Insights into a Deep Learning Workflow for Automatic Detection of Seismic Signals at Mount Etna Volcano

OPEN ACCESS

Andrea Carducci[*1], Ornella Cocina[2], Mariangela Sciotto[2], Andrea Cannata[2,3], Serafina Di Gioia[4], Alessandro Vuan[1], Angela Saraò[1], Ken Tanaka Hernández[1,5], and Monica Sugan[1]

Abstract

We present a workflow for the automatic detection of seismic events in volcanic areas, leveraging deep learning (DL) tools for the picking and association of seismic phases. Specifically, we employ PhaseNet and EQTransformer via the SeisBench open-source platform to perform automated picking of *P*- and *S*-wave arrivals from continuous seismic waveforms. To evaluate performance, we systematically test several pretrained models available within SeisBench using seismic data from Mount Etna (Italy), encompassing both volcano-tectonic (VT) and long-period (LP) events recorded from January 2019 to June 2020. Each model configuration is quantitatively assessed through statistical analyses, comparing results against existing seismic catalogs to identify the most effective approach for the Etna region. Our findings show that models trained on volcano-specific datasets significantly outperform those based on tectonic-only data. To further classify VT and LP signals, we apply a frequency-based algorithm using a modified frequency index, which effectively distinguishes VT, LP, low spectrum, mixed, and unclassified type events based on their spectral characteristics. This classification demonstrates strong agreement with the reviewed catalogs. Overall, the proposed workflow performs well in the complex setting of Mount Etna, highlighting the potential of pretrained DL models for semi-automated volcanic monitoring. At the same time, our results underscore the need for site-specific model validation and expert oversight to ensure reliable outcomes.

Cite this article as Carducci, A., O. Cocina, M. Sciotto, A. Cannata, S. Di Gioia, A. Vuan, A. Saraò, K. Tanaka Hernández, and M. Sugan (2026). Insights into a Deep Learning Workflow for Automatic Detection of Seismic Signals at Mount Etna Volcano, *Seismol. Res. Lett.* **XX**, 1–18, doi: 10.1785/0220250431.

Supplemental Material

Introduction

Volcano monitoring is an ever-evolving task that plays a key role in our understanding of the various phases of volcanic activity and risk assessment (Scarpa and Tilling, 1996; Pallister and McNutt, 2015). One of the most critical components of this process is seismic monitoring, which requires acquisition and rapid processing of large volumes of data from dense seismic networks (Matoza and Roman, 2022, and references therein). In addition, seismic records from volcanic areas can exhibit complex waveforms that are often influenced by lithological heterogeneities, ambient noise, and resonance effects (e.g., Kumagai and Chouet, 2000; Chouet and Matoza, 2013). This complexity poses a significant challenge for seismologists who must accurately classify volcanic signals.

Machine learning (ML) techniques have emerged as powerful tools for assisting human analysts in identifying and classifying complex seismic patterns. Several approaches have successfully applied supervised and unsupervised ML algorithms to volcanic waveform classification, yielding promising results (e.g., Falsaperla *et al.*, 1996; Ohrnberger, 2001; Del Pezzo, 2003; Langer *et al.*, 2003; Alasonati *et al.*, 2006; Orozco *et al.*, 2006; Benitez *et al.*, 2007; Esposito *et al.*, 2008; Ibs-von Seht, 2008; Giacco *et al.*, 2009; Ibáñez *et al.*, 2009; Langer *et al.*, 2009; Köhler *et al.*, 2010; Hammer *et al.*, 2012; Malfante *et al.*, 2018; Abed *et al.*, 2026). More recently, deep learning (DL) models have been introduced for automated seismic phase-picking and signal-type classification (e.g., Ross *et al.*, 2018; Zhu and Beroza, 2019; Mousavi

1. National Institute of Oceanography and Applied Geophysics, OGS, Trieste, Italy, https://orcid.org/0000-0002-5598-4831 (AC); https://orcid.org/0000-0001-5536-1717 (AV); https://orcid.org/0000-0001-5734-5702 (AS); https://orcid.org/0009-0003-2730-3925 (KTH); https://orcid.org/0000-0002-1247-3193 (MS); 2. Istituto Nazionale di Geofisica e Vulcanologia (INGV), Sezione Osservatorio Etneo, Catania, Italy, https://orcid.org/0000-0003-1856-830X (OC); https://orcid.org/0000-0001-8711-1392 (MS); https://orcid.org/0000-0002-0028-5822 (AC); 3. Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Sezione di Scienze della Terra, Università di Catania, Catania, Italy; 4. International Centre for Theoretical Physics – ICTP, Trieste, Italy, https://orcid.org/0000-0002-3375-1850 (SDG); 5. Università degli Studi di Trieste, Trieste, Italy

*Corresponding author: acarducci@ogs.it

et al., 2020; Fee *et al.*, 2025), demonstrating high accuracy in both tectonic and volcanic environments.

SeisBench (Woollam *et al.*, 2022) is a modern open-source framework that facilitates the application of DL models to seismic waveform data. It provides access to multiple pretrained models, including those trained on both tectonic and volcano-specific datasets (e.g., Volpick: Zhong and Tan, 2024). Although these models have demonstrated strong generalization across various tectonic environments, their scalability and reliability when applied to local volcanic datasets remain underexplored. In this study, we examine how the nature of the training dataset influences DL model performance in a volcanic setting, with particular attention to model architecture, training data type (tectonic vs. volcanic), and waveform complexity under the noisy conditions typical of volcanoes. We benchmark several pretrained DL models, within the SeisBench framework, by applying them to continuous seismic waveforms recorded by the permanent network at Mount Etna (Italy) from January 2019 to June 2020. The workflow spans phase picking, event association, and event classification, and is optimized for high-performance computing (HPC) environments.

Our primary aim is to identify the DL picking models that maintain robust performance throughout the detection pipeline. We focus specifically on the detection of volcano-tectonic (VT) earthquakes and long-period (LP) events. VT earthquakes, typically associated with brittle fractures due to stress changes (e.g., Chouet and Matoza, 2013; Matoza *et al.*, 2014), exhibit well-defined *P* and *S* phases and broad frequency content, similar to tectonic signals. In contrast, LP events are usually associated with fluid pressurization processes, such as resonance of fluid-filled cracks or conduits triggered by pressure transients in the fluid, thermal stresses, volatile pressures, or rapid bubble formation during magma ascent (e.g., Chouet *et al.*, 1994; Chouet, 1996; Chouet and Matoza, 2013; Matoza *et al.*, 2015). The detection of LP events in the Mount Etna area is usually performed using stations located in the summit area of the volcano (Cannata *et al.*, 2013) because they capture shallow events beneath the central craters of the volcano (e.g., Patanè *et al.*, 2013; Cannata *et al.*, 2015; Sciotto *et al.*, 2022).

The different source mechanisms produce different spectral signatures: VT earthquakes typically have high frequencies and clear phase onsets, making them easier to detect, whereas LP events have lower frequency content, emergent phase onsets, and no distinct *S* phases, making them more difficult for phase pickers to detect (Zhong and Tan, 2024, and references therein). Despite the characteristics described earlier, the spectral content does not necessarily reflect the source mechanism, and for process discrimination, moment-tensor and single-force source representations should be considered (e.g., Saraò *et al.*, 2010, 2023; Matoza and Roman, 2022).

Given these spectral differences, we implement a frequency-based classification algorithm using a modified frequency index (FI) to automate the categorization of events as VT, LP, low spectrum (LS), mixed or unclassified types. In this study, the LS class corresponds to events with predominantly low-frequency spectral content and no detectable contribution from the summit stations in the FI calculation, which are not classified as LP. Mixed-type events occur when FI metrics give conflicting indications, so they cannot be distinctly classified. Events belong to the unclassified category when the FI cannot be calculated.

We then evaluate the performance of both the DL-based workflow and the FI classification scheme by comparing the resulting catalogs with reference catalogs for the Etna region.

Data

The waveform dataset used in this study was recorded by 38 seismic stations operated by the Istituto Nazionale di Geofisica e Vulcanologia (INGV), Sezione Osservatorio Etneo (OE), covering the period from January 2019 to June 2020. All stations are equipped with broadband, three-component seismometers, recording in real time at a sampling rate of 100 Hz. These stations are mainly located in the Etna volcano area, with a few situated in the surrounding sectors (Fig. 1). Four stations in the summit area of the volcano (ECNE, ECPN, EPDN, and EPLC) are used for the detection of LP events (Cannata *et al.*, 2013). To provide a robust reference for benchmarking and comparison with the results obtained in this study, we used two independent seismic event catalogs provided by INGV-OE: the VT earthquake catalog and the LP event catalog, both restricted to events located within the bounding box shown in Figure 1.

The VT catalog contains 4243 VT earthquakes, including origin times and hypocentral parameters, and manually labeled *P* and *S* picks (Alparone *et al.*, 2020; Barberi *et al.*, 2020). Events were localized using the Hypoellipse algorithm (Lahr, 1999). Figure S1, available in the supplemental material to this article, provides the distribution of the local magnitude ($0.2 \le M_L \le 4.1$) along with the horizontal and vertical errors. The 99th percentile localization errors are within 1.90 km horizontally and 2.60 km vertically. Hypocentral depth ranges from −1.60 to 38.24 km, with positive values indicating hypocenters below sea level and negative values indicating elevation above sea level.

The LP catalog contains 65,068 events triggered by the summit stations during the same period (Sciotto *et al.*, 2022; Cannata *et al.*, 2025), hereafter referred to as the unfiltered catalog. From this dataset, 30,741 localized events were selected, and an additional filter using an amplitude threshold was applied to ensure higher reliability, resulting in what we hereafter call the filtered catalog.

This dual-catalog approach ensures that both VT and LP event families are well represented, enabling robust benchmarking of the DL-based workflow across signal types with distinct spectral and temporal characteristics.

Methodology

We optimized a distributed workflow designed to automate seismic phase detection, association, and classification, using ML.

