Innovation for earth quake prediction model using python:

Earthquake Prediction Model with Machine Learning:

It is well known that if a disaster occurs in one region, it is likely to happen again. Some regions have frequent earthquakes, but this is only a comparative amount compared to other regions.

So, predicting the earthquake with date and time, latitude and longitude from previous data is not a trend that follows like other things, it happens naturally.

I will start this task to create a model for earthquake prediction by importing the necessary python libraries:

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt
```

```
data = pd.read_csv("database.csv")

data.columns
```

```
Index(['Date', 'Time', 'Latitude', 'Longitude', 'Type', 'Depth', 'Depth Error',
       'Depth Seismic Stations', 'Magnitude', 'Magnitude Type',
       'Magnitude Error', 'Magnitude Seismic Stations', 'Azimuthal Gap',
       'Horizontal Distance', 'Horizontal Error', 'Root Mean Square', 'ID',
       'Source', 'Location Source', 'Magnitude Source', 'Status'],
      dtype='object')
```

Now let's see the main characteristics of earthquake data and create an object of these characteristics, namely, date, time, latitude, longitude, depth, magnitude:

```
data = data[['Date', 'Time', 'Latitude', 'Longitude', 'Depth', 'Magnitude']]

data.head()
```

	date	Time	Latitude	Longitude	Depth	Magnitude
0	01/02/1965	13:44:18	19.246	145.616	131.6	6.0
1	01/04/1965	11:29:49	1.863	127.352	80.0	5.8
2	01/05/1965	18:05:58	-20.579	-173.972	20.0	6.2
3	01/08/1965	18:49:43	-59.076	-23.557	15.0	5.8
4	01/09/1965	13:32:50	11.938	126.427	15.0	5.8

Since the data is random, so we need to scale it based on the model inputs. In this, we convert the given date and time to Unix time which is in seconds and a number. This can be easily used as an entry for the network we have built:

```
import
datetime
            import time


            timestamp = []
            for d, t in zip(data['Date'], data['Time']):
                try:
                    ts = datetime.datetime.strptime(d+' '+t, '%m/%d/%Y %H:%M:%S')
                    timestamp.append(time.mktime(ts.timetuple()))
                except ValueError:
                    # print('ValueError')
                    timestamp.append('ValueError')
            timeStamp = pd.Series(timestamp)
```

```
data['Timestamp'] = timeStamp.values

final_data = data.drop(['Date', 'Time'], axis=1)

final_data = final_data[final_data.Timestamp != 'ValueError']

