to reconstruct specific signals at each sensor and report all those as “inliers”. We keep almost the same setup, and only change the input data dimensions to account for more complex noise conditions. However, we followed the suggestion of the reviewer and trained a single auto-encoder for two stations and compared the results, which confirmed our choice to make the detector station-specific (Supplementary Figure 2). This figure also allowed us to evaluate the performance of just applying the STA/LTA detector, which corresponds to using a correlation threshold of 1.0 to calculate recall and precision, i.e., selecting all STA/LTA detections as outliers. This results naturally in the highest recall, but with very low precision values around 0.2.

L150-L160: Here I am a bit confused, maybe the phrasing is not so clear to me. I understand that the correlation coefficient (CC) is a reasonable choice here and the CC considers phase and amplitude. Though, it seems to me that RMS would also perform in a reasonable way and take also into account phase and amplitude information. Moreover, the outliers in Figure 3 have larger STA/LTA values than the non-outlier, so amplitude seems to be important.

For the given examples, the outlier had indeed larger STA/LTA. However, this is not necessarily the case for all signals of interest we want to pick up (see new example added in the supplementary material). We agree that construction loss or RMS is an optional choice for triggering outliers, and that it also includes phase information. However, we found that using RMS does not work well. The new Supplementary Figure S4 shows the RMS vs. CC. which clearly shows that a correlation threshold is a much better choice for catching confirmed blasts compared to an RMS threshold. We added this to the text.

L163: How did you find out that the dimension of your auto-encoder is optimal for that task? Is there a way to evaluate the performance with changing dimensions?

We use recall and precision with respect to our ground true data. See answer above and new figures.

L168: It could be interesting to show a histogram of the CC values for both stations. There you could draw your thresholds as a line and estimate the percentage of outliers on each station. I assume that they should be in a similar range. Maybe you can add that to Figure 9.

The new Supplementary Figure 4 created to compare correlation coefficients and RMS now shows the distribution of CC values with respect to the threshold. We also added a histograms for OSLN2 and EKBG1 as suggested (Figure S5).

L200: How does the KerasTuner work? Is it doing a grid-search? What are the ranges of hyperparameter? Maybe that is also information for the appendix, but it is important for validating the training phase as a reader.

Information on which parameters were optimized for the blast detector was indeed missing. We optimize the filter length while the number of filters in each of the 5 convolutional layers of the AlexNet model was kept constant. The input parameters of KerasTuner are either ranges of hyperparameters (here: filter length) or different options (here: max vs. average pooling) that are tested by searching the parameters space. KerasTuner can iterate over all possible hyperparameter combinations (grid search) or use another optimization algorithm (we use a Bayesian algorithm from KerasTuner options) where the objective function is the classification accuracy. We added this information in the text.

L202: Do you shuffle the data? What is the dataset size here? Is it the 1870 blasts + 1870 noise samples? Especially compared to the number of parameters which are trained inside the CNN.

Data are randomly split into training and test set, thereby being shuffled (info added). The data size is 1,272 blasts and 1,272 noise examples. This number is given later in the section “Results of blast classifier”. The reason for not using all 1,870 signals for training is that we wanted to simulate a workflow where we only train with the previously detected blasts from the outlier detector, excluding those that we only identified after screening all STA/LTA detections. As discussed in the paper, ideally this data set should be larger for training models having a lot of parameters. Nevertheless, the model performance encouraged us that our approach is valid despite of a small training data set.

L211-213: Here it is not clear to me what processing is happening. The processing was station-dependent for the auto-encoder. Is it now the same case? How does the catalog look like? Does it contain the seismic data from all stations? Only the event labels in relation to the neighborhood of the city? How many labels do you have?

This was not clearly described indeed. The blast classifier is only trained with data from a single station. The reason for this choice was explained above. We tested different models using blasts samples from multiple stations (one model is included now in the supplementary material) as well as multiple channel waveforms from several stations combined as input, but with limited success, most likely due to the limited observation period. Hence, we decided to focus on the single station classifier in this publication. We added the missing information.

Figure 5,6,7: The size of the star indicates the potential ground truth radius. Where does that come from?

