Anonymous Referee #3 (Report #1)

The proposed work focuses on the development of a framework for building-level earthquake damage assessment. High-resolution SAR data are used as input together with building-related datasets and earthquake-related intensity maps for the classification of the building damage states. The paper is well-written and the presence of a complete repository with data and codes to evaluate the proposed methods is of great value.

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.

The adopted ML technique for the implementation of the classifier has been justified by looking into the literature where similar solutions have been successfully adopted and compared with other approaches. Due to the presence of significative differences between the proposed method, in particular in terms of combined input data, and the previous works, other ML techniques should be used for comparison. This aspect should be mentioned in the manuscript.
>>> In lines 182-190 of the revised manuscript, we explain why deep-learning techniques were ruled out for the current study. In lines 191-202, we also mention having compared other ML techniques in the preliminary stage of the study and the reason for settling on the Random Forest classifier and Histogram-Based Gradient Boosting classifier for the remainder of the study.

The selection of an ML technique includes the definition of multiple hyperparameters that can be modified/optimized to maximize the system's performance. This point has not been discussed at all in the work. Even if default hyperparameters have been used, they have to be presented and future optimization solutions can be considered.
>>> We have included the following text in the revised manuscript to discuss the treatment of hyperparameters, starting at line 203.
The next step involves tuning of the hyper-parameters of the chosen classifier algorithms, where hyper-parameters are the model parameters that are not directly learnt during the training phase. Probst et al. (2019) provide a thorough overview of the hyper-parameters and tuning strategies for the random forest algorithm. Random forest algorithms have three main hyper-parameters, including the number of trees in the forest, the node size, and the number of features sampled when looking for the best split for a node. The number of features sampled at each split is set to the square root of the number of predictor variables, which Probst et al. (2019) indicate as a reasonable value for low-dimensional classification problems. Optimal values for the number of trees in the forest and the node size are obtained through an exhaustive grid search strategy, to pick the combination of hyper-parameter values that result in the best cross validation score. For all other hyper-parameters which have less of an impact on the model performance, we use the default values provided by the software package scikit-learn (Pedregosa et al., 2011).

Even if the datasets are unbalanced, proper strategies have been considered to overcome this problem. However, in Fig. 6 it is evident that the learned classifier is significantly biased in the prediction of low-number labels. This aspect needs to be further discussed and potential solutions to overcome this problem should be proposed. The reported results for a multi-class classifier are barely useful, is it a problem of the method or mainly due to the available data? The possibility to compare the obtained results with other baseline methods could have answered in part to this question.
>>> The observation that the learned classifier tends to be biased in the prediction of lower damage grades is true, and this is particularly evident in the case of the Puebla earthquake results in Figure 6b. As mentioned in the manuscript, there aren't many previous studies that have attempted multi-class damage classification for earthquake building damage that we can compare our results with.

Mangalathu et al. (2020) report that in their attempt to classify building damage from the 2014 South Napa earthquake into three damage classes, the random forest algorithm was correctly able to identify only 12.5% of the red-tagged buildings. Using the 2016 Kumamoto earthquake as a case study for binary damage classification, Bai et al. (2017) obtained a prediction accuracy of 38.9% for identifying damaged buildings when multi-temporal post-event SAR images were used along with the K-Nearest Neighbours learning algorithm. Lanteri et al. (2017) report that for the 24 August 2017 Central Italy earthquake, the Copernicus Emergency Management Service's damage grading maps for the event, made by comparing pre- and post-event optical satellite images, correctly predicted 18.85% of the highly damaged or completely destroyed buildings in the affected areas. In comparison to these previous studies, the prediction accuracy for the highest damage grade in the current study ranges from 42% to 47% for the 2020 Zagreb, 2020 Puerto Rico, and 2015 Gorkha earthquakes. We contend that these results are still quite useful, as they are able to correctly identify over 40% of the heavily damaged or collapsed buildings. We have also included the following text in the revised manuscript, beginning at line 355:

From Figure 6, we also observe that the true-positive prediction rates for the intermediate damage grades are lower than those for the no-damage and highest damage grades for the 2015 Gorkha earthquake and the 2020 Puerto Rico and Zagreb earthquakes. We believe that this partly stems from the fact that the existing damage scales that are widely used for field surveys of building damage, such as EMS-98 do not map directly to information available through earth observation data, particularly for the lower damage grades. For instance, the first three damage grades for reinforced concrete structures according to EMS-98 involve increasing levels of cracking in the beams and columns or partition and infill walls, and buckling of the reinforcement rods. Unless this kind of damage results in debris caused by excessive spalling of the concrete cover or partial collapse of infill walls that is visible outside the structure, these damage levels as defined in EMS-98 may be challenging to identify from EO data alone. Dell'Acqua and Gamba (2012) and Cotrufo et al. (2018) both 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, and merits further research.

