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 red in the annotated version).

Reviewer 1

The overall quality of the preprint (general comments)
The overall quality of the preprint is good. The topic of testing machine learning models for heuristic building damage assessment is significant for the science community. The research is well structured and explained. The authors made an effort to test a large number of different experiment scenarios. There are minor deficiencies in the paper, mostly related to the nature of the dataset, not the method.

Thank you for your positive comments.

(1) Not so long ago, this paper would probably deserve to be published as is (with some technical corrections). But, since the research shows similar methods and results compared to previously published papers, this reviewer believes that the authors should make an additional effort to demonstrate the added value of this research to the body of knowledge. There is not much more to do regarding experiments, but adding a Discussion chapter is an opportunity to improve the paper.

Thank you very much for this comment. We have added a discussion chapter in the manuscript highlighting the added value of this study.

"Previous studies have aimed to test a machine learning framework for seismic building damage assessment (e.g., Mangalathu et al., 2020; Roeslin et al., 2020; Harirchan et al., 2021; Ghimire et al., 2022). They evaluated various machine learning and data balancing methods to classify earthquake damage to buildings. However, these studies (Mangalathu et al., 2020, Roeslin et al., 2020, Harirchan et al., 2021) had limitations such as limited data samples, damage classes, and building characteristics limited to a spatial coverage and range of seismic demand values. Ghimire et al. (2022) also used a larger building damage database, but did not investigate the importance of input features as a function of damage levels and did not compare machine learning with conventional damage assessment methods.
Our study aims to go beyond previous studies by testing advanced machine learning methods and data resampling techniques using the unique DaDO dataset collected from several major earthquakes in Italy. This database covers a wide range of seismic damage and seismic demands of a specific region, including undamaged buildings. Most importantly, this study reveals the importance of input features according to the degrees of damage and finally compares the machine learning models with a classical damage prediction model (Risk-UE)."

Individual scientific questions/issues (specific comments)

(2) It appears that most ML-based articles have trouble with inconsistent datasets and imbalanced recorded damage distributions, obtaining similar results which are scientifically acceptable but not impressive. The Discussion chapter cannot solve ML-related issues in earthquake damage assessment, but the authors can present their views on limitations, opportunities, advances, future work or the way forward based on their findings.

Advances:
The contribution of this paper is not only to reproduce what we know but also to highlight the importance of the input feature changes with the damage grade: this is new!

This point is added in the discussion section as:
" 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.
"

Limitations:
The prediction of seismic damage by machine learning remains until now tested on geographically limited data. The damage distribution is strongly influenced by region-specific factors such as construction quality and regional typologies, implementation of seismic regulations and hazard level. Therefore, machine learning-based models can only work well in regions with comparable characteristics and a host-to-target transfer of these models should be studied.

These points are added in the discussion section as:
" The prediction of seismic damage by machine learning remains until now tested on geographically limited data. The damage distribution is strongly influenced by region-specific factors such as construction quality and regional typologies, implementation of seismic regulations and hazard level. Therefore, machine learning-based models can only work well in regions with comparable characteristics and a host-to-target transfer of these models should be studied.
However, integrating data from a wider range of input features and earthquake damage from different regions, relying on a host-to-target strategy, could help achieve a more natural balance of data sets and lead to less biased results."

Opportunities:
In recent years, there has been a proliferation of open building data, such as the OpenStreetMap-based dynamic global exposure model and building damage dataset after an earthquake (such as DaDO). We must therefore continue this paradigm shift initiated in 2015 by Riedel et al. which consisted in identifying the exposure data available and as certain as possible, and in finding the most effective relationships for estimating the damage, unlike conventional approaches which proposed established and robust methods but relying on data not available and therefore difficult to collect. The global dynamic exposure model will make it possible to meet the challenge of modelling exposure on a larger scale on available data, using a tool capable of integrating this large volume of data. Machine learning methods are one such rapidly growing tool that can aid in exposure classification and damage prediction by leveraging readily available information. It is therefore necessary to continue in this direction in order to evaluate the

performance of the methods and their pros and cons for maximum efficacy of the prediction of damage.
This point is added in the discussion section.