Construction blasts at the Fornebu metro line are reported to the public through a web portal (https://nabovarsling.no/fornebubanen). We added the missing reference in the data availability section. The precise locations are not provided at the web portal, only that a blast occurred at a certain time inside a radius of about 100 m. Hence, we scale the stars in the figure so that the symbol size approximately includes this location uncertainty (info added). In the case of the Losby quarry blasts, we simply use the known quarry location as ground true and the extent of the quarry as symbol size. In addition, we know that this quarry usually has a blast at 12:00 (not daily though). Therefore, although the signal location can be a bit off due to lack of resolution, we can be almost certain that a particular signal originates from that quarry.

L232: Here you mention 1272 locatable outliers triggered at OSLN2 and EKBG1. In line 214 you mention 1870 located blasts? Aren't they based on the outliers from the two stations? I get a bit confused with the type of events and how they were detected and how they relate to each other. I think it would be helpful for the reader to provide a table with the different event types and how they were detected.

The blast classifier was trained with blasts identified by the outlier detector (1,272), but it was able to classifier more blasts which were not detected by the outlier detector. Our "ground-true" used for evaluation is the reference data set which includes 1,870 blast manually identified by screening all locatable STA/LTA detections. We clarified this in the text.

L244: Interesting to see a pause during the school holidays in summer. It seems counter-intuitive to me since the summer times in northern countries are the times used for constructions (at least that's what I see on German roads for example).

It might sound quite exceptional that a whole country goes into silent mode each Easter and July (Christmas is maybe not so surprising), but this is indeed the case in Norway, including most construction works. The only exceptions are upgrades at existing railway tracks which are preferably done during holidays.

L248: „reviewed reference data in the background of Figure 8b" – so they are all STA/LTA detections? Again, a summary table of all catalogs and datasets would be helpful for the reader.

Yes, the reviewed event list originates from all detected and locatable events after running an STA/LTA detector. We feel that an additional table is not needed since it would only include three numbers. Instead, we clearly state in the text (see for example conclusions) that there are three catalogs:

Reference STA/LTA manually screened: 1,870 blasts.

Outlier detection: retrieved 1,271 blasts.

Blast classifier: retrieved 1,385 – 1,627 blasts (depended on threshold).

Table 2: Why do you use balanced accuracy? I though the blast/not blast classes are balanced. However, in L289 you say that the data is unbalanced. I assume that your imbalance concerns the location of blasts and not the blast vs non-blast class. Is that correct? If yes, then you could use the classic accuracy and rephrase L289 to make it clear.

The reviewer is right. What we show is plain accuracy because noise and blasts are already balanced. The imbalance only concerns blast observations on different stations and blast location. We changed this.

L283: It could be interesting to show a few examples of the detections which cannot be located.

We added some examples in supplementary Figure S6.

L284/285: „...the actual number of false classifications is negligible.“ - I would argue that this statement depends on the type of application. What are the implication of a false classification for the downstream smart-cities-applications?

This is a good question. If the goal is early warning in case of unusual events (accidents, attacks) and a seismic monitoring system should automatically alert the city authorities and the public, any false detection should be avoided. One way to ensure this could be to combine seismic alerts with other data sources available to the city authorities. Since there were no unusual events of public interest in Oslo during our project, we focus here on ongoing construction activity. Here the goal for a smart-city solution could be simply to provide the public with real-time information on a public dashboard or a mobile app. If citizen felt shaking, they can check if this was related to one of the many construction sites in Oslo. A false detection would hence not be a big issue. We added this discussion.

L310-317: These statements are interesting but could you provide some references to that (e.g. features of CNN represent spectral representation, auto-encoders mimic Fourier transforms). Moreover, why do you think your method is superior to a Fourier-transform based approach in that task? I don’t see suggestions/evidence in the manuscript that your method performs better. Did you try that?

