IL NUOVO CIMENTO **45 C** (2022) 82
DOI 10.1393/ncc/i2022-22082-y

COMMUNICATIONS: SIF Congress 2021

Machine learning analysis of a local seismic network in Mt. Amiata (Italy)

S. GAVIANO([1])([2])

([1]) *Dipartimento di Scienze della Terra, Università di Firenze - Firenze, Italy*
([2]) *INGV, Sezione di Pisa - Pisa, Italy*

received 16 February 2022

Summary. — Since March 2016 a small network of 11 seismic stations, deployed by Istituto Nazionale di Geofisica e Vulcanologia (INGV), has recorded about 1000 earthquakes in the southern part of Mt. Amiata. The continuous seismic waveforms are reprocessed with phase recognition pickers based on machine learning (ML) algorithms trained with global datasets of local earthquakes to get a more comprehensive earthquake catalog. This new catalog is compared with the already available events picked and located manually to assess the performance of ML-based analysis. The manually detected earthquakes are then used to assemble a dataset suitable for ML analysis. In a later stage, we investigate how the automatic detection performance could be further enhanced with specific training of the ML pickers with data coming from the INGV network (INSTANCE dataset) and from the local network itself.

1. – Introduction

The recognition of P and S waves arrival times (phase picking) is crucial for the detection and location of earthquakes. Until now, different algorithms have been developed in order to automate the process, but they face different limitations that make human review of the data still mandatory. An example of an automatic method is the short term average to the long term average (STA/LTA). It bases the detection on the ratio between the energy computed on a short time period with respect to the one over a longer time period. This method works for high signal to noise ratio, but also produces many false positives. Another example of an automatic method is the template matching technique. Based on the fact that the activation of the same fault generates alike shaped signals, the template matching computes crosscorrelation between the data and the waveforms of events already cataloged (templates) to reveal similarities [1]. This approach can identify an order of magnitude or more of new events than classical methods. Its weak point is its blindness to never seen before shapes and, for example, cannot be implemented in areas

that have never been monitored and have been previously inactive, or to large magnitude earthquakes.

The increased number of sensors installed in the last years and the progressively increasing rate of acquisition determined an exponential growth of the amount of seismic data to analyze. Methods already available deal with limitations no longer affordable in order to handle these data efficiently.

A novel approach is offered by Deep Learning (DL), a new branch of Machine Learning (ML) that spread in the last few years. These data-driven methods are based on a succession of layers of progressively higher level of representation that are learned via models called neural networks (NN) [2]. Once exposed to a great amount of high quality examples (training dataset), these algorithms are able to find the statistical structure and learn the rules that govern the input data. A key role in the training success is due to the non-linearity introduced by the activation functions of the layers. If only linear transformations are possible, the final result of the operations performed over the stack of layers would still implement a linear transformation. In this way, no benefit would be gained by the multiple layer representation. The learning process is controlled by a loss function that evaluates the distance between the current and the expected output. The value obtained is then used as a feedback to update the model parameters. Once the training is over, the networks are able to infer original answers on new data. The model performance is strongly influenced by the architecture (*i.e.*, the type and succession of the different layers) and by the dataset they have been trained with.

Examples of the tasks commonly implemented by these algorithms are classification, regression and sequence prediction.

Two of the latest developed DL codes that perform picking are the Generalized Phase Detection (GPD) [3] and EQTransformer (EQT) [4]. Their neural networks are deeply different and have been trained on different datasets, which make it difficult to compare their performances.

This article presents a test of the two NN on the continuous data recorded by 11 stations of a local seismic network deployed in an area characterized by low magnitude seismicity and low signal to noise ratio. Results are then compared against the detections available in a referee catalog reviewed by expert analysts. The data collected by the network and the picks detected by the seismologists are then used to create a small benchmark dataset that follows the SeisBench standardization guidelines [5]: the AMIATA dataset.

2. – Data overview and the AMIATA dataset

The data processed in this paper have been recorded at Mount Amiata, in the province of Siena (Italy). In that area, at a depth of about 4–8 km, lays a seismic marker known as K-horizon (which corresponds to the 405°C isotherm) that highlights the top of the brittle/ductile transition. It is inferred that its seismic reflectivity is due to highly pressurized fluids [6]. The site also hosts two active geothermal fields for power production, in Bagnore and Piancastagnaio. These plants continuously extract fluids from the crust and are active since the early 1960s [7].

