ANNALS OF GEOPHYSICS, 66, 5, 2023; doi:10.4401/ag-8935
OPEN ACCESS

Random Forest based estimate to assess the damages of future earthquakes: preliminary results

Federica Di Michele[*,1], Enrico Stagnini[2], Donato Pera[3], Roberto Aloisio[2,4], Pierangelo Marcati[2]

(1) Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Milano, Milano, Italy
(2) Gran Sasso Science Institute (GSSI), L'Aquila, Italy
(3) Department of Information Engineering, Computer Science and Mathematics, University of L'Aquila, Coppito, L'Aquila (Italy)
(4) INFN-Laboratori Nazionali del Gran Sasso, Assergi (AQ), Italy

Article history: received December 30, 2022; accepted February 15, 2024

Abstract

In this paper we present a case study where the Random Forest (RF) Classifier, has been used to estimate the damage to buildings caused by a (possible) future earthquake, starting from the data of past earthquakes. This preliminary work is based on the Shakedado dataset, containing information on buildings and ground shaking parameters for the six major earthquakes that occurred in Italy between 1981 and 2012. We perform the following two conceptual experiments:

E1. Assume that Emilia seismic sequence has just ended and the data from the other major earthquakes that have occurred in the past (L'Aquila, Pollino and Irpinia) are available. We calculate the damage level for each building in the Emilia dataset.

E2. Assume that the Pollino seismic sequence has just ended and the data from the other major earthquakes with comparable magnitude (L'Aquila, Emilia) are available. We calculate the damage level for each building in the Pollino dataset.

Both training and test datasets contain only masonry buildings located within 10 km of the main shock of each sequence. The results demonstrate the ability of the RF algorithm to discriminate between light/no and medium/severe damaged building, with a good accuracy especially for E1.

Keywords: Earthquake; Artificial Intelligence; Random Forest; Building Damages; Risk Mitigation

1. Introduction

Italy is one of the European countries with the highest level of siesmic risk, in fact, in the last 50 years it has been hit by many strong earthquakes, which have caused victims, but also enormous economic and social damage. Earthquakes are known to be unpredictable, and prevention is currently the only way to reduce their impact on infrastructures, buildings, and more in general, on human life. Therefore, seismic risk mitigation requires an interdisciplinary approach that necessarily starts with an accurate knowledge of the territory and its geological and architectural characteristics. Recent decades have seen a sharp increase in the use of artificial intelligence (AI)

techniques for both disaster assessment and post-disaster management. One of the first applications of a ML-supervised algorithm to classify building damages has been reported in [Mangalathu et al., 2020], where the authors analyzed the damage pattern after the 6.0 Mw earthquake that struck the cities of Napa, American Canyon, and Vallejo in August 2014. The working dataset contained 2276 buildings grouped into severe, moderate, and light damage levels. Four different classifiers were used to assess the damage, K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA), Decision Trees (DT), and Random Forest (RF). The latter has achieved better results reaching an accuracy of 66%. In Roeslin et al. [2020] a similar approach has been applied to the Mexico City urban area, hit by a strong earthquake (Mw = 7.1) in 2017. The dataset contained only 340 datapoints divided between negligible/light and moderate/heavy damaged buildings. Four algorithms have been employed: Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest. As in [Mangalathu and Jeo, 2018] RF gets the best score, reaching an accuracy of 65%. In [Yerlikaya-Özkurt and Askan, 2020], the authors consider a set of buildings damaged by three seismic events: Erzincan (1992, Mw = 6.6), Dinar (1995, Mw = 6.2) and Duzce (1999, Mw = 7.2). The target variable was the building damage level, assigned as light, moderate and severe/collapse. The analysis was conducted using the Classification and Regression Trees algorithm [Breiman et al., 2017].

More recently similar techniques have been applied to the 2009 L'Aquila earthquake [Di Michele et al., 2022; Di Michele et al., 2023(b); Di Michele et al., 2021]. Analysis was conducted using a dataset, built during the projects Open Data Ricostruzione [GSSI, 2019(a)] and Open Data L'Aquila [GSSI, 2019(b)], containing after cleaning 2532 buildings. Also, in this case, the RF ensures the best performace. Other interesting results are available in the literature, see for example [Mangalathu and Jeon, 2018, 2020; Mangalathu, Jeon and DesRoches, 2018; Harirchian et al., 2021; Kourehpaz and Molina Hutt, 2022].

In recent years, numerous datasets have been collected and made available by the scientific comunity. Among these, we mention the Da.DO project [Dolce et al., 2019], collecting data on tens of thousands of buildings damaged by strong earthquakes in Italy since 1980 (the year of the Irpinia earthquake). Based on that work, Faenza et al. introduced a new dataset called *ShakeDado* [Faenza et al., 2020] combining building information provided by the platform Da.DO and data from Shakemaps [Michelini et al., 2020]. *ShakeDado* contains data from six earthquakes, occurred in Italy between 1980 and 2012, with different magnitudes ($\geq$ 5 Mw) and not homogeneus geologic characteristics. Each building in the dataset is described through 34 features. Among these, number of stories, average year of construction, structural materials, seismic classification year of the Municipality, and Vs30 are independent of the seismic event under consideration. The other features are the maximum value (intended between all events of the same seismic sequence) of the intensity in MCS scale, PGA, PGV, SA at 0.3s, 1.0s, and 3.0s, reported according to the last ShakeMap release [Michelini et al., 2020]. For each of these features uncertainty, magnitude of the earthquake associated with the maximum values, distance between the eartquake source and the considered building are reported, together with the way in which this distance has been calculated (R_{JB} and R_{epi} symbolize the distances calculated with respect to the fault plane and with respect to the epicenter, respectively).

The aim of this work is to answer the following question:

'Is it possible to use data from past earthquakes to forecast future damage scenarios?'

It is certainly an important, but also very complex task. Earthquakes of magnitude greater than 5.5 Mw are, fortunately, rare events and each one is different from the others, in terms of magnitude, hypocenter depth, slip distribution, focal mechanism etc. On the other hand, subsurface characteristics and construction technologies can also be significantly different from one city to another, even if we consider the same country. Therefore, a preliminary step, aimed to make the datasets as homogeneous as possible, is mandatory. Here we use the supervised ML, however other approaches, based on transfer learning and semi-supervised techniques, are also possible and will be the subject of further studies. It is also pointed out that it is beyond the scope of this paper to conduct a comparative analysis of various supervised ML techniques, but we focus on the Random Forest algorithm, an ensemble method based on decision trees (DT) that has been shown to provide satisfactory results for classification tasks in many previous similar case studies [Mangalathu and Jeon, 2018; Roeslin et al., 2020; Di Michele et al., 2023].

The paper is organized as follows. In Section 2 we describe the datasets and the ML workflow employed for our analysis. In Section 3 we introduce our conceptual experiments, and we analize the results. Finally, in the last section some comments and remarks are provided.

2. Material and Methods

As mentioned above, our working dataset is ShakeDado [Faenza et al., 2020], containing information about six seismic sequences, namely Irpinia (1980), Umbria Marche (1997), Pollino (1998), Molise (2002), L'Aquila (2009), Emilia (2012). The events of each sequence with a magnitude $\geq$ 5.0 Mw are reported in Table 1. Before starting our analysis, we underline that this dataset, as available on the platform Da.D.O, is not georeferenced and we refer to [Faenza et al., 2020] for the spatial distribution of the buildings.

We can group the data into two subsets. The first group contains the *'one shock'* earthquakes, and the second one the *'multi-shocks'* events, namely complex seismic sequences characterized by multiple events of comparable magnitude. Pollino earthquake surely belong to the first group. Also, Irpinia 1980 and L'Aquila 2009 can be considered in "such a way" *one-shock* events because the main shock has in both cases magnitude much higher than the others. The Umbria Marche (1997) and Molise (2002) seismic sequences contain multiple events with comparable magnitudes and therefore belong to the 'multi-shocks' group. The 2012 earthquake in Emilia deserves a separate discussion. In this case, two events with similar intensity (6.1 Mw and 6.0 Mw) were recorded, but all the damaged buildings, included in our dataset, are close to the epicenter of the 6.0 Mw event, and far enough from the epicenter of the 6.1 Mw earthquake. We will therefore include the 6.0 Mw event the one *'one-shocks*' category.

Epicenters of the four one-shock earthquakes are displayed in Figure 1, where we also highlighted the radius of 10 km from each epicenter.

For all the earthquakes the damage rating is assigned over six levels from no-damage (D0) to heavy damage/collapsed (D5), namely:

- D0: no damage
- D1: light damage
- D2: moderate damage
- D3: medium damage
- D4: serious damage
- D5: heavy damage or collapse.

This classification is too detailed for the purposes of this work and has rarely been used for AI-based applications.

We consider a binary division of damage classes as follows:

- D0-D1: from no to light damage, is renamed as D-NL
- D2-D5: from moderate to heavy damage, is renamed as D-MH.

