Reviewer 1

Overall, a very well written paper that will appeal to earthquake engineers active in building damage assessment and the wider audience interested in the use of novel technologies (InSAR) and machine learning. Nevertheless, some points would benefit from additional explanations/clarifications. My comments are provided below.

We would like to thank the reviewer for highly constructive comments on the article. The suggestions are highly appreciated. Our responses to the comments are provided below.

Comment 1

Lines 175-176: "The SAR-derived DPMs published by the ARIA project are used as the primary remote-sensing proxy to identify surface-level changes that are potentially attributable to earthquake-induced building damage."

Could you please clarify how you interpret the correlation between surface-level changes and earthquake-induced building damage? Would it be possible to have seismic building damage without significant change in the ground surface level? If so, how would you proceed to remotely detect building damage using InSAR or EO?

==

In previous events where earthquake DPMs were generated by the ARIA team, such as the 2015 Gorkha earthquake in Nepal and the 2016 Central Italy earthquakes, the DPMs were found to have good correlation with the actual damage mapped in field surveys [Yun et al. (2015), Sextos et al. (2018)]. However, landslides and rockfall can also lead to surface-level changes, and so can changes in vegetation and other phenomena such as building construction. Thus, while attempting to detect building damage, care needs to be exercised to limit the focus of the DPMs to locations where buildings are known to exist. The time-span between the acquisition of the pre-event and post-event images can also have a considerable impact on the potential false positives. The closer the 'before' and 'after' bracket the event, the fewer the false positives that are likely to be observed.

It is also certainly possible to have seismic building damage without observing significant change in the ground surface level. Damage to internal walls or columns that may have severely compromised the structural integrity of the building without causing externally visible damage or collapse may not be detectable through remote sensing. Storey drifts of 2% for braced steel structures or concrete shear wall structures may be technically classified as 'collapsed' (eg. FEMA 356), but such drift levels might be smaller than the level detectable with the 1–3 m spatial resolution offered by the current generation of SAR sensors. Dong and Shan (2013) provide a good review of previous studies that investigated the relationship between building appearance in remote sensing data and building damage grades. In general, they concluded that the higher damage states like complete collapse are more detectable through remote sensing data, but lower damage states are more challenging to detect. With the advent of commercial SAR satellite constellations like Capella Space and ICEYE, sub-meter SAR imagery is becoming available, and some of these deficiencies should be addressable.

In recognition of these limitations in the SAR and EO datasets, in our study we have proposed to incorporate other variables (such as building attributes or the expected macroseismic intensity at the location of the building) to mitigate these issues. We have updated this section of the manuscript to clarify these limitations, and how we propose to address them.

==

Comment 2

Lines 185-188: “The problem presented is one of multi-class classification, and two ensemble machine-learning classification algorithms are employed—the Random Forest classifier for the cases involving only numeric features, and Histogram-Based Gradient Boosting classifier for the cases which also involve categorical features amongst the selected building attributes.”

Could you please explain why you selected the random forest algorithm and histrogram-based gradient boosting classifier? Did you try any other algorithms? How did they perform?

==
In a preliminary phase, we compared different algorithms that permit multiclass classification, including support vector machines, k-nearest neighbours, Naive Bayes, and Random Forest. Since the problem of damage classification typically involves highly imbalanced datasets, where the buildings in "no damage" state dominate the buildings in all other damage states, often by multiple orders of magnitude, all of the above classifier algorithms tended to overlearn the label with the higher number of training examples (i.e., "no damage"). The Random Forest algorithm was eventually selected for the study as it allows for the assignment of weights to the training examples. The training examples in each damage class were then weighted in inverse proportion to the class frequencies observed in the input data, in order to better handle the class imbalance in the input damage datasets. The Histogram-Based Gradient Boosting classifier was preferred in the cases where categorical features were present amongst the selected building attributes, in addition to purely numerical features. This was because the Histogram-Based Gradient Boosting classifier provides native categorical support, which helps avoid one-hot encoding to transform categorical features as numeric arrays. We have included this explanation in the revised version of the manuscript.
==

Comment 3

Lines 188-189: “The models are trained with a 70% subset of the available data, and then the best-fit models are tested against the 30% hold-out subset.”

Could you please explain why you chose a 70%/30% for the training and testing set? Did you try 80%/20% for example? Was 70%/30% giving the best performance?