Future work:
These points are added in the discussion section as:
"Future work will therefore have to address several key issues that have been discussed here but that need to be further investigated. For example, the weight of the input features varies according to the level of damage, but one can question the systematization of this observation whatever the dataset and the feature considered. The efficiency of the selected models and the management of imbalance data remain to be explored, in particular by verifying regional independence. Taking advantage of the increasing abundance of exposure data and post-seismic observations, the balance of input data and observed damage levels could be solved by aggregating datasets independent of the exposure and hazard contexts of the regions, once the host-to-target transfer of the models has been resolved."

Here are some topics which the authors could discuss in the additional chapter:

(3) Highlight the differences and stated advances - the use of oversampling and the conditioned importance of structural features related to damage states. What is the implication? The abstract and conclusion lack some numerical research highlights.

Some numerical research highlights will be added in the abstract and conclusion.

Oversampling methods can penalize imbalanced feature distribution but may introduce bias in damage prediction from overfitting.
Conditioned structural feature importance in damage states suggests that damage is not a linear combination of features but depends on damage grades.

This comment is added in the discussion section as:
"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.

In addition, the distribution of damage is often imbalanced, impacting the performance of machine learning models by assigning higher weights to the features of the majority class. However, data balancing methods like random oversampling can reduce bias caused by imbalanced data during the training phase, but they may also introduce overfitting issues depending on the distribution of input and target features. Thus, integrating data from a wider range of input features and earthquake damage from different regions, relying on a host-to-target strategy, could help achieve a more natural balance of data sets and lead to less biased results."

(4) How do the authors explain obtaining similar results for significantly different combinations of methods, set sizes, features and target classes? Why do we all get similar results? What is the way forward? Should we dismiss machine learning, or should we improve something? Are we possibly missing a key feature to include in post-earthquake surveys? Which one?

We are obtaining similar results for different combinations of methods, set sizes, features, and target classes because in seismic risk analysis the most uncertain component is given by hazard not by damage related to vulnerability. Without deeper analysis, it is not possible to conclude definitely on this, but post-earthquake observation always revealed the same trends between damage and features, while ground motion, fault rupture, slip duration, etc. question always seismologists! It is definitively not a scientific approach but our personal findings.
In addition, there is also a possibility that some key features characterizing hazard or vulnerability (not yet identified) are overlooked. Instead of dismissing machine learning methods, further investigation should be conducted to explore these issues in greater detail.

This comment is addressed in the discussion section as:
"The machine learning model trained and tested on the DaDO dataset resulted in similar damage prediction accuracy values reported in existing literature using different models and datasets with different combinations of input features, which might suggest that the uncertainty related to building vulnerability in damage classification may be small, while the primary source of uncertainty may be from the hazard part (such as ground motion, fault rupture, slip duration, etc.).."
"Finally, key input features (still not yet identified) describing hazard or vulnerability may be unexplored, and incorporating them into the models may improve the accuracy of damage classification."

(5) The authors had the unique advantage of using multiple earthquake datasets from the same region. That is maybe the key difference compared to other papers – the dataset covered the whole MSI range while single earthquakes provide only a fraction. And yet, similar results were obtained? Why?

We obtained comparable damage prediction results between datasets covering a wide range of MSI values and those including only a subset of MSI values following the data balancing techniques. This may be because the wider MSI value dataset provides more naturally balanced training data for the model to learn and penalizes the skewed distribution of target features.

This comment is addressed in the discussion section as:
"In addition, the distribution of damage is often imbalanced, impacting the performance of machine learning models by assigning higher weights to the features of the majority class. However, data balancing methods like random oversampling can reduce bias caused by imbalanced data during the training phase, but they may also introduce overfitting issues depending on the distribution of input and target features. Thus, integrating data from a wider range of input features and earthquake damage from

different regions, relying on a host-to-target strategy, could help achieve a more natural balance of data sets and lead to less biased results.”