There are essentially two reasons for this choice. Firstly, when a strong earthquake occurs, once the first phase of the emergency has passed, it is essential to assess the damage suffered by the buildings, identifying which buildings are not or light damaged and it is, therefore, immediately usable. The standard procedure requires that each building will be inspected by a civil infrastructures technician who certifies its safety level. This is often a very time-consuming and expensive approach. On the other hand, in seismic risk planning, it is very important to know which buildings can be considered safe, and which ones are instead more vulnerable. In the next future, properly optimized AI tools will likely contribute to building damage assessment and risk mitigation.

Earthquake	Magnitude (Mw)	Data	Buildings
Irpinia	6.9	1980-11-23	38095
	5.0	1980-11-24	
	5.0	1980-11-25	
Umbria Marche	5.7	1997-09-26	6980
	6.0	1997-09-26	
	5.2	1997-10-03	
	5.4	1997-10-06	
	5.2	1997-10-12	
	5.6	1997-10-14	
	5.0	1998-03-21	
	5.1	1998-04-03	
Pollino	5.6	1998-09-09	3966
Molise	5.7	2002-10-31	14110
	5.7	2002-11-01	
L'Aquila	6.1	2009-04-06	52678
	5.1	2009-04-06	
	5.1	2009-04-06	
	5.1	2009-04-07	
	5.5	2009-04-07	
	5.4	2009-04-09	
	5.2	2009-04-09	
	5.0	2009-04-13	
Emilia	6.1	2012-05-20	1866
	5.1	2012-05-20	
	5.2	2012-05-20	
	6.0	2012-05-29	
	5.5	2012-05-29	
	5.5	2012-05-29	

Table 1. Earthquakes with Mw > 5.0 for the six seismic sequences included in ShakeDaDO [Faenza et al., 2020+].

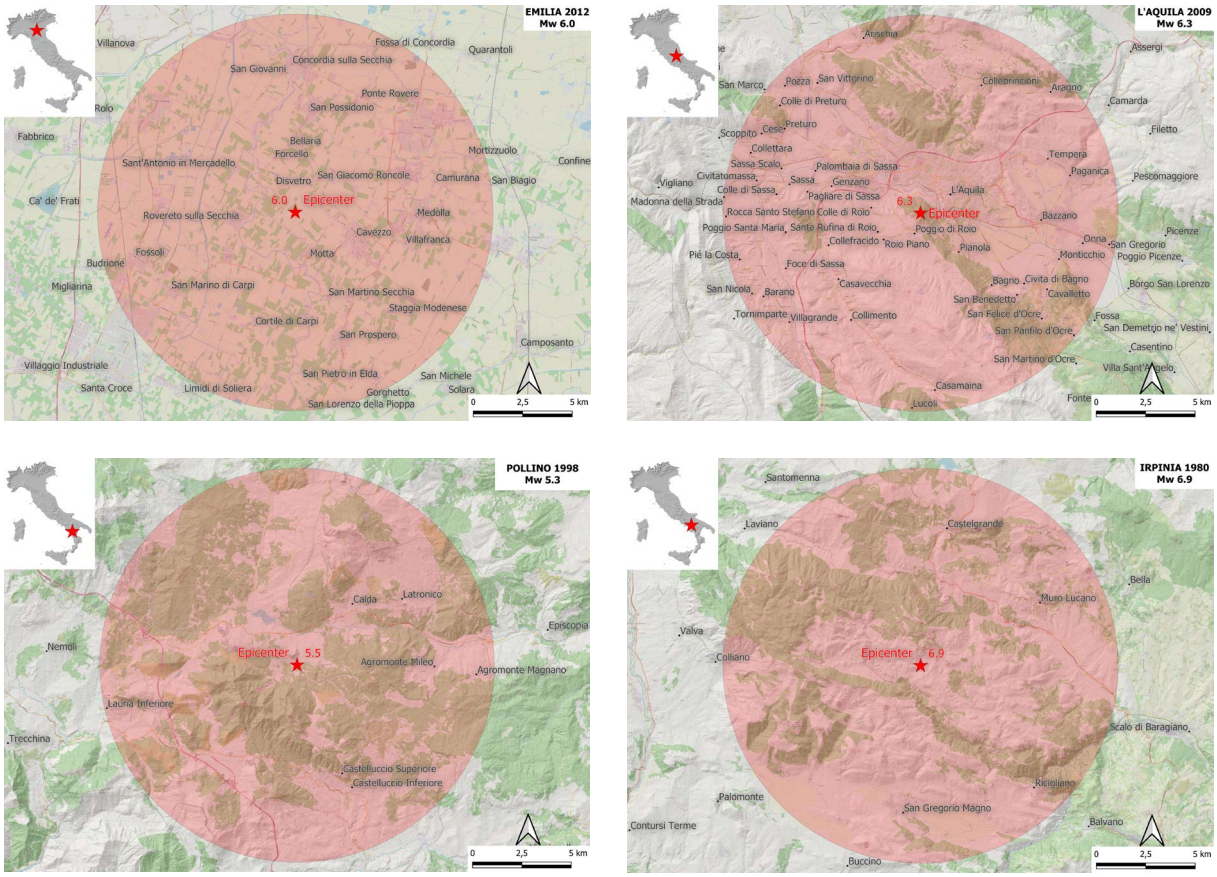

Figure 1. Epicenters of four shock earthquakes considered in this study. Highlighted in red the area of 10 km radius from each epicenter (red star). For the position of the buildings with respect to the epicenter we refer to ShakeDaDO [Faenza et al., 2020+].

3. Machine Learning Tools

Within this paper, to classify the damaged buildings, we employ the Random Forest (RF) algorithm introduced in [Ho, 1995]. It is a very robust tool, able to prevent overfitting, which can be used for several purposes: classification, regression, and features importance evaluation.

The functioning of the RF classifier is based on the Decision Trees (DT), DT uses the training dataset to establish a set of rules that are employed to rank the elements of the test and validation datasets [Raschka and Mirjalili, 2017; Pereira and Borysov, 2019]. Like a real tree, a DT starts from the root, where the dataset is complete, and then splits it along the branches according to a given characteristic f to maximize the Information Gain (IG) [Raschka and Mirjalili, 2017]:

$$IG = (D_s, f) = I(D_i) - \sum_{j=1}^{m} \frac{N_s}{N_i} I(D_s)$$

Roughly speaking the IG is the difference between the impurity $I(D_s)$ of the starting node of s, containing N_s elements, and the sum of that of the j child nodes, weighted by the fraction of elements each of them contains (N_s/N_j). Impurity can be calculated using several criteria such as:

Gini Impurity

$$I_G = 1 - \sum_{1=1}^{c} p\,(i|t)$$

Entropy $$I_H = \sum_{1=1}^{c} p\,(i|t)\log_2 p(i|t)$$

Classification error $$I_E = 1 - \max(p(i|t))$$

where $p(i|t)$ is the probability that an element belongs to the class i for an assigned node t. If $p(i|t) = 0$, then all the data-points belong to the same class. The process of subdivision of the parent node to the child nodes continues until the maximum depth n, fixed *a priori*, is reached. The RF is an ensemble model based on DTs. In the standard approach, k sub-samples of n elements are selected with replacements from the training set. In this way, the k decision trees built on each subset are weakly correlated. Each of the trees provides a classification label for some of the elements of the data set and the algorithm selects the most popular label using the so-called *majority vote*.

A suitable ML workflow for classification purposes reads as follows:

- dataset acquisition, preparation, and cleaning
- model(s) choise and hyper-parameters optimization
- perfomace evaluation.

Here, dataset preparation will be described in the next section and references therein. Roughly speaking we have two datasets: one, containing information about the past earthquake, is used for taining and validation. The second, which includes data from Emilia (or Pollino), is used for testing the model.

Regarding the choice of method, we use RF, which has shown better performance than other classifiers in solving similar problems. However, we also compared the performance of three classifiers Random Forest, Decision Tree (DT) and k-nearest neighbors (KNN), using the default values for the hyperparameters. The results will be reported in the following sections.

Optimizing hyperparameters is a crucial step in an ML workflow. There are three main possible approaches: *manual optimization*, *grid search* and *random search*. In the first case, it is the analyst who selects the hyperparameters values based on his own experience and dataset knowledge, but this rarely provides an sub-optimal set. More efficient is the *grid search*, where the model is trained on a grid of hyperparameters values selected by the user. If the grid is dense enough, this method can provide at least a sub-optimal set, but the calculation time could be very long. The best approach is usually the *random search*, where the analyst selects a range of values within which the system randomly selects the values. This strategy maximizes the probability of finding at least a sub-optimal hyperparameters set.

Among this work we employ the package RandomizedSearchCV of scikit-learn, where Randomized Search is combined to the cross-validation technique useful for reducing the overfitting, as explained in [Raschka and Mirjalili, 2017]. Each model was optimized using the training dataset, over a space of more than 300 hyperparameters combinations, and the selected value are reported in the Appendix.