The low background seismicity is monitored by a small local network of 12 stations (TV) deployed by INGV and operative from March 2016 until December 2020. 8 sensors are Guralp CMG-6TD 30 s and work in the interval (1–100) Hz; the others are broadband working in the interval (0.033–100) Hz (Lennarz L1-3D $T = 1$ s). Other 15 stations of the Italian National Network (IV) are installed nearby. Expert analysts' revision of

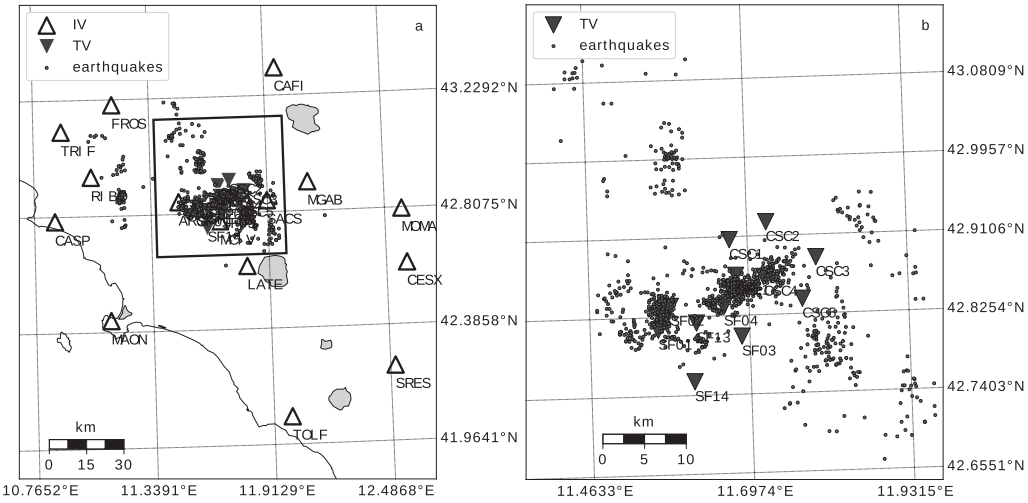

Fig. 1. – (a) Map of the locations of the earthquakes epicenters (circles) available in the referee catalog reviewed by skilled analysts (from 03-2016 to 01-2020). Downward triangles represent the seismic stations of the microseismic network (TV). Upward triangles represent seismic stations of the Italian National Network (IV). Panel (b) represents a magnification of the area present in the square in (a).

the seismograms reports 1515 earthquakes whose epicenters are shown in figs. 1(a) and (b). Seismicity is clustered into distinct focal volumes and two of them are close to the production wells [7]. The Local Magnitude of the events ranges between ($-0.2 \leq M_L \leq 3.6$) with a mean value of $\overline{M_L} = 0.6$ and a standard deviation of $\sigma = 0.5$.

In the time interval 03-2016 to 01-2020 earthquakes occurred at a rather constant rate of 1 event/day. Occasionally, small sequences of tens of earthquakes, lasting 1–3 days, occur and are not clearly tight to a clear mainshock-aftershock sequence. This could suggest that the geothermal exploitation plays a major role in the earthquake-generation process. However, until now, no other independent further data are available for a comparison and this relationship may not be investigated.

2˙1. *Amiata dataset.* – The development and comparison of DL models can be a difficult task. The quality of the training data is crucial to maximize the performance and the identification of the strengths and weaknesses of different algorithms is challenging, if they have been trained on different datasets. Therefore, the necessity of benchmark repositories emerges, designed with standard structures and made up of high quality data representing the greatest variability of earthquake features. A recent effort toward standardization is SeisBench (The Seismology Benchmark Collection), a toolbox that defines some guidelines to collect, analyze and compare different datasets, by unifying the application of machine learning pipelines to seismic data [5].

Some reference datasets already available are the Italian seismic dataset for machine learning (INSTANCE), made up of a collection of 1.2 million of noise and local earthquake waveforms recorded in the Italian peninsula [8], and the Stanford Earthquake Dataset (STEAD), that gathers 1.2 million seismic signals of regional earthquakes recorded all over the world [9].