final_data.head()
```

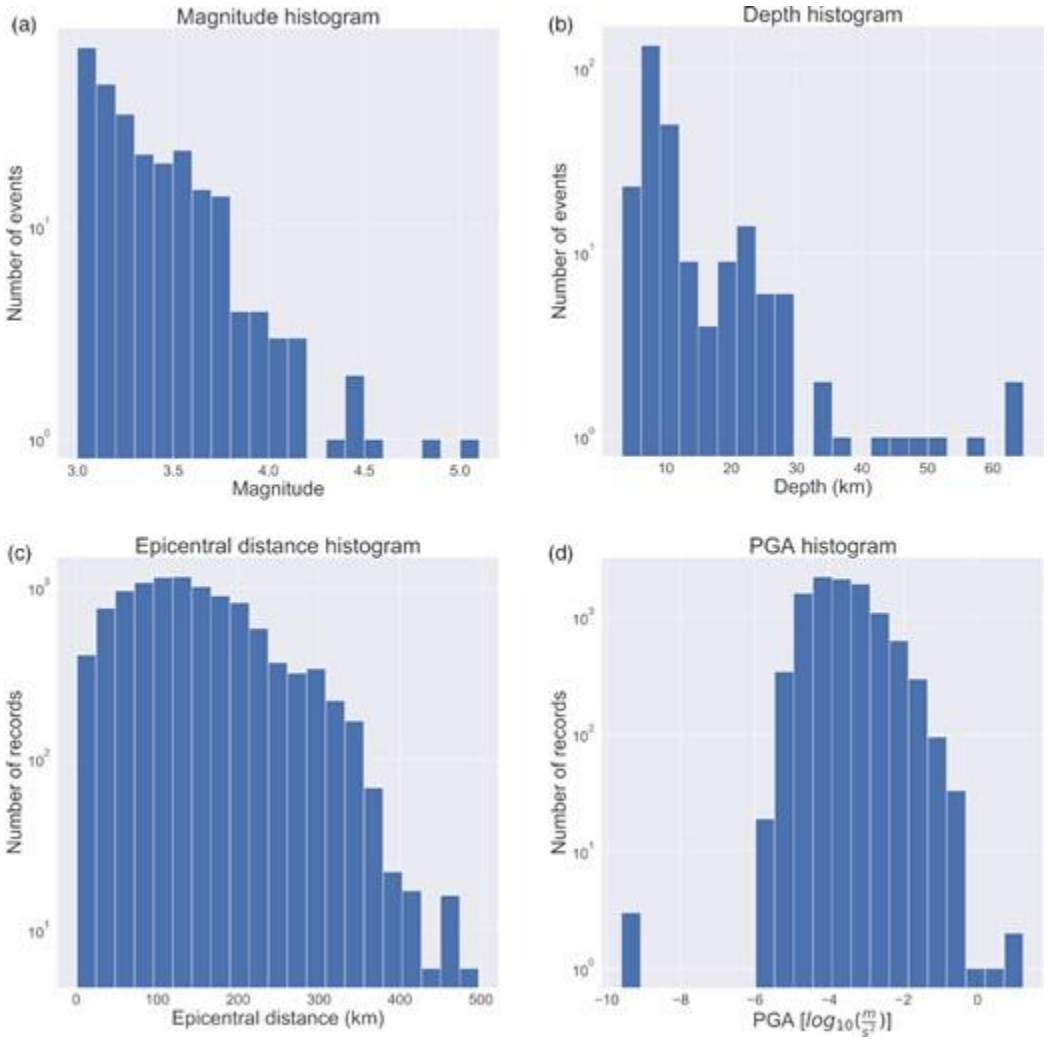
	Latitude	Longitude	Depth	Magnitude	Timestamp
0	19.246	145.616	131.6	6.0	-1.57631e+08
1	1.863	127.352	80.0	5.8	-1.57466e+08
2	-20.579	-173.972	20.0	6.2	-1.57356e+08
3	-59.076	-23.557	15.0	5.8	-1.57094e+08
4	11.938	126.427	15.0	5.8	-1.57026e+08

DATA COLLECTION:

The input data from central western Italy (hereinafter denominated the CW data set) consist of three-component waveforms of 256 earthquakes with magnitude $2.9 \leq M \leq 5.1$ (Fig. 1a), from 39 stations. The earthquakes occurred between 2013 January 1 and 2017 November 20. The earthquake depths range from 3.3 to 64.7 km (Fig. 1b). The stations and the earthquakes are located in the area bounded by latitude [41.13°, 46.13°] and longitude [8.5°, 13.1°] (Fig. 2), with epicentral distances ranging from 10 to 498 km (Fig. 1c). The area

overlaps slightly with the area of central Italy used in J2020 from which the pre-trained model architecture is taken. To avoid possible data leakage from the pre-trained model, events that were used in that study were excluded from our data set. We chose the stations having the largest number of events recorded while making sure that there is an acceptable spatial distribution of the stations to cover more area with stations that are close to earthquake epicentres to simulate possible early warning use (more details about the stations in Supporting Information Table S1). The stations belong to the IV, GU and MN networks.

Figure 1.



Histograms of the selected 256 earthquakes: (a) earthquake-magnitude distribution, (b) earthquake-depth distribution, (c) epicentral distance distribution and (d) distribution of PGA [$\log_{10}$(m s^{-2})]

Figure 2.

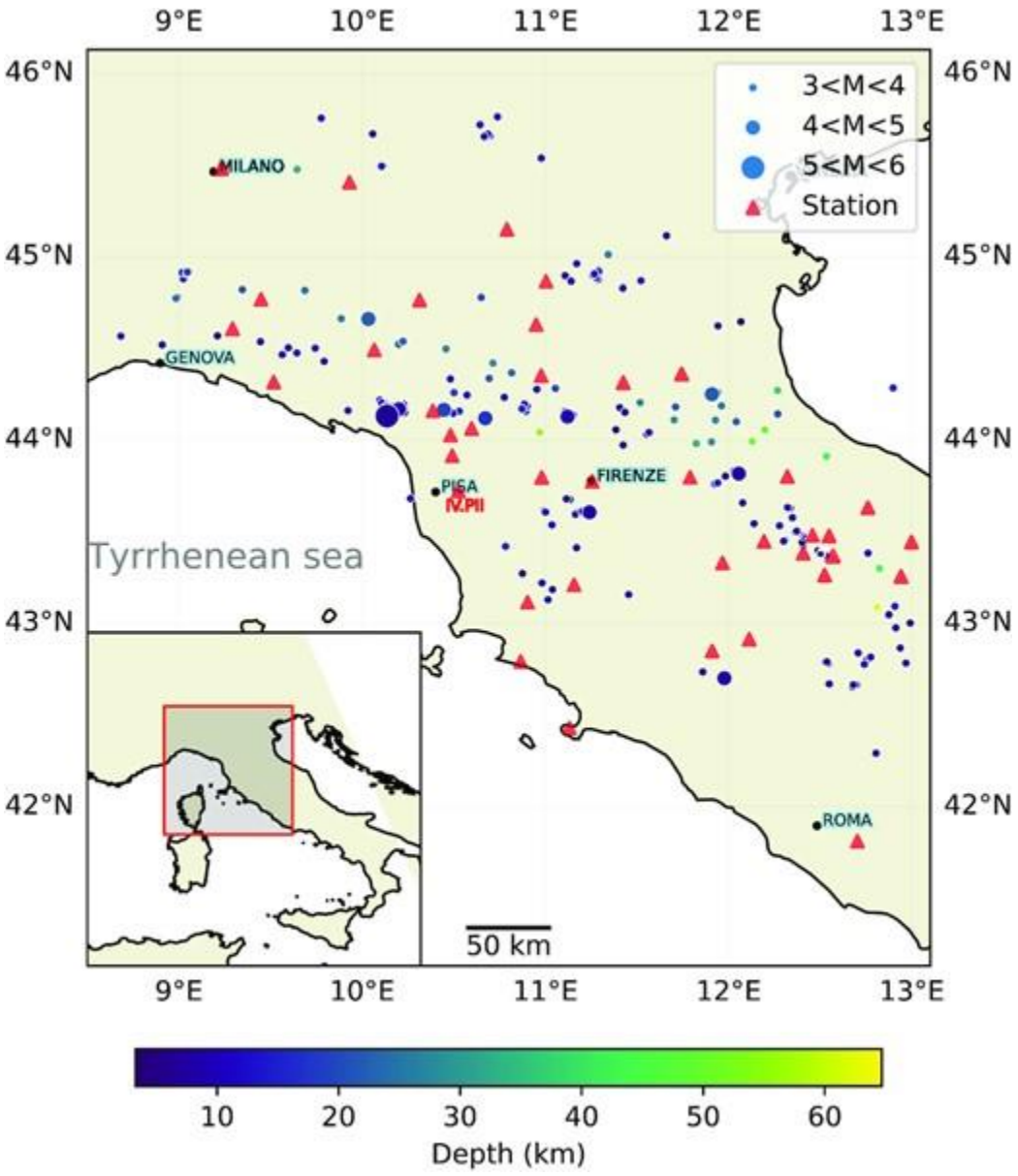


Table 1.

PGA residual R statistics for the four experiments on CW data set.

Experiment	Median	Mean	STD
No TL	−0.061	−0.029	0.508
CI → CW	−0.048	−0.019	0.474
STEAD → CI → CW	−0.039	−0.019	0.479
STEAD →CI → CW + add. data	−0.037	−0.013	0.449