Given the preliminary nature of this work and the dataset size, we include in the optimization tool just few hyperparameters. Indeed, the RF algorithm, as implemented in scikit-learn contains numerous hyperparameters can be adjusted to optimize the performance. Among them, the main one is certainly the n_estimators, which has a default value of 100 and controls the number of decision trees that are grown in the forest. Other hyperparameters considered in the optimization are the max_depth, which assigns the maximum depth of each tree, and the criterion i.e., the functional used to quantify the impurity and thus the quality of the split. For all the other tunable hyperparameters we refer to the user documentation of scikit-learn available on (https://scikit-learn.org/). We finally remark that, if no parameter value is set, the algorithm assumes default values.

The last step of our workflow is the model evaluation, usually provided by the accuracy, namely the ratio of correctly assessed buildings to the total number of them:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

From now on TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives. Other useful quantities are, recall, precision, and f1 score defined respectively as

$$\text{Recall} = \frac{TP}{TP + FP}$$

$$\text{Precision} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = \frac{TP}{TP + \frac{FN + FP}{2}}$$

Roughly speaking F1 can be read as the armonic mean of precision and recall.

4. Dataset Preparation

In this work, we analyze the 'one-shock' dataset. For each building, in addition to the damage level, other 33 features are assigned [Faenza et al., 2020]. However, for the purpose of this paper, we just account for 11 predictive variables. There are two main reasons which justify this choice. Firstly, we preferred to use parameters that have already considerd in other studies, in order to better evaluate the obtained results. Moreover due the preliminary nature of this study, the complexity and the non-homogeneity of the available data, we preferred to limit the number of input variables to reduce the risk of overfitting. The 11 features selected for this study are: Macroseismic Intensity (I_MCS), Peak ground velocity (PGV), Peak ground acceleration (PGA), Spectral Acceleration at 0.3 s (SA 0.3 s), Spectral Acceleration at 1.0 s (SA 1.0 s), Spectral Acceleration at 3.0 s (SA 3.0 s), Shear wave velocity at 30 m (Vs30), Average Year of Construction, Number of Storeys, Age classification Code, Year of Seismic Classification, and Damage level. For the ground motion parameters and macroseismic intensity, the maximum value recorded during the seismic sequence is reported for each building. In order to make the data coming from the various earthquakes more homogeneous, we remove the data points for which the maximum values of PGA, PGV, IMCS, SA 0.3 s, SA 1.0 s, and SA 3.0 s are not reached for the main event for both the Irpinia and the L'Aquila earthquakes and we restrict our analysis on the buildings that are located within a radius of 10 km from the main shock epicenter. Furthermore, we only consider masonry buildings, which are the majority of the buildings available in the original datasets.

Therefore we are ready to perform the following conceptual exercises:

E1. Assume that Emilia seismic sequence has just ended and the data from the other major earthquakes occurred in the past (L'Aquila, Pollino and Irpinia) are available. We calculate the damage level for each building in the Emilia dataset.

E2. Assume that the Pollino seismic sequence has just ended and the data from the other major earthquakes with comparable magnitude (L'Aquila, Emilia) are available. We calculate the damage level for each building in the Pollino dataset

For E1 our hypothesis is completely reasonable, in fact, the earthquake in Emilia is the most recent of those considered in this study. In the case of E2, on the other hand, the Pollino earthquake occurred in 1998, before the earthquakes in L'Aquila and Emilia. This, while not influencing the feasibility of our study a priori, should be properly considered in the discussion of the results. In this case data from Irpinia earthquake is not included in the training set, because the magnitude of the main shock is much greater than that recorded in Pollino (6.9 *versus* 5.6 Mw). In the following, we separately analyze the two and we compare the obtained results.

4.1 Experiment E1

First, we need to construct the set for training/validation and test procedures. For this purpose, we introduce the dataset $E1_{TV}$ obtained merging data from Irpinia 1980, Pollino 1998, and L'Aquila 2009 earthquakes. $E1_{TV}$ is composed of 34795 buildings of which 19395 (56%) with no-light damage and 15400 (44%) with medium-heavy damage.

Test dataset is named $E1_{Test}$, and in this experiment, coincides with the Emilia 2012 set. It contains 1496 buildings, divided into 878 (59%) D-NL and 618 D-MH (41%) buildings.

As a part of our preliminary analysis, we report in Figure 2 the feature importance score, obtained using the RF tool, for the two datasets $E1_{TV}$, on the left side and $E1_{Test}$, on the right side. In both cases *feauture_importances* tool as implemented in scikit-learn it is used setting *n_estimator* equal to 300 and the *test_size* ratio equal to 0.20. Let's just specify the information provided for $E1_{Test}$ cannot be used for model selection and hyperparameters optimization, but it will help us in the interpretation of the results.

According to Figure 2, the Average Year of Construction is the variable that most affects the level of damage for the training-test dataset. These results are confirmed by the confusion matrix displayed in the Appendix. The other characteristics have comparable scores except for the Year of Seismic Classification, which we will not consider in the subsequent analysis. A similar trend is also recorded for the test dataset but, of course, we don't know this in advance, because the damage level is not available in our conceptual exercise.

We have ten characteristics available a priori for both training and test dataset, that can be successfully used to calibrate and optimize the model. Six feature refers to the earthquake itself: PGV, PGA, SA 0.3 s, SA 1.0 s, SA 3.0 s, MCS intensity. Their distribution is shown in Figure 3-4, for both $E1_{TV}$ and $E1_{Test}$, in the left and right column, respectively. The other four characteristics refer to the building and to the geological proprieties of the subsoil (Vs30) and they are reported in Figure 5 for $E1_{TV}$ (left column) and $E1_{Test}$ (right column). In each picture also the damage level is displayed, the red part of each bin represents the portion of the building which has been seriously damaged, and the green one is those with no-low damage. It is immediately evident that the various features have different distributions between $E1_{TV}$ and $E1_{Test}$ (except perhaps for the number of stories).

Although the values distribution are visibly different, for each feature (except SA 1.0 s), the range characterizing $E1_{Test}$ is well populated also in the train-validation dataset. In other terms, the test dataset seems well described by training. This is a purely qualitative observation, which should be better quantified on more extensive datasets, to define the limits and the application of this technique.

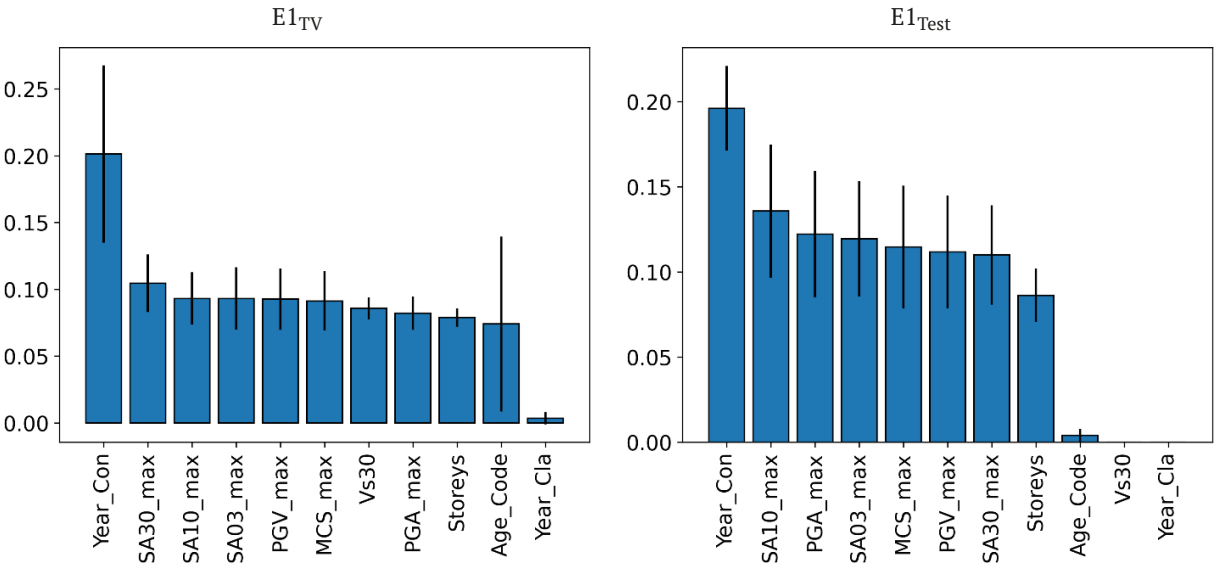

Figure 2. Feature importance score for the training dataset $E1_{TV}$ and the test set $E1_{Test}$, with standard deviation, obtained using the RF algorithm.

The choice of the 2012 earthquake as the test dataset has several advantages that certainly affect the obtained results. As already mentioned, it is the most recent of the four considered earthquakes and this makes the experiment more realistic. On the other hand, the magnitude of the Emilia earthquake is intermediate compared to the other events and is close to that of L'Aquila, a relatively recent event, which provides most of the data points in $E1_{TV}$.