The discussion of the obtained results is limited to the list of the classification performance obtained, further discussions should be reported. For instance, for the Gorkha earthquake, the presence of multiple building attributes was considered the main cause for the higher performance obtained. This claim can be verified by training a new model excluding this information and evaluating the performance drop. In general, the application of a sensitivity analysis could be mentioned as a useful method to better understand the role of the different input features in the evaluation of the output class.

>>> We have taken this suggestion into consideration and trained a new model for the 2015 Gorkha earthquake where the building attributes were excluded from the input feature vector. We have included the following text in the discussion section, starting from line 333

We observe that for the 2015 Gorkha earthquake, for which multiple building attributes are available for both damaged and undamaged buildings, the prediction accuracies for both binary and multi-class classification are significantly higher when compared to the earthquakes where fewer or no building attributes are available for use as input features. In order to understand the impact of including the building attributes on the performance of the classifier, we also trained the ML model for this earthquake without using any of the building attributes and limiting the input feature vector to the MMI and DPM values alone. The precision and recall for all damage grades are lower for this reduced model compared to the results reported for the full model in Table 2 and Table 3, for both multi-class classification and binary classification respectively. The recall score for the "Destruction" damage grade drops from 0.47 for the full model to 0.31 for the reduced model in the multi-class classification task, and from 0.73 to 0.45 in the binary classification task. Similarly, the balanced accuracy score drops from 0.36 to 0.20 in the multi-class classification task, and from 0.82 to 0.59 in the binary classification task. These results clearly demonstrate the importance of including the additional building attributes in the analysis. A partial dependence analysis of the damage grade on the non-location building attribute variables for this event indicates that the building age has an impact on the damage grade, with older buildings being related to higher damage.

Anonymous Referee #4 (Report #2)

The manuscript is focused on the use of ML classification where different inputs are used such as shaking maps, the SAR derived DPMs map and building information with the aim of semi-automated building damage assessment due to earthquakes.
General comments:
1.The manuscript is highly focused on the description of DPMs which is a product not coming from the revised work. This section could be summarized into the data and materials.
In addition, it could be preferable to remove the section background and to summarize several of the info provided in the section 2.2 Machine learning in building damage assessment into the introduction section.
2.The data and materials could be presented in a more separated way in my opinion.
>>> We appreciate these helpful comments concerning potential improvements to the structure of the manuscript. Accordingly, we have condensed the portion of the text describing the DPM product, and now it appears under the 'Input data' subsection. The 'Background' section has also been removed, and the review of previous literature on the use of machine learning along with earth observation data for building damage assessment has been subsumed into the introduction section. The 'Input data', 'Data processing', and 'Study areas' are now presented in separate subsections in the revised manuscript.

3.In addition, the authors could present and thus discuss the problem of the different spatial resolution of the inputs…indeed, the spatial resolution of the DPMs map (which if I correctly understood is 30m) could be feasible for some buildings only… the spatial resolution of the shaking map is not discussed…
The authors could highlight that this method could be applied only to part of the urban areas where the input spatial resolution fit the average dimension of the buildings….the spatial resolution could represent a limit of this methodology
>>> We have now included the spatial resolution of the ShakeMap product (Line 146 of the revised manuscript), this is typically available on a 1km grid spacing. Each pixel of the DPM measures approximately 30m across. With the present resolution, the proposed method is indeed more likely to be useful for detecting large damaged buildings, damaged building aggregates, and damaged dense building blocks, more than damage to isolated smaller buildings. We have included a short note regarding this point in the discussions section at line 374.

4. Finally, some additional specific comments are following reported.
Please notice that line numbers refer to the manuscript “nhess-2022-125-manuscript-version3.pdf” I suppose is the clean version of the manuscript after round 1 revision
Line 142
While this study focused on implementing and testing this framework for earthquake related damage, the proposed framework adopts a modular approach…
Probably this sentence could be…
While the above mentioned studies focused on implementing and testing this framework for earthquake related damage, the proposed framework adopts a modular approach…

>>> The suggested change has been made in the revised manuscript, and now appears at line 114.

