Comments from first revision:

1. Even though the topic of the research is clearly defined, the objectives are not sufficiently explained. Why should we explore ML models for damage assessment of building portfolios? What are the limitations of traditional/existing methodologies (e.g., Risk-UE)? Lines 54-58 mention the challenges in developing exposure models, which are true regardless of the damage assessment methodology. Finally, is the purpose of the manuscript to only demonstrate the benefits of ML models in this field or to use the developed heuristic model in other regions and future seismic events as well?

Authors' response: As noted in the manuscript, we believe in the manner of Riedel et al. 2015 that we need to change the way we look at exposure models because of the abundance of data and methods to explore them. While the Risk-UE, FEMA, GNDT methods rely on defined and validated models for a set of data characterizing the structures, we reverse the process by evaluating the available exposure data and test whether these data are sufficient to assess risk. For this reason we believe that it is necessary to test machine learning methods as an alternative to classical methods, and directly on damage prediction without going through the vulnerability. Again, because more and more post-seismic data collect information on building features and damage levels, without direct information on vulnerability.

Defining the exposure model is a challenge for damage prediction with Risk-UE because this information may not be readily available, and collecting it during emergencies could be too time-consuming and expensive. On the other hand, machine learning models can provide damage estimates more cost-effective way, as readily available data can be used to develop relationships between building features and damage. Furthermore, these models can help to discover new relationships, incorporate large amounts of data (e.g., global dynamic exposure models), and formally consider uncertainty. In addition, machine learning methods allow us to change the paradigm by proposing a heuristic approach to damage prediction based on available data. This approach was already mentioned in Riedel et al. (2015).

For the moment, the objective of this paper is to assess the efficiency of the machine learning methods, the distribution of the input data (imbalance issue) and the efficiency of the prediction. Before applying this model to other regions, analyses will have to be proposed, such as host-to-target adjustments by changing the region and thus the construction portfolio or the nature of earthquakes.

We add a discussion section in the new version of the manuscript.

I disagree with the authors' narrative. By any means, there is no abundance of publicly available building exposure data and that is exactly why most of the comprehensive available exposure models use census and cadastral databases (e.g., European Seismic Exposure model 2020 – Crowley et al. 2020). Nevertheless, recently, new tools have become available or developed (e.g., OpenStreetMap, ML models coupled with satellite images) which can significantly improve the development and refinement of exposure models. In any case, the topic of this manuscript is not the development of exposure models using ML models (e.g., Pelizari et al. 2021).

According to the authors' view, it is challenging to define an exposure model for damage prediction using a traditional methodology such as Risk-UE. However, the input exposure and structural data required by the proposed heuristic model are way too specific (e.g., floor area,

regularity, ground slope condition, roof type, position of building, etc.), much more difficult to collect for building portfolios and finally almost never available prior to the occurrence of seismic events and the completion of data collection and damage assessment surveys. Thus, such model is too cumbersome to use for seismic risk assessment, while the exposure models used by traditional methodologies mentioned by the authors are more convenient to use for seismic risk assessment of building portfolios.

Finally, the exposure model is one of the components of a seismic risk model. The link of exposure with vulnerability and geo-refenced seismic intensity/ground motion is crucial for the development and verification of seismic risk models; and it is not discussed or addressed throughout the manuscript.

2. Lines 93-94: Why did the authors consider damage data from seven earthquakes and not the entire DaDO database? Typically, ML models benefit from the use of large datasets.

Authors' response: Yes, you are right. Our choice is purely arbitrary, having chosen the earthquakes having led to the most observations and among the most "famous" (in our opinion, purely speculative) of the DaDO database.

I could understand if the authors did not select data from relatively old events which might be affected by data quality issues and could be tricky and troublesome to use. An arbitrary selection of events is not a solid argument in my opinion, especially because ML models benefit considerably from the use of large datasets in the training phase. The authors should either clearly justify their selection of past events in the manuscript or use the entire DaDO database.

3. The input parameter Building location in terms of latitude and longitude is irrelevant given that the latitude and longitude of the epicentre of the earthquake is not used. Why the authors did not use the epicentral/hypocentral or source-to-site distance instead? As a consequence, the importance of Lat and Long in Figure 5 is misleading.

Authors' response: Since the epicentral distance for all the buildings is not available in DaDO, we prefer to choose the lat/long data, in order to bring out an effect linked to the position of the building in the urban area. For example, we have highlighted (not discussed here because to be specified) the location of buildings in the oldest areas as being the most vulnerable, in connection with the organization of Italian cities (and in a broader sense, European). It should not be forgotten that the location of earthquakes is imprecise (unlike the lat/long of buildings) which can lead to bias. Finally, the distance is integrated in the definition of the seismic demand in the form of macroseismic intensity.

These data are therefore interesting to explore but certainly not essential to our study. We wish to keep them in order to also show the impact of the location of the structures in the urban area.