I compiled the AMIATA dataset collecting the data recorded by the dense network and the nearby stations of the National Seismic Network. The low magnitude hand labelled events, characterized by low SNR and recorded in an area of continuous fluid extraction, represent a high quality reference set, suitable for the creation of a benchmark repository. The suite follows Seisbench guidelines and gathers a total of 13721 3-component velocity seismograms sampled at 100 Hz and 30 s long. Waveforms have a mean value of the SNR of (6 ± 7) dB. Together with the waveform dataset, 70 metadata parameters provide information about the earthquake source, the waveform characteristics, the recording station and other derived quantities. For each earthquake a mean of 9 station data are available. Since the dataset is probably not big enough for a full training it can be useful for transfer leaning, offering the possibilities of using the waveforms, the metadata, or a combination of the two data items to skill up the network abilities to recognize the features of these kind of events.

2˙2. *Dataset preparation.* – AMIATA is made up of an HDF5 file that assembles the earthquake waveforms and a csv file that collects the corresponding metadata. The structure of both follows the guidelines of standardization of SeisBench [5].

The catalog compiled by expert analysts and the continuous waveforms recorded by the 27 stations are available on a local INGV server, as well the stations information. The 30 s of each 3-component waveform contain only one earthquake and have been selected in order to randomly start between 5 and 2 s before the P arrival time if available (between 15 and 12 s before S arrival otherwise). The waveforms are preprocessed in order to fill the gaps through interpolation to avoid missing data, remove mean and linear trend, highpass filtering above 0.4 Hz and cosine tapered.

All sensors are already oriented along N-S and E-W directions and no rotation of the horizontal component is necessary.

70 metadata parameters are derived or computed on the 3-component waveforms. They provide a global view of the characteristics of the data and can be used for data selection or as labels in the training process. They give information about the source, such as its location, the origin time and its magnitude; the traces, by the extraction or calculation of a collection of parameters such as maximum amplitude and SNR; the path from the source to the station (the traveltime and the epicentral distance) and the station (name, network, channels and location coordinates).

Figure 2 shows the distributions of different metadata representing the earthquake sources.

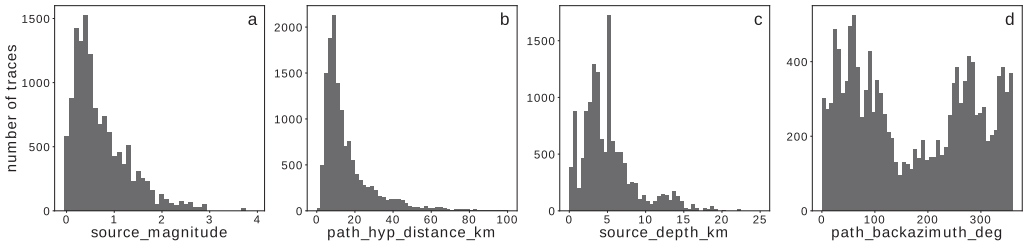

Fig. 2. – Distributions of the local magnitude M_L, the hypocentral distance (km), the depth (km) and of the backazimuth of the earthquakes (deg) of the AMIATA dataset.

3. – GPD and EQT comparison: seismic sequence analysis

The Gutenberg-Richter empirical power law [10] states a relationship between earthquake magnitude and frequency (1),

$$\log_{10} N = a - bM, \tag{1}$$

where N is the number of earthquakes with a magnitude greater or equal to M, a is the total number of earthquakes and b, also called *b-value* and variable between 0.8 and 1.2, is characteristic of each region and represents the relative number of large earthquakes compared with the small ones. For example, according to this formula, for each magnitude 3 earthquake, we expect 10 of magnitude 2 and 100 of magnitude 1. In order to properly describe the seismic activity of an area the detection of low magnitude earthquakes is paramount. These events are usually characterized by low values of SNR which makes them more difficult to recognize with respect to bigger ones. The lowest magnitude to which a network or a catalog includes all the earthquakes represents its completeness M_C [10]. In 2010, the completeness of the Italian National Seismic catalog was $M_C = 2.5$ (with some regional limitations), as reported by the Italian Civil Protection [11].

This paper presents a test of the performances of EQT and GPD on continuous data recorded in an area characterized by low magnitude seismicity ($M_L < 3.6$).

