The authors would like to thank the reviewers for the comments on this manuscript. We provide herein our response to the reviewer's comments, which have been taken into account for improving the quality of our manuscript.

In the following, we provide (in bold character) point-by-point answers to each of the Reviewers' comments, and indicate the related changes that we have made to the revised manuscript (in blue in the annotated version after second round of review and in red after the first round of review).

Reviewer 2:

Reviewer's Response: The paper can be published after the following minor change in reference to my comment #7: Please could you add this specification: "(with the exclusion of the 2016-2017 Central-Italy earthquake for which data processing is ongoing)"

Thank you for your comment. This comment is addressed in the manuscript as:

"The Database of Observed Damage (DaDO, Dolce et al., 2019) is accessible through a web-GIS platform and is designed to collect and share information about building features, seismic ground motions and observed damage following major earthquakes in Italy from 1976 to 2019 (with the exclusion of the 2016-2017 Central-Italy earthquake for which data processing is ongoing)."

Reviewer 3:

Thank you very much for your comments.

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?

Author's 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.

Reviewer's response: 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.

Author's response: We disagree with the reviewer on the abundance of data. Contrary to what the reviewer says, there is an abundance of data characterizing urban areas (not vulnerability nor exposure, just urban areas), whether from national censuses, cadastral databases, OpenStreetMap, etc. I don't see why this abundance of data should be opposed to national census and cadastral data. The proof of this is ESRM20, which has enabled a Europe-wide study to be carried out. Nevertheless, a careful reading of the Crowley et al. report reveals strong assumptions about some factors describing structures, associating them with capacity or fragility curves. These assumptions simply reflect the fact that, at this scale, some information are not available, and the information that is generally required to assess the seismic response of structures is not directly available. While this type of approach requires very precise (and often unavailable) information, or uses strong assumptions (e.g. the level of seismic design) to associate each structure with generic models of behaviour (see, for example, the extreme variability of these fragility curves), what we propose here is to start from the information available and see to what extent this information can predict damage. We can even imagine other factors that could be used to predict damage. It's even possible to imagine other urban factors (e.g. based on

GHS urban shape or other satellite imagery) that provide a large amount of data. It is then possible to test relationships (like Riedel et al) between these factors and predictions of vulnerability or seismic damage. This is the aim of this study, which we describe as fully as possible in the introduction.

The input parameters we test are all available in databases or even from satellite images etc...

Finally, our aim is to predict damage. If we find a relationship between building characteristics, ground motion (hazard) and damage, why go through vulnerability: it's the very principle of AI-based methods to explore relationships to be discovered. That's why we don't discuss vulnerability in this manuscript, as it's not the focus of the study. We then believe that this is the novelty of this study and the aim of developing AI-based methods for this field, as also suggested and tested by many others authors (see references).

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.

Author's 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.

Reviewer's response: 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.

Author's response : As indicated in the manuscrit, we consider these earthquakes to contain sufficient data for training (see chapter 4.3.2, in which we further reduce the data set for training and ultimately give the same accuracy metrics). These events, in our opinion, represent Italian earthquakes (region and magnitude range) and we could have tested the whole set of events but

this would not have changed the results since even a subset of the data provide maximum quality results.

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.

Author's 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.

Reviewer's response: First of all, it is a trivial task to calculate epicentral or source-to-site distance (e.g., JoynerBoore 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.

Author's response: The heuristic model is being tested on the building features available in the DaDO dataset, we do not pretend to think that this model can be deployed everywhere. Specific studies to test it on other regions will be carried out. In our dataset, we have data that we are testing (the principle of machine learning methods). We can conclude that the location (lat, long) of buildings is important, and we give an explanation for the fact that the position of a building in a city partly reflects its class (once again, this is the advantage of using ML-type methods), for example by distinguishing (indirectly) the old center from the external suburbs. The effect linked to the position of the building is certainly mainly due to the urban shape we've known for a long time: historic centers surrounded by more recent neighbourhoods, and therefore different buildings. So it's not the "position" directly that influences (we don't understand the link with rural/urban??) but the position in the urban shape, which certainly and indirectly reflect the basic feature of the buildings. This is what we explain in the manuscript to interpret the importance of these features.

Yes, we totally agree, there are uncertainties! Macroseismic intensity is a quantitative estimate of ground motion (according to EMS98) considering the damage. Our study does not pretend to solve all the questions, but provides elements of understanding according to the data that we have, by analyzing to what extent the precision of the results is satisfactory or not. Because we consider the microseismic intensity, sure these microseismic intensity reflect the depth, distance and magnitude of the earthquake. But we did not add these parameters to the machine.

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.

Author's 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."

Reviewer's response: 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.

Author's response: The distribution of number of stories in NF1 is 11.01%, 28.11%, 40.01% for 1, 2, and 3 storeys, respectively, in the 2009-L'Aquila earthquake building damage dataset (which was used to study features importance). Our conclusion relates to this dataset. We have never claimed that our conclusions are valid whatever the dataset. We are not saying that these parameters are not important: we are saying that they are not the most important ones in our heuristic model using our dataset and using the

process/tools to test the important features. Hence the added sentence in the previous version.

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.

Author's 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.

Reviewer's response: 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.

Author's response: 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. We are not considering the spatial correlation of the IM.

As explained, we didn't test other parameters, as it makes much more sense to use intensity maps for this kind of study deriving empirical or heuristic models (see paper by Wald et al. for example). We could debate the performance of the PGA and the acceleration response spectra on the effectiveness of these parameters compared to others but this point is not the main focus of this study and could be tested in further study.

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?

Author's 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.

Reviewer's response: 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.

Author's response: In our reply, the fact that we want to treat Mw and distance as parameters is linked to the desire to see their impact and therefore their independence or conditional dependence on the response of the structures (i.e., sufficiency in the sense proposed by Luco and Cornell). This point is absolutely not the aim of this article. We could also look at all the source parameters, all the site condition parameters, or even the strain level of each region, but no, we haven't done that because of the missing data in most of the case. We believe that these points are out of the scope of this study

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

Author's 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.

Reviewer's response: 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.

Author's response: In the new version, we add:

For the regression-based machine learning models, DG is converted into a continuous variable as tested by Ghimire et al. (2022): first, the damage grades were ordered and considered as a continuous variable ranging between 0 (DG0) and 5 (DG5). Because the regression model outputs a real value between 0 and 5 and not an integer, 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 integer.

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.

Author's 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.

Reviewer's response: 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.

Author's response: Sorry but we are not sure to understand the question but we try to provide addition elements. The performance metrics given are about testing. Some papers in the references focused only on testing performance metrics, few on training but really without providing comparison. In Ghimire et al. 2022, the performance metrics were also

computed for validation and test sets with the same values. We decided that this is not a key issue to deal with at this stage of the study presented here.

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.

Author's 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).

Reviewer's response: 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.

Author's response: We use k-fold cross-validation (with 10-fold) (a common process for hyperparameter tuning (for hyperparameters provided in Table 3, which are recognized to be the most important parameters in literatures for these machine learning methods, very classical method). So, by this way, from training set, we randomly select a % for training and % for testing, for different combination of hyperparameters (through randomized grid search) and we select the optimum evaluated in terms of performance metrics on testing. The values of hyperparameters are provided in Table 3, they are very similar to what observed in Ghimire et al. (2022).

We addressed this comment in the manuscript as:

"The hyperparameters indicated in Tab. 3 were chosen after tests performed by Ghimire et al. (2022)."

"Table 3. Summary of optimized hyperparameters parameters, accuracy A_T and quantitative statistical error values for the regression-based and classification-based machine learning methods in the test set. The parameters are the hyperparameters chosen for the machine learning models (the other model parameters not mentioned here are the default parameters in the Scikit-learn documentation (Pedregosa et al., 2011)). The best accuracy and error values are indicated in bold. The optimum hyperparameters were selected thanks to k-fold cross-validation (with 10-fold), by randomly select a % for training and % for testing, for different combination of hyperparameters and the optimum evaluated in terms of performance metrics on testing is finally selected."

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.

Author's 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."

Reviewer's response: 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.

Author's response: We do discuss this as suggested by the reviewer. This study does not pretend to answer all question but provide some information on the use of ML for damage assessment. We suggest a few reasons to interpret over-fitting or under-fitting but at this stage of the study, we have no other interpretations. We have already tested several ML methods, several solutions to solve imbalance data, we choose several metrics to evaluate the performance and selected the most optimal hyper-parameters. Other reasons can be explored but we don't have the information for that.

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.

Author's 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.

Reviewer's response: 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.

Author's response: As pointed by the reviewer throughout his review, we are not dealing with vulnerability because our aim is to estimate damage from features, without going through vulnerability thanks to AI-based methods. For this reason, we believe that this is a new (or at least a different) paradigm.

With basic features and easily available (see comment 1), the accuracy of the damage prediction is the same as Risk-UE model: so we really consider that the benefit is real and effective.

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.

Author's response: In ESRM2020, the data used to calculate the risk was available (I don't believe the data was specifically collected by the group in charge of ESRM20) and is thus named already available, the difference being that we don't use vulnerability functions. The data used herein are the available data, without the need to make all the assumptions such as those mentioned by the reviewer, or to go and collect other information for the application of the original Risk-UE method. For this reason, we insist

on the fact that even if the features are basic (really basic) and the damage scale also simplified (green-yellow-red), the accuracy of the results is not better nor worse than by other methods. Testing will be necessary, but this is a new (or at least a different) paradigm that we are exploring.