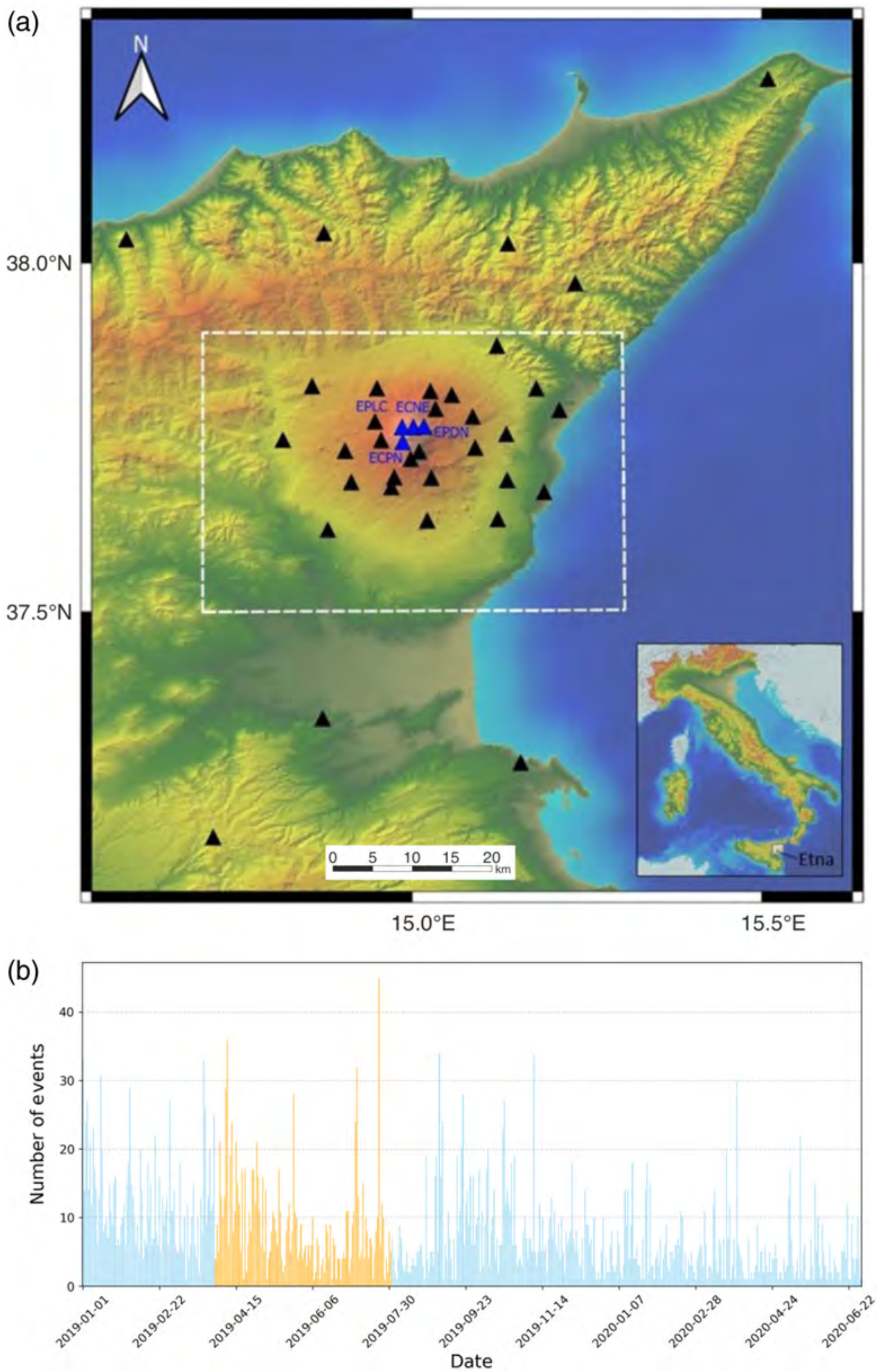


Figure 1. (a) Map of the seismic stations used in this work (triangles). Blue triangles indicate summit stations. The white dashed box outlines the boundaries of the Etna area as defined by INGV-OE for monitoring purposes. The bottom-right inset shows the location of Mount Etna within the Italian territory. (b) Daily histogram of the volcano-tectonic (VT) earthquakes in the Etna area (Istituto Nazionale di Geofisica e Vulcanologia [INGV] - OE VT catalog); the April–July 2019 test time frame is evidenced in orange. The color version of this figure is available only in the electronic edition.

The process consists of three primary stages: (1) *P-S* phase picking from input waveforms, using the open-source tool SeisBench (Woollam *et al.*, 2022) and a set of pretrained ML-based picker models (Zhu and Beroza, 2019; Mousavi *et al.*, 2020); (2) association of the ML-predicted phases into seismic events using the Gaussian Mixture Model Association (GaMMA; Zhu *et al.*, 2022); and (3) event classification based on spectral characteristics using an FI approach (e.g., Buurman and West, 2010).

This pipeline was developed in Python and optimized for use on an HPC platform (CINECA Supercomputing Centre, SuperComputing Applications and Innovation Department, 2024) with multi-graphic processing unit (GPU) support, enabling accelerated processing of large seismic datasets. The overall speedup for total runtimes (input time + computing time + output time) reached up to ~4.5 when using multicore and multi-GPU configurations compared to nonparallelized execution. A flowchart summarizing the complete workflow is shown in Figure 2. Additional scripts include benchmarking tools for evaluating performance across key steps, filtering associated events, and automatically labeling them according to selected volcanic event classes.

Phase picking with SeisBench models

We used two state-of-the-art DL models available in SeisBench: PhaseNet (Zhu and Beroza, 2019) and EQTransformer (Mousavi *et al.*, 2020). Each model was evaluated across multiple pretrained

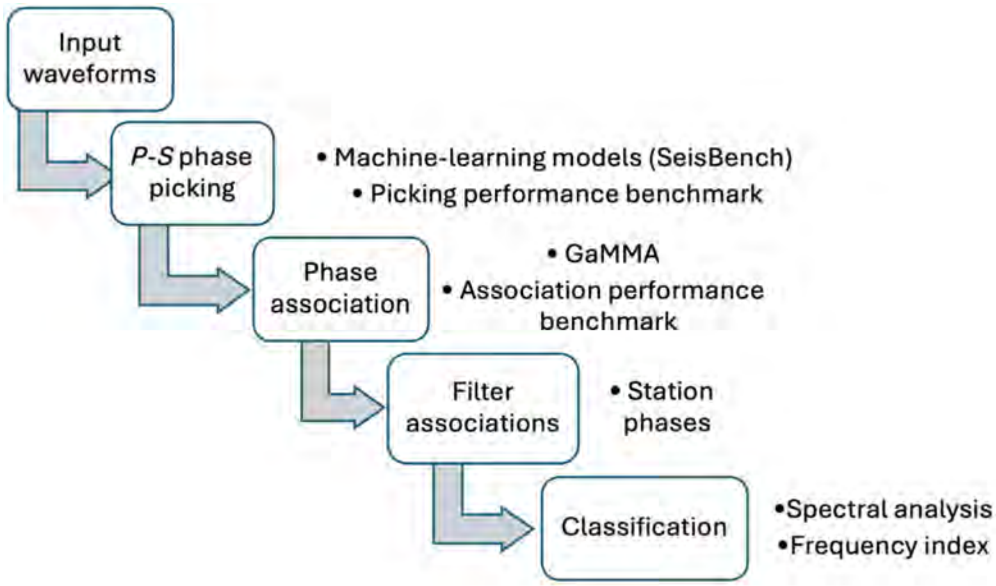


Figure 2. Flowchart of the optimized workflow for the automatic detection of seismic events in volcanic areas. GaMMA, Gaussian Mixture Model Association. The color version of this figure is available only in the electronic edition.

configurations (Table S1), which differ in training datasets, source–receiver distance regimes, and event types (tectonic or VT). Regarding the latter distinction, we consider datasets that mainly include seismic events related to brittle fractures as tectonic, as they represent typical tectonic environments and feature events with broad spectral content. Examples of these datasets are the STanford EArthquake Dataset (STEAD; Mousavi *et al.*, 2019), INSTANCE (Michelini *et al.*, 2021), the Southern California Earthquake Data Center (SCEDC) (2013), and the Northern California Earthquake Data Center (NCEDC) (2014) datasets. On the other hand, datasets compiled by balancing high-frequency earthquakes with the inclusion of LP events are referred to as VT, as they are representative of volcanic environments characterized by the presence of mixed signals. An example of such a dataset, used in this work, is the Volpick dataset (Zhong and Tan, 2024). To ensure relevance to the local volcanic setting of Mount Etna, we excluded models trained on teleseismic or distant regional datasets, following the spatial-domain selection criteria proposed by Münchmeyer *et al.* (2022). Only models trained on local datasets (0–150 km source–receiver distance) with abundant, well-distributed *P* and *S* picks were considered. In total, 10 combinations of pickers and pretrained models were included in our benchmark. The tested configurations included: PhaseNet Original (NCEDC, 2014) and EQTransformer Original (Mousavi *et al.*, 2019), PhaseNet and EQTransformer trained on INSTANCE (Michelini *et al.*, 2021), SCEDC (2013), STEAD (Mousavi *et al.*, 2019), and Volpick VT dataset (Zhong and Tan, 2024). Picks were extracted using a user-defined probability threshold. Lower thresholds increased the number of picks but also introduced more false detections. We tested probability thresholds starting from 0.2.

To identify the most effective phase-picking models, we first conducted a benchmark using seismic waveforms recorded over four months, between April and July 2019 (see Fig. 1b). This time frame is characterized by intense seismic activity in the INGV-OE VT catalog, resulting in a larger number of cataloged events than in other periods. This increases the probability of recording robust numbers of *P-S* phases in the Etna dataset, making it ideal for the first testbed of the 10 model configurations. Only the top-performing configurations from this initial evaluation were selected for subsequent application to the full 1.5-yr dataset and for event association.

Phase association

We associated the predicted *P* and *S* picks with seismic events using GaMMA (Zhu *et al.*, 2022), a Bayesian Gaussian Mixture Model (BGMM) framework. Unlike traditional grid-based associators (e.g., Weber *et al.*, 2007; Friberg *et al.*, 2010; Patton *et al.*, 2016), GaMMA treats phase association as a probabilistic clustering problem, incorporating expected travel times using a 1D velocity model. We used the same Etna-specific 1D velocity model (based on Hirn *et al.*, 1991 and Patanè *et al.*, 1994) as used in the INGV-OE VT catalog (Table S2). Travel times were calculated using an Eikonal solver. Associations were performed using the BGMM approach, with the Density Based Spatial Clustering of Applications with Noise (DBSCAN) clustering method (Ester *et al.*, 1996) used to divide long sequences of picks into segments. The minimum number of samples in a neighborhood for a point to be considered a core point was set to 5, which allowed the DBSCAN method to find dense clusters and thus ensured the robustness of the association. The maximum time between two picks for one to be considered as a neighbor of the other was calculated based on the distribution of stations and the velocity of *P* waves in the area. The oversampling factor to determine the number of initial clusters was set to 5, a value that balances the high number of predicted picks from the model configurations and the number of recording stations. The covariance prior factor of time residuals was set to 5 because higher values of this factor could determine the splitting of single events into multiple events with the same origin time. An event was accepted with a minimum of five associated phases ($P \geq 3$ and $S \geq 2$).

DL pickers typically generate a very high number of picks, especially when low-probability thresholds are used (e.g., Zhu and Beroza, 2019; Mousavi *et al.*, 2020). Although this reflects their enhanced sensitivity, many of these automatically predicted phases may still represent false detections.

To improve the reliability of automated event association, we require each event to be supported by a minimum number of stations with both *P*- and *S*-phase picks ("*P*&*S* stations"). Setting a minimum number of *P*&*S* stations helps balance the need to minimize false positives (FPs) with maintaining adequate sensitivity for event detection across the network, thereby improving the robustness of the automated association process.

Spectral classification

To distinguish between VT and LP events, we implement an automated classification method based on the FI. This parameter captures the dominant frequency content of seismic signals by comparing spectral amplitudes in two characteristic bands: a high-frequency band, prevalent in VT events, and a low-frequency band, typically prominent in LP events. Building on standardized FI-based approaches (Buurman and West, 2010; Matoza *et al.*, 2014; Song *et al.*, 2023; Zhong and Tan, 2024), we enhance sensitivity by calculating the FI as the logarithmic ratio of the maximum spectral amplitudes, rather than the mean values, within the selected frequency bands for each trace (see Text S1).

These frequency intervals were calibrated by averaging the stacked spectral amplitudes of waveforms representative of both VT earthquakes and LP events, following the approach of Buurman and West (2010). A trace is retained for FI calculation only if it meets a quality check based on signal-to-noise ratio (SNR $\geq$ 2.5).

For each detected event, FI values were calculated and averaged from up to four stations selected among the six closest to the epicenter. To ensure robust and unbiased FI calculations, we implemented an adaptive filtering workflow based on SNR (see Text S1 and Table S3). Each event was then classified into one of five categories LP, LS, VT, mixed, and unclassified based on the following criteria derived from the FI analysis: FI value (sign of the average log-ratio); FI-neg (the fraction of component FIs that are negative); and FI size (the number of components contributing to the average FI, serving as a measure of classification robustness).

Among the events characterized by a low-frequency content, the presence of components from summit stations was used as an additional criterion to qualify LP events. Summit stations (installed above 2500 m.a.s.l.) are typically the most sensitive to LP activity at Mount Etna, where LP sources are generally located at very shallow depths (1–3 km a.s.l.) within the volcanic edifice (e.g., Patanè *et al.*, 2013; Cannata *et al.*, 2018; Sciotto *et al.*, 2022). To the best of our knowledge, no deep LP events have been documented at Mount Etna, although more energetic LPs can also be recorded by more distant stations in the network.

Events with low spectral content that did not satisfy this spatial criterion were grouped in a class named LS.