Before proceeding with the analysis, we compare the performance of three classifiers (Random Forest, Decision Tree (DT) and k-nearest neighbors (KNN)) using the hyperparmaters default values. We specify, that this preliminary analysis as well as parameter optimization described below, was performed working exclusively on the $E1_{TV}$ dataset,

appropriately divided into a subset of training and one of validation. As expected, the best accuracy of 0.71 is obtained by RF. The other two classifiers achieved slightly lower accuracy: 0.69 for DT and 0.68 for KNN. This justifies the choice of using random forests for this study. Finally, before proceeding to the optimisation of the hyperparameters, we test the sensitivity of the RF method with respect to the features. To this end, we compare the performance of the algorithm on 3 different datasets. The first one is $E1_{TV}$, the other two contains the same datapoint as $E1_{TV}$, but the three and the five best scoring features reported Figure 2. In both cases the total accuracy remains quite high (0.70). This is quite expected in fact much of the information is contained in the Average Year of Construction. For our poupouse, not having the Test dataset theoretically available, we account for all available variables except the Year of Seismic Classification.

For the hyperparameters optimization we employ the RandomizedSearchCV of scikit-learn. $E1_{TV}$ is divided into train set, which contains the 90% of the data points, and validation set which contains the remaining data points. The ratio between D-NL and D-MH is maintained in this procedure. The model is optimized over a space of 80 combination. The best hyperparameters combination (n_estimator=4157, max_depth=278, criterion=entropy) is used for damage prediction on the Emilia data set, for which the damage distribution is a priori, unknown. The predicted damage level is then compared with that available on Da.D.O and the classifier performance is evaluated in terms of accuracy, recall and precision (see Table 2).

	Precision	Recall	f1-score	Support
D-NL	0.72	0.58	0.64	878
D-MH	0.53	0.68	0.60	618
Accuracy			0.623	1496

Table 2. Performance of the best-fit model on the test set $E1_{Test}$.

The accuracy obtained on the test dataset set is 0.623, comparable with the available literature [see for example Roeslin et al., 2020].

For sake of completeness results coming from five different tests (the best one plus four randomly selected) are listed in the Appendix together with the set of hyperparameters. We observe how the hyperparameters that maximizes the accuracy for the train and validation dataset (which we recall is the only one available a priori for our conceptual experiment), does not coincide with the maximum accuracy in the test dataset, which could also have a significantly different distribution, however we note as the model's performance remains quite stable for all five tests, this confirms the robustness of the developed forecasting model.

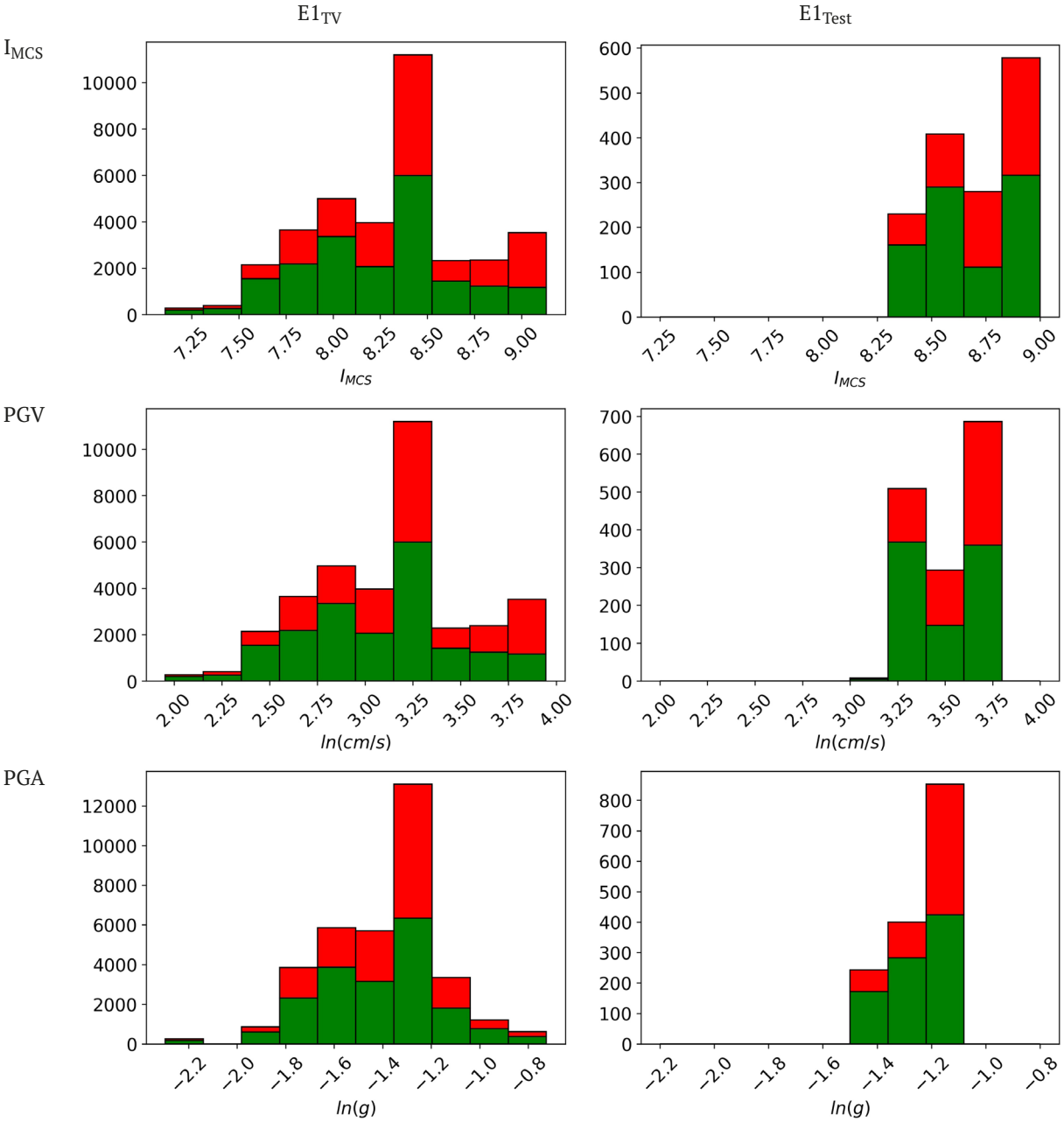

Figure 3. Values of the MCS intensity, PGV and PGA for $E1_{TV}$ (left side) and $E1_{Test}$ (right side). Red part of each bin represents the seriously damaged buildings and the green the no-low damage buildings.

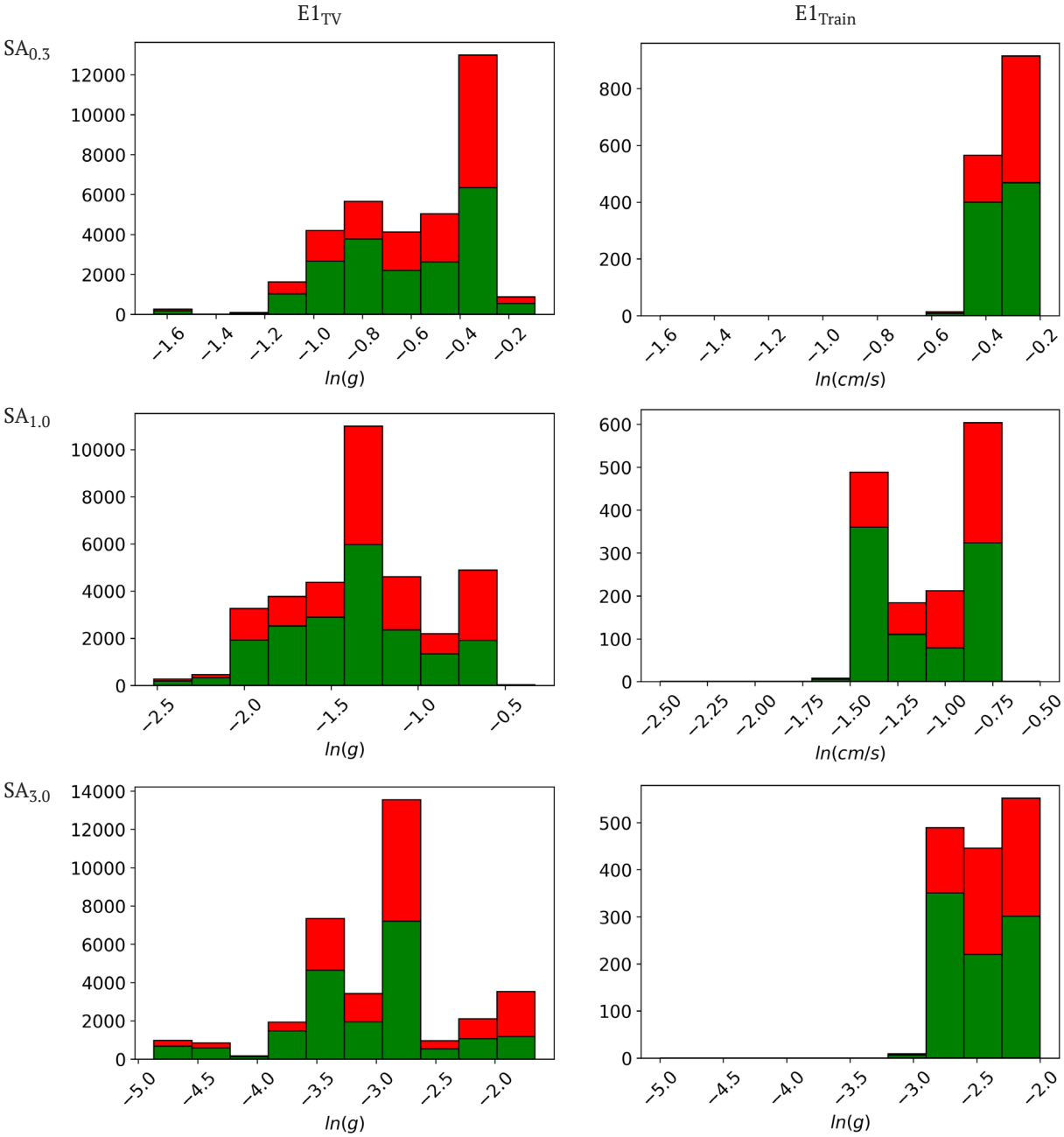

Figure 4. Values of the SA 0.3 s, SA1.0 s and SA30 s for $E1_{TV}$ (left side) and $E1_{Test}$ (right side). Red part of each bin represents the seriously damaged buildings and the green the no-low damage buildings.