(6) How do the authors evaluate the usefulness of the research and model implementation for new earthquakes? How to implement the model without the class of undamaged buildings? Without them, what will the model tell local authorities - that all buildings are damaged? Why should they use machine learning and not the traditional Risk-UE method? What are the benefits?

Machine learning frameworks can provide reasonable earthquake damage estimates using readily available features. We can either use models trained on post-earthquake datasets in regions with similar design and hazard characteristics to estimate potential damage during future earthquakes. We can collect samples from future earthquake damage and use a representative sampling framework proposed by Stojadinović et al. (2021) to train models for damage predictions.

If the dataset lacks undamaged buildings, traffic-light-based damage classification using machine learning models could be a solution because heavily damaged buildings can be classified as red, moderately damaged buildings as yellow, and non to slightly damaged buildings as green.

On the one hand, 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 using a 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).

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.”

(7) How about transferability? The authors should explain the sentence addressing transferability (lines 599-601) in more detail. Obviously, there are differences between regions regarding code implementation or the human impact on construction quality (Turkey earthquake).

As explained above, machine learning models for predicting earthquake damage are based on limited data from specific regions. We know that the damage distribution is significantly influenced by factors such as construction quality, typologies, implementation of seismic regulations, and seismic hazards. Thus, these models may only work well in regions with similar characteristics.

This comment is addressed in the discussion section as:
"The prediction of seismic damage by machine learning remains until now tested on geographically limited data. The damage distribution is strongly influenced by region-specific factors such as construction quality and regional typologies, implementation of seismic regulations and hazard level. Therefore, machine learning-based models can only work well in regions with comparable characteristics and a host-to-target transfer of these models should be studied. In addition, the distribution of damage is often imbalanced, impacting the performance of machine learning models by assigning higher weights to the features of the majority class. However, data balancing methods like random oversampling can reduce bias caused by imbalanced data during the training phase, but they may also introduce overfitting issues depending on the distribution of input and target features. Thus, integrating data from a wider range of input features and earthquake damage from different regions, relying on a host-to-target strategy, could help achieve a more natural balance of data sets and lead to less biased results. "

Technical corrections
(1) There is a need for some technical corrections, highlighted in the attached file. The authors should carefully check the paper for unnecessary long phrases, missing articles or spelling. For example:
Line 6 needs better wording regarding "six models" (possibly - six models were considered: regression- and classification-based machine learning models, each using random forest, gradient boosting and extreme gradient boosting).

Thank you for the comments they are taken into account in the revised manuscript.
"Six models were considered: regression- and classification-based machine learning models, each using random forest, gradient boosting and extreme gradient boosting."

(2) There seems to be some redundant text in Lines 173-179.
In fact, we were explaining how the damage grades were used for classification (works on discrete labels) and regression (works on continuous values) machine learning methods.
"To develop the heuristic damage assessment model, the damage grades are considered as the target feature. The damage grades are discrete labels, from DG0 to DG5. Three most advanced classification and regression machine learning algorithms were selected: random forest (RFC) and regression (RFR) (Breiman, 2001), gradient boosting classification (GBC) and regression (GBR) (Friedman, 1999), and extreme gradient boosting classification (XGBC) and regression (XGBR) (Chen and Guestrin, 2016)."
==================================

Reviewer 2

General comments
The article describes the application of machine learning models to a heuristic method for post-earthquake damage assessment, applied to observed damage data after seismic events that affected Italy. The topic is interesting and worthy of investigation but in my opinion there are some points that need to be clarified before its publication. The tables and figures are clear and complete. The writing is fluent. The bibliography is extensive, there are only a few corrections to be made mentioned later.

Thank you very much for your positive comments.