Based on the available literature, the LS class could also include VT events characterized by low frequencies (1–5 Hz), emergent *P*-wave onsets, and poorly identifiable *S* phases. Similar features have been described for VT-B events (Wassermann, 2012) and interpreted as the effect of strong scattering during wave propagation through shallow crustal layers. At Mount Etna, where, as in other volcanic areas, shallow seismogenic structures are present, low-frequency VT events have also been documented and related to seismic-wave attenuation processes occurring in the uppermost crust (depth < 1 km; Patanè and Giampiccolo, 2004). Consequently, the LS class, without implying a specific source mechanism, includes all seismic events characterized by low-frequency content not belonging to the LP class.

Based on these criteria, each event was classified as follows:

- Pure LP if FI < 0 and FI-neg $\geq$ 0.5 and at least one of the summit stations in the INGV-OE seismic network (ECNE, ECPN, EPDN, and EPLC) recorded a component that contributed to the FI calculation of the event, indicating that spectral energy is concentrated in the low-frequency band and the event is spatially bound to the summit sector of the Etna area.
- LS if FI < 0 and FI-neg $\geq$ 0.5 and no component from summit stations contributed to the FI calculation of the event.
- Pure VT if FI $\geq$ 0 and FI-neg < 0.5, consistent with dominant high-frequency content.
- Mixed (FI $\geq$ 0 and FI-neg > 0.5 or FI < 0 and FI-neg < 0.5) in all other cases, where frequency characteristics could not clearly distinguish between the different classes.
- Unclassified if no FIs were calculated for any components or stations.

From this point forward, we refer to these categories (VT, LP, LS, mixed, and unclassified), as the event labels assigned through the combined FI and FI-neg classification scheme. In addition, the FI size parameter was used in the classification, as a larger number of contributing components (higher FI size) indicates greater confidence and robustness in the assigned label.

These labeled events were then compared with the available INGV-OE catalogs to evaluate the performance of the classification scheme.

Results

We evaluated various model configurations using the four-month (April–July 2019) data subset and a range of probability

TABLE 1

Metrics for Phase Picking Against Manual VT Phases from Pretrained Models, for Probability (Prob.) Thresholds ≥ 0.2 (Four-Month Test)

Model	Prob. ≥	*P* Recall (%)	*S* Recall (%)	TP *P*	TP *S*	True *P*, Pred *S*	True *S*, Pred *P*	FN *P*	FN *S*	FP *P*	FP *S*
PhaseNet INSTANCE	0.2	86.64	65.79	9,115	2,104	15	11	1,391	1,083	288,453	78,918
PhaseNet Original	0.2	95.53	80.93	10,051	2,588	24	2	446	608	663,289	743,700
PhaseNet SCEDC	0.2	83.99	34.55	8,837	1,105	260	47	1,424	2,046	1,377,506	492,615
PhaseNet STEAD	0.2	71.89	74.30	7,564	2,376	217	2	2,740	820	62,913	570,504
PhaseNet Volpick	0.2	94.94	80.71	9,989	2,581	52	5	480	612	1,228,179	983,164
EQTransformer INSTANCE	0.2	90.09	68.61	9,478	2,194	17	10	1,026	994	240,177	34,024
EQTransformer Original	0.2	57.55	60.44	6,055	1,933	31	2	4,435	1,262	48,674	39,831
EQTransformer SCEDC	0.2	91.47	28.92	9,624	925	231	14	666	2,259	1,908,564	500,093
EQTransformer STEAD	0.2	69.33	70.86	7,294	2,266	24	2	3,203	930	49,024	93,706
EQTransformer Volpick	0.2	95.85	80.05	10,084	2,560	43	8	394	630	1,342,183	1,293,393

P recall (%), recall score of predicted *P* phases; *S* recall (%), recall score of predicted *S* phases; TP *P*, true positives with phase *P*; TP *S*, true positives with phase *S*; true *P*, pred *S*, manual *P* phases with corresponding predicted *S* phases; true *S*, pred *P*, manual *S* phases with corresponding predicted *P* phases; FN *P*, false negatives with phase *P*; FN *S*, false negatives with phase *S*; FP *P*, false positives with phase *P*; FP *S*, false positives with phase *S*; and VT, volcano-tectonic.

thresholds starting from 0.2 to assess picker performance against the *P*- and *S*-phase arrivals listed in the INGV-OE VT catalog. The best-performing configurations were then applied to the full 1.5-yr dataset for phase picking. Event association was performed using filtering criteria based on the number of *P*&*S* stations. Detected events were subsequently classified according to their associated FI values. This integrated strategy enabled comprehensive benchmarking of model performance and robust validation of workflows within the complex seismic environment of Mount Etna.

Evaluation of phase-picking performance

To identify the best-performing phase-picking models, we benchmarked them against VT seismic waveforms recorded during the period of elevated activity in April–July 2019 (see Fig. 1b). We applied 10 pretrained configurations (combinations of PhaseNet and EQTransformer with different training datasets) to this subset and evaluated their outputs against manually labeled *P* and *S* picks. Picks were matched within a maximum time tolerance of 0.5 s, in accordance with previous benchmarks (e.g., Münchmeyer *et al.*, 2022).

In the following, the manual picks linked to the INGV-OE VT catalog are referred to as "true" picks, whereas the picks recognized and classified by DL models are referred to as "predicted" picks.

Evaluation metrics (Text S2) include, among others, true positive (TP, *P*, or *S*), false positive (FP, *P* or *S*), false negative (FN, *P*, or *S*), recall, precision, and F1 score (e.g., Powers, 2011), as well as the distribution of time differences between predicted and manual VT picks. Recall values were calculated separately for the *P* and *S* picks across different probability thresholds. The reference dataset is limited to well-characterized VT events, which have been manually revised to enable a robust evaluation of Recall. This subset provides a reliable measure of the model's ability to accurately detect *P*- and *S*-phase picks associated with VT events. However, seismic recordings at Mount Etna contain many additional signals not represented in the reference catalog. These include events occurring outside the spatial domain of the catalog (Fig. 1) and other event types, such as LP events (which can outnumber VT events by at least an order of magnitude) and VT-B events, which may also exhibit *P*- and *S*-phase arrivals. When the model detects such signals, they are counted as FPs. As a consequence, the reported FP and derived precision or F1-score are strongly influenced by the characteristics of the VT reference catalog rather than by actual model misdetections. Therefore, both precision and F1-score are systematically underestimated in this study. Recall remains the most robust and interpretable metric for evaluating model performance in this context, whereas precision and F1-score are reported for completeness in the supplemental material but should be interpreted with caution.

Among the 10 models evaluated, three configurations demonstrated superior performance when a probability threshold of at least 0.2 was applied (see Table 1). Specifically, EQTransformer Volpick achieved recall rates of 95.85% for *P*-phase picks and 80.05% for *S*-phase picks. Similarly, PhaseNet Volpick yielded recall values of 94.94% for *P* phases and 80.71% for *S* phases. The original PhaseNet configuration also performed strongly, with recall rates of 95.53% for *P*-phase and 80.93% for *S*-phase detection.

Figures S2 and S3 show the distributions of time difference between the true and predicted picks for the four-month analysis period, using PhaseNet and EQTransformer, respectively. The distributions are shown as a function of probability thresholds, separately for *P* and *S* phases, with bin widths of 0.02 s. Notably, the EQTransformer Volpick and PhaseNet Volpick models exhibit more symmetric distributions around zero, particularly for *P* phases, indicating better temporal alignment. Figures S4 and S5 show precision and F1-score for the same four-month analysis period and different probability thresholds, for completeness.

Furthermore, statistical descriptors (see Table S4) show that EQTransformer Volpick has mean time differences closest to zero for both *P* and *S* phases (*P*: −0.003 s and S: −0.012 s), followed by PhaseNet Volpick (*P*: 0.006 s and S: −0.004 s). PhaseNet Original shows higher mean time differences (*P*: −0.023 s and *S*: −0.075 s), suggesting a systematic temporal bias, with predicted picks slightly delayed relative to the true ones. For all three models considered, the recall scores consistently exceed 80% for probability thresholds ≥ 0.2 (Table 1).

Table S5 also shows that up to probability thresholds ≤ 0.5, these three models still show recall values of at least ∼65% in both *P*- and *S*-phase distributions, whereas the other models obtain significantly lower scores at the same probability levels.

Based on the benchmark, we selected the three top-performing models with recall for *P* and *S* picks above 80% and applied them to the full dataset spanning January 2019 to June 2020. Table S6 summarizes the recall scores for each configuration at the same probability threshold (P ≥ 0.2).

The results confirmed the preliminary findings: all three models maintained high *P*- and *S*-phase recall scores with minimal performance degradation when scaling up. The recall values at threshold 0.2 closely match those reported in Table 1 at the same probability levels, confirming the preliminary test observations. Figure 3 shows the time-difference distributions for the full dataset. The shapes and skewness are again similar to the four-month sample (Figs. S2 and S3): the *P* picks are more symmetrical around zero difference than the *S* picks, with the latter showing slight skewness toward positive values for the Volpick configuration. The negative shift observed for the PhaseNet Original configuration is confirmed.

An abrupt reduction in the number of detected picks occurs when the probability threshold increases from 0.2 to 0.3 (Table S6). Importantly, this transition results in only a marginal decrease in recall—less than 1.2% for *P*-phase picks and 2.2% for *S*-phase picks. This small reduction in recall demonstrates that nearly all true phase arrivals are still retained at the higher threshold, whereas a substantial portion of low probability, likely spurious picks are excluded. By setting the threshold at ≥0.3, false detections that might otherwise result in erroneous or artifactual associations during the subsequent event-building process can be effectively reduced. This higher threshold maintains sufficient sensitivity for robust phase detection while substantially improving the precision of the picks used in event association.

Phase association accuracy with GaMMA

We used GaMMA to associate the predicted picks of the three top-performing models into events using a minimum pick probability of 0.3. The resulting catalogs were compared with the INGV-OE VT catalog (Table 2).

A maximum spatial tolerance of 4 km in epicentral distances was adopted, approximately twice the horizontal uncertainty of manual VT location (see Fig. S1). This tolerance accounts for the fact that absolute locations derived from the GaMMA associator are being compared with manual locations. The maximum allowed origin-time difference was set to 1.5 s. All manual and predicted earthquakes are considered only within the study area (Fig. 1).

Comparison with the INGV-OE VT catalog (Table 2) shows that all models achieved recall rates above 84%, with EQTransformer Volpick performing best (86.50%), followed by PhaseNet Volpick (84.99%), and PhaseNet Original (84.31%). Missed events refer to true VT events that, according to our matching strategy, were not associated with any detected events. Some of these may still be present in the original catalog, but with mismatched origin times or locations. These missed events are sparsely distributed across the study region. As shown in Table 2, the EQTransformer Volpick model performs best in terms of the recall score. To reduce the number of potential FPs, we applied a filtering criterion based on the minimum number of *P&S* stations with a probability value ≥ 0.3. This filtering step significantly improved the results while maintaining a recall above 81% (Fig. 4) with a minimum of three *P&S* stations. After filtering, the number of FPs decreased markedly for the Volpick models (Table 3), and the spatial distribution of the remaining events became more consistent and closely matched that of the manual VT event map (Fig. 5). This confirms the effectiveness of the filtering strategy in retaining high-confidence detections. Figures 4 and 5 show that using a minimum of three *P&S* stations provides the best balance between maintaining a robust recall score and achieving a spatial distribution consistent with the manual VT catalog. Further refinements to the filtering strategy are required, as the distribution of events in Figure 5 does not reflect the seismicity distribution at Etna as reported in the literature (Firetto Carlino *et al.*, 2022; Sciotto *et al.*, 2022; and references therein).