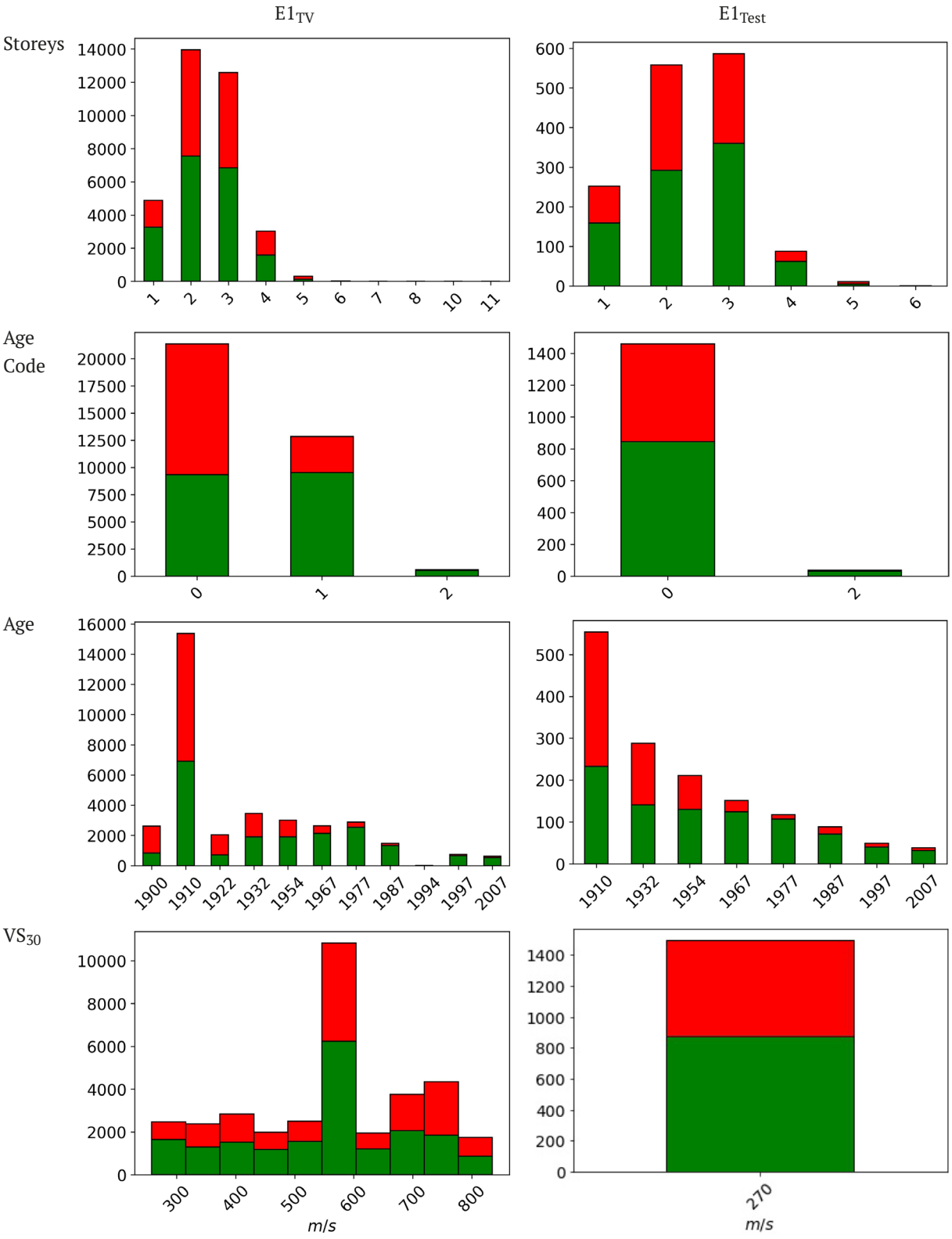

Figure 5. Building features distribution for $E1_{TV}$ (left side) and $E1_{Test}$ (right side). Red part of each bin represents the seriously damaged buildings and the green the no-low damage buildings.

4.2 Experiment E2

Let us introduce a new dataset named $E2_{TV}$ obtained merging data from L'Aquila, and Emilia earthquakes and the test set, named $E2_{Test}$, corresponding to the Pollino earthquake. As before $E2_{TV}$ is used to train/validate the model and $E2_{Test}$ to test it. $E2_{TV}$ contains 28485 buildings having, among them 15584 (58%) are classified as D-NL and 11901 (42%) as D-MH. The percentage of buildings with only minor damage or no damage is higher here in the previous experiments, as data from the earthquake in Irpinia, the one with the greatest magnitude, are not included. $E2_{Test}$ contains 1563 buildings, divided into 1077 (69%) D-NL and 486 (31%) D-MH buildings.

The feature importance score is reported in Figure 6, for both $E2_{TV}$ and $E2_{Test}$. As for the previous experiment, the Year of Seismic Classification is the variable that least influences the damage in the training set, it is not included in the subsequent analysis. Similar conclusions can be drawn by looking at the correlation matrix reported in the Appendix. However, in this case, two distributions are markedly different, much more than in the case of dataset E1. This is not surprising, since this event is characterized by a much lower magnitude (5.6 Mw versus 6.0-6.1 Mw) and occurred 20 years before the other two.

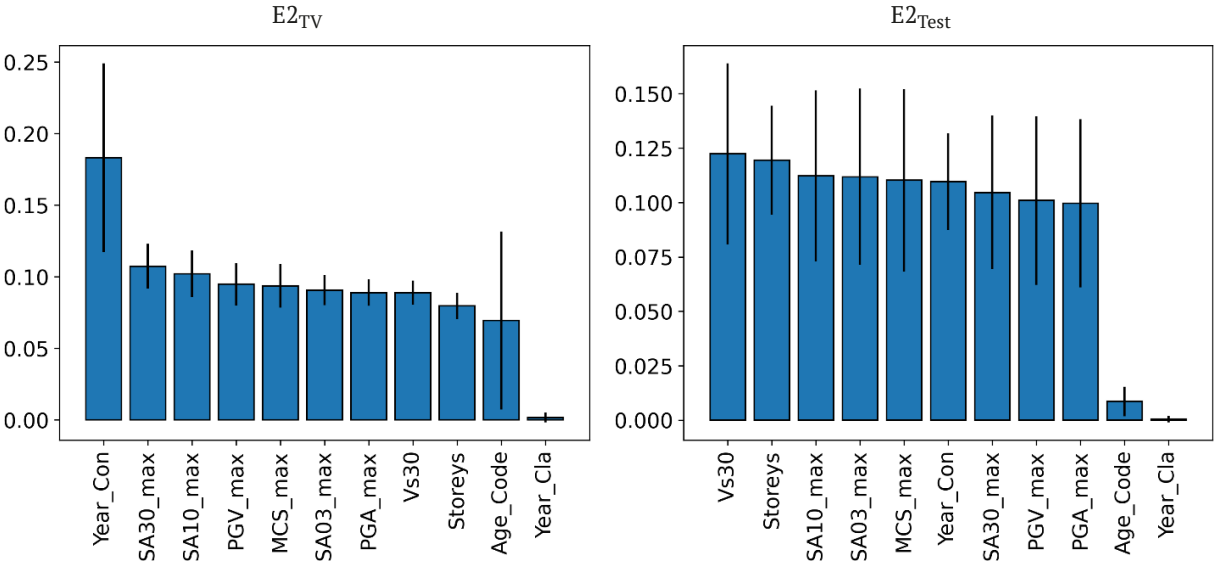

Figure 6. Feature importance score with standard deviation for the training dataset $E2_{TV}$ and the test set $E2_{Test}$, obtained using the RF algorithm.