Line 190
The authors refer 2 “Thus, we clip out parts of the DPM that are outside of built-up areas, based on building footprint maps and land-use maps.”
How the authors managed the high inaccuracies of SAR derived map geometry?
Also, since SAR products come from a lateral view it is well know that even if a terrain correction method is applied its radiometry content could be highly compromised (i.e. the real and imaginary parts), leading to an incorrectness of the coherence data which is the main info used to produce the DPMs maps (this effect could be more relevant for cities built in mountain areas).
Is in the DPMs metadata an info about the data accuracy (indeed, coherence reliability and spatial accuracy)? The authors could check for it and if present discuss this within the manuscript.
>>> The SAR single-look complex (SLC) images are processed using the InSAR Scientific Computing Environment (ISCE) processor developed by NASA-JPL (see Yun et al., 2015; Tay et al., 2020). The ISCE toolset (Rosen et al., 2012; Rosen et al., 2018) makes use of available information about the precise orbits of the satellite(s), atmospheric delay models, and a digital elevation model (DEM) to yield a well-aligned geocoded image. It also handles terrain correction or removal of topographically induced phase variations. The DPMs produced by Yun et al. (2015) for the 2015 Gorkha earthquake showed good correlation even in mountain areas with independent damage analyses by the National Geospatial-Intelligence Agency and the United Nations Institute for Training and Research’s United Nations Operational Satellite Applications Programme. Nevertheless, coherence difference (COD) alone is indeed less effective in places with vegetation growth, agricultural activities, snowfall or rainfall, where the COD may not be due to earthquake induced damage. The DPM metadata does not typically include quantitative information about its coherence reliability and spatial accuracy. However, the readme file included with each DPM does mention that the each pixel of the DPM measures about 30 meters across and that the DPM may be less reliable over vegetated areas.

Figures 6
maybe I am in wrong but i am not sure the label’s numbers (i.e 1,2,3…) were explained in the text….i suppose these refer to a damage level…
in addition, it is not clear to me way the authors are using color scales for the confusion matrices…if in the main diagonal higher values refer to higher accuracies, all other cases are omission or commission errors…
finally, the confusion matrices are not in deep discussed….supposing that the authors used a progressive labeling with the damage level (i.e. 0= no damage and 5 (or 3) = max damage) probably there are some reasons why the puebla case study shows the highest false positives… maybe it is related to the average dimension of the buildings if compared to the input’s spatial resolution?
>>> The label descriptions have been added to the figure captions. The labels indeed refer to progressive damage grades. For the Puebla event, we believe the lower prediction accuracy stems primarily from the limited size of the damage dataset available for training, compared to the other three events. The average dimension of the buildings doesn't seem to be the driving factor, as we are focusing on the damages sustained in the highly urban CDMX district where the average building size is of comparable dimension to the DPM pixel size. We have included the following paragraph in the discussions section:
While four key building attributes were also available for the damaged buildings for the 2017 Puebla earthquake, the non-availability of the same for the undamaged buildings meant that a complete dataset with building attributes could not be used for the training of the ML model. Of the four events considered in this study, the 2017 Puebla event had the smallest building damage dataset available for training the ML model. Only 219 buildings in Mexico city had a complete set of building attributes and damage labels and were also covered by the DPM and ShakeMap layers, as compared to thousands or hundreds of thousands of buildings for the other three events. While the Random

Forest classification model performs well for this event in the training phase, the trained model fails to correctly identify even a single partially collapsed or totally collapsed building in the test set. Further attempts at reducing the potential overfitting of the model to the limited training data subset by adjustments to the model hyperparameters did not lead to any noticeable improvements in prediction accuracy for the event.

References cited in this response

Lanteri, L., Pispico, R., & Cremonini, R. (2017). Two case histories of EMS data application: Earthquake in Central Italy and flooding in Piedmont region. Copernicus EMS Mapping User Workshop 2017. Ispra, Italy.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Machine Learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.4018/978-1-5225-9902-9.ch008

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), 1–15. https://doi.org/10.1002/widm.1301

Roseu, P. A., Gurrola, E., Sacco, G. F., & Zebker, H. (2012). The InSAR scientific computing environment. 9th European Conference on Synthetic Aperture Radar, 730–733.

Rosen, P. A., Gurrola, E. M., Agram, P., Cohen, J., Lavalle, M., & Riel, B. V. (2018). The InSAR Scientific Computing Environment 3.0: A Flexible Framework for NISAR Operational and User-Led Science Processing. 2018 IEEE International Geoscience and Remote Sensing Symposium, 4901–4904.

Tay, C. W. J., Yun, S.-H., Chin, S. T., Bhardwaj, A., Jung, J., & Hill, E. M. (2020). Rapid flood and damage mapping using synthetic aperture radar in response to Typhoon Hagibis, Japan. Scientific Data, 7(1), 1–9. https://doi.org/10.1038/s41597-020-0443-5