Classification of events via FI

To classify events as VT, LP, LS, mixed, and unclassified, we employed an FI-based approach. To calibrate the FI frequency bands, we analyzed the mean amplitude spectra from a subset of about 100 manually selected VTs and 100 LP events (Fig. 6), with high SNR and high-quality waveforms. The selected VT earthquakes are at depths of −1 to 10 km, whereas the selected LP events are at depths of −3 to 0 km. As shown in Figure 6a,b,

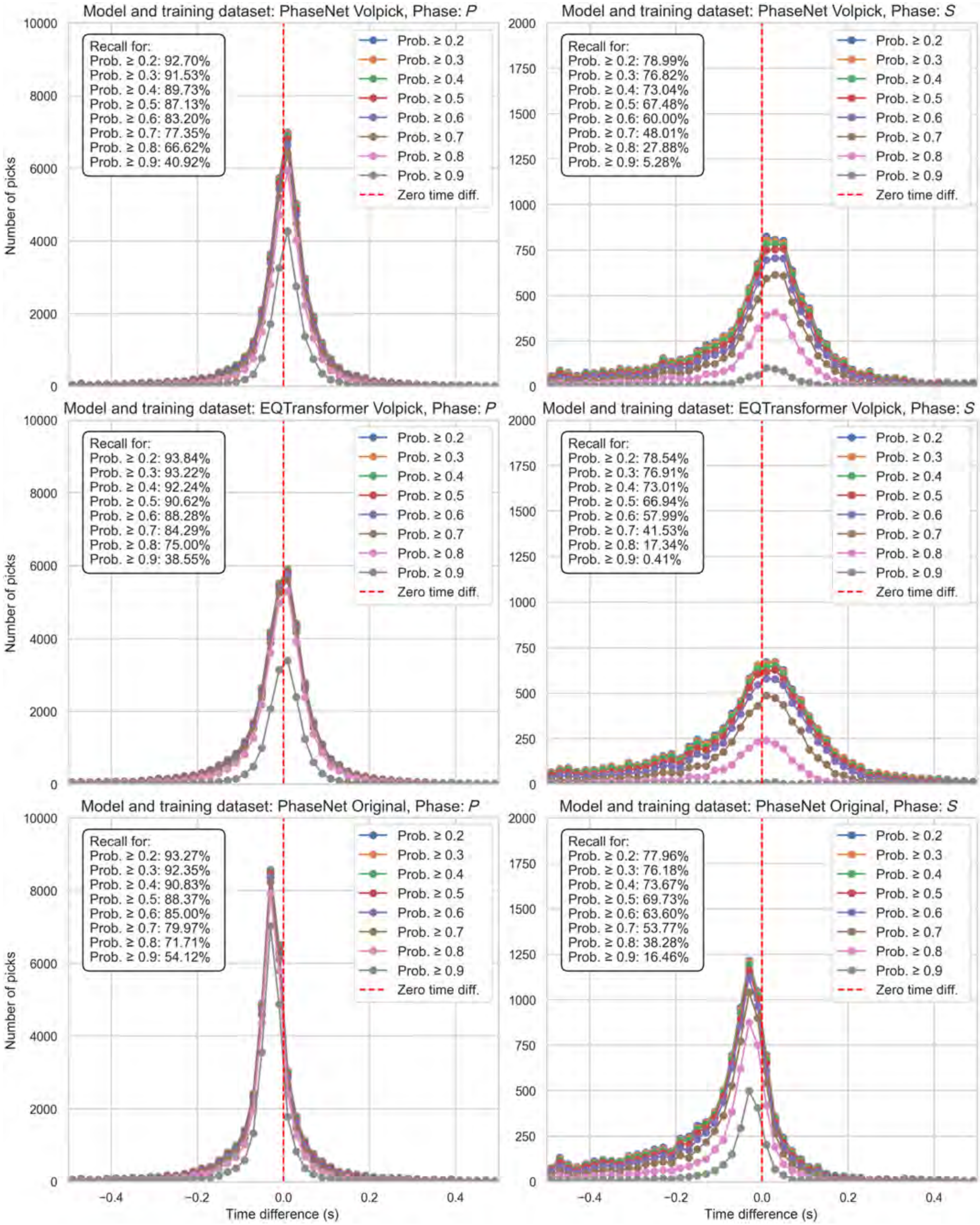


Figure 3. Time difference distributions between manual and predicted VT phases (1.5-yr dataset) for models, and training datasets: PhaseNet Volpick, EQTransformer Volpick, and PhaseNet Original. Recall scores for each probability threshold are shown in the upper-left corner of each panel. The color version of this figure is available only in the electronic edition.

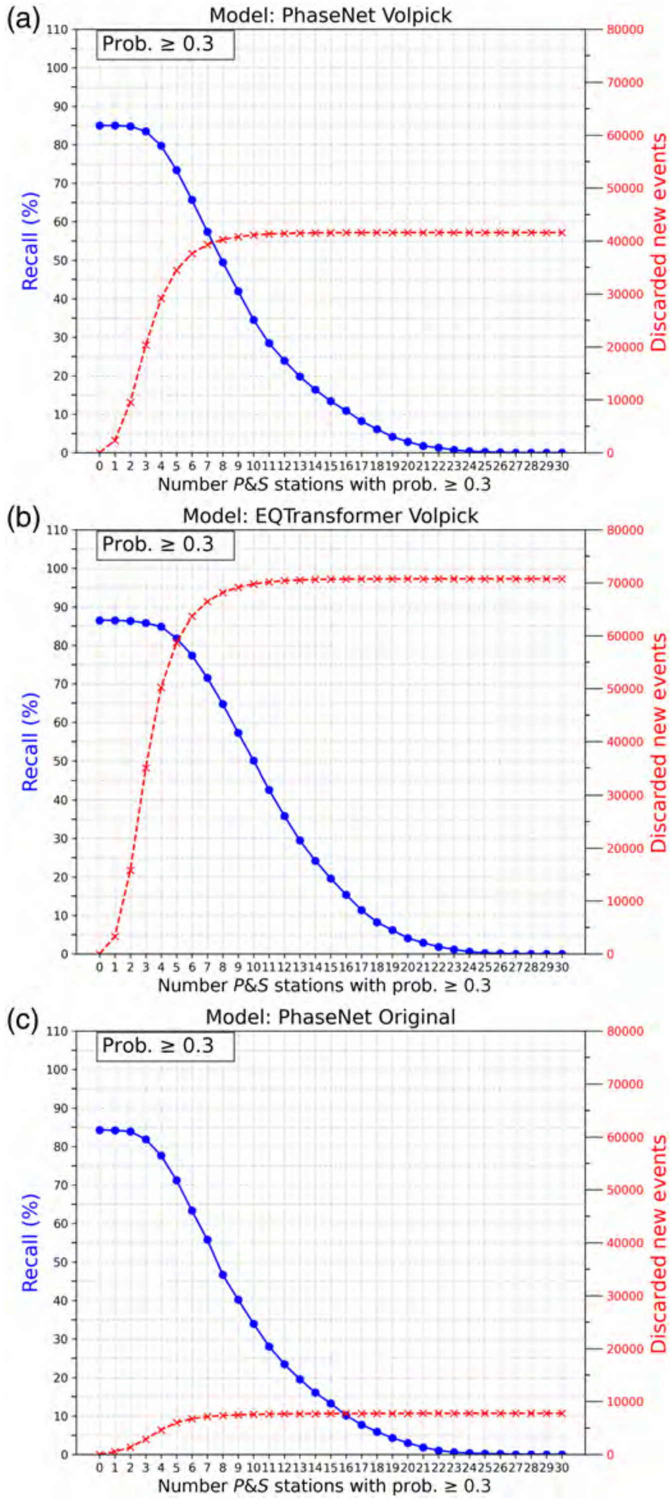


Figure 4. Recall scores relative to the true VT catalog and number of newly detected events discarded in selected models against the number of *P&S* stations with probability thresholds ≥ 0.3. (a) PhaseNet Volpick, (b) EQTransformer Volpick, and (c) PhaseNet Original. The color version of this figure is available only in the electronic edition.

TABLE 2
Metrics for the Earthquakes Predicted by Selected Models, Referred to the True VT Catalog (January 2019–June 2020), Probabilities ≥ 0.3

Model	**TP**	**FP**	**FN**	**Recall (%)**	**Misses (%)**
PhaseNet Volpick	3,607	41,604	637	84.99	15.01
EQTransformer Volpick	3,671	70,745	573	86.50	13.50
PhaseNet Original	3,578	7,779	666	84.31	15.69

TP, true positive events; FP, false positive events; FN, false negative events; recall (%), recall scores relative to the true volcano-tectonic (VT) catalog. Misses (%) indicate the percentage of true VT events that are missed by the models (i.e., FN events).

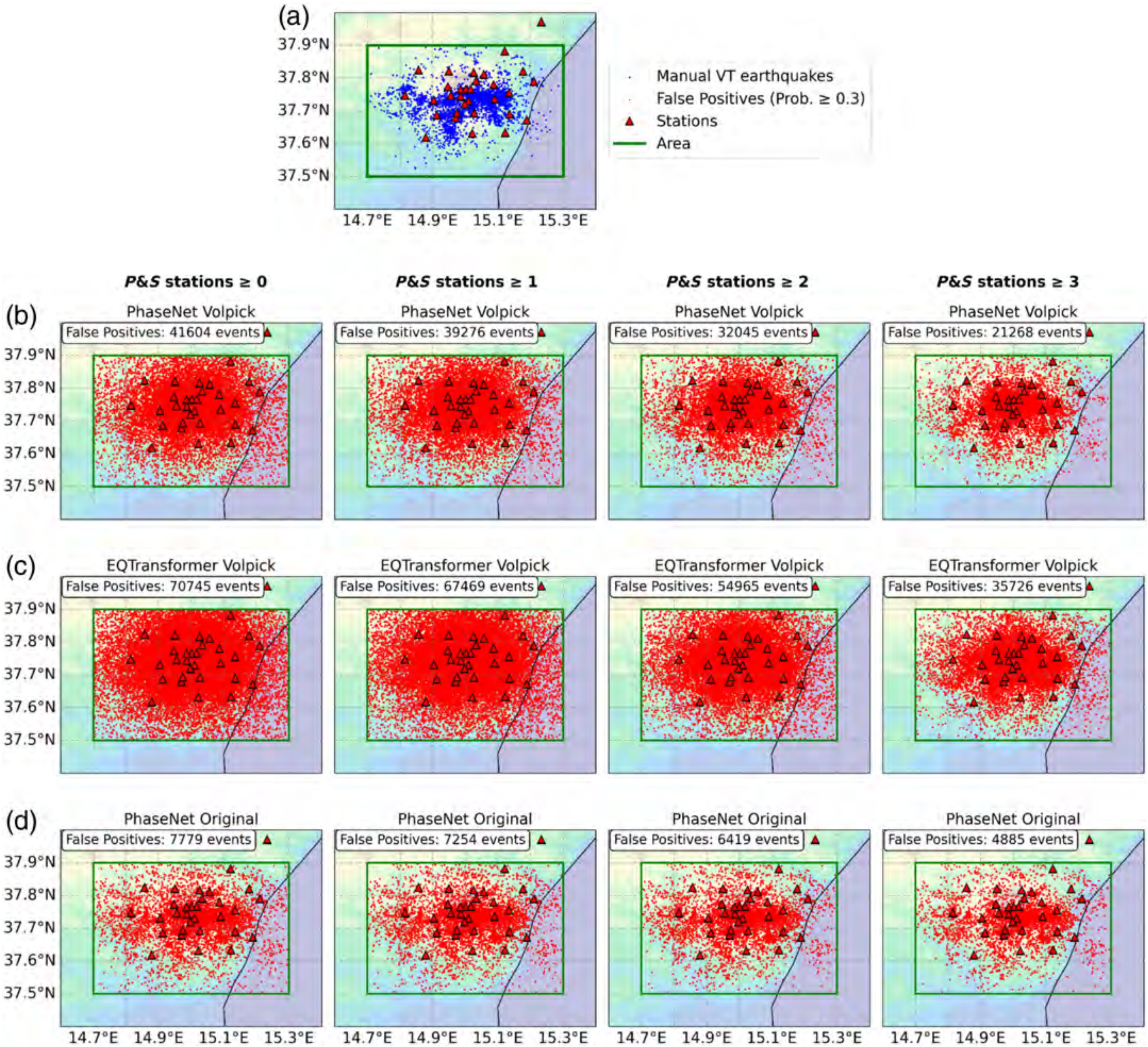
TABLE 3
Metrics of Associated VT Earthquakes and New Events from Different Models Referred to the True VT Catalog, After Filtering Associations with at Least Three *P&S* Stations

Model	**TP**	**FP**	**Recall (%)**	**Discarded FP**
PhaseNet Volpick	3,543	21,268	83.48	20,336
EQTransformer Volpick	3,641	35,726	85.79	35,019
PhaseNet Original	3,474	4,885	81.86	2,894

TP, true positive earthquakes; FP, false positive events; recall (%), recall scores relative to the true volcano-tectonic (VT) catalog; and discarded FP, new FP events with at least three *P&S* stations.

the horizontal (N and E) components consistently exhibit higher amplitude than the vertical (Z) component. Therefore, only horizontal FI values were used in subsequent analyses. For the FI computation, we consider a 6 s wide signal around each *P*-wave pick, starting 1 s before the *P* phase. Therefore, the FI of any event is computed as the average of FI values from all valid north–east components associated with the event. Based on the stacked spectra (Fig. 6c), we defined the characteristic low-frequency band as 0.5–3.0 Hz (associated with LP events) and the high-frequency band as 3.0–20.0 Hz (associated with VT events). Although subjectively defined (Buurman and West, 2010), these bands provide a reliable basis for FI calculation.