3˙1. *Pickers overview.* – Among the DL algorithms for picking released until now, two of the ones that show better performances are the Generalized Phase Detection (GPD) developed by Ross in 2018 [3] and EqTransformer (EQT) developed by Mousavi in 2020 [4]. The DL algorithm performance can be evaluated through their Precision and Recall defined as in equations 2, where t_p, f_p and f_n are, respectively, the true positives, the false positives and the false negatives.

$$Precision = \frac{t_p}{t_p + f_p} \qquad Recall = \frac{t_p}{t_p + f_n} \tag{2}$$

Precision represents the fraction of correct detections to all detections that have been made, and *Recall* is the ratio between the correct detections and all the detections that should have been made [3]. GPD accounts a $Precision > 99\%$ and a *Recall* between 96% and 99% for most threshold choices on the validating dataset, which means that noise is rarely labeled as seismic phases, but sometimes seismic phases are classified as noise [3]. EQT performance on a test set of 113 k waveforms, of both quakes and noise, presents both a *Precision* and a *Recall* of 99% [4]. However, since the two NNs have been trained and tested on different datasets, it is difficult to compare their performances.

The two models have deeply different architectures: the Generalized Phase Detection (GPD) is a convolutional neural network (CNN) made up of 4 convolutional (CONV1D) and 2 fully connected layers. The sequence of CONV1D takes as input a window of 4 s, 3-component ground velocity waveform and provides a progressively higher representation of the input by the extraction of representative features. The output of these layers is then concatenated into a vector and used as input of the fully connected layers that translate it into three forecast probabilities expressing the possibility for the input window to host noise, a P wave or an S wave. The detection of a phase is declared if the forecast probability is higher than a threshold defined by the user. The network has been trained

and validated on 4.5 million 3-component, hand labeled, data sampled at 100 Hz, of local earthquakes recorded by the Southern California Seismic Network ($-0.81 < M_L < 5.7$) (more information about the dataset are available in [3]).

EqTransformer (EQT) has a more complex architecture: the input is a three-channel ground velocity 60 s long waveform that feeds a succession of convolutional and residual layers with two BLSTM and an LSTM. Through a global view of the input, a first attention mechanism selects the areas that may be earthquake signals. These slides are then analyzed by other two attention mechanisms to recognize P and S wave arrivals. For each time instance of the input, the algorithm calculates a forecast probability associated to the presence of earthquakes, P and S waves: the detection is declared if the corresponding probability is equal or greater than the threshold defined by the user. EQT was trained on a subset of STEAD [9], a collection of 1.2 million hand-labeled, 60 s, 3 component waveforms of regional earthquakes from all over the world, whose magnitudes range between –0.5 and 7.9 (small earthquakes with $M_L < 2.5$ comprise the majority of the data set [9]).

3˙2. *Sequence analysis*. – In this section I test the performance of the two algorithms on a small sequence of 56 earthquakes occurred between December 16 and 31 2018. Figure 3(a) shows the seismicity rate between 2018-12-14 and 2019-01-03. The sequence presents two bursts of 13 and 19 earthquakes, with a duration of 10 and 7 hours, respectively, occurred between 2018-12-20T19:17:00 and 2018-12-22T06:55:00. The M_L of the events ranges between $-0.10 \leq M_L \leq 1.06$ and the epicenters lie inside the TV network (fig. 3(b)).

I run the pickers on the data of 11 stations of the TV network and compare the automatic detections with the 987 events already available in the reference catalog. Before feeding the networks, I preprocess the input data by demeaning, linear detrending and by filtering in the interval of (3–20) Hz for GPD and (1–45) Hz for EQT, according to the authors guidelines. I select a threshold of 0.95 for GPD and 0.01 for both phases and detections for EQT.

I compare the automatic picks with the reference ones by calculating the number of matches within $\Delta T = 0.2$ s for different thresholds. I choose this ΔT value by taking the sum of the maximum error associated to the picks of the reference catalog (0.1 s), and the uncertainty of GPD detections (0.1 s) resulting from the step width of the sliding window (EQT picking errors are comparable with the GPD ones). Figure 4 shows the

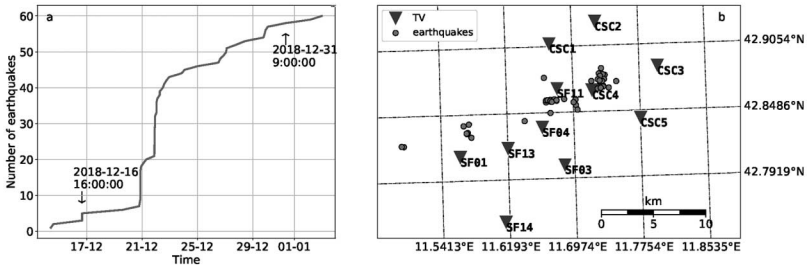

Fig. 3. – (a) Seismicity rate between 2018-12-14 and 2019-01-03. The two arrows indicate the start and the end of the time period of the sequence investigated (b) Location of the earthquakes of the sequence that happened between 2018-12-16T16:00:00 and 2018-12-31T09:00:00. Triangles indicate the stations of the microseismic network (TV).