We acknowledge that our statement about the nature of features learned by CNNs was a bit too speculative. We do not know for sure that the latent features represent the spectral coefficient in some form. This was simply a hypothesis based on the fact the convolutional filters of different length extract features from the input waveforms. There are certainly more features (e.g., waveform polarization) that can be among the latent representation of the input waveforms. Nevertheless, if for the sake for argument we assume that the auto-encoder would find the spectrum to the most efficient way to encode waveform data, there would be no need for a deep learning model. The new auto-encoder models with different number of latent features we added now (see above) include one where the latent feature number is the same as the number of input data samples (3xT). If performs wore than our preferred model with 2xT latent features. Hence, if the hypothesis that the

3xT model learns to mimic the Fourier spectrum is true, the 2xT model would also outperform the Fourier transform. However, given that all this is a bit too hypothetic, we decided to remove this part from the discussion.

L336/337: Could you provide a relative number of false alarms here?

This was not clearly explained. If an event is locatable, it is a potential event of interest since it is not clearly defined what an outlier is. In that sense, our automatic outlier detector combined with the Rg-wave based localization does not produce false alarms. Any locatable event can be interest. Having said this, there are rare instances when the location is just based on randomly associated noise bursts mistaken for Rg waves at different stations from a distinct event. These would be real false alarms but are very rare. We encountered this in less than 1 in 100 locatable outlier events on average. On the other hand, we can also evaluate the outlier detector with respect to our reference blast detector before attempting an event localization, which sorts out many false alarms but also unlocatable blasts. In this case, the false alarm rate corresponds to the precision values given in the new supplementary figures as introduced above.

Line-specific phrasing points:

L46: Another interesting argument would be that seismic data provide a completely new sensory perspective compared to visual or audio data, introducing different type of information.

Added.

L70: This sounds like that you will present clustering strategies in the manuscript which is not the case. Maybe remove the word „cluster“ or replace it with „groups“ or „localized event clusters“.

Good point. We rephrased.

L73: Maybe name „surface wave“ directly „Rayleigh wave“ to exclude other types of surface waves, which you do not consider in this study.

Changed.

L116: I would add the reference of Valentin 2010 to show that this idea of data compression with auto encoders exists since a while.

Unfortunately, we were not able to find this reference. There is a paper from the year 2018 but we are not sure if this is the one the reviewer suggested. We would appreciate a DOI or a complete citation.

Figure 5 captions: There is a typo in the last phrase of the caption.

Thank you for spotting this typo. Corrected.

Figure 9 and b: what is „event STA/LTA“? Is that the STA/LTA threshold you set to 4 earlier? If yes, I would suggest using the same term.

The event STA/LTA is the maximum STA/LTA ratio for a particular event signal, not the threshold. We clarified.

Response of Reviewer A

This paper aims to investigate seismic events generated by construction activities in Oslo, Norway, focusing on blasts from tunnel construction and underground water storage facilities. Employing low-cost seismic sensors deployed between 2021 and 2023, the study develops a prototype automatic urban seismic monitoring system. By integrating machine learning techniques, including outlier detection and classification using Convolutional Neural Networks (CNNs), the research endeavors to efficiently detect, locate, and classify these urban seismic events, thereby contributing to enhanced infrastructure monitoring and public safety measures in urban environments.

While the paper is well-written and clearly explains its findings, there are opportunities to enhance its impact by incorporating additional points for discussion and analysis. The corresponding comments are given below.

We would like to thank the reviewer or his helpful suggestions.

1. The abstract and introduction need to clarify the significance of the study's topic. For instance, explaining the potential damage caused by construction blasts to structures in urban areas, and how such events can impact public safety and infrastructure integrity, would enhance the reader's understanding of the importance of the research.

Thank you for the suggestion. We added these points in the introduction with a few more references, addressing the effect of blasting on structures.

2. The paper should elaborate on the importance of employing AI models, particularly auto-encoders and Convolutional Neural Networks (CNNs), in urban seismic monitoring. Providing examples of how these models contribute to the accurate detection and classification of seismic events, thereby enabling timely response and mitigation efforts, would strengthen the discussion.

We added a few more references of previous works on using deep learning in urban monitoring, signals clustering in particular. We also elaborated a bit more on the benefit of using deep learning instead of traditional seismic monitoring methods.