The three component, 3C, waveform data were downloaded using the INGV FDSN web services for HN* (acceleration) and HH* and EH* (velocity) channels, where * ∈ [E, N, Z]. The data were processed to remove the instrument response, velocity data were differentiated to acceleration and, if necessary, the data were resampled to 100 Hz. For $M < 4$ earthquakes, the HH and EH channels were used after differentiation and for earthquakes with $M \geq 4.0$ the HN channels were used. For certain stations and for some earthquakes, the waveform data were completely missing and we chose to replace them with zeros, as in J2020. This data selection and processing follows the criteria outlined in J2020.

The target variables consisted of the IMs associated with each recording: PGA, PGV and SA at 0.3, 1 and 3 s periods, which were extracted from windows that started 5 s before the predicted P-arrival and ended 10 s after the predicted S-arrival (please note that this window size is not related to the window size of the CNN input used in the experiments and mentioned in the later chapters of the study). For stations that had no data, the IMs were calculated using the USGS ShakeMap software using the latest

configuration for Italy (Michelini et al. 2020) to ensure no missing output data (target variables). Since we are using a fixed number of stations for the input, we must always provide an array of fixed size (cf. J2020). However, sometimes input data is missing, in this case we use an array of zeros to fill the missing data (Garcia-Laencina et al. 2010), which is a natural way to employ the station dropout technique (Kriegerowski et al. 2019). We have a fixed set of outputs of the model for which we need to provide training and validation target data. Although using the ShakeMap approximations introduces further assumptions into the model, we consider it the best way to fill the missing data because we want our model to provide approximations for a site even when the input data are missing (which could be important for e.g. EEW purposes). This resulted in the following composition of target values: 91.4 per cent were observed, while 8.6 per cent were calculated