We then calculated FI and FI-neg for the ~200 manually selected events. The resulting FI-based classification yielded a bimodal distribution that clearly separates pure VTs from pure LPs (Fig. 6d), consistent with previous studies (Buurman and West, 2010; Matoza *et al.*, 2014; Song *et al.*, 2023; Zhong and Tan, 2024).



We then applied the same classification procedure to events predicted by EQTransformer Volpick, PhaseNet Volpick, and PhaseNet Original (see Text S1 and Table S3).

The labeled LP events are filtered using spatial criteria specific to the Etna area. As detailed in Text S1, only events located within 5 km of at least one summit station are considered. This additional seismological constraint is consistent with LP source locations obtained with conventional methods in the area (Cannata *et al.*, 2013, 2018; Sciotto *et al.*, 2022) and further reduces the chances of erroneous predicted detections in the ML output.

All three models reproduced the bimodal structure, with Volpick-trained models detecting up to ~25 times more LP events than PhaseNet Original (Fig. 7), highlighting the benefit of the volcano-specific training.

Figure 5. (a) Map of the manual VT earthquakes for comparison; below: maps of new false positive (FP) events associated with (b) PhaseNet Volpick, (c) EQTransformer Volpick, and (d) PhaseNet Original, with increasing minimum number of *P&S* stations from left to right. The color version of this figure is available only in the electronic edition.

The temporal evolution of labeled events is shown in Figure 8, which also includes unclassified events. As the minimum FI size increases from left to right, classification robustness improves. All models show a drop in detection rates after September 2019.

In particular, we observe that:

PhaseNet Original (Fig. 8a) closely follows the cumulative curve of true detected VT events up to September 2019, after

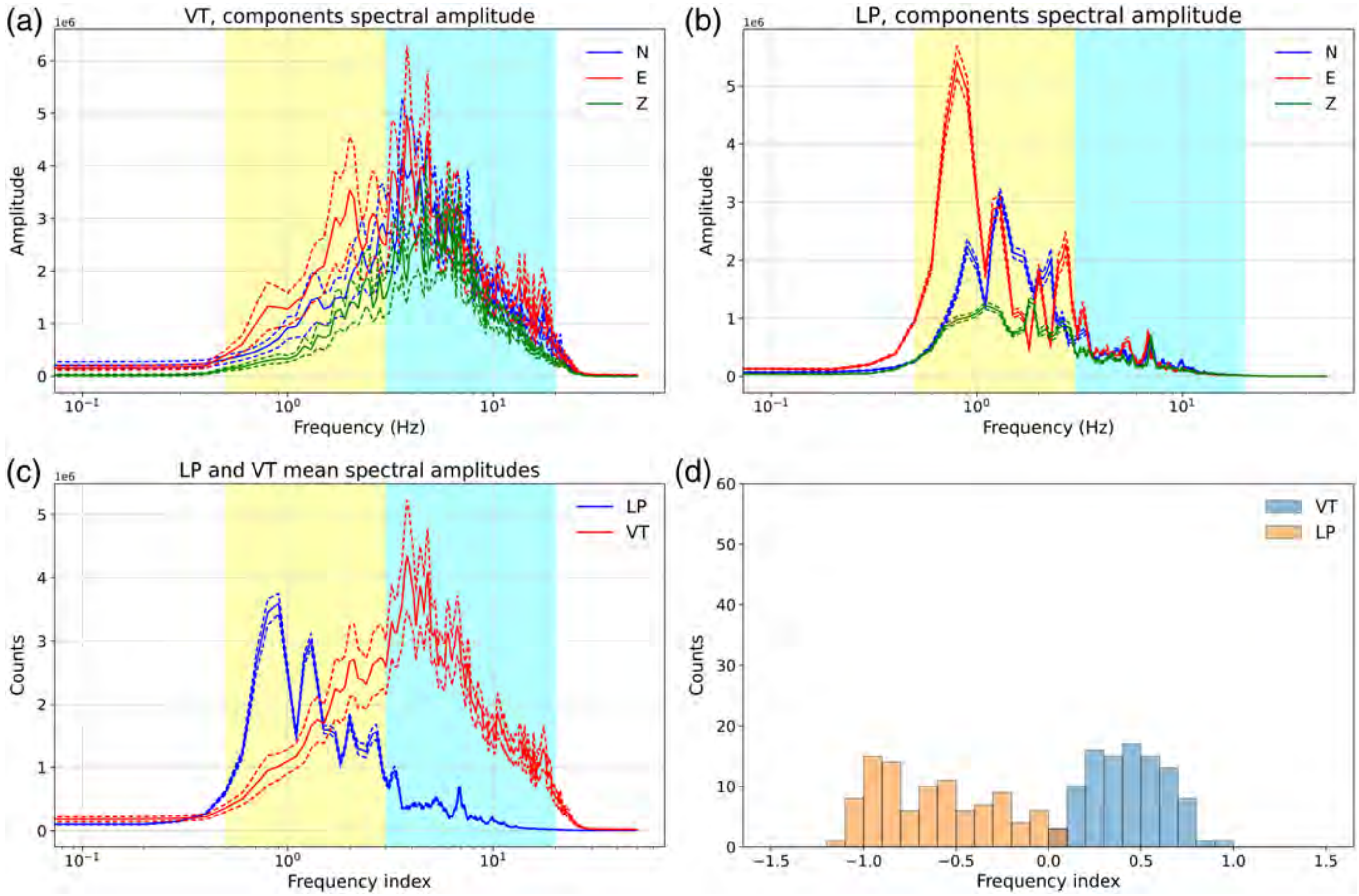


Figure 6. (a) Mean amplitude spectra of N-E-Z components from selected VT earthquakes; (b) mean amplitude spectra of N-E-Z components from selected long-period (LP) events. The yellow band indicates the low-frequency range (0.5–3.0 Hz) associated with LP events, whereas the light-blue band is the frequency band (3.0–20.0 Hz) associated with VT events. (c) Mean stacked spectral amplitudes of selected VTs and LPs; and (d) histogram of pure VTs and LPs. The color version of this figure is available only in the electronic edition.

which a decline in productivity is observed. Increasing the minimum FI size beyond 3 further reduces the number of classified events. The model also shows a slight increase in LS event detection after September 2019, whereas LP detection remains very limited, consistent with its training on tectonic-only datasets. Increasing the minimum FI size beyond 3 further reduces the number of LP labeled events.

PhaseNet Volpick (Fig. 8b) detects more VTs and significantly more LPs than the original model. The cumulative LS curve increases sharply during the first three months of 2019, more than doubling the number of VTs, before stabilizing and following a trend similar to VTs. A decline in VT and LP detections after September 2019 is again observed, mirroring the trend of PhaseNet Original.

EQTransformer Volpick (Fig. 8c) achieves the highest number of VT and LP detections among the three models. Predicted VT and LS events exceed the true VTs, although LP detection still under-represents both filtered and unfiltered true LPs after September 2019, showing a drop similar to the one seen in the other models.

As an example, Figure 9 shows the spatial distributions of events predicted by EQTransformer Volpick using the associator-provided locations. The predicted VTs align spatially with the manual VT catalog. LS and mixed events remain consistent with the network's station distribution. In contrast, LP events cluster near the summit stations. Analogous plots for the other models are provided in Figures S6 and S7. Compared to Figure 5, these results highlight how filtering choice, guided by prior knowledge of the area's seismic characteristics, plays a critical role for reliable event classification and location. The filtering strategy, specifically designed to limit false associations of LP events, combines adaptive frequency-band selection, multiple SNR checks, and spatial constraints based on the distribution of summit and peripheral stations (Text S1). This systematic refinement reduces spurious classifications, strengthens the robustness of the catalog, and improves the physical interpretability of seismic patterns, as evident in the clearer and more geologically consistent organization of VT and LP seismicity in Figure 9.

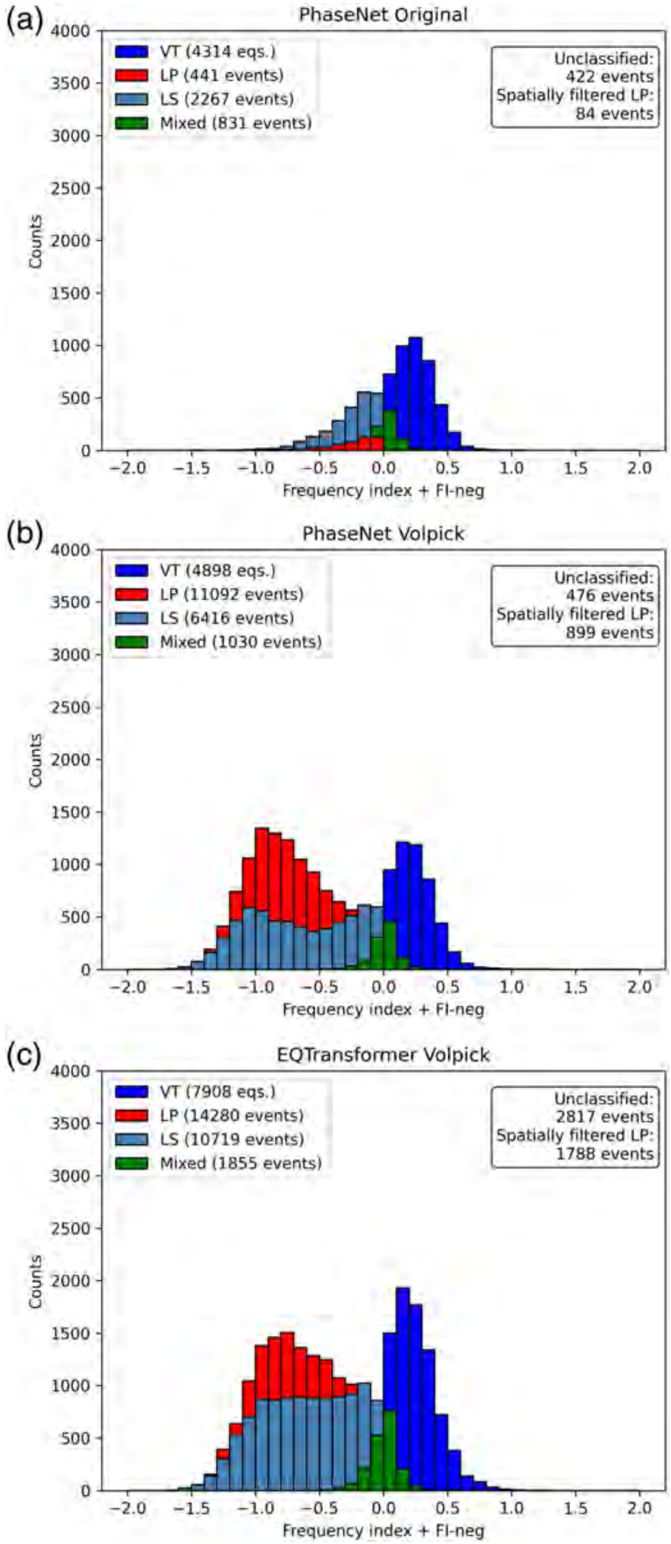


Figure 7. Histograms of labeled events according to frequency index (FI) + FI-neg from (a) PhaseNet Original, (b) PhaseNet Volpick, and (c) EQTransformer Volpick. Numbers of unclassified and spatially filtered LP events are shown in the boxes in the right corner of each plot. The color version of this figure is available only in the electronic edition.