==

There is certainly a tradeoff between using more samples for training the algorithm versus reserving sufficient samples for the test set. Using a higher fraction of the available data for training can result in overfitting. Previous empirical studies have demonstrated that using 70-80% of the data for training and reserving 20-30% of the data for testing yields optimal results in terms of improving the accuracy of the model while minimizing the tendancy for overfitting [see Gholamy et al. (2018) for instance]. The decision to choose a 70%/30% split for the training and testing set (say, over an 80%/20% split) was ultimately driven by the paucity of 'collapse' labels in the damage datasets, particularly for the 2017 Puebla event where reserving only 20% of the dataset for testing would leave very few 'collapse' labels in the test set to evaluate the accuracy of the fitted model.
We have added this explanation in the revised manuscript, with references that also justify this approach.
==

Comment 4

Lines 211-213: “Another important reason to undertake a binary damage classification exercise is that it permits the aggregation of building damage datasets from different events into a larger training pool.”

Could you please clarify what you understand under a “larger training pool” for a machine learning model?
The paucity of building damage data that can be used for training is one of the main challenges affecting machine learning models for damage prediction. Different countries use different methodologies and different damage scales to assess building damage following earthquakes. While the definition of the lower damage grades might differ considerably between different scales, collapse is often consistently defined. Thus, if the focus is restricted to identifying collapsed buildings from non-collapsed buildings, a wider set of events from the region can be used train the model, given that the training labels in this case coming from different events will be consistently defined. We have included this information in the revised version of the manuscript.

Lines 325-329: “Cross-regional training datasets will also help greatly improve the performance of these models for earthquakes in new regions previously unseen by the model. By expanding the datasets used to train the ML damage classification models, we can transfer the learning from regions with more damage data availability to data sparse regions. Cross-regional training is also critical as it will ultimately make such damage classification models more robust as they can be more confidently applied to future disasters, which may affect regions the model has not been trained on.”

Could you comment on the performance of a ML model applied to region on which it has not been trained on? When creating a “larger training pool”, could you please explain how you would capture regional specificities?

This reviewer is of the opinion that each location/region has specificities (e.g., construction practices, seismic setting, ground conditions) and thus has concerns regarding the applicability of a "one-fits-all" ML damage prediction model.

==

The location/region-specific concerns expressed by the reviewer are well appreciated. One of the eventual promises of the framework described in this paper is to be able to predict damage using InSAR data even for locations that aren't present in the training data. Ideally, region-specific damage detection models could be developed that take into consideration input features that are region-specific. Alternatively, region-specific or location-specific characteristics could be encoded as additional input features to a global remote-sensing based damage detection model. For instance, one of the inputs in the proposed methodology is the ground shaking intensity map (ShakeMap) generated by the US Geological Survey, which does take into consideration local site conditions, albeit through a proxy measure (Vs30). The tectonic setting is also taken into account implicitly in the derivation of the ShakeMap, as the choice of the ground motion model used to predict the ground shaking intensities in the affected area depends on the tectonic region type. If information about building construction types is available, this can be encoded as a categorical input feature, as was done for the 2015 Gorkha and 2017 Puebla examples in this study. Other researchers such as Moya et al. (2018) have also attempted to incorporate fragility functions into a machine learning

framework. The concerns raised by the reviewer are also valid for the current state-of-practice in earthquake risk assessment, where ground motion models or empirical fragility / vulnerability models that have been derived for a particular region where sufficient data are available, are routinely employed for damage / loss assessment in other regions (ideally sharing similar characteristics) where not enough data are available. We have included a discussion about this limitation and potential source of bias in the concluding remarks.
==

Comment 5

Lines 320-323: " The training of the machine learning models happens prior to the disaster event, and the trained model can be deployed for damage detection following an earthquake as soon as the pre-event building inventory, ShakeMap, and DPM become available ".

Could you please clarify the process? Why does this information only appear in the conclusion? Did you try to train a ML model for a region prior to a disaster and test the ML model after the earthquake event?

==
This statement is meant to depict how the proposed framework would work in a real-time post-event damage assessment environment. Within the scope of the current study, we unfortunately did not come across building-level damage data from multiple events within the same country or geographic region that could be used to train a ML model for the region using data from previous events prior to a disaster and test the ML model after the subsequent earthquake event. The phrasing of the sentence has been improved to better convey the intention, to: "*The training of the machine learning models would have been undertaken prior to the disaster event, and the trained model can then be deployed for damage detection following an earthquake as soon as the pre-event building inventory, ShakeMap, and DPM become available*"
==

Comment 6

Lines 342-343: “Code availability. The Python code and Jupyter notebooks used for the analysis are available at https://github.com/gemsciencetools/eodamage-detection under the GNU Affero General Public License (v3.0).”