using ShakeMap. We have calculated the first P arrival times on the stations from the theoretical travel times using the Java TauP Toolkit by Crotwell et al. (1999) that calculates theoretical travel times and paths as implemented in the Obspy Python library (Krischer et al. 2015) and the ak135 velocity model (Kennett et al. 1995). We performed a visual check of several waveforms and it was found that the theoretical travel times calculated were satisfactory to the purpose of this study.
We use two other data sets for pre-training the models for TL. The first data set is the J2020 central Italy data set (CI hereinafter) consisting of 915 earthquakes recorded by 39 stations with the same sampling rate and target labels as provided in this study. CI has essentially the same structure as the data set from this study except for a different set of recording stations.

The other data set used consists of a globally distributed set of local earthquake waveforms STEAD (Mousavi et al. 2019). This data set is much larger than CI, and it provides 3C single station earthquake waveforms, that is the data set structure is different from the data set of this study. In particular, the waveforms provided by STEAD are not recordings from a fixed set of stations. This essentially means that we cannot use STEAD to pre-train a CNN model that has the same, multistation, architecture for IM prediction that we use in this study although they can be used in the first layers of our CNN model as explained below. STEAD has a sampling rate of

100 Hz and the amplitude units of data counts. The maximum epicentral distance is 350 km. We used only the earthquake waveforms with magnitude $M \geq 3$ (the criterion also used for preparing the data set for this study) providing 106 245 waveforms from STEAD. Detailed information about STEAD can be found in the respective article.
We also checked other available data sets, such as LEN-DB (Magrini et al. 2020). However, we chose STEAD because it is a global data set having the data sampled using the same sampling rate as the CI and CW data (100 Hz).