Specific comments
The most significant problem in my opinion is not taking into account the uninspected buildings that would increase the number of buildings that reach the DG0 damage level. Without taking these buildings into account, there is a non-real distribution of damage that is amplified by the application of the "random oversampling" method. It is also not clear to me how this oversampling method is applied; it should be explained in more detail. Is it artificially "adding" buildings that reach the damage levels above DG0 in order to have the same number of buildings reaching the different damage levels? If so, the information regarding the percentage of buildings of a typology that reach a specific damage level is lost. This point needs to be clarified.

Thank you very much for the comment, here is our response, which was also included in the manuscript between line numbers 129 to 134.
The data on building damage from earthquake survey other than Irpinia earthquake damage survey mostly includes damaged buildings. This is because the data was collected based on requests for damage assessments after the earthquake event (Dolce et al. 2019). The damage information in DaDO database is still relevant for testing the machine learning models for heuristic damage assessment. Mixing these datasets to train machine learning models can lead to biased outcomes. Therefore, the machine learning methods were developed on the other earthquake's dataset excluding Irpinia dataset, and the Irpinia earthquake dataset was used only in the testing phase. Thus, the objective is to test machine learning for predicting damage according to the building's features and intensity measures, and the missing data is not a concern of this study.

The random oversampling method is a classical method to balance the data in the training set to reduce the bias due to the overrepresented number of a class of target features. This does not correspond to an artificially "adding" buildings way. In fact, all data points from majority and minority training sets are used. Additionally, instances are randomly picked, with replacement, from the minority training set till the desired balance is achieved, adding the same minority samples might result overfitting, thereby reducing the generalization ability of the classifier (Dubey et al., 2014). In the end, machine learning represents the "relation" between feature and target, and the test is to assess the target knowing input features. So, by this oversampling approach, the information regarding the percentage of buildings of a typology that reach a specific damage level is not lost.

Other observations for consideration are given below:

1) line 45-47: it says that the damage is given by the combination of seismic hazard, exposure and vulnerability/fragility but it does not explain what these three elements are

Seismic hazard, exposure, and vulnerability/fragility are well-known components of risk analysis. For the audience of this journal (NHEES), we believe that providing further information on these concepts is not necessary: we introduced them as general information only, and the main aim of this study is on machine learning methods for damage assessment.

2) line 48-49: the phrase from "For" to "scenario" is useless unless explained

For risk analysis, several options are available to characterize the hazard: frequency-based, intensity-based or scenario-based (see for example definition in Crowley et al. ESRM20/EFEHR 2020 reference). The scenario-based risk assessment consists in defining a scenario earthquake in terms of magnitude, location, fault etc. and to predict the associated ground motion.

We added explanation in line 49: For scenario-based risk assessment, damage and related consequences are computed for a single earthquake defined in terms of magnitude, location, and other seismological features.

3) vulnerability/fragility is mentioned but the difference between these two elements is not explained

Vulnerability and fragility functions are quite classical for the audience of NHESS's special issue on seismic hazard and risk, and they are not a critical aspect of this study. They are two ways to represent the expected (probable) damage for a given intensity measure of the ground motion.

4) line 60-61: "superior computational efficiency, easy handling of complex problems, and the incorporation of uncertainties" regarding the use of artificial intelligence applied to seismic risk assessment. These are strong statements that should be justified or reported in the conclusions with appropriate explanation

This statement is justified by the list of references (previous studies) published and listed in the manuscript: we do not state this, we refer to publications that state this in lines 62-65.

5) line 71: DaDO database is cited without saying it is Italian data (it is said later but it would be appropriate to say it here too)

We addressed this comment in the manuscript by adding "in Italy" after the Database of Observed Damage (DaDO).