The dataset contains ten characteristics, six referring to the earthquake itself, PGV, PGA, SA 0.3 s, SA 1.0 s, SA 3.0 s, IMCS. Their distribution is shown in Figure 7-8 ($E2_{TV}$ and $E2_{Test}$ left and right panels respectively). We notice how the ranges of values that characterize PGV, SA10, and SA30 in the test dataset are not well represented in the TV set, and this certainly worsens the performance of the forecasting model. In Figure 9 the building features are reported, in this case, the ranges of values are reasonably similar between the two datasets, although not perfectly overlapping. We can therefore state as, although there is a good homogeneity between the building typologies and the subsoil properties among the considered earthquakes, the seismic events used to train the model are not suitable for reproducing the damage due to Pollino earthquake. We will see later that this observation will be confirmed in the test phase of the AI-based model.

As in the previous case, we test three different classifiers on $E2_{TV}$. The best accuracy is obtained using RF (0.71) slightly less for DT (0.70) and for KNN (0.67). A sensitivity analysis in also reported for sake of completeness. We compare the performance of the RF algorithm on 3 different datasets. The first one is $E2_{TV}$, the other two contains the same datapoint as $E2_{TV}$ but just the three and five features, corresponding to the best scoring features in Figure 6. The results are like those already observed for E1. There is a slight and not substantial decrease in accuracy (0.70 for both the reduced datasets), motivated by the fact that much of the damage information is contained in RandomizedSearchCV is employed to select the best hyperparameters set. $E2_{TV}$ is divided into a

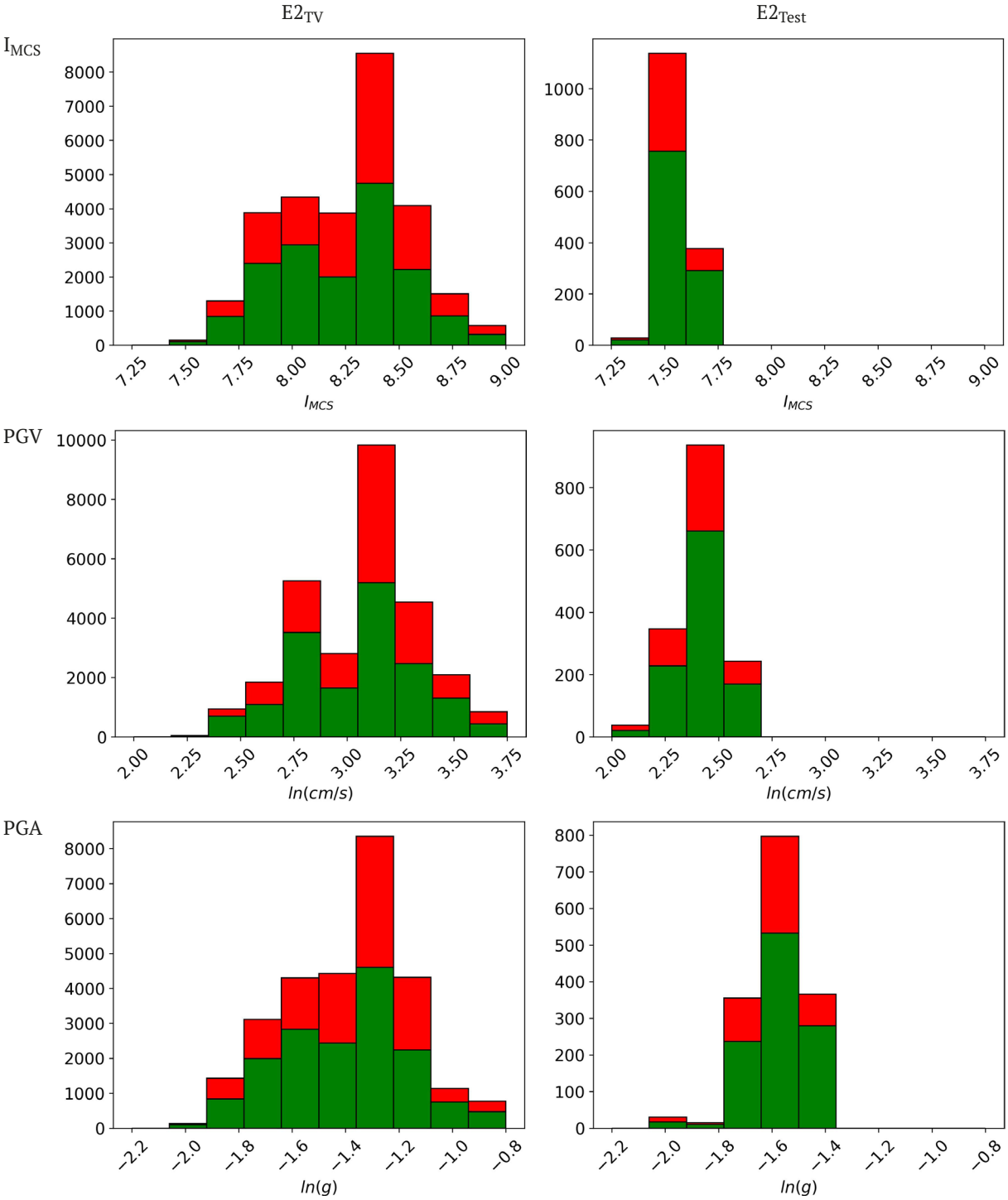

Figure 7. Values of the MCS intensity, PGV and PGA for $E2_{TV}$ (left side) and $E2_{Test}$ (right side). Red part of each bin represents the seriously damaged buildings and the green the no-low damage buildings.

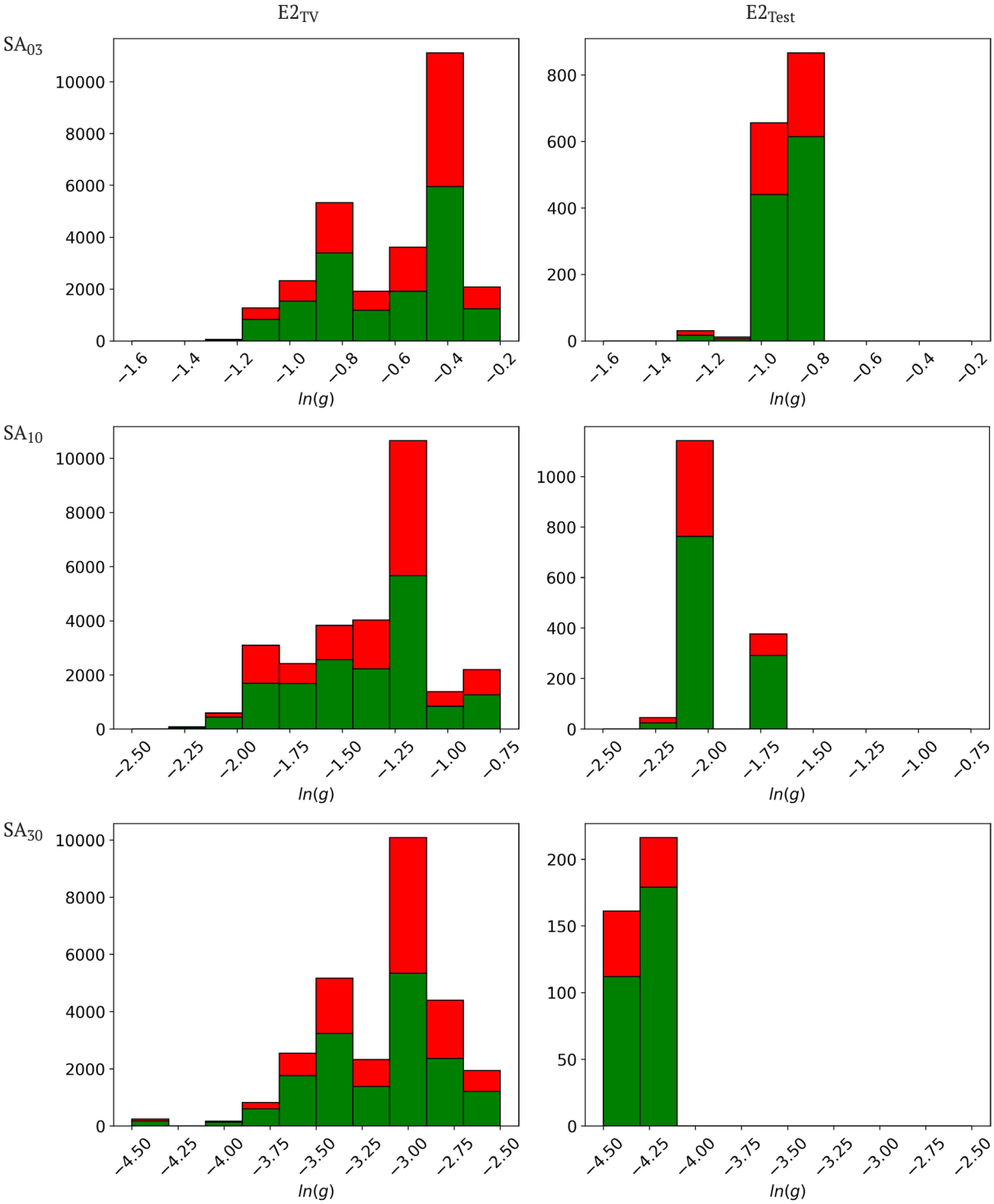

Figure 8. Values of the SA 0.3 s, 1.0 s and 3.0 s $E2_{TV}$ (left side) and $E2_{Test}$ (right side). Red part of each bin represents the seriously damaged buildings and the green the no-low damage buildings.

