scientific reports

Check for updates

OPEN

Single-station analysis of Campi Flegrei (Italy) seismic signals using multiscale entropy and unsupervised learning

Alberico Grimaldi[1], Ortensia Amoroso[1]✉, Silvia Scarpetta[1], Vincenzo Convertito[2], Ferdinando Napolitano[1], Giovanni Messuti[1], Paolo Capuano[1], Lucia Nardone[2], Guido Gaudiosi[2] & Danilo Galluzzo[2]

Seismic activity in volcanic regions such as Campi Flegrei (Italy) provides essential insights into subsurface dynamics and potential hazards. However, high background noise and continuous data volume challenge event detection and classification. Here, we apply a Self-Organizing Map (SOM) approach, combined with Linear Predictive Coding (LPC), STA/LTA ratios, and Multiscale Entropy (MSE), to analyze single-station seismic data. The method successfully identifies uncatalogued events and anomalies associated with fumarolic tremor, and reveals temporal relationships between clustering variation, CO_2 emissions, and rainfall, suggesting environmental modulation. To assess the real-time applicability, the trained SOM was used on independent data from early 2025, confirming its ability to detect tremor intensification and anticipate a major local earthquake (Md 4.4). These results highlight the potential of entropy-based unsupervised learning for rapid seismic characterization and continuous volcano monitoring.

Keywords Single-station analysis, Campi Flegrei (Italy), Unsupervised learning, Seismic monitoring of a volcano, Multiscale entropy

The Campi Flegrei volcanic complex, located in southern Italy, is a large, near-circular collapse caldera approximately 12 km in diameter, encompassing densely populated urban areas including Naples, Pozzuoli, and Cuma, with a combined population of over 2 million residents. Following its last eruption in 1538 (Monte Nuovo eruption), Campi Flegrei underwent centuries of subsidence. However, since the 1950s, the caldera has displayed episodic unrest, marked by significant ground uplift, seismic activity, and geochemical changes[1]. Major uplift episodes occurred from 1968 to 1972 and from 1982 to 1985, the latter accompanied by approximately 16,000 earthquakes[2,3]. Although the 1990s brought a period of relative calm, seismicity, ground deformation, and degassing activity resumed in the early 2000s, indicating renewed volcanic activity[4,5]. In response to increasing deformation and seismic activity, the Italian Civil Protection raised the alert level of Campi Flegrei to "yellow" in December 2012, signaling an elevated monitoring status due to the potential for heightened volcanic hazards[6]. Since 2014, additional ground uplift, shallow seismic events, and increased CO_2 emissions from hydrothermal sites such as Pisciarelli and Solfatara have underscored the caldera's persistent unrest[7]. Among these indicators, the Pisciarelli hydrothermal area has shown the most pronounced changes, with the formation of new fumarolic vents and increased shallow seismicity. In 2010, to better monitor these changes, a seismic station was installed near the main fumarole to continuously record fumarolic tremor, which has proven valuable as a proxy for hydrothermal fluid dynamics and CO_2 emissions[8]. In the past two years, the Campi Flegrei caldera has seen an increase in seismic activity, with a total of 9146 seismic events[9]. These include both low-magnitude earthquakes with duration magnitude (Md) larger than 0 and shallow seismic swarms primarily concentrated beneath the Solfatara and Pisciarelli areas[10]. Since early 2024, 73 seismic swarms have been recorded. Two notable swarms occurred in February 2025 and March 2025. The February swarm was characterized by more than 1000 events occurring in five days (February 15–19), while the March 2025 swarm comprised 66 events, including the largest magnitude earthquake, Md 4.6, ever recorded in the area. This activity correlates with ongoing ground deformation that reached about 140 cm since 2005, and increased degassing (about 5kton CO_2 per day), further highlighting the persistent unrest within the caldera. To enhance the efficiency of volcano monitoring systems,

[1]Department of Physics E.R. Caianiello, University of Salerno, Fisicano, 84084 Salerno, Italy. [2]Osservatorio Vesuviano, Istituto Nazionale di Geofisica e Vulcanologia, Naples, Italy. ✉email: oamoroso@unisa.it

it is increasingly necessary to complement traditional techniques—carried out by expert operators—with artificial intelligence-based approaches. These advanced methods can provide fast and accurate identification of various types of events, such as low-energy phreatic explosions, long-period earthquakes, landslides, and volcanic tremor. The integration of artificial intelligence (AI) into monitoring workflows has the potential to improve the timely detection of signals that may indicate a change in the state of the volcano, thus supporting risk mitigation actions related to both volcanic and seismic hazards, being the latter of main concern during bradiseismic crisis[11]. Supervised and unsupervised machine learning methodologies are being applied in highly hazardous tectonic and induced seismic areas[12] and volcanoes for detection, analysis, and interpretation of complex patterns. Supervised methodologies, such as convolutional neural networks (CNNs), have been used for automatic seismic event classification and P-wave detection, yielding high accuracy in discriminating between noise and events[13,14]. Unsupervised approaches, such as clustering via Self-Organizing Maps (SOMs) or Gaussian Mixture Models (GMMs), have proven effective for detecting anomalies and uncovering patterns in volcanic tremor and seismic signals[15,16]. One of the main challenges in monitoring the Campi Flegrei area is the fast and reliable automatic classification of seismic signals. As mentioned above, the Campi Flegrei area is mainly characterized by low-energy shallow seismicity, often occurring as seismic swarms, localized in restricted areas of the caldera and poorly observable by all the stations of the seismic monitoring network, also due to the high background noise. For this reason, in some cases, using a small number of seismic stations to characterize such events is unavoidable. This restrictive aspect directs the research activity towards the development of new methodologies of analysis based on a reduced number of seismic signals. The proposed study has the potential to contribute to increasingly rapid and accurate identification of relevant events in noisy time series, addressing these challenges and enhancing monitoring capabilities. We propose a combination of an unsupervised neural network and different data encoding techniques to distinguish between seismic noise and seismic transients in continuous data collected from a single station located within the Pisciarelli hydrothermal area, utilizing distinct feature-based encoding to improve classification. A central aspect of our study is the use of Multiscale Entropy (MSE) for features' extraction. Originally designed to analyze physiological and biological signals[17,18], the Multiscale Entropy quantifies the information content of a signal across multiple time scales. Here, we demonstrate its strong potential as a feature for characterizing seismic signals. We develop a strategy based on Self-Organizing Maps (SOMs), an unsupervised learning technique that facilitates visualization and clustering of complex datasets while preserving their topological relationships.