First of all, it is a trivial task to calculate epicentral or source-to-site distance (e.g., Joyner-Boore distance) for any number of locations given their coordinates and the coordinates of the epicentre. Secondly, the incorporation of epicentral or source-to-site distance might considerably benefit the performance of the ML models. Once again, the use of longitude and latitude of the buildings only is irrelevant because it is conditioned on the "unknown" epicentre of the earthquake. As a consequence, the heuristic model cannot be used for other events apart from the ones in the DaDO database.

What do the authors mean by “an effect linked to the position of the building in an urban area”? The factors that contribute to the dynamic response and damage of buildings are uncorrelated and independent from their position in an urban or rural area. The crucial factors/parameters among others, are the earthquake magnitude, source-to-site distance, site conditions, ground slope and topography, material of construction, structural attributes, etc. How does the position of a building in an urban area affect the ground motion, dynamic response and structural damage? Are the authors referring to the effect of the buildings’ location in an urban area in a post-disaster situation?

Moreover, every empirical or analytical dataset used in the development and verification of seismic risk models is affected by uncertainty and bias to a certain degree. For instance, the same argument brought by the authors can be brought to the development of every ground motion model the past decades. There is an abundant of information regarding the epicentre of past events (e.g., publications, national and European databases, etc.) and can be very easily accessed in platforms such as USGS. Finally, macroseismic intensity is a qualitative measure of the severity of ground motion and expected building damage, hence the epicentral/ hypocentral distance is rather important for the estimation of macroseismic intensity.

4. Observing the data distribution of Figure 2, it is clear that the wide majority of the buildings (85%) are one-storey. Therefore, the input parameters Height of building, Number of storeys and Regularity in terms of elevation are not relevant for the training of the ML models. In general, these structural parameters are crucial for the seismic response and vulnerability of buildings, thus I believe the authors should address this issue.

Authors’ response: No, the NF1 category corresponds to 0-3 storey. What we observe in our database is that statistically (not building by building) these 3 parameters are not the most important. This does not mean that for a specific building they are not, it means that in our database, these 3 parameters do not mainly contribute to the distribution of observed damage.

In the learning phase can only explore information contained in the training dataset. However, compared to other studies, the performance of the machine learning methods used is comparable, which tends to confirm the importance of the parameters considered here. We add a sentence in the discussion.

“Moreover, the machine learning methods only train on the data available in the learning phase, that reflects the building portfolio in the study area. The importance of the features contributing to the damage could thus be modulated, and would require a host-to-target adjustment for the application of the model to another urban zone/seismic region.”

My apologies, I was misled by the notation NF1. What is the distribution of number of storeys in the NF1 category? If the wide majority it is indeed 1-2 storeys, then it makes perfect sense why these three input parameters are not statistically important. In any case, these three parameters are generally important for the seismic response and potential damage of buildings.

5. Considering the above observation, did the authors test the employment of the recorded/median PGA instead of/along with MSI? Potentially, the performance of the heuristic model could be improved and outperform traditional approaches.

Authors’ response: No we use only MSI and yes, it is of course possible to use other intensity measure for ground motion (such as PGA) but as suggested by Cua et al. (2010), macroseismic

intensities represent the spatially-distributed ground motion and are more effective in communicating ground motion levels in relation to human experiences and incurred losses.

I completely disagree that MSI represents the spatially distributed ground motion more effective than ground motion IMs. This is exactly why the widest majority of ground motion and vulnerability models the past decade use PGA, spectral acceleration, spectral displacement, etc. The authors should explore the use of PGA instead of or along with MSI for a subset of the DaDo database for which PGA is available.

7. Lines 142-143: Did the authors test the importance of other parameters provided by USGS, such as M_w and hypocentral depth of the main events?

Authors' response: No we did not test other parameters. However, we did some tests (not presented here) on the magnitude: training for a magnitude range and testing on other magnitude range, or training and testing on the same magnitude earthquakes. Of course, we do not have enough examples to validate the results but without more efforts dedicated to this issue, no clear trends were observed. The question of Mw and hypocentral depth is finally (and indirectly) linked with the sufficiency of the IM (in the sense proposed by Luco and Cornell) and this was not tested here.

I understand that the limited number of events is cumbersome for the employment of M_w and hypocentral depth as input parameters. Nonetheless, the aspects of sufficiency and efficiency are completely out of the scope of the present study and more importantly this is not the reason for incorporating Mw as an input parameter; but rather as a parameter to express the earthquake's severity. The efficiency and sufficiency metrics (e.g., Luco and Cornell 2005) are used to assess the performance of IMs and seismic response models (based on PBEE principles). The former is measured by the standard deviation of the residuals and the latter by the conditional independence of the predicted response on M_w and distance. They are indeed essential metrics in the development of damage prediction, fragility and vulnerability models. On the other hand, the proposed heuristic model is a classification based model and these metrics cannot be calculated.

8. Lines 175-176: It is not clear how the DG is converted into a continuous variable for the regression ML models.

Authors' response: As stated in the manuscript, the method to convert DG into a continuous variable is given in a previous paper Ghimire et al., 2022. Regression models were considered with the damage grade as a continuous variable ranging between 1 (DG1) and 5 (DG5). Because the regression model outputs a real value between 1 and 5 and not a label, we rounded the output (real number) to the nearest integer to plot the confusion matrix. However, the error matrices were computed without rounding the model outputs to the nearest whole integer.