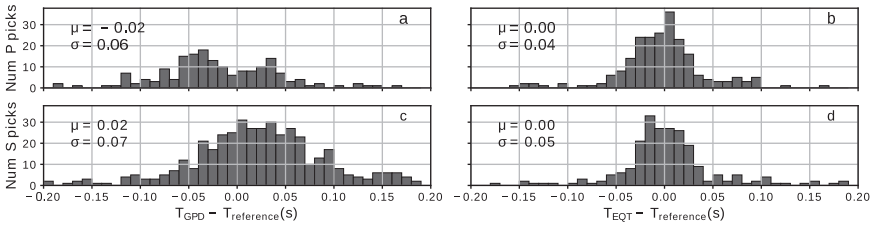

Fig. 4. – Time difference between automatic picks matching reference ones within 0.2 s. Results are shown for GPD ((a), (c)), and EQT ((b), (d)). The thresholds considered are 0.95 for GPD, 0.01 and 0.1 for EQT earthquake and phase detections.

TABLE I. – *Mean value calculated over the 11 stations of the TV network of the number of P and S detections. The first column is filled with the results extracted by the reference catalog (human). The detections of GPD and EQT are enumerated for different phase thresholds, for GPD; earthquake and phase thresholds for EQT.*

		GPD				EQT		
	Human	0.95	0.99	0.995	0.998	0.01 0.1	0.1 0.1	0.2 0.1
P	41	2261	428	198	71	85	76	60
S	44	12972	3935	2231	987	89	80	64

ΔT distributions for the P and S picks for both the algorithms: the plots are obtained using a threshold of 0.95 for GPD, and 0.01 and 0.1 for the earthquake and phase forecast probability thresholds of EQT.

In table II are the fractions of the automatic picks matching the ones of the reference catalog for the two algorithms and for different threshold values (the results obtained running EQT with earthquake and phase detection thresholds 0.01 are equal to the ones with 0.01 and 0.1 thresholds).

Figure 5 shows three examples of run of EQT with thresholds 0.2 for earthquakes and 0.1 for phase detections. The first three rows of each column represent the three

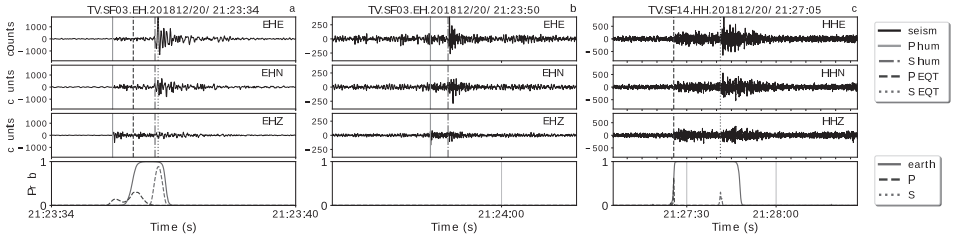

Fig. 5. – Examples of EQT performance with earthquake and phase probability detection thresholds of 0.2 and 0.1. The first three rows in each column represent the three-component velocity seismograms(in counts). Solid and dashdot vertical lines indicate P and S arrivals recognized by human analysts. Dashed and dotted lines are P and S detections of EQT. In the fourth row, the three lines represent the prediction probabilities associated with the earthquake (solid line), the P (dashed line) and the S (dotted line) phase detections.

component velocity seismograms. The vertical lines indicate the detection of the P and S phases performed by humans (solid and dashdotted lines), and the predictions of EQT (dashed and dotted lines for P and S arrivals, respectively). The fourth row shows the forecast probabilities associated to the presence of an earthquake, of a P or an S-wave in the seismograms. Figures 5(a) and (b) represent the data of two events reported in the reference catalog: EQT has successfully recognized the first, but not the second one. The data in fig. 5(c), recorded by SF14, show a new earthquake detected by EQT and not previously identified by humans. Figure 6 shows the performance of GPD on the same data of fig. 5. Most of the GPD predictions are false positives for all the three cases.