Discussion

Performance evaluation on the four-month sample dataset identified PhaseNet Volpick, EQTransformer Volpick, and PhaseNet Original as the top-performing configurations among the 10 tested. During inference of the whole 1.5-yr dataset, all three achieved high-VT pick-level recall for probability thresholds ≥ 0.3 (>91% for *P* picks, >76% for *S* picks; Table S6, Fig. 3) with VT event-level recall above 84% after association (Table 2). Applying a *P*&*S* ≥ 3 station filter effectively reduced FPs while maintaining ≥81% recall (Table 3), confirming that simple network-level filtering preserves high detection sensitivity even with large numbers of automatic picks.

The comparison between PhaseNet Original and Volpick highlights the impact of training data on performance. Although their picking and event recall scores are nearly identical (Table S6 and Table 2, Fig. 3), PhaseNet Original exhibits a systematic timing offset, consistent with previous studies (e.g., Sugan *et al.*, 2023; Poggiali *et al.*, 2025). At the event level, after FI classification, Volpick identifies more new events than PhaseNet Original, including 11,092 LPs (vs. 441), 4898 VTs (vs. 4314), and three times more LS events (6416 vs. 2267; Fig. 7). This improvement reflects the advantage of Volpick's more diverse, volcano-specific training set, consistent with Zhong and Tan (2024).

Model architecture also plays a key role. EQTransformer Volpick slightly outperforms PhaseNet Volpick in true VT detection, respectively with 86.50% versus 84.99% of recall before filtering (Table 2) and 85.79% versus 83.48% after *P*&*S* filtering (Table 3). It remains the most productive after FI classification, labeling 7908 VTs, 10,719 LS events, 1855 mixed events, and 14,280 LPs (Fig. 7c). Its hybrid convolutional neural network (CNN)-transformer architecture likely improves sensitivity to weak onsets and noisy conditions, enhancing both picking and classification robustness.

However, all models show a marked decline in detection and classification after September 2019 (Fig. 8), coinciding with increased volcanic tremor (1–5 Hz) at Etna (Sciotto *et al.*, 2022). Volcanic tremor overlaps with VT and LP frequency bands, lowering SNR and reducing picking and FI-based classification confidence. EQTransformer exhibits the smallest performance drop, up to FI size = 3 (Fig. 8c).

Figure 9 shows the spatial distribution of FI-based labeled VT, LP, LS, and mixed events using the associator-provided locations, which are not intended as precise event localizations. LP detections cluster near summit stations, whereas LS and mixed events remain widely distributed. The combined spatial- and FI-based filters effectively determine good consistency between labeled VT and the locations of VT manual catalogs, calibrating the ML output according to the local features. Nevertheless, as previously observed (Fig. 8), during high-tremor periods, more events are shifted to unclassified or mixed categories, highlighting FI's limitations under noisy conditions. This work focuses on phase picking, association, and classification. Detailed event localization is not addressed.

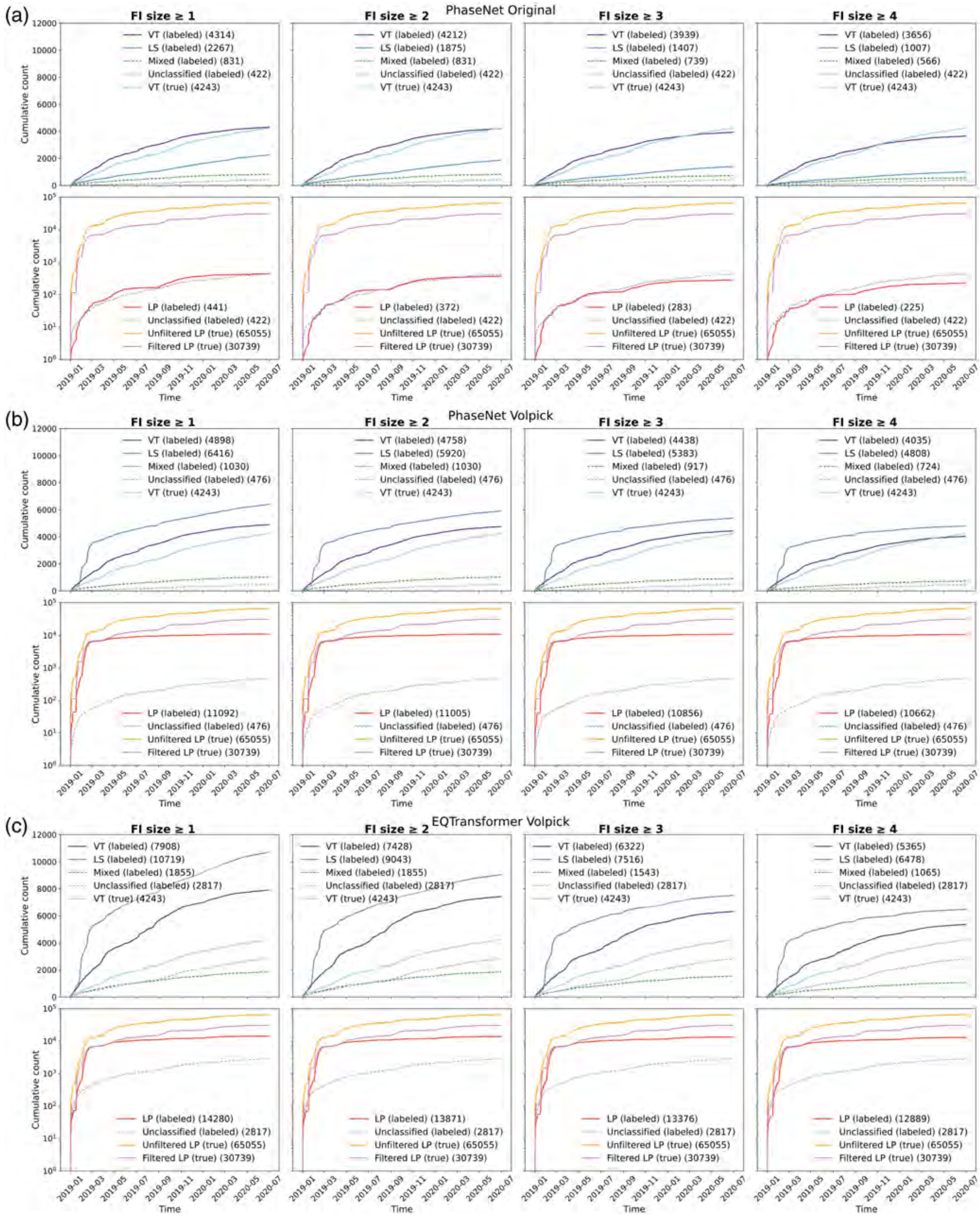


Figure 8. Cumulative counts of labeled event classes (VT, low spectrum [LS], LP, mixed, and unclassified) over time, obtained from the workflow for three model configurations: (a) PhaseNet Original, (b) PhaseNet Volpick, and (c) EQTransformer Volpick. Results are shown alongside true VT and LP events (both filtered and unfiltered), with the total number of events in each group indicated in parentheses. For each configuration, panels from left to right represent increasing minimum FI size thresholds. The color version of this figure is available only in the electronic edition.

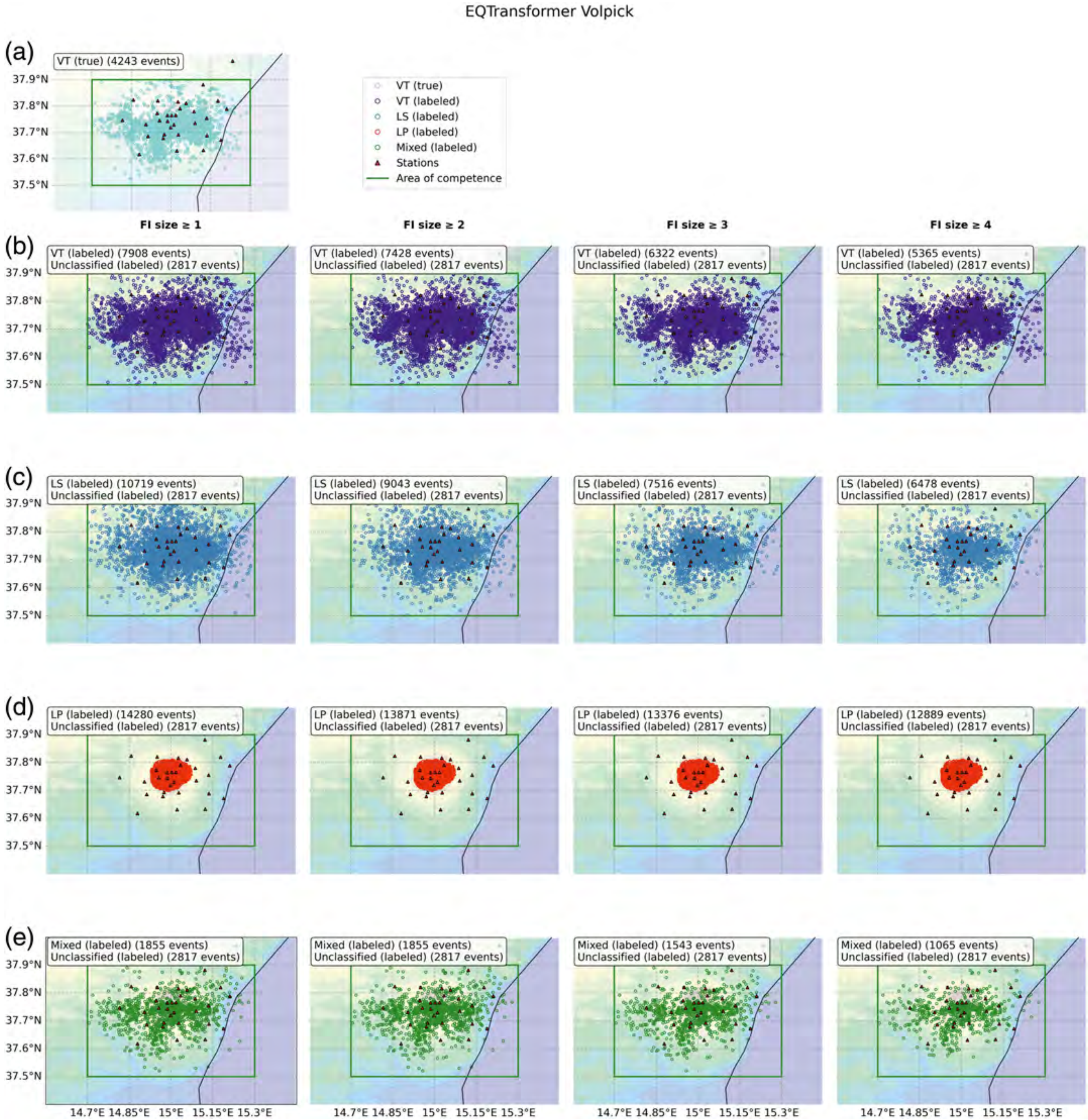


Figure 9. Maps of (a) true VT events, (b) labeled VT earthquakes from EQTransformer Volpick, (c) labeled LS events from EQTransformer Volpick, (d) labeled LP events from EQTransformer Volpick, and (e) labeled mixed events from EQTransformer Volpick. In panel (a), the number of true VT events is shown in the upper-left corner; in panels (b–e), the numbers of labeled and unclassified events are shown in the upper-left corner, and FI size increases from left to right. The color version of this figure is available only in the electronic edition.