“With more than 10,000 samples compiled, the Database of Observed Damage (DaDO) in Italy, platform of the Civil Protection Department, developed by the Eucentre Foundation (Dolce et al., 2019),”

6) line 82-82: sentence from "By" to "assessment" is unclear, explain better

According to psychology, the heuristic technique is any approach to solving a problem that employs a practical method that is not guaranteed to be optimal, perfect, or rational but is nevertheless sufficient for reaching an immediate, short-term goal or approximation. Where finding an optimal solution is impossible or impractical, heuristic methods can be used to speed up the process of finding a satisfactory solution. Heuristics can be mental shortcuts that ease the cognitive load of making a decision. In our case, heuristics have been proposed to explain how people can make decisions, come to judgments, and solve problems. These rules typically come into play when people face complex problems or incomplete information and reduce the stress for making decisions when decision-makers face uncertainties. The cognitive load related to making a decision is then reduced when uncertainties are explicitly considered. For example, in machine learning: the result is not certain, and the data are incomplete but the heuristic model provides the best results with a clear representation of the uncertainties.

We have modulated this sentence in the manuscript as:
“By analogy in psychology, this procedure can reduce the cognitive load associated with uncertainties when making decisions based on damage assessment, by explicitly considering the uncertainties in the assessment, being aware about the incompleteness of the information and the accuracy level to make a decision.”

7) line 91: it says that DaDO has observed damage data for major earthquakes in Italy. Specify the time range of data collected as there have been other earthquakes in Italy for which there are no data in DaDO for different reasons (very old earthquakes for which data were not being collected and more recent such as the 2016-2017 earthquake for which data are being processed).

The time range (1976-2019) of data collected is added 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.”

8) line 103: specify that the scale from DG0 to DG5 is EMS98

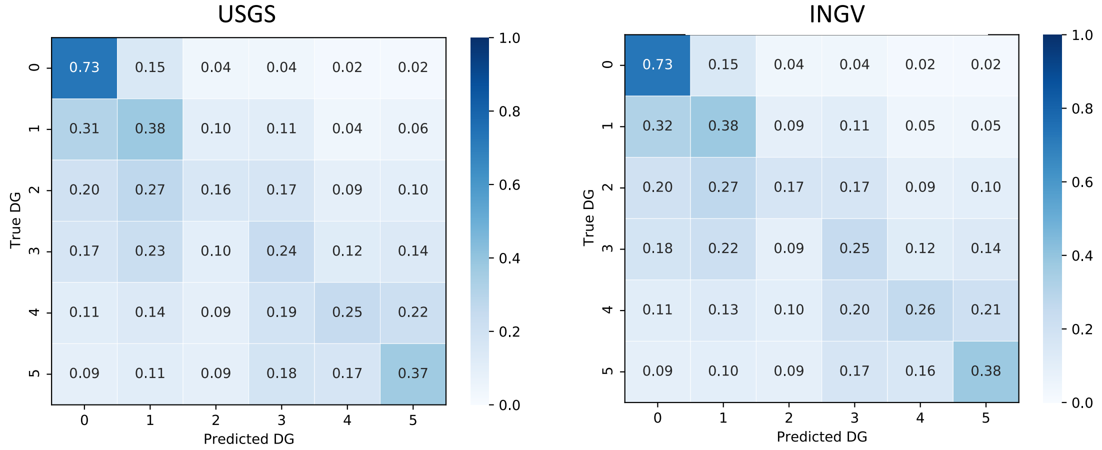
EMS-98 is added in the manuscript as:
“The converted EMS-98 damage grade (DG) ranges from damage grade DG0 (no damage) to DG5 (total collapse).”

9) line 127-129: mention that there are mostly damaged buildings in the database but do not explain how take into account this element i.e., the fact that the buildings in the database are not all those in the municipalities considered, there are also the buildings that have not been inspected that we can assume have zero damage.

We learn the relation between the input feature and the target values in training phase to develop machine learning models. Once this relation is defined, we test on the rest of the dataset: the question developed here is not to assess the whole amount of buildings but rather to develop a machine that (for example) could be applied to the whole number of buildings. We answered this question in the first comment.