Results and discussion

Study area, dataset treatment and learning strategy

The present study focuses on the Pisciarelli hydrothermal area within the Campi Flegrei caldera (Fig. 1), a highly active volcanic region in southern Italy, characterized by intense degassing, shallow seismicity, and frequent hydrothermal activity. In 2023, the Campi Flegrei area experienced intense seismic activity, peaking between August and September. Throughout 2023, continuous seismic data were recorded by the temporary station V0102, operated by the Istituto Nazionale di Geofisica e Vulcanologia (INGV-Osservatorio Vesuviano) and located about 50 meters from the Pisciarelli fumarole. Moreover, RSAM at V0102 exhibited a marked increase in September (Fig. S1). To build the dataset employed for the analysis, the vertical component of the continuous recording was segmented into non-overlapping 1-minute windows. Each window contains 12,000 samples for V0102. The choice of a 1-minute window for segmenting the continuous recordings, also adopted in the work of Esposito et al.[20], represents a suitable compromise, balancing the characteristic time scales of various phenomena such as fumarolic tremors, individual seismic events, and seismic swarms. In this work, only the vertical component of the recording is utilized, since the characteristic tremor of the fumarole is mainly polarized along this direction[8]. The resulting dataset spans from June 9 to November 12, 2023, yielding 224,640 waveform segments for V0102 alone. Additional datasets were collected from two nearby temporary stations, PESG and RENG, located about 500-600 meters from the fumarole-mud pool system. The same segmentation procedure was applied to the PESG and RENG data, covering slightly different but still overlapping time periods due to distinct operational timelines. Finally, data recorded between January and May 2025 from the CPIS seismic station located near the Pisciarelli mud pool were utilized to assess the efficacy of the SOM map that was trained on V0102 data. This map was evaluated on a novel dataset that had not been previously observed but was recorded in close proximity (about 20 meters away). The analysis approach relies on extracting features that capture different characteristics of the 1-minute waveforms. We used three complementary coding techniques: Linear Prediction Coding (LPC), which encodes the envelope of the spectrum of the waveform using the calculated coefficients; Short-Time Average to Long-Time Average (STA/LTA), which highlights transient amplitude variations; and Multiscale Entropy (MSE), a metric originally developed for physiological signal analysis[17] that quantifies signal complexity across multiple temporal scales. Each encoding yields a feature vector representing the seismic signal in a compact, information-rich form suitable for unsupervised machine learning analysis. The encoded datasets, composed of feature vectors derived from each 1-minute waveform, are employed to train Self-Organizing Maps (SOMs), an unsupervised neural network algorithm designed for clustering and visualization of high-dimensional data. This technique provides a compact yet powerful tool to explore the variability within the seismic recordings collected at Pisciarelli, as we aim to identify recurrent patterns in the seismic waveforms, detect seismic events, and isolate anomalous recordings. Figure 2 provides a summary overview of the adopted strategy. Panel (a) shows an example of a 15-minute-long continuous waveform. The first step of the preprocessing consists in segmenting the continuous waveform into 1-minute windows, as indicated by the red box. Each segmented waveform is then translated into three types of features, shown in panel (b): Linear Prediction Coefficients (LPC), STA/LTA ratio, and Multiscale Entropy (MSE). Finally, as shown in panel (c), the extracted features are used to train the Self-Organizing Map (SOM), which performs unsupervised clustering of the input vectors. The resulting map is analyzed to compute a clustering index, which highlights temporal variations in the similarity of signal features