3. It would be beneficial to discuss the degree of damage caused by construction blasts to structures. Providing examples or case studies illustrating the varying levels of damage, categorized based on established scales such as EMS-98 grades 1 to 5, would offer insight into the potential impact on urban infrastructure.

We rephrased and now refer to newly added references for assessment of blasting on structures. Since we do not have examples from our study area, we feel we cannot discuss the potential damage of blasting in our case study in depth. Our focus and expertise are on event detection, and not on civil engineering. However, we acknowledge that this would be a worthwhile additional study given the major construction activity in Oslo.

4. Exploring the reasons behind higher-level incidents, such as specific construction activities or geological factors, would provide valuable context for understanding the variability in blast severity and its implications for urban planning and risk management.

We already mentioned quick clay mobilization and instances where the construction blast yield was higher than anticipated. However, we feel it would go beyond the scope of the study to explore the reasons for the latter due to lack of information provided by the construction companies.

The practical implications of the study should be elucidated, highlighting how the developed model can be utilized by stakeholders such as municipalities, governmental organizations, and urban planners for proactive monitoring of construction activities and assessing their impact on urban environments.

We added a new paragraph in the discussion section elaborating on how the results of our study can be utilized by stakeholders.

5. Emphasizing the importance of outlier detection in improving the prediction accuracy of the developed model would enhance the discussion on the effectiveness of the proposed methodology.

We added more statements on why we choose outlier detection instead of clustering. We also provide more information (new figure) to facilitate model comparison.

6. Discussing the possible implications of the developed method, such as its potential application in other urban areas or its integration into smart city solutions, would broaden the scope of the study and highlight its broader societal impact.

The new paragraph in the discussion section includes more information on how the results can be utilized in smart city solutions.

7. Addressing limitations beyond the issue of unbalanced data, such as the adaptability of developed neural network models to different geographical regions or potential biases in the training data, would provide a more comprehensive assessment of the study's constraints.

Adaption of our approach to other regions required re-training of the models, as usual when adapting machine learning to new areas. We believe that the auto-encoder-based outlier detector and the CNN blast classifier are robust methods to be used for other urban networks. The limitations are only present with respect to the data available. The system would require a station network with sufficient resolution for location. The outlier detector does not require a large data set. However, the blast classifier would of course require rather frequent blasting to gather enough events for training the CNN. We added this to the already existing discussion about limitations of our methodology.

8. Explaining how the developed method can be practically implemented and whether modifications are necessary for use in different locations would offer valuable guidance to potential users and address concerns regarding the model's generalizability.

We added these issues to the discussion section.

9. Comparing the developed model with existing literature on earthquake detection and monitoring in the discussion section would provide context for evaluating its performance and potential advancements in urban seismic monitoring techniques.

To our knowledge an auto-encoder outlier detector does not exist for seismic event monitoring. Other deep learning methods (event classification and phase detectors) share similarities with our blast classifier but cannot be simply applied to our data set. For example, PhaseNet (a popular deep-learning phase picker) needs P and S waves which are not observed for most blasts. To our knowledge no study exists which focuses on detecting Rg waves-dominated signals in cities. Therefore, a direct comparison is not possible. We do however provide a comparison with the most simple event detector: the STA/LTA trigger. We now also discuss a potential alternative method from deep leaning, which is deep clustering. However, we feel application and comparison to our data set is beyond the scope of this study.

10. Incorporating a comparison with similar studies in the literature, if available, would provide a benchmark for evaluating the novelty and effectiveness of the proposed methodology, enhancing the paper's overall contribution to the field.

We already mentioned similar studies in the introduction, and now also added unsupervised clustering to discuss and compare an alternative approach. We clearly state why we think that clustering is not optimal for outlier detection in our case. We could not identify other studies which would allow for a straight-forward application and direct comparison with our results. Our objective is too specific (local blasts in high noise environment) for any openly available ML methods for seismic event detection to be applied.

11. Addressing the above-mentioned questions in the paper, particularly in the introduction, discussion, and conclusion sections, would enhance the clarity, relevance, and impact of the research findings.

We addressed the above-mentioned questions in the introduction, discussion, and conclusions sections.