10) line 142-144: it is said that in DaDO there are MSI values provided by USGS ShakeMap tool but the intensity values in DaDO are MCS and were calculated by INGV (National Institute of Geophysics and Volcanology). Also, are the intensities coming from the "ShakeDaDO" database being considered? If yes, cite this database with the correct reference. If no, it might be useful to consider it (Faenza L., Michelini A., Crowley H., Borzi B., Faravelli M (2020) ShakeDaDO: A data collection combining earthquake building damage and ShakeMap parameters for Italy, Artificial Intelligence in Geosciences, Volume 1, 2020, Pages 36-51, doi.org/10.1016/j.aiig.2021.01.002.)

In this study, the data was selected from the DaDO database because it contains more comprehensive information on buildings compared to ShakeDaDO (e.g., only two types of structural materials were present in ShakeDaDO: Ca and Mu). We selected MSI values from the USGS ShakeMap to the DaDO database to test a machine learning-based damage prediction framework that can be applied globally. However, we also conducted a test using intensity values from both INGV and USGS and observed that both intensities yielded similar results in damage prediction (shown in below figure).



11) line 297: for completeness, I suggest to mention the value of ADG for DG3

Value of A_{DG} for DG3 is added in the manuscript.
"In the confusion matrix (Fig. 3d: RFC, Fig. 3e: GBC, and Fig. 3f: XGBC), the accuracy A_{DG} values also show higher model efficacy for the lower DGs (86% for DG0 and 39% for DG1) and lower efficacy for the higher DGs (5%, 23%, 12% and 17% buildings correctly classified in DG2, DG3, DG4 and DG5, respectively)."

12) lines 327-335: explain more about these 4 methods

These are common methods in machine learning. We remind key concepts in our manuscript to avoid redundancy with other publications referenced in our manuscript.

We hope that the reviewer will agree with this way that consists in referencing previous works and limiting thus the number of pages of this manuscript.

13) lines 371-375: it is not clear the sentence from "Notes" to "areas"

We want to say that incorporating the latitude and longitude of a building as input features in a machine learning model may indirectly account for the impact of local geology in building response and the spatial distribution of vulnerability (buildings in the old town are relatively vulnerable than those in new urban areas).

We change this sentence in the manuscript.
"Note also that the importance score associated with the location feature can indirectly capture variations in local geological properties and the spatially distributed vulnerability associated with the built-up area of the L'Aquila-2009 portfolio (e.g., the distinction between the historic town and more modern urban areas)."

14) traffic-light system: the introduction of this method for comparison in my opinion is not very significant. It is said that it was used during post-earthquake emergency situations but in my experience in Italy it is not. Instead of this traffic-light system, I suggest to consider the damage levels as they are directly present in the Aedes forms i.e. DG0, DG1, DG2+DG3, DG4+DG5, in addition to the five damage grades of the EMS98.

Right, in Italy, damage was classified using the Aedes form into four categories (DG0, DG1, DG2+DG3, DG4+DG5). We used a traffic-light system based on severity damage (DG0+DG1, DG2+DG3, DG4+DG5) in order to regroup DG by severity, i.e. in the same spirit as for emergency classification, with the idea to be able to classify rapidly the severity of damage using ML. As observed in our result, TLS-based scale provides a better classification than considering EMS98 scale, because of a lot of mis-classification (even in the field) between two consecutive DG (mentioned lines 414-420) as:

"The efficacy of the heuristic damage assessment model using TLS-based damage classification indicates that classifying damage into three classes is much easier for the machine compared with the six-class classification system (EMS-98 damage classification). This is also observed during damage surveys in the field, which sometimes find it hard to distinguish the intermediate damage grades, such as DG2 and DG3, or DG3 and DG4."

15) Explain better how this machine learning method can be used in the post-event phase. Could a recorded value such as PGA be used for hazard instead of macroseismic intensity?