Accurate localization requires dedicated methods, which vary by event type. At Etna, VT, events are typically located using first-arrival phase techniques (e.g., Hypoellipse with Lahr, 1999), whereas LP locations are determined via a grid search that integrates semblance and amplitude distribution with a constant single-layer velocity model (Cannata *et al.*, 2013). Therefore, although event localization and detailed depth constraints are not included in this study, the applied methodologies demonstrate that the general classification of events remains robust. These findings confirm that integrating deep-learning pickers with the GaMMA association method and FI classification provides a solid baseline for automated volcano-seismic monitoring. Although fully automated workflows cannot yet replace expert oversight and manual validation, the integration of ML with human supervision represents a major advancement in the detection and characterization of volcanic seismicity (Sciotto *et al.*, 2022). ML models markedly expand the ability to identify and quantify seismic populations, and their effectiveness can be further improved through systematic validation on a per-volcano basis, accounting for local station networks, geological structures, and noise conditions.

Despite these advances, LP detection remains challenging, and FI alone cannot fully resolve classification ambiguities during high volcanic tremor. Effective implementation requires volcano-specific model training, local calibration, and sustained expert oversight. In this framework, ML enables high-throughput event characterization while maintaining human supervision as a critical component for reliable hazard assessment. Future integration of more advanced spectral analysis methods may further enhance event-type discrimination and reduce ambiguities in automated classifications of volcanic seismicity (e.g., Zali *et al.*, 2024; Abed *et al.*, 2026).

Conclusions

We present a distributed workflow that leverages DL models pretrained on global datasets, to detect and classify seismic events at Mount Etna. The region's comprehensive, high-quality seismic catalogs make Etna an exemplary test bed for evaluating and benchmarking advanced detection and classification methods, offering insights transferable to other volcanic regions. By applying and benchmarking several DL models over 1.5 yr of data, we found that configurations trained on the Volpick dataset, especially EQTransformer Volpick, consistently achieved superior performance. These models reliably differentiated VT, LP, LS, and mixed events using a modified FI for spectral classification, with event distributions matching VT manual catalogs up to the onset of energetic volcanic tremor. However, episodes of increased volcanic tremor amplitude can reduce model accuracy by obscuring VT and LP signals.

Our results underscore the importance of volcano-specific training datasets, such as Volpick, in enhancing the performance and generalizability of DL models. EQTransformer Volpick, among others, stands out and demonstrates considerable promise for real-time monitoring, event association, and seismic signal classification in volcanic environments. Although automated methods show remarkable potential, their optimal application can be further enhanced through expert supervision and site-specific validation to ensure reliability and interpretability. In this respect, combining ML and human oversight represents a decisive step toward more robust, adaptive, and scalable real-time volcanic monitoring. Looking ahead, incorporating more advanced spectral analysis techniques may further improve event-type discrimination and reduce ambiguities in automated classifications.

Data and Resources

All data used in this article came from published sources listed in the references. The supplemental material for this work includes additional explanations on the methods (Text S1 and Text S2), tables (Tables S1–S6), and figures (Figures S1–S7).

Declaration of Competing Interests

The authors acknowledge that there are no conflicts of interest recorded.

Acknowledgments

This study was carried out within the RETURN Extended Partnership and received funding from the European Union Next-GenerationEU (National Recovery and Resilience Plan [NRRP], Mission 4, Component 2, Investment 1.3 – D.D. 1243 2/8/2022, PE0000005). This work was also partially funded by the NRRP Project TeRABIT (Terabit network for Research and Academic Big data in Italy - IR0000022 - PNRR Missione 4, Componente 2, Investimento 3.1 CUP I53C21000370006) in the frame of the European Union-NextGenerationEU funding. This research was also supported by the Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (OGS) and Consorzio Interuniversitario del Nord-Est per il Calcolo Automatico (CINECA) under the high-performance computing–Training and Research for Earth Sciences (HPC-TRES) program, Awards Numbers 2024-05 and 2025-01. The authors acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer. Monica Sugan designed the study. Andrea Carducci prepared the data for the analysis and performed the deep learning (DL) workflow. Mariangela Sciotto, Ornella Cocina, and Andrea Cannata provided the volcano-tectonic (VT) and long-period (LP) catalogs and seismic waveforms. All the authors analyzed the results. The first article draft was written by Andrea Carducci, Monica Sugan, Angela Saraò, and Alessandro Vuan. All authors contributed to the evaluation and discussion of the results and the writing of the final article. Finally, the authors thank the editor and the anonymous reviewers for their constructive comments, which helped improve the article.

References

Abed, W., Z. Zali, M. Sciotto, O. Cocina, A. Cannata, M. Picozzi, P. Martínez-Garzón, A. Vuan, A. Saraò, and M. Sugan (2026). Hidden patterns in volcanic seismicity: Deep learning insights from Mt. Etna's 2020–2021 activity, *Sci. Rep.* **16,** 6155, doi: 10.1038/s41598-026-36677-x.

Alasonati, P., J. Wassermann, and M. Ohrnberger (2006). Signal classification by wavelet-based hidden Markov models: Application to seismic signals of volcanic origin, in *Statistics in Volcanology*, H. M. Mader, S. G. Coles, C. B. Connor, and L. J. Connor (Editors), The Geological Society of London on behalf of The International Association of Volcanology and Chemistry of the Earth's Interior, 161–173, The Geological Society of London, London, doi: 10.1144/IAVCEI001.13.

Alparone, S. C., G. Barberi, S. D'Amico, G. Di Grazia, F. Ferrari, E. Giampiccolo, V. Maiolino, A. Mostaccio, C. Musumeci, A. Scaltrito, *et al.* (2020). Mt. Etna Revised and Concise Seismic Catalog from 1999 (EtnaRCSC) [Data set], *Istituto Nazionale di Geofisica e Vulcanologia (INGV)*, doi: 10.13127/ETNASC/ETNARCSC.

Barberi, G., G. Di Grazia, F. Ferrari, E. Giampiccolo, V. Maiolino, A. Mostaccio, C. Musumeci, A. Scaltrito, M. Sciotto, G. Tusa, *et al.* (2020). Mt. Etna Revised Seismic Catalog from 2020 (EtnaRSC2020), *Istituto Nazionale di Geofisica e Vulcanologia (INGV)*, doi: 10.13127/ETNASC/ETNARSC2020.

Benitez, M. C., J. Ramirez, J. C. Segura, J. M. Ibanez, J. Almendros, A. Garcia-Yeguas, and G. Cortes (2007). Continuous HMM-based seismic-event classification at deception Island, Antarctica, *IEEE Trans. Geosci. Remote Sens.* **45,** no. 1, 138–146, doi: 10.1109/tgrs.2006.882264.

Buurman, H., and M. E. West (2010). Seismic precursors to volcanic explosions during the 2006 eruption of Augustine Volcano, in *The 2006 Eruption of Augustine Volcano, Alaska*, J. A. Power, M. L. Coombs, and J. T. Freymueller (Editors), U.S. Geological Survey Professional Paper 1769, U.S. Geological Survey, Menlo Park, California, 41–57.

Cannata, A., G. Di Grazia, M. Aliotta, C. Cassisi, P. Montalto, and D. Patanè (2013). Monitoring seismo-volcanic and infrasonic signals at volcanoes: Mt. Etna case study, *Pure Appl. Geophys.* **170,** 1751–1771, doi: 10.1007/s00024-012-0634-x.

Cannata, A., G. Di Grazia, M. Giuffrida, S. Gresta, M. Palano, M. Sciotto, M. Viccaro, and F. Zuccarello (2018). Space-time evolution of magma storage and transfer at Mt. Etna volcano (Italy): The 2015–2016 reawakening of Voragine crater, *Geochem. Geophys. Geosys.* **19,** doi: 10.1002/2017gc007296.

Cannata, A., G. Di Grazia, and M. Sciotto (2025). Catalog of long period seismic events recorded at Mt. Etna during the time interval 2019-2020 (LPSEC2019_2020), *Istituto Nazionale di Geofisica e Vulcanologia (INGV)*, doi: 10.13127/ETNA/ELPC/2019_2020.

Cannata, A., G. Spedalieri, B. Behncke, F. Cannavò, G. Di Grazia, S. Gambino, S. Gresta, S. Gurrieri, M. Liuzzo, and M. Palano (2015). Pressurization and depressurization phases inside the plumbing system of Mount Etna volcano: Evidence from a multiparametric approach, *J. Geophys. Res.: Solid Earth* **120,** no. 9, 5965–5982, doi: 10.1002/2015jb012227.

Chouet, B. A. (1996). Long-period volcano seismicity: Its source and use in eruption forecasting, *Nature* **380,** no. 6572, 309–316, doi: 10.1038/380309a0.

Chouet, B. A., and R. S. Matoza (2013). A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption, *J. Volcanol. Geotherm. Res.* **252,** 108–175, doi: 10.1016/j.jvolgeores.2012.11.013.

Chouet, B. A., R. A. Page, C. D. Stephens, J. C. Lahr, and J. A. Power (1994). Precursory swarms of long-period events at Redoubt Volcano (1989–1990), Alaska: Their origin and use as a forecasting tool, *J. Volcanol. Geotherm. Res.* **62,** nos. 1/4, 95–135, doi: 10.1016/0377-0273(94)90030-2.

CINECA Supercomputing Centre, SuperComputing Applications and Innovation Department (2024). LEONARDO: A Pan-European Pre-Exascale Supercomputer for HPC and AI applications, *J. Large-Scale Res. Facil.* **8,** A186, doi: 10.17815/jlsrf-8-186.

Del Pezzo, E. (2003). Discrimination of earthquakes and underwater explosions using neural networks, *Bull. Seismol. Soc. Am.* **93,** no. 1, 215–223, doi: 10.1785/0120020005.

Esposito, A. M., F. Giudicepietro, L. D'Auria, S. Scarpetta, M. G. Martini, M. Coltelli, and M. Marinaro (2008). Unsupervised neural analysis of very-long-period events at Stromboli volcano using the self-organizing maps, *Bull. Seismol. Soc. Am.* **98,** no. 5, 2449–2459, doi: 10.1785/0120070110.

Ester, M., H. P. Kriegel, J. Sander, and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise, *Second Int. Conf. on Knowledge Discovery and Data Mining (KDD'96)*, Proceedings of a Conference Held August 2-4, 1996, 226–331.

Falsaperla, S., S. Graziani, G. Nunnari, and S. Spampinato (1996). Automatic classification of volcanic earthquakes by using multi-layered neural networks, *Nat. Hazards* **13,** no. 3, doi: 10.1007/bf00215816.

Fee, D., D. Tan, J. Lyons, M. Sciotto, A. Cannata, A. Hotovec-Ellis, T. Girona, A. Wech, D. Roman, M. Haney, and S. De Angelis (2025). A generalized deep learning model to detect and classify volcano seismicity, *Volcanica* **8,** no. 1, 305–323, doi: 10.30909/vol/rjss1878.

Firetto Carlino, M., L. Scarfì, F. Cannavò, G. Barberi, D. Patanè, and M. Coltelli (2022). Frequency-magnitude distribution of earthquakes at Etna volcano unravels critical stress changes along magma pathways, *Commun. Earth Environ.* **3,** no. 1, 68, doi: 10.1038/s43247-022-00398-6.

Friberg, P., S. Lisowski, I. Dricker, and S. Hellman (2010). Earthworm in the 21st century, *EGU General Assembly 2010*, 2–7 May 2010, Vienna, Austria, 12,654 pp.

Giacco, F., A. M. Esposito, S. Scarpetta, F. Giudicepietro, and M. Marinaro (2009). Support vector machines and MLP for automatic classification of seismic signals at Stromboli volcano, *Front. Artif. Intell. Appl.* **204,** 116–123, doi: 10.3233/978-1-60750-072-8-116.

Hammer, C., M. Beyreuther, and M. Ohrnberger (2012). A seismic-event spotting system for volcano fast-response systems, *Bull. Seismol. Soc. Am.* **102,** no. 3, 948–960, doi: 10.1785/0120110167.