3˙3. *Results and discussion.* – The number of detections retrieved by GPD and EQT differ significantly (table I): on one side, EQT picks twice as much P and S waves than those listed in the reference catalog. On the other side, the number of detections of GPD varies significantly as a function of the threshold. In particular, for a threshold of 0.95 and with respect to the reference catalog, it picks 54 and 295 times more P and S waves respectively, but it also finds more false positives. These values show that GPD is more sensitive than EQT. This is probably due to the length of the GPD input window (4 s). The double attention mechanism approach of EQT evidently generates less false positives. It also becomes apparent the importance of selecting the appropriate threshold. In order to better understand which value to choose, I analyzed the number of automatic picks matching the ones in the available reference catalog within $\Delta T = 2$ s. The results reported in table II show strong variations between the fractions of the picks retrieved for the stations by both algorithms that, apparently, are not related to the different sensors, but may be related to the installation site. GPD seems to better recognize S phases than P, EQT shows a better balance, as a consequence of its detection approach. Generally, low number of matches obtained by one algorithm do not imply similar values for the other (for example P detection on SF04 and CSC4): some feature of the signal, such as the SNR or artificial disturbances, may influence differently the predictions on the two platforms. The performance of the pickers appear to vary between stations, for example GPD retrieves the 0.96 of the S human picks on SF14 with a threshold of 0.998, and the 0.44 with SF03. Higher threshold usually guarantee more selection and, in order to find a balance between true and false positives, a different value may be selected for each station.

Concerning the quality of the matching picks, figs. 4(a) and (c) show the distributions of the time differences between the automatic and the corresponding human picks. In

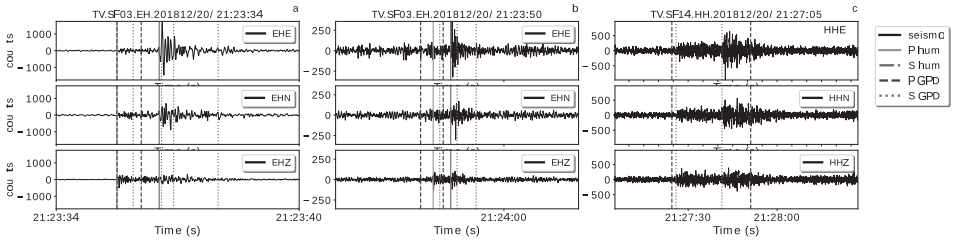

Fig. 6. – Examples of GPD performance with phase probability detection thresholds of 0.99. The three rows in each column represent the three-component velocity seismograms (counts). Solid and dashdot vertical lines indicate P and S arrivals recognized by human analysts. Dashed and dotted lines are P and S detections of GPD.

figs. 4(a) and (c) we can see that GPD often anticipates the detection of P phases and postpones the detections of the S. Results obtained selecting higher thresholds (but not shown in this paper) do not differ significantly, suggesting that differences may depend on other reasons as, for example, the filtering band of the input signals or the training. EQT distributions shown in figs. 4(b) and (d) are narrower and centered around 0 as its pick times are more in accord with the catalog ones.

The examples in figs. 5 and 6 and the data in tables I and II show the performances of the two algorithms on the seismicity of Mount Amiata, characterized by earthquake with low M_L and SNR (∼6 dB). On one side, EQT is able to retrieve a mean fraction of 0.48 of the human detections (depending on the threshold), but is also robust in the prediction of unseen events with fewer false positives with respect to GPD. GPD appears not to perform as well: the fraction of human detections retrieved is of 0.23 for a threshold of 0.99 with a higher number of false positives. The mean value of the fraction of S is higher (0.82), but also the number of false detections. As the ground truth of the area is

TABLE II. – *Fraction of automatic P and S picks that match the ones in the reference catalog for different stations and DL pickers. For each algorithm are showed the values for different thresholds of forecast probabilities: the phase threshold for GPD, both earthquake and phase detection thresholds for EQT.*