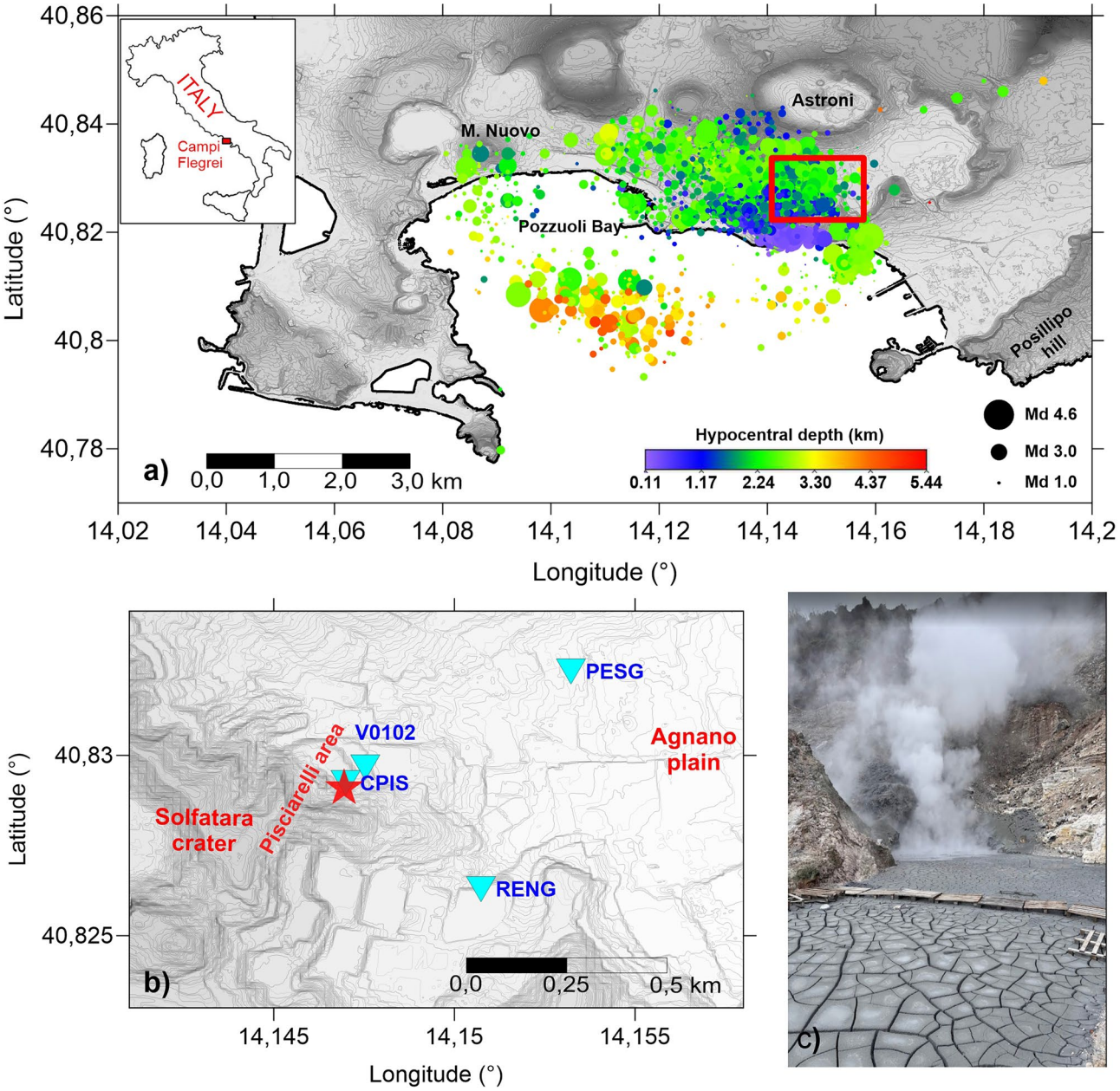

Fig. 1. Study area and monitoring stations in the Campi Flegrei (Italy) caldera. (**a**) Map of seismic events recorded in the Campi Flegrei area, color-coded by hypocentral depth. (**b**) Zoomed view of the Pisciarelli sector and location of the seismic stations used in the analysis: V0102, CPIS (both near the fumarole), PESG, and RENG. (c) Field photograph of the Pisciarelli fumarole and associated mud pool, showing the highly active hydrothermal system. Topographic data are based on the digital terrain model from Vilardo et al.[19]. The figure was created using commercial software Surfer version 22 (Golden Software, LLC) (https://www.goldensoftware.com/products/surfer).

throughout the observation period. To compare and validate detected seismic events, we refer to the INGV seismic catalogue[10], which covers the study period reporting 2,028 earthquakes occurring in the area confined within latitudes 40.797–40.840 and longitudes 14.083–14.170. These events exhibit duration magnitudes from −0.76 to 4.2, and depths between 0.18 km and 4.5 km. Environmental data, including daily rainfall (from the San Marcellino meteorological station), temperature, and CO_2 flux measurements at Pisciarelli, were integrated in the analysis to explore potential relationships with seismic activity. Although the meteorological station is located about 10 km from V0102, recent studies confirm the validity of its rainfall data for the Pisciarelli site[21].

Change in the functionality of the V0102 seismic station

The first SOM network is trained using input vectors consisting of 40 points representing the 40 coefficients computed using the LPC algorithm. A SOM map with 4×4 units or nodes, each displayed as a hexagon, was

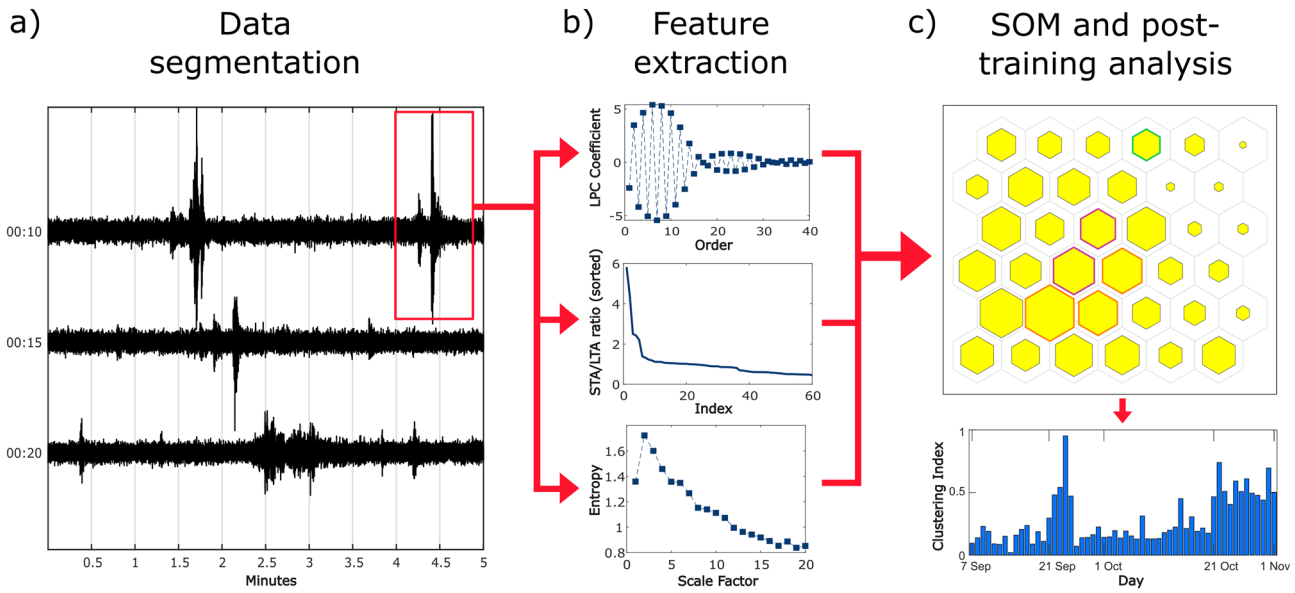

Fig. 2. Summary of the adopted processing workflow, integrating data preprocessing, feature extraction, and unsupervised clustering via Self-Organizing Maps (SOMs). (**a**) An example of a 15-minute continuous waveform recorded at V0102 station is shown, from which individual 1-minute windows are extracted (highlighted in the red box). (**b**) For each 1-minute waveform, three features are computed: Linear Prediction Coefficients (LPC), STA/LTA ratio, and Multiscale Entropy. These features were selected to capture the spectral, energy, and signal complexity, respectively. (**c**) The extracted features are used to construct the input vectors for the SOM training. The resulting SOM map (top) illustrates the organization and density of similar waveform patterns, while the clustering index (bottom) quantifies the temporal variations in the feature distribution.