Hirn, A., A. Nercessian, M. Sapin, F. Ferrucci, and G. Wittlinger (1991). Seismic heterogeneity of Mt Etna: structure and activity, *Geophys. J. Int.* **105,** no. 1, 139–153, doi: 10.1111/j.1365-246X.1991.tb03450.x.

Ibáñez, J. M., C. Benítez, L. A. Gutiérrez, G. Cortés, A. García-Yeguas, and G. Alguacil (2009). The classification of seismo-volcanic signals using Hidden Markov Models as applied to the Stromboli and Etna volcanoes, *J. Volcanol. Geotherm. Res.* **187,** nos. 3/4, 218–226, doi: 10.1016/j.jvolgeores.2009.09.002.

Ibs-von Seht, M. (2008). Detection and identification of seismic signals recorded at Krakatau volcano (Indonesia) using artificial neural networks, *J. Volcanol. Geotherm. Res.* **176,** no. 4, 448–456, doi: 10.1016/j.jvolgeores.2008.04.015.

Köhler, A., M. Ohrnberger, and F. Scherbaum (2010). Unsupervised pattern recognition in continuous seismic wavefield records using self-organizing maps: Unsupervised seismic pattern recognition, *Geophys. J. Int.* **182,** no. 3, 1619–1630, doi: 10.1111/j.1365-246x.2010.04709.x.

Kumagai, H., and B. A. Chouet (2000). Acoustic properties of a crack containing magmatic or hydrothermal fluids, *J. Geophys. Res.* **105,** no. B11, 25,493–25,512, doi: 10.1029/2000jb900273.

Lahr, J. C. (1999). HYPOELLIPSE: A computer program for determining local earthquake hypocentral parameters, magnitude, and first-motion pattern, *U.S. Geol. Surv. Open-File Rept. 79-431*, doi: 10.3133/ofr79431.

Langer, H., S. Falsaperla, M. Masotti, R. Campanini, S. Spampinato, and A. Messina (2009). Synopsis of supervised and unsupervised pattern classification techniques applied to volcanic tremor data at Mt Etna, Italy, *Geophys. J. Int.* **178,** no. 2, 1132–1144, doi: 10.1111/j.1365-246X.2009.04179.x.

Langer, H., S. Falsaperla, and G. Thompson (2003). Application of Artificial Neural Networks for the classification of the seismic transients at Soufrière Hills volcano, Montserrat, *Geophys. Res. Lett.* **30,** no. 21, 2003GL018082, doi: 10.1029/2003GL018082.

Malfante, M., M. Dalla Mura, J. I. Mars, J. Métaxian, O. Macedo, and A. Inza (2018). Automatic classification of volcano seismic signatures, *J. Geophys. Res.: Solid Earth* **123,** no. 12, doi: 10.1029/2018jb015470.

Matoza, R. S., and D. C. Roman (2022). One hundred years of advances in volcano seismology and acoustics, *Bull. Volcanol.* **84,** 86, doi: 10.1007/s00445-022-01586-0.

Matoza, R. S., B. A. Chouet, P. B. Dawson, P. M. Shearer, M. M. Haney, G. P. Waite, S. C. Moran, and T. D. Mikesell (2015). Source mechanism of small long-period events at Mount St. Helens in July 2005 using template matching, phase-weighted stacking, and full-waveform inversion, *J. Geophys. Res.: Solid Earth* **120,** no. 9, 6351–6364, doi: 10.1002/2015jb012279.

Matoza, R. S., P. M. Shearer, and P. G. Okubo (2014). High-precision relocation of long-period events beneath the summit region of Kīlauea Volcano, Hawai'i, from 1986 to 2009, *Geophys. Res. Lett.* **41,** no. 10, 3413–3421, doi: 10.1002/2014gl059819.

Michelini, A., S. Cianetti, S. Gaviano, C. Giunchi, D. Jozinović, and V. Lauciani (2021). INSTANCE – the Italian seismic dataset for machine learning, *Earth Syst. Sci. Data* **13,** no. 12, 5509–5544, doi: 10.5194/essd-13-5509-2021.

Mousavi, S. M., W. L. Ellsworth, W. Zhu, L. Y. Chuang, and G. C. Beroza (2020). Earthquake transformer—An attentive deep-learning model for simultaneous earthquake detection and phase picking, *Nat. Commun.* **11,** no. 1, 3952, doi: 10.1038/s41467-020-17591-w.

Mousavi, S. M., Y. Sheng, W. Zhu, and G. C. Beroza (2019). STanford EArthquake Dataset (STEAD): A global data set of seismic signals for AI, *IEEE Access* **7,** 179,464–179,476, doi: 10.1109/ACCESS.2019.2947848.

Münchmeyer, J., J. Woollam, A. Rietbrock, F. Tilmann, D. Lange, T. Bornstein, T. Diehl, C. Giunchi, F. Haslinger, D. Jozinović, *et al.* (2022). Which picker fits my data? A quantitative evaluation of deep learning based seismic pickers, *J. Geophys. Res.: Solid Earth* **127,** no. 1, e2021JB023499, doi: 10.1029/2021jb023499.

Northern California Earthquake Data Center (NCEDC) (2014). Northern California earthquake data center, UC Berkeley Seismological Laboratory, Dataset, doi: 10.7932/NCEDC.

Ohrnberger, M. (2001). Continuous automatic classification of seismic signals of volcanic origin at Mt. Merapi, Java, Indonesia, *Ph.D. Thesis.*

Orozco, M., M. E. García, R. P. W. Duin, and C. G. Castellanos (2006). Dissimilarity-based classification of seismic signals at Nevado Del Ruiz Volcano, *Earth Sci. Res. J.* **10,** 57–66.

Pallister, J., and S. R. McNutt (2015). Synthesis of volcano monitoring, in *The Encyclopedia of Volcanoes*, H. Sigurdsson (Editor), Academic Press, Amsterdam, The Netherlands, 1151–1171, doi: 10.1016/B978-0-12-385938-9.00066-3.

Patanè, D., and E. Giampiccolo (2004). Faulting processes and earthquake source parameters at Mt. Etna: state of the art and perspectives, in *American Geophysical Monograph*, A. Bonaccorso, S. Calvari, M. Coltelli, C. Del Negro, and S. Falsaperla (Editors), Volcano Laboratory, Mt. Etna, Vol. 143, 167–189, doi: 10.1029/143GM11.

Patanè, D., A. Aiuppa, M. Aloisi, B. Behncke, A. Cannata, M. Coltelli, G. Di Grazia, S. Gambino, S. Gurrieri, M. Mattia, and G. Salerno (2013). Insights into magma and fluid transfer at Mount Etna by a multiparametric approach: A model of the events leading to the 2011 eruptive cycle, *J. Geophys. Res.: Solid Earth* **118,** 3519–3539, doi: 10.1002/jgrb.50248.

Patanè, D., E. Privitera, F. Ferrucci, and S. Gresta (1994). Seismic activity leading to the 1991–1993 eruption of Mt. Etna and its tectonic implications, *Acta Vulcanol.* **4,** 47–55.

Patton, J. M., M. R. Guy, H. M. Benz, R. P. Buland, B. K. Erickson, and D. S. Kragness (2016). Hydra—The National Earthquake Information Center's 24/7 seismic monitoring, analysis, catalog production, quality analysis, and special studies tool suite, *U.S. Geol. Surv. Open-File Rept. 2016-1128*, 28 pp., doi: 10.3133/ofr20161128.

Poggiali, G., L. Chiaraluce, Z. E. Ross, W. Zhu, and C. Marone (2025). Fault geometry and source mechanics of the Altotiberina Fault system from a high-resolution machine-learning earthquake catalog, *Bull. Seismol. Soc. Am.* **115,** no. 5, 2181–2201, doi: 10.1785/0120250072.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to Roc, informedness, markedness & correlation, *J. Mach. Learn. Technol.* **2,** no. 1, 37–63.

Ross, Z. E., M.-A. Meier, E. Hauksson, and T. H. Heaton (2018). Generalized seismic phase detection with deep learning, *Bull. Seismol. Soc. Am.* **108,** no. 5A, 2894–2901, doi: 10.1785/0120180080.

Saraò, A., O. Cocina, E. Privitera, and G. F. Panza (2010). The dynamics of the 2001 Etna eruption as seen with full moment tensor analysis, *Geophys. J. Int.* **181,** 951–965, doi: 10.1111/j.1365-246x.2010.04547.x.

Saraò, A., L. Moratto, E. Giampiccolo, and O. Cocina (2023). Moment magnitude for earthquakes in the area of the Etna volcano, *Geophys. J. Int.* **234,** 2520–2534, doi: 10.1093/gji/ggad257.

Scarpa, R., and R. I. Tilling (1996). *Monitoring and Mitigation of Volcano Hazards*, Springer Berlin Heidelberg, Berlin, Heidelberg, doi: 10.1007/978-3-642-80087-0.

Southern California Earthquake Data Center (SCEDC) (2013). Southern California Earthquake Data Center, Caltech, Dataset, doi: 10.7909/C3WD3XH1.

Sciotto, M., A. Cannata, G. Di Grazia, and P. Montalto (2022). Volcanic tremor and long period events at Mt. Etna: Same mechanism at different rates or not? *Phys. Earth Planet. In.* **324,** 106850, doi: 10.1016/j.pepi.2022.106850.

Song, Z., Y. J. Tan, and D. C. Roman (2023). Deep long-period earthquakes at Akutan Volcano from 2005 to 2017 better track magma influxes compared to volcano-tectonic earthquakes, *Geophys. Res. Lett.* **50,** no. 10, e2022GL101987, doi: 10.1029/2022gl101987.

Sugan, M., L. Peruzza, M. A. Romano, M. Guidarelli, L. Moratto, D. Sandron, M. P. Plasencia Linares, and M. Romanelli (2023). Machine learning versus manual earthquake location workflow: testing LOC-FLOW on an unusually productive microseismic sequence in northeastern Italy, *Geomat. Nat. Hazards Risk* **14,** no. 1, 2284120, doi: 10.1080/19475705.2023.2284120.

Wassermann, J. (2012). Volcano seismology, *New Manual of Seismological Observatory Practice 2 (NMSOP2)*, 77 pp., doi: 10.2312/GFZ.NMSOP-2_CH13.

Weber, B., J. Becker, A. Hanka, A. Heinloo, M. Hoffmann, T. Kraft, D. Pahlke, J. Reinhardt, and H. Thoms (2007). SeisComP3 - automatic and interactive real time data processing, *Geophysical Research Abstracts*, Vol. 9, 09219, General Assembly European Geosciences Union (EGU).

Woollam, J., J. Münchmeyer, F. Tilmann, A. Rietbrock, D. Lange, T. Bornstein, T. Diehl, C. Giunchi, F. Haslinger, D. Jozinović, *et al.*

(2022). SeisBench—A Toolbox for Machine Learning in Seismology, *Seismol. Res. Lett.* **93,** no. 3, 1695–1709, doi: 10.1785/0220210324.

Zali, Z., S. M. Mousavi, M. Ohrnberger, E. P. S. Eibl, and F. Cotton (2024). Tremor clustering reveals pre-eruptive signals and evolution of the 2021 Geldingadalir eruption of the Fagradalsfjall Fires, Iceland, *Commun. Earth Environ.* **5,** no. 1, 1, doi: 10.1038/s43247-023-01166-w.

Zhong, Y., and Y. J. Tan (2024). Deep-learning-based phase picking for volcano-tectonic and long-period earthquakes, *Geophys. Res. Lett.* **51,** no. 12, e2024GL108438, doi: 10.1029/2024gl108438.

Zhu, W., and G. C. Beroza (2019). PhaseNet: A deep-neural-network-based seismic arrival time picking method, *Geophys. J. Int.* **216,** 1, 261–273, doi: 10.1093/gji/ggy423.

Zhu, W., I. W. McBrearty, S. M. Mousavi, W. L. Ellsworth, and G. C. Beroza (2022). Earthquake phase association using a Bayesian Gaussian mixture model, *J. Geophys. Res.: Solid Earth* **127,** no. 5, e2021JB023249, doi: 10.1029/2021jb023249.

Manuscript received 11 December 2025
Published online 18 June 2026