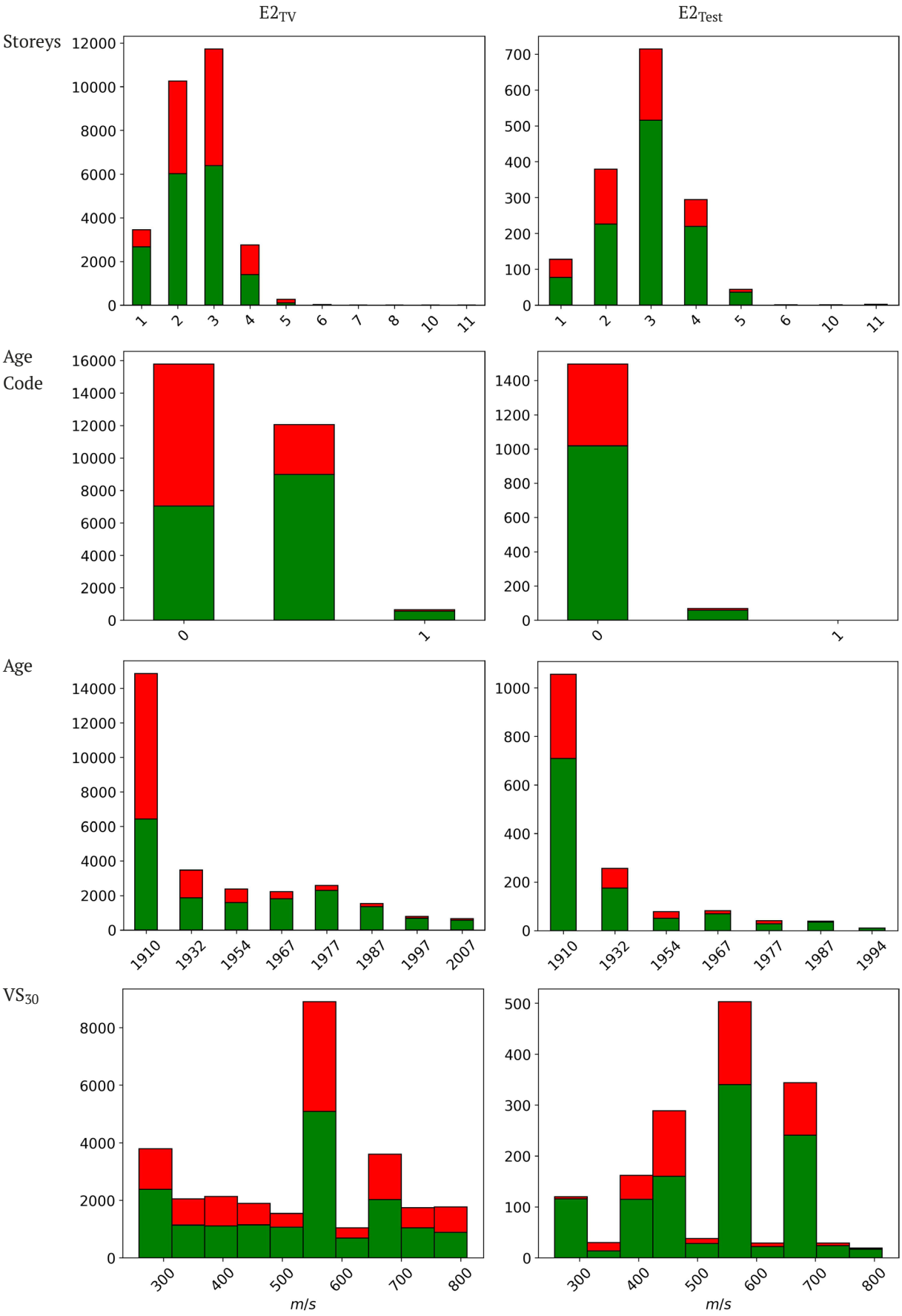

Figure 9. Building feature distribution for $E2_{TV}$ (left side) and $E2_{Test}$ (right side). Red part of each bin represents the seriously damaged buildings and the green the no-low damage buildings.

training set and a validation set, containing 90% and 10% of the data points respectively. The model is optimized over a space of 80 combinations. The best combination of hyperparameters (n_estimator=2373, max_depth=13, criterion=entropy) is used to train $E2_{TV}$ and to assess the damage on the Pollino dataset. Results are reported in Table 3. Other four hyperparameters combination are reported, for sake of completeness, in the Appendix. As expected for E2 the results are not satisfactory, although the total accuracy is high, reaching 0.66. The performance of the classifier on the D-MH subset is completely wrong, with a precision of less the 30%. The classifier fails to distinguish heavily damaged from slightly damaged buildings, classifying most data points as D-NL. Similar results are also obtained with other combinations of hyperparameters, confirming the stability of the proposed method.

	Precision	Recall	f1-score	Support
D-NL	0.69	0.93	0.79	1077
D-MH	0.27	0.06	0.10	486
Accuracy			0.66	1563

Table 3. Performance of the best-fit model on the test set $E2_{Test}$.

5. Discussion and conclusions

In this paper, we use Shake Da.D.O to estimate the damage to buildings after a strong earthquake using information available from other events. In more detail, we consider two cases, named E1 and E2. In the first one, we use data from Irpina, Pollino, and L'Aquila earthquakes to train a RF algorithm and discriminate buildings with no-light and medium-high damage. The model, suitably optimised, was used to assess the damage caused by the 2012 Emilia earthquake. The results obtained are quite satisfactory and in line with the extensive literature available. We note that in this case the data in the test and training datasets were quite similar in the sense that the ranges of values assumed by the 10 features in the test dataset were well represented in the training dataset. In contrast, for the second experiment E2, the value ranges of the ground motion parameters in the test dataset were not well populated in the training and validation datasets. For this reason, the classifier is unable to assess the correct damage level. For both datasets, we also evaluated the impact of each characteristic on the level of damage. In general, we found that the most relevant feature is the year of construction, except for the Pollino dataset, where Vs30 seems to play the main role. This difference for sure worsens, the results of the classifier for the E2 experiment.

While this work does not give an exhaustive and definitive answer to the question 'is it possible to use data from past earthquakes to *predict* future damage scenarios using AI?', it does provide some important indications of a possible methodology, as follows:

1) the data must be well-constrained a priori. For example, the range of values of each characteristic of the test dataset must be well populated in the training set.
2) the mangnitude considered must be as close as possible and the data homogeneous in terms of distance from the epicenter and construction typology.
3) The hyperparameter optimization, now based on total accuracy, does not appear to be the best choice, it seems more suitable for the purpose of this paper to test alternative approaches such as maximizing recall for D_NL buildings. For the test E1 the buildings labeled as D-NL have a precision 0.72 and recall 0.58. These results suggest a discrete, but still accettable, number of false positives that should better to minimize during a seismic risk assessment activity. On the contrary for the D-MH building the precision is 0.53 and the recall is 0.68, corresponding to a larger number of false negative than false positive. For test E2, the results provided by the best set of hyperparameters are not satisfactory at all, especially for the D-MH building (recall equal to 0.06). In this case it would be better to provide an average result obtained using different hyperparameter values, thus minimizing the possibility of using a set of hyperparameters that is *too unbalanced*.

4) hyperparameters values, optimized on the training-validation set, are used to classify data on the test set. Observing the results, it seems better to use an ensemble of models and apply an appropriate target value selection mechanism such as a majority voting. However, in this case the main problem is that the optimization is performed on the train dataset and not on the test dataset whose damage level in unknow a priori. In fact, the accuracy we detect is significantly lower on the test dataset than on the validation dataset. This cannot be considered a limitation of the applied methodology but is an inherent uncertainty in the problem we posed: in fact, the two datasets contain information related to different events. Certainly the perfomances will improve with the addition of new data, such as those related to the Amatrice earthquake, in the available databases.

Another interesting approach to provide an answer to our main question could be the combined use of high-performance computing techniques and AI tools. In particular, HPC techniques have recently been used to generate synthetic seismograms of scenarios earthquakes [Paolucci et al., 2021, Di Michele et al., 2021, Di Michele et al., 2022, Pera et al., 2023]. So AI and HPC can be used to generate data-driven risk scenarios, using simulated ground motion data referring to possible future as input instead of interpolated data generated for high seismic risk regions.

Acknowledgements. This work was partially supported by the GSSI "Centre for Urban Informatics and Modelling" (Italian Government – Presisdenza Consiglio dei Ministri – CUIM project delibera CIPE n. 70/2017).
This work has been also partially funded by the European Union – NextGenerationEU under the Italian Ministry of University and Research (MUR) National National Centre for HPC, Big Data and Quantum Computing CN_00000013 – CUP: E13C22001000006.