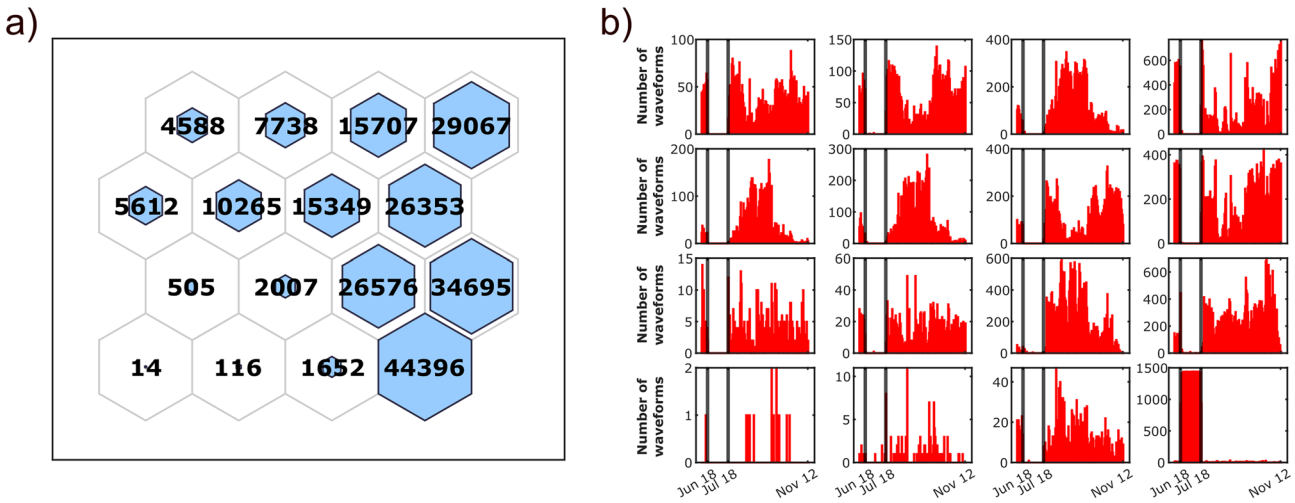

Fig. 3. (**a**) The 4×4 SOM trained using 40 LPC coefficients as encoding features. Each hexagon represents a node of the map, and the corresponding number represents the number of waveforms that fall in each node. The size of the colored hexagons scales with the number of waveforms that fall in each node; (**b**) distribution of waveforms throughout the days of analysis for each map node. The height of the columns indicates the number of signals recorded on a specific day that fall into the node of the map. The two black vertical lines indicate June 18 and July 18.

used for clustering the V0102 dataset. The sizes of the blue hexagons in Fig. 3a represent the number of 1-minute traces that fall into each node. These are called node hits. Traces that fall into the same node form a family. We have individually analyzed all families corresponding to the 16 SOM nodes and, for each, we focus on the temporal histograms of the cumulative number of events per day, as shown in Fig. 3b.

The black vertical lines in each temporal histogram of Fig. 3b mark two pivotal days: June 18 and July 18, 2023. The clear separation in the temporal distribution of the recorded waveforms suggested by the SOM map

highlights an unexpected division of the waveforms into two periods: the timespan between June 18 and July 18, whose waveforms are collected in the bottom right node of the map, and the period outside this interval. Subsequent visual inspection, guided by SOM clustering results, revealed the presence of anomalous waveforms on the key dates of June 18 and July 18. The characteristics of these signals suggest a possible anthropogenic origin, potentially linked to human activities such as infrastructure operations or other external disturbances. Further investigation confirmed that a station malfunction occurred on June 18 that was subsequently repaired on July 18. These occurrences, highlighted by unsupervised learning, underscore the utility of the technique in distinguishing recording differences and evaluating dataset quality.

Uncatalogued seismic events detection

Using LPC coefficients as input features proved to be effective in isolating outliers in the dataset, but no other significant clusters were observed. The STA/LTA ratio, on the other hand, is well-known in seismic monitoring as one of the main tools for seismic event detection. Finally, MSE carries information about waveform complexity, which is a less commonly estimated parameter for waveforms in seismic monitoring. The second self-organizing map training aims to evaluate the effectiveness of using both STA/LTA and MSE values simultaneously as input features, with the goal of proving that the two measures, when coupled, carry complementary information that can improve the dataset analysis. Each 1-minute waveform is encoded using 60 STA/LTA values and 20 MSE values. This analysis spans from July 20 to November 12, post-maintenance of the seismic station V0102 (on July 18). The size of the map, shown in Fig. 4, is set to 6×6 nodes, allowing data to be more widely distributed in the map.

Post-training analysis revealed a distinct region on the SOM map, where a group of nodes in the upper right corner displayed larger inter-node distances compared to the rest of the map, as indicated by the darker-colored hexagons in Fig. 4b. To better understand the relationships between nodes, we applied Ward's hierarchical clustering to the SOM prototypes[22]. The resulting clusters (Fig. S2a) were spatially compact, reflecting the consistency of the learned topological structure and confirming that similar nodes were effectively grouped in close proximity by the SOM. Two clusters highlight the nodes in the upper right corner, with the single node in the corner forming a separate cluster by itself. To quantify the clustering quality, we computed the silhouette score[23], which ranges from -1 to 1 and measures how well each data point fits within its assigned cluster. Values near 1 denote well-separated, cohesive clusters, while values near -1 indicate poor assignment. The accompanying silhouette plot (Fig. S2b) illustrates the clustering structure. The resulting silhouette score of $s = 0.453$ indicates good cluster cohesion and separation, supporting the robustness of the unsupervised strategy. Finally, to evaluate the effectiveness of the SOM map in identifying seismic events, we computed, for each node, the percentage of data corresponding to earthquakes reported in the revised INGV catalogue[10]. As depicted in Fig. 4c, the nodes in the upper right corner of the map contain a high percentage of earthquakes recorded during the analyzed period, highlighting the potential of the SOM map for event detection applications. Nodes in cluster C1 of Figure S2a exhibit the largest percentage. Notably, the data associated with node of event cluster C1 include some uncatalogued seismic events (Figs. 4d and S3a).

Daily clustering variation

The SOM map algorithm is advantageous in that it allows for the projection of only a portion of the dataset or an entire new dataset onto the trained map. Thus, it is possible to exhibit the distribution of each day's recordings on the map and to search for a potential connection between daily seismic activity and environmental factors. Figure 5 shows data from three days of analysis projected on the entire SOM map, while Figure S4 shows the distribution of waveforms throughout the days of analysis for each map node, color-coded by month of analysis. In the present study, we quantify the narrowness of the distribution of single-day hits on the map, around the node with the maximum number of single-day hits. Starting from the projection of the waveforms of a single day on the SOM map, we define a clustering index I which measures how tightly clustered the single-day hits are on the map (more details in "Methods"). A clustering index of 1 indicates that the recordings for a single day are grouped in one node, while if the recordings are uniformly distributed across all nodes of the map, the clustering index value is 0.

The bar chart shown in Fig. 6b represents the daily variation of the clustering index for the analyzed period. The plot shows four main peaks in the period from July 20 to October 21, followed by a quite uniform increase in the following period. The comparison of the distribution of the clustering index with the rainfall shown in Fig. 6a suggests a possible relationship between the two parameters. The observed co-variation can be explained by taking into account the proximity of the analyzed station to the Pisciarelli fumarolic-mud pool, which is characterized by peculiar fumarolic tremors whose amplitude has been connected to the degassing of the fumarole-mud pool system[1]. This is confirmed by the plot of the CO_2 flux daily variation shown in Fig. 6d. A similar plot, also comparing fumarole temperature with the collected data and trends, is shown in Fig. S5.

Nodes that correspond to the highest clustering peaks are highlighted in Fig. 7 with a specific colour for each month. A closer inspection of the waveforms collected in the marked nodes reveals the presence of the fumarolic tremor (Fig. S3C). Interestingly, the map shows that clustering occurrences in August, October, and November are nearby, but September's cluster is in a different area of the map, which suggests that the recordings of fumarolic tremors might be triggered by or related to different phenomena in the area. It must be noted that the clustering of fumarolic tremors in the map is driven by the Multiscale Entropy values, thus suggesting that complexity over different time scales can be a piece of valuable information to extract from seismic waveforms to properly characterize such signals, typical of the fumarole-mud pool system.

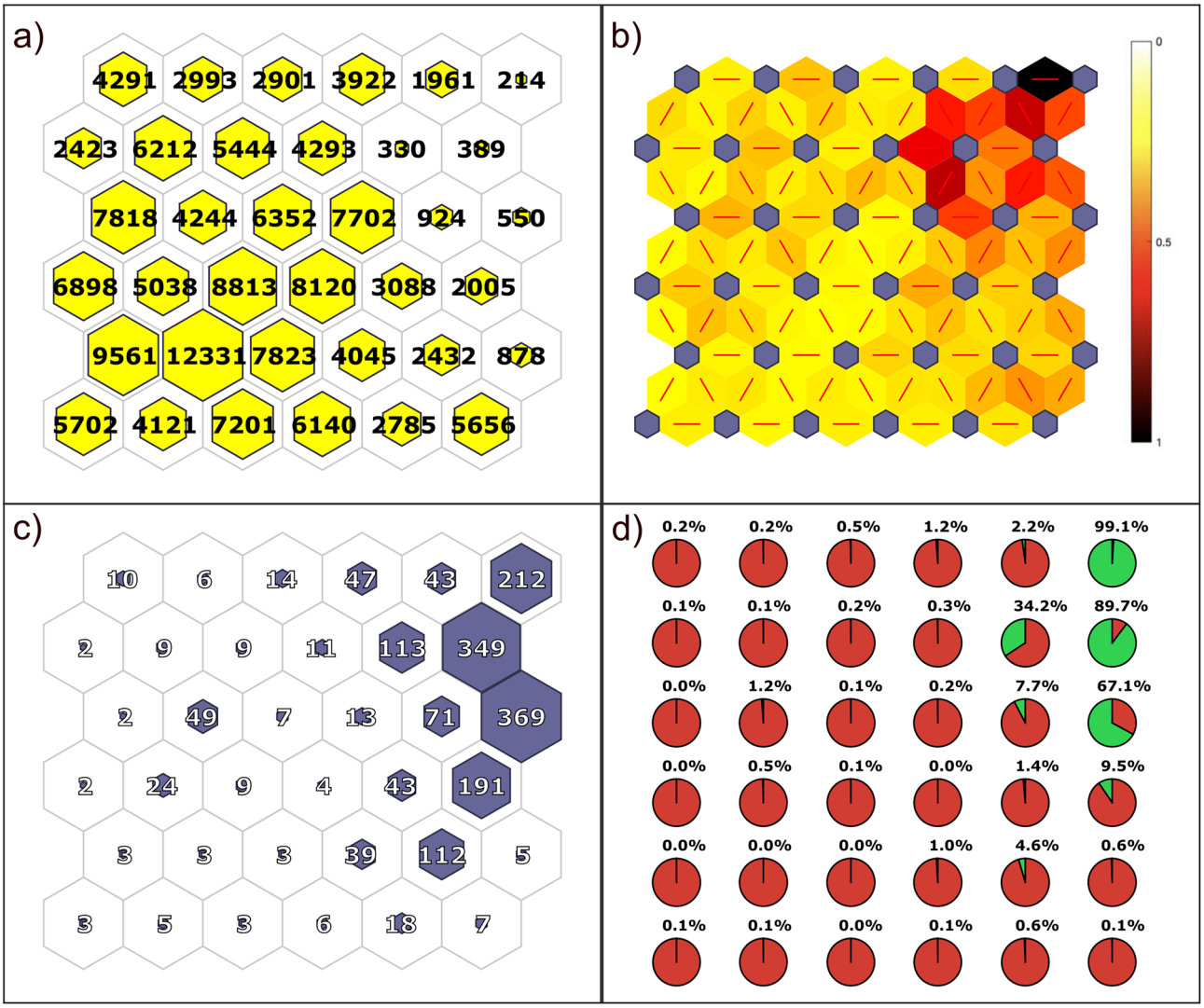

Fig. 4. The 6×6 SOM trained using STA/LTA ratio and Multiscale Entropy as encoding features for seismic waveforms of the restricted Dataset (July 20 – November 12). (**a**) Hits on the 6×6 map. Each hexagon represents a node of the map. The size of the colored hexagons scale with the number of traces that fall in each node. (**b**) Normalized euclidean distances between prototypes of nodes in the SOM map. The colorbar indicates that greater distances are associated with darker colors. (**c**) Hits of the 1-min signals containing a catalogued event in the 6×6 map; (**d**) Pie charts for each node showing the percentage of catalogue events to the total number of waveforms in the node at the top of the graph.

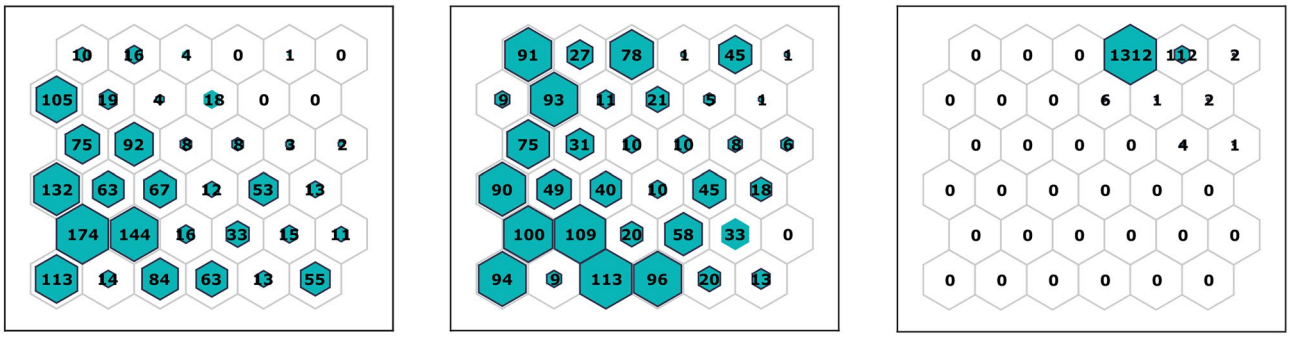

Fig. 5. Projection of the seismic signals recorded on September 6, September 12, and September 24 on the 6×6 SOM Map, from left to right. The corresponding clustering index I is 0.22, 0.09, and 0.95, respectively.

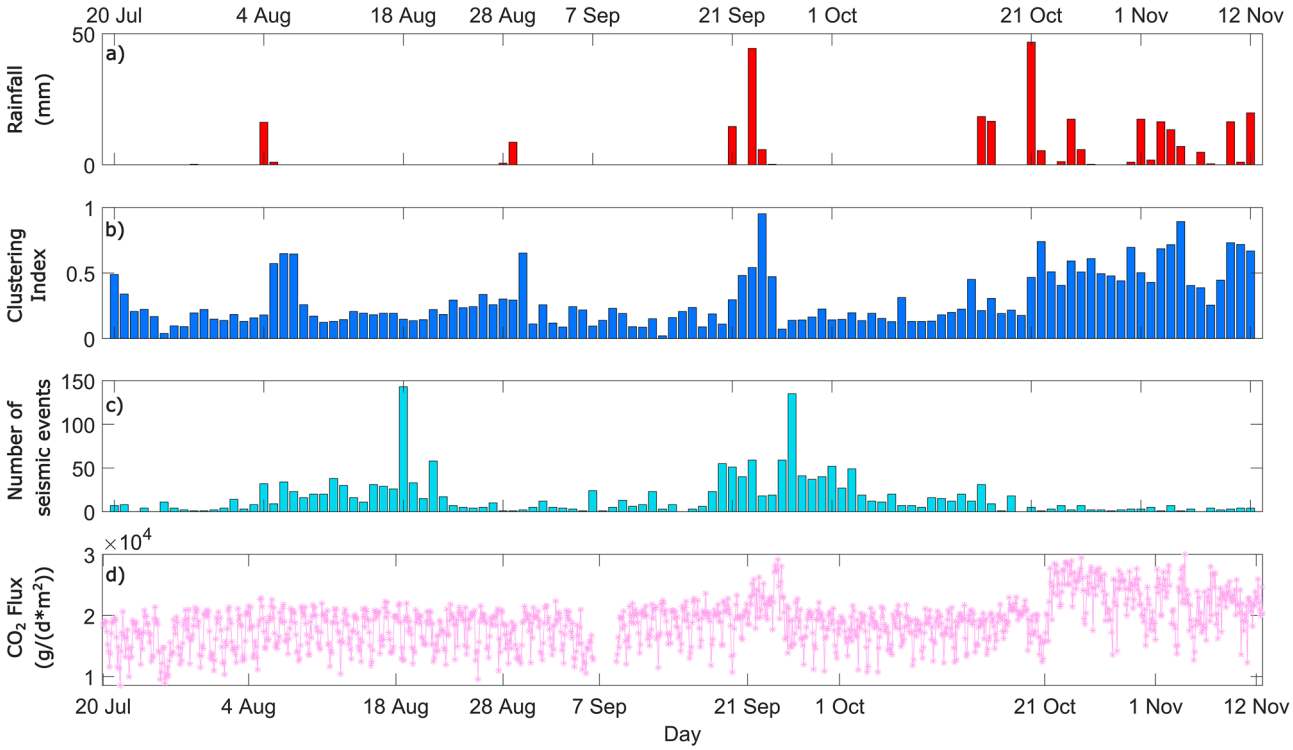

Fig. 6. Four bar charts comparing rainfall millimeters, clustering index of the map, and the daily number of seismic events reported in the catalogue, along with measurements of CO_2 flux (measured in g/(d m^2)) during the days of analysis for the V0102 seismic station dataset.

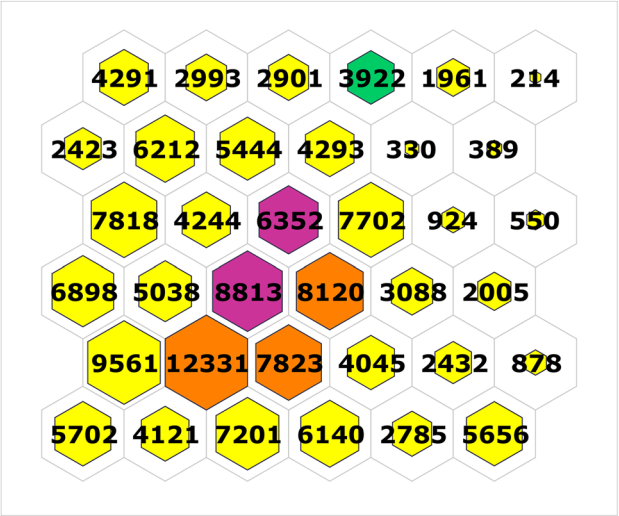

Fig. 7. Hits on the 6×6 map. Each hexagon represents a node of the map. The size of the coloured hexagons scales with the number of traces that fall in each node. The nodes in a different colour than yellow emphasize the existence of a high clustering index (larger than 0.6) in a certain month: hot pink is for August, green is for September, and orange is for October and November.

SOM-based analysis and clustering evolution for PESG and RENG datasets

A similar SOM-based protocol was followed for the analysis of the seismic datasets recorded by the PESG and RENG seismic stations, respectively. The employed PESG dataset was recorded from July 20, 2023, to October 24, 2023, while the RENG spans from August 25 to October 20. The daily evolution of the clustering index for the PESG dataset did not exhibit any significant peaks, as shown in Fig. 8. The clustering index did not exceed the value of 0.3 for any day of the analysis, indicating the absence of notable clusters on the SOM map. On the other

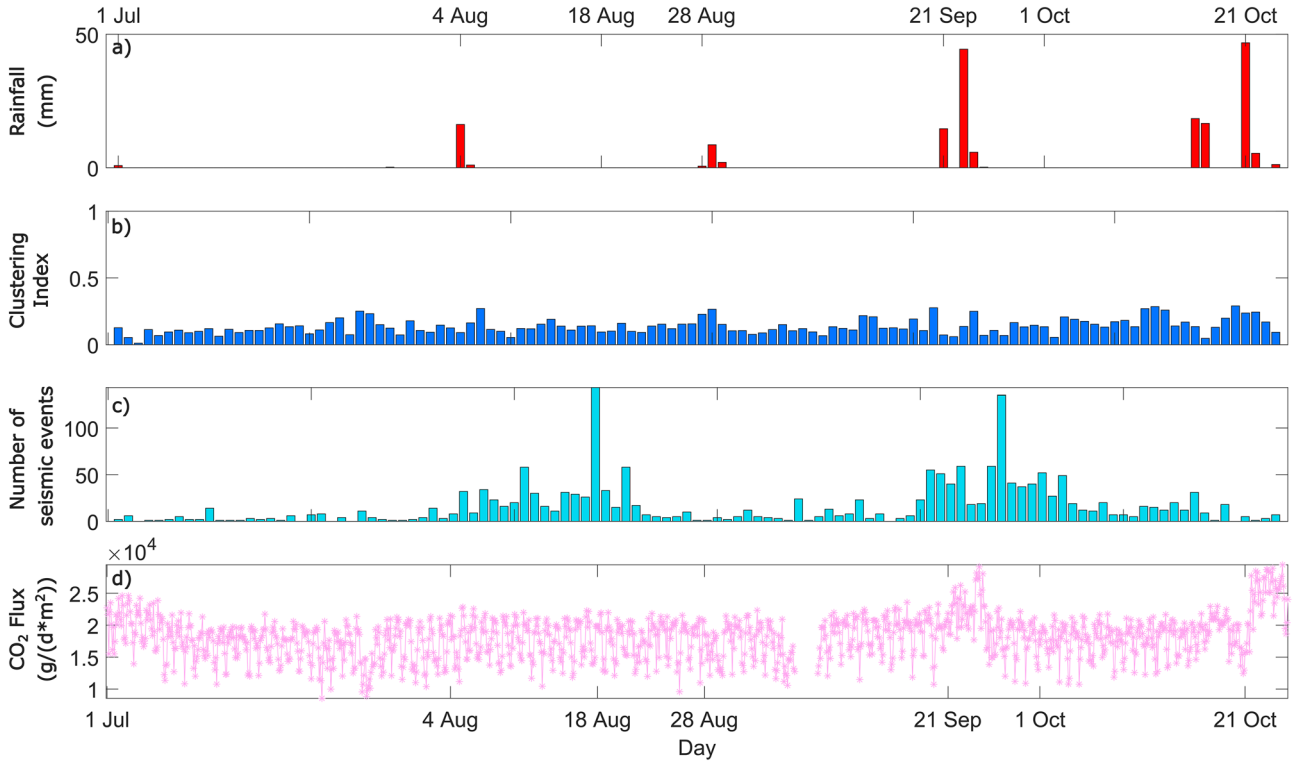

Fig. 8. Four bar charts comparing rainfall millimeters, clustering index of the map, the daily number of seismic events reported in the catalogue, and the measurements of CO_2 flux during the days of analysis for the PESG seismic station dataset.

hand, the daily clustering variation computed for the RENG dataset exhibited a peak on September 24, as shown in Fig. 9. A comparable peak is also evident in the V0102 seismic analysis for the same day (see Fig. 6), hinting at the capacity of the RENG seismic station to effectively record fumarolic tremor data. However, it should be noted that the temporal coverage of the RENG dataset is limited, and a more in-depth analysis will require a larger dataset. The difference between the two results, specifically the absence of a fumarolic tremor-cluster in the PESG map, may be attributed to several factors. These include the relative distance from the fumarole-mud pool system (approximately 600 meters from the PESG station and 500 meters from the RENG station), given that the fumarolic tremor phenomenon is highly localized and the signal attenuates very quickly with the distance[1]. Additionally, differences in wave propagation paths and sensor installation conditions (buried sensor not too close to sources of anthropogenic noise (RENG) vs sensor installed on the floor in a house (PESG)) may contribute to the discrepancies observed.

Real-time application potential: insights from the CPIS dataset (Jan–May 2025)

To further validate the application potential of the proposed technique, we projected new data recorded from January to May 2025 onto the SOM map trained on V0102 data from July 20, 2023, to November 12, 2023. Since the V0102 station was not operational in 2025, we used data from the nearby CPIS station instead. This period is of particular interest because it coincided with a peak in hydrothermal degassing accompanied by a CO_2 flux, along with a temperature increase recorded in the Pisciarelli area. These phenomena are reflected in the signal as a progressive increase in RSAM during the same period. Since the V0102 waveforms were sampled at 200 Hz and the CPIS waveforms at 100 Hz, the V0102 data were downsampled to 100 Hz before training the SOM map. We then repeated the procedure previously performed on the V0102 dataset using the downsampled signals. This approach ensures consistency between datasets and enables the projection of new CPIS signals onto the map trained on V0102 data. We verified that both the SOM map and the clustering index remained substantially unchanged after variation in the sampling rate. This is because the input vectors of the SOM are not significantly influenced by frequency components above 50 Hz (the Nyquist frequency for 100 Hz sampling) due to strong attenuation at higher frequencies. This comparison is shown in Figure S6. Therefore, we used the map developed from the downsampled V0102 data recorded in 2023 to analyze the 2025 data from the CPIS station, without an additional training phase. Figure 10 presents the results of the analysis, where the blue bars indicate the clustering index computed for the CPIS data projected onto the V0102 map. Unlike in the previously analyzed periods, there is no evident co-variation between the clustering index and rainfall, as shown. The most interesting result is that the high clustering index values that persist nearly without interruption from early April to May 13, 2025, occur concurrently with the significant, continuous rise in RSAM values during this period. These values also reflect persistently elevated maximum temperatures and increased hydrothermal tremor activity. This pattern

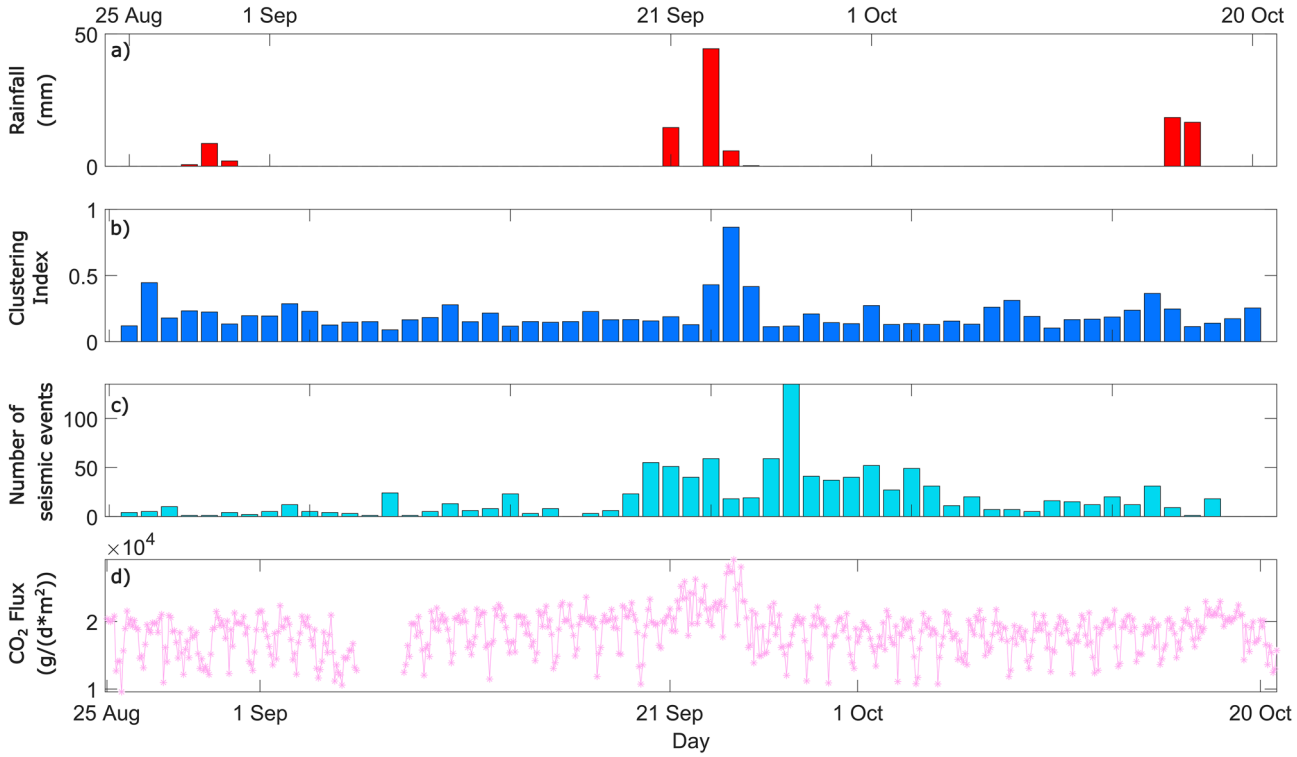

Fig. 9. Four bar charts comparing rainfall millimeters, clustering index of the map, the daily number of seismic events reported in the catalogue, and the measurements of CO_2 flux during the days of analysis for the RENG seismic station dataset.