I think it would be beneficial for the reader to include a short paragraph briefly explaining how the DG is converted into a continuous variable.

10. Lines 195-196: Were the reported metrics throughout the manuscript obtained from the training or testing datasets? I believe it is important to clarify this.

Authors' response: Sorry we do not understand this question/comment. The metrics ADG and AT, and MAE and MSE are given in the section results and discussion. They are only presented here.

Please allow me to rephrase and elaborate my question. Did the authors use a testing dataset in each case to assess the models' architecture and hyperparameters values? If yes, the reported metrics throughout the manuscript corresponds to the performance of the ML models on the training or the testing dataset? If no, I believe it is fundamental that the authors provide the way of fine-tuning the models, and avoiding over- and under- fitting, without the use of testing or validation subsets.

11. Essential information is missing from the manuscript regarding the optimization of the hyperparameters presented in Table 3. How did the authors fine-tune the models? How was under- and over- fitting prevented? In particular, Random-forest and XGBC models are prone to overfitting.

Authors' response: The hyperparameters were tuned in the training dataset using cross validation method and the other hyperparameters not mentioned in Tab. 3 are the default parameters in the Scikit-learn documentation (Pedregosa et al., 2011).

The author should explain exactly which hyperparameters were tuned and how (i.e., under which objective). Did the authors use a k-fold cross validation? Why it is not mentioned anywhere in the manuscript?

Related to comment 11: Are the reported metrics obtained from the training dataset or calculated as the average from the cross validation datasets? This is the most important process in the training of ML models along with the data selection, so it cannot be missing from the manuscript. Finally, the default values of the hyperparameters in the Scikit-learn library are supposed to be tuned to optimize the performance of the ML models for a given problem in hand. Their impact is significant and cannot be neglected by any means. I believe this is an essential part that is completely missing from the manuscript.

12. Chapter 4.3.1: A very long discussion of the results is included, which the reader can interpret by observing the figures. However, the fact that a large number of predictions are underestimated is only mentioned in line 503 and it is not discussed. This finding needs to be elaborated and explained, as it may be related to comment 11.

Authors' response: The underestimation may be a consequence of the choice of the machine learning models, their implementation or the features considered. The interest of machine learning is also to have a relevant representation of the errors and limits of these methods. We will add this point in the revised manuscript.

"as certainly a consequence of the choice of ML models, their implementation (including imbalance issues), the distribution of input and target features considered, or all. The interest of machine learning model is also to have a relevant representation of the errors and limits of these methods."

I believe an important part of the discussion is to investigate this and it is missing from the manuscript. Possibly, the main reason for this underestimation is the under- or over- fitting of the heuristic model, given that the authors used the default hyperparameters values.

13. Lines 555-557 & 597-599: Based on the results and this conclusion, there is no benefit of employing XGBC over the traditional approach of RISK-UE. The authors should provide justification for this important finding and elaborate on the potential benefits of ML models over RISK-UE.

Authors' response: Machine learning frameworks can provide reasonable earthquake damage estimates using readily available features. machine learning allows us to change the paradigm by proposing a heuristic approach to damage prediction based on available data. This approach was already mentioned in Riedel et al. (2015). This is the major advantage of such methods, because defining the exposure model is a challenge for damage prediction with Risk-UE: this information may not be readily available, and collecting it could be too time-consuming and expensive. On the other hand, machine learning models can provide damage estimates using a more cost-effective way, as readily available data can be used to develop relationships between building features and damage. And finally, these models can help to discover new relationships, incorporate large amounts of data (e.g., global dynamic exposure models), and formally consider uncertainty. In addition,

This comment is addressed in the discussion section as: "The machine learning models achieved comparable accuracy to the Risk-UE method. In addition, TLS-based damage classification, using red for heavily damaged, yellow for moderate damage, and green for no to slight damage, could be appropriate when the information for undamaged buildings is unavailable during model training."

"Indeed, it is worth noting that the importance of the input features used in the learning process changes with the degree of damage: this indicates that each feature may have a contribution to the damage that changes with the damage level. Thus, the weight of each feature does not depend linearly on the degree of damage, which is not considered in conventional vulnerability methods.

Firstly, in order to propose a paradigm shift, strong evidence is required to suggest that a new proposed methodology should be adopted over the traditional ones. In general, I agree with the authors' view that ML models have already shown promising advances in the field of seismic hazard and risk. I also agree that such models can help to discover nonlinear patterns in large datasets and incorporate various sources of uncertainty. However, the findings of this study suggest otherwise; meaning that there is no benefit of employing the proposed heuristic model over the Risk-UE methodology.

Secondly, why is the heuristic model a more cost-effective way to estimate seismic damage and risk of building portfolios over the Risk-UE methodology? In terms of computational efficiency? Required input data? Once again, the "readily available" exposure data used as input parameters of the heuristic model are way more difficult to collect, curate and homogenize for building portfolios in comparison to the data required by other models, such as the European Seismic Exposure Model 2020.