		GPD				EQT		
	Station	0.95	0.99	0.995	0.998	0.01 0.1	0.1 0.1	0.2 0.1
P Picks	SF14	0.85	0.67	0.61	0.43	0.70	0.59	0.54
	SF13	0.33	0.13	0.08	0.06	0.58	0.52	0.44
	SF11	0.23	0.11	0.06	0.04	0.34	0.32	0.26
	SF04	0.07	0.02	0.02	0.02	0.50	0.41	0.39
	SF03	0.54	0.38	0.27	0.10	0.27	0.27	0.27
	SF01	0.55	0.45	0.30	0.15	0.75	0.60	0.55
	CSC5	0.26	0.06	0.03	0.0	0.37	0.34	0.29
	CSC4	0.16	0.02	0.02	0.0	0.59	0.57	0.51
	CSC3	0.50	0.26	0.24	0.16	0.55	0.47	0.42
	CSC2	0.34	0.17	0.14	0.07	0.34	0.34	0.31
	CSC1	0.46	0.22	0.14	0.04	0.40	0.30	0.26
	mean	0.39	0.23	0.17	0.10	0.49	0.43	0.37
S Picks	SF14	0.96	0.96	0.96	0.96	0.64	0.55	0.49
	SF13	0.95	0.95	0.93	0.89	0.62	0.53	0.45
	SF11	0.92	0.82	0.71	0.59	0.33	0.27	0.25
	SF04	0.68	0.46	0.22	0.06	0.46	0.40	0.38
	SF03	0.73	0.65	0.50	0.44	0.29	0.29	0.29
	SF01	0.90	0.90	0.86	0.86	0.71	0.57	0.52
	CSC5	0.94	0.94	0.92	0.86	0.39	0.36	0.31
	CSC4	0.96	0.96	0.88	0.63	0.61	0.57	0.51
	CSC3	0.90	0.90	0.90	0.88	0.50	0.48	0.35
	CSC2	0.67	0.67	0.64	0.58	0.27	0.27	0.24
	CSC1	0.81	0.80	0.76	0.67	0.31	0.28	0.24
	mean	0.86	0.82	0.75	0.67	0.47	0.42	0.38

not available, it is not possible to quantify the real number of false positives. However, in the future, a better selection may be obtained associating and locating the automatic picks, as the phases concurring to the determination of physically plausible hypocenters are true.

Both the algorithms have been trained also on earthquakes with magnitudes in the same range as Amiata's, however results may vary depending on the balance of the training datasets with respect to the earthquakes features and sites characteristics. Transfer learning may improve the performance on the methods through further training on data that may have specific features, such as low SNR or proper peculiarities of an area. These are some of the reasons why I created the Amiata dataset and I look forward to using it in the application of transfer learning to EQT.

* * *

Data used in this article have been provided by INGV, Sezione di Pisa. I am grateful to Damiano Biagini and Marco Capello for the field work in Amiata. I thank Rebecca Bruni, Michele D'Ambrosio, Carlo Giunchi, Davide Piccini and Gilberto Saccorotti for providing phase picks and earthquake catalogs. I would also like to thank Alberto Michelini who provided some useful notebooks.

REFERENCES

[1] Gibbons S. J. and Ringdal F., *Geophys. J. Int.*, **165** (2006) 149.
[2] Chollet F., *Deep Learning with Python* (Manning) 2018.
[3] Ross Z. E., Meier M. A., Hauksson E. and Heaton T. H., *Bull. Seismol. Soc. Am.*, **108** (2018) 2894.
[4] Mousavi S. M., Ellsworth W. L., Zhu W. *et al.*, *Nat. Commun.*, **11** (2020) 3952.
[5] Woollam J., Münchmeyer J., Tilmann F., Rietbrock A., Lange D., Bornstein T., Diehl T., Giunchi C., Haslinger F., Jozinović D., Michelini A., Saul J. and Soto H., *Seismol. Res. Lett.*, **93** (2022) 1695.
[6] Brogi A., *Int. J. Earth Sci.*, **97** (2008) 677.
[7] D'Ambrosio M., Giunchi C., Piccinini D., Bruni R. and Saccorotti G., *Seismicity in the geothermal area of Mt. Amiata Volcano (Italy), EGU General Assembly 2020, Online, 4–8 May 2020, EGU2020-9942*, https://doi.org/10.5194/egusphere-egu2020-9942.
[8] Michelini A., Cianetti S., Gaviano S., Giunchi C., Jozinović D. and Lauciani V., *Earth Syst. Sci. Data*, **13** (2021) 5509.
[9] Mousavi S. M., Sheng Y., Zhu W. and Beroza G. C., *IEEE Access*, **7** (2019) 179464.
[10] Shearer P. M., *Introduction to Seismology*, Second edition (Cambridge University Press) 2009, p. 245.
[11] Schorlemmer D., Mele F. and Marzocchi W., *J. Geophys. Res. Solid Earth*, **115** (2010) B04308.