Appendix

Confusion matrices

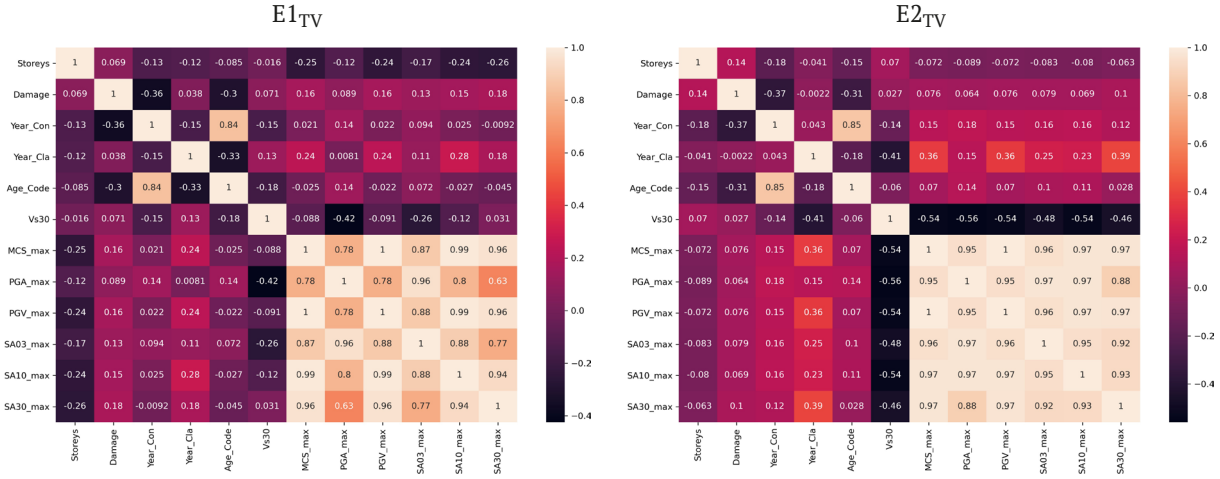

Figure A1. Correlation matrix for the dataset $E1_{TV}$ and $E2_{TV}$.

For sake of completeness, we report, in the following tables, the performance of RF classier using four different set of hyperparameters among the 80 tested in the optimization procedure. Together to the result getting for the optimal set (in bold on the first line of each table). The grid for the optimization has been carried out as follow:

n_estimators = [int(x) for x in np.linspace(1000,3000, num=100)]

criterion = ['entropy','gini']

max_depth = [int(x) for x in np.linspace(10,200, num=100)]

For each experiment the result in terms of total accuracy, recall and precision are weakly dependent on the model hyperparameters, confirming the stability of the employed workflow.

E1	Precision					Recall					Accuracy					Support
D-NL	**0.72**	0.76	0.72	0.76	0.76	**0.58**	0.59	0.59	0.62	0.57						878
D-MH	**0.53**	0.55	0.54	0.57	0.55	**0.78**	0.68	0.68	0.73	0.74						618
Total											0.62	0.62	0.63	0.67	0.65	1496

E2	Precision					Recall					Accuracy					Support
D-NL	**0.69**	0.66	0.66	0.66	0.68	**0.93**	0.60	0.66	0.65	0.91						1077
D-MH	**0.27**	0.27	0.26	0.26	0.27	**0.06**	0.33	0.26	0.27	0.8						486
Total											0.66	0.51	0.53	0.53	0.65	1563

References

Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone (2017). Classification and Regression Trees, Routledge.

GSSI (2019). Open Data L'Aquila, https://www.opendatalaquila.it/.

GSSI (2019). Open Data Ricostruzione, https://opendataricostruzione.gssi.it/home.

Di Michele, F., E. Stagnini, D. Pera, B. Rubino, R. Aloisio, A. Askan and P. Marcati (2023). A Machine Learning Tool for Damage Classification: The Case of L'Aquila 2009 Earthquake, Natural Hazards, in press.

Di Michele, F., O. Giannopoulou, E. Stagnini, D. Pera, B. Rubino, R. Aloisio, A. Askan and P. Marcati (2023). Machine Learning for damage classification, risk mitigation and post earthquake management 7th International Conference on Earthquake Engineering and Seismology (7ICEES).

Di Michele, F., J. May, D. Pera, V. Kastelic, M. Carafa, C. Smerzini … and P. Marcati (2022). Spectral element numerical simulation of the 2009 L'Aquila earthquake on a detailed reconstructed domain, Geophys. J. Int., 230, 1, 29-49.

Di Michele, F., D. Pera, J. May, V. Kastelic, M. Carafa, A. Styahar, B. Rubino, R. Aloiso and P. Marcati (2021). On the possible use of the not-honoring method to include a real thrust into 3D physical based simulations, In: 21st International Conference on Computational Science and Its Applications (ICCSA), Cagliari, Italy, 2021, 268-275, doi: 10.1109/ICCSA54496.2021.00044.

Dolce, M, E. Speranza, F. Giordano, B. Borzi, F. Bocchi, C. Conte, A. Di Meo, M. Faravelli and V. Pascale (2019). Observed Damage Database of Past Italian Earthquakes: The Da. DO Webgis, Boll. Geofis. Teor, Appl., 60, 2.

Faenza, L., A. Michelini, H. Crowley, B. Borzi and M. Faravelli (2020). ShakeDaDO: A Data Collection Combining Earthquake Building Damage and Shakemap Parameters for Italy, Artif. Intell. Geosci. 1, 36-51.

Harirchian, E., V. Kumari, K. Jadhav, S. Rasulzade, T. Lahmer and R.R. Das (2021). A Synthesized Study Based on Machine Learning Approaches for Rapid Classifying Earthquake Damage Grades to Rc Buildings, Applied Sci., 11, 16, 7540.

Ho, T.K. (1995). Random Decision Forests, In Proceedings of 3rd International Conference on Document Analysis and Recognition, 1, 278-82. IEEE.

Kourehpaz, P. and C. Molina Hutt (2022). Machine Learning for Enhanced Regional Seismic Risk Assessments, J. Struct. Engin. 148, 9, 04022126.

Mangalathu, S. and J.-S. Jeon (2018). Classification of Failure Mode and Prediction of Shear Strength for Reinforced Concrete Beam-Column Joints Using Machine Learning Techniques, Engineering Structures, 160, 85-94.

Mangalathu, S. and J.-S. Jeon (2020). Regional Seismic Risk Assessment of Infrastructure Systems Through Machine Learning: Active Learning Approach, J. Struct. Engin., 146, 12, 04020269.

Mangalathu, S., J.-S. Jeon and R. Des Roches (2018). Critical Uncertainty Parameters Influencing Seismic Performance of Bridges Using Lasso Regression, Earthq. Engin. Struct. Dyn., 47, 3, 784-801.

Mangalathu, S., H. Sun, C.C. Nweke, Z. Yi and H.V. Burton (2020). Classifying Earthquake Damage to Buildings Using Machine Learning, Earthquake Spectra 36, 1, 183-208.

Michelini, A., L. Faenza, G. Lanzano, V. Lauciani, D. Jozinović, R. Puglia and L. Luzi (2020). The New Shakemap in Italy: Progress and Advances in the Last 10 Yr, Seism. Res. Lett., 91, 1, 317-33.

Paolucci, R., C. Smerzini and M. Vanini (2021). BB-SPEEDset: A validated dataset of broadband near-source earthquake ground motions from 3D physics-based numerical simulations, Bull. Seism. Soc. Am., 111, 5, 2527-2545.

Pera, D., F. Di Michele, E. Stagnini, B. Rubino, R. Aloisio and P. Marcati (2023). Numerical Simulations of 1461 and 1762 San Pio delle Camere (L'Aquila) Earthquakes Using 3D Physic-Based Model, Lecture Notes in Computer Science, 14111 LNCS, 549-565, doi: 10.1007/978-3-031-37126-4_35.

Pereira, Câmara F. and S.S. Borysov (2019). "Machine Learning Fundamentals." In Mobility Patterns, Big Data and Transport Analytics, 9-29. Elsevier.

Raschka, S. and V. Mirjalili (2017). Python Machine Learning: Machine Learning and Deep Learning with Python. Scikit-Learn, and TensorFlow, Second Edition Ed.

Roeslin, S., Q. Ma, H. Juárez-Garcia, A. Gómez-Bernal, J. Wicker and L. Wotherspoon (2020). A Machine Learning Damage Prediction Model for the 2017 Puebla-Morelos, Mexico, Earthquake, Earthquake Spectra, 36, 2_suppl., 314-39.

Yerlikaya-Özkurt, F., A. Askan (2020). Prediction of Potential Seismic Damage Using Classification and Regression Trees: A Case Study on Earthquake Damage Databases from Turkey, Nat. Hazards 103, 3, 3163-80.

***CORRESPONDING AUTOTHOR: Federica DI MICHELE**,
Istituto Nazionale di Geofisica e Vulcanologia, Milano (Italy)
e-mail: federica.dimichele@ingv.it