GitHub repo accessed by the reviewer on 29 May 2022

Data and code were found for the Gorkha, Puebla, and Zagreb earthquakes, as well as the central Italy earthquakes. However, no folder/code related to the Puerto Rico earthquake could be found.

==
The folder containing the data and code related to the Puerto Rico earthquake has been added to the repository.
==

****

Reviewer 2

The paper describes a framework for building damage classifications after an earthquake that combines InSAR data, high-resolution building inventory data and earthquake ground shaking intensity maps. The classification is performed using two different strategies (multi-class and bynary classification) and two different machine learning algorithms.

I am a researcher in the field of computer science, particularly in machine learning. Therefore, this review will focus on issues related to machine learning techniques used to solve a problem in the area of natural disasters.

The article is well structured and organised. It is not difficult to understand the main ideas of the work done and it is well written. The work described represents a very valid proposal but the original contributions are minor. Anyway, if the proposed framework proposed is well evaluated and validated, for me, it can be accepted for publication. However, as the paper is, in terms of validation, does not follows the standards and requirements of this journal.

The authors claim innovation in the use of the Multi-class damage grade classification using InSAR data. I have nothing against but the results presents are weak to validated the method. I think authors should present results that clearly show the benefits of the strategy used when compared to other techniques proposed. The multi-class strategy seems to me very important but the results are weak and are not compared. Also, the selection building damage categories is a critical point, regarding the urgency of action. This should be more discussed and explained in the paper.

Deep learning techniques are been used to solve many problems in many fields and presenting good results or better than the previous methods. I found it very strange that the authors would talk about recent advances in machine learning and then not use deep learning techniques for classification.

I understand that in some situations it can be difficult to use deep learning techniques, particularly when there is not much data available for training. If that's the case, I think the authors should present results demonstrating this. Also, the data augmentation methods should be considered.

==

The comments made by the reviewer are well taken. We would like to emphasize that the focus of the article is more about the integration of multiple input datasets including InSAR data for building damage detection, rather than on machine learning itself.

As far as we are aware, multi-class damage classification at the building level using SAR data has not been attempted before, so we are unfortunately unable to compare our results with any existing literature, and this was precisely one of the reasons we pursue this study. It was also our intention to propose an open framework and data that other earthquake engineers and risk modellers could use for rapid loss assessment.

Deep learning techniques are more suited for structured data (images, audio, text) with large sample sizes, while the datasets used in this study are tabular (each row representing one building unit) and are small or medium sized (typically of the order of 1,000–10,000 samples). While it's true that for tabular data, both decision forest based approaches and deep neural networks could be used, as the reviewer acknowledges, deep learning techniques perform better with large sample sizes, which was unfortunately not the case for this study, given the limited availability of building-level damage datasets and the limited number of labelled damaged buildings within each dataset. Grinsztajn et al. (2022) conclude that for medium sized tabular data (~10,000 samples), tree-based models outperform deep learning methods, with much less computational cost. Similarly, Xu et al. (2021)

also conclude that forests perform better than deep neural nets for tabular data with small sample sizes.

The reviewer's comment about the selection of appropriate building damage categories is apt. Dell'Acqua and Gamba (2012) highlight the need to develop damage scales specific to earth observation based damage assessments, possibly tied to existing damage scales that are widely used for field surveys of building damage (such as EMS 98). Cotrufo et al. (2018) propose a building damage assessment scale tailored for optical satellite imagery and aerial imagery. However, a similar damage scale tailored for InSAR based building damage assessment is still lacking. The revised manuscript discusses these limitations and potential areas for future research in the concluding remarks.

References cited in this response

Cotrufo, S., Sandu, C., Giulio Tonolo, F., & Boccardo, P. (2018). Building damage assessment scale tailored to remote sensing vertical imagery. European Journal of Remote Sensing, 51(1), 991–1005. https://doi.org/10.1080/22797254.2018.1527662

Dell'Acqua, F., & Gamba, P. (2012). Remote sensing and earthquake damage assessment: Experiences, limits, and perspectives. Proceedings of the IEEE, 100(10), 2876–2890. https://doi.org/10.1109/JPROC.2012.2196404

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? https://arxiv.org/abs/2207.08815

Xu, H., Kinfu, K. A., LeVine, W., Panda, S., Dey, J., Ainsworth, M., Peng, Y.-C., Kusmanov, M., Engert, F., White, C. M., Vogelstein, J. T., & Priebe, C. E. (2021). When are Deep Networks really better than Decision Forests at small sample sizes, and how? https://arxiv.org/abs/2108.13637v4