Yes, it is possible to use other intensity measure for ground motion (such as PGA) but as suggested by Cua et al. (2010), macroseismic intensity represent the spatially-distributed ground motion and is more effective in communicating ground motion levels in relation to human experiences and incurred losses.
In the post-event phase, machine learning is not useful for damage assessment building-by-building on the field!

Machine learning frameworks can provide reasonable earthquake damage estimates using readily available features at regional scale. We can either use models trained on post-earthquake datasets in regions with similar construction and hazard characteristics

to estimate potential damage during future earthquakes. We can also collect samples from future earthquake damage and use a representative sampling framework proposed by Stojadinović et al. (2021) to train models for damage predictions.

This comment is addressed in the discussion section as:
Most importantly, this study highlights the importance of input features according to the degrees of damage and finally compares the machine learning models with a classical damage prediction model (Risk-UE). 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.
In recent years, there has been a proliferation of open building data, such as the OpenStreetMap-based dynamic global exposure model (Schorlemmer et al., 2020) and building damage dataset after an earthquake (such as DaDO). We must therefore continue this paradigm shift initiated by Riedel et al. (2014, 2015) which consisted in identifying the exposure data available and as certain as possible, and in finding the most effective relationships for estimating the damage, unlike conventional approaches which proposed established and robust methods but relying on data not available and therefore difficult to collect. The global dynamic exposure model will make it possible to meet the challenge of modelling exposure on a larger scale on available data, using a tool capable of integrating this large volume of data. Machine learning methods are one such rapidly growing tool that can aid in exposure classification and damage prediction by leveraging readily available information. It is therefore necessary to continue in this direction in order to evaluate the performance of the methods and their pros and cons for maximum efficacy of the prediction of damage.
”

Technical corrections

Below my observations:

1) line 70: in the text there are these abbreviations not found in the bibliography: MINVU, 2021; MTPTC, 2010; NPC, 2015
Addressed.

2) line 287: AT for GBR is 0.50 but in the text it is listed as 0.49.
Addressed.

3) line 288 and 294: Fig. 2 is mentioned instead it is Fig. 3.
Addressed.

4) Reference: there are 3 papers absent in the text: Ghimire et al. 2021, Riedel and Gueguen 2018, Seo et al. 2012.

Addressed.

Thank you for your interesting remarks and comments.

================================

Reviewer 3

General comments:

The article explores the use of advanced machine learning algorithms for post-earthquake heuristic damage assessment of buildings using a subset of the Italian DaDO database. The topic is very interesting and, in my opinion, quite important for the improvement of existing methodologies in earthquake scenario simulations and seismic risk analysis of building portfolios. The authors considered an extensive literature and, overall, the research is well presented and the writing is good, although I think some parts are unnecessary long.

Thank you for these positive remarks

Undoubtedly, the manuscript includes a significant amount of work related to the training and evaluation of the ML models using an innovative approach to tackle known issues in the development of ML models for damage assessment. However, I believe that discussion is missing in several key topics and the authors should consider a few additional aspects.

Specific 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?

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.

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.

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.

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.

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.

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 so 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.

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."

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.

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.

6. Lines 129-131: A justification is missing regarding why the inclusion of the Irpinia-1980 dataset in the training can lead to biased outcomes. Why is it only relevant for testing the models?

As stated in the manuscript and directly related to the imbalanced issue: The Irpinia-1980 building damage portfolio was constructed using the specific Irpinia-1980 damage survey form, while the AeDES damage survey form was used for the others. The Irpinia-1980 dataset will therefore be analysed separately. The data on building damage from earthquake surveys other than Irpinia earthquake damage survey mostly includes damaged buildings. This is because the data was collected based on requests for damage assessments after the earthquake event (Dolce et al. 2019). The damage information in the DaDO database is still relevant for testing the machine learning models for heuristic damage assessment. Mixing these datasets to train machine learning models can lead to biased outcomes. Therefore, the machine learning methods were developed on the other earthquake dataset excluding the Irpinia dataset, and the Irpinia earthquake dataset was used only in the testing phase.

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?

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.

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

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.

9. Lines 184-185: This is not true. It entirely depends on the ML algorithm and the training process. For example, in the case of artificial neural networks, the eigenvalues of the training dataset used by some common optimization algorithms have a considerable impact.

As said in this manuscript, the presence of correlated features does not impact the overall performance of these machine learning methods, as tested in Ghimire et al., 2022. For that reason, no specific data cleaning methods were applied to the DaDO database. Related to neural network, we have no idea and we did not test these methods.

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.

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.

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.

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).

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.

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."

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-EU. The authors should provide justification for this important finding and elaborate on the potential benefits of ML models over RISK-EU.

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.

14. Lines 580-581: The XGBC model is not optimal, it just performed slightly better than the other models.

Ok we modify the sentence by "the most efficient model for this dataset" because it is not only slightly when looking at the results (errors and uncertainties).

15. Lines 599-600: From which results did the authors draw this conclusion? Do similar building portfolios refer to primarily one-storey buildings?

No, this is related to the Risk-UE model comparison, since Risk-UE was also developed for a given port-folio and applied to others, assuming the same buildings characteristics. For EMS98, Risk-UE, GNDT, FEMA etc… methods, never the host-to-target adjustment is discussed but the same question is. In our study, we mention conditionally that machine learning could provide a reliable estimate, but supposing the same portfolio, and also the same earthquake characteristics for example, since we observe that the importance of the features change with the DG.

We smooth the sentence by: "…building portfolios, after host-to-target adjustment."

A new section Discussion will be added, mentioned this point.

"The prediction of seismic damage by machine learning remains until now tested on geographically limited data. The damage distribution is strongly influenced by region-specific factors such as construction quality and regional typologies, implementation of seismic regulations and hazard level. Therefore, machine learning-based models can only work well in regions with comparable characteristics and a host-to-target transfer of these models should be studied."

Minor edits:

1. Line 39: A reference is missing for this interesting information.

We add Silva et al. 2019

2. Line 46: What do the authors mean by necessary damage?

To assess, we need damage assessment. But "necessary" is not necessary. We remove this word here.

3. Line 127: This sentence does not read well; I suggest to rephrase it.

Difficult to rephrase it because it means what it means. A tentative: "Building damage data from earthquake surveys other than the Irpinia earthquake damage survey primarily include damaged buildings."

4. Line 132: Replace "methods" with models and "earthquake's" with earthquakes'.

done

5. Lines 177-178: This sentence is a repetition of the one in lines 173-175. I suggest to just mention that the same ML algorithms were used for regression and classification.

The new paragraph "To develop the heuristic damage assessment model, the damage grades are considered as the target feature. The damage grades are discrete labels, from DG0 to DG5. Three most advanced classification and regression machine learning algorithms were selected: random forest (RFC) and regression (RFR) (Breiman, 2001), gradient boosting classification (GBC) and regression (GBR) (Friedman, 1999), and extreme gradient boosting classification (XGBC) and regression (XGBR) (Chen and Guestrin, 2016). A label (or class) was thus assigned to the categorical response variables (DG) for the classification-based machine learning models. For the regression-based machine learning models, DG is converted into a continuous variable to minimize misclassifications (Ghimire et al., 2022)."

6. Lines 208-209: MAE and MSE are acronyms. Replace the words "average" with mean.

Right! We replace.

7. Line 299: Replace "Summary of optimized input parameters" with Summary of optimized hyperparameters. The term input parameters refers to the input variables (e.g., MSI, Building age, etc.).

Totally Right! We replace.

8. Line 407 & 591: Replace "machine" with "machine learning model".

Done

Citation: https://doi.org/10.5194/nhess-2023-7-RC3

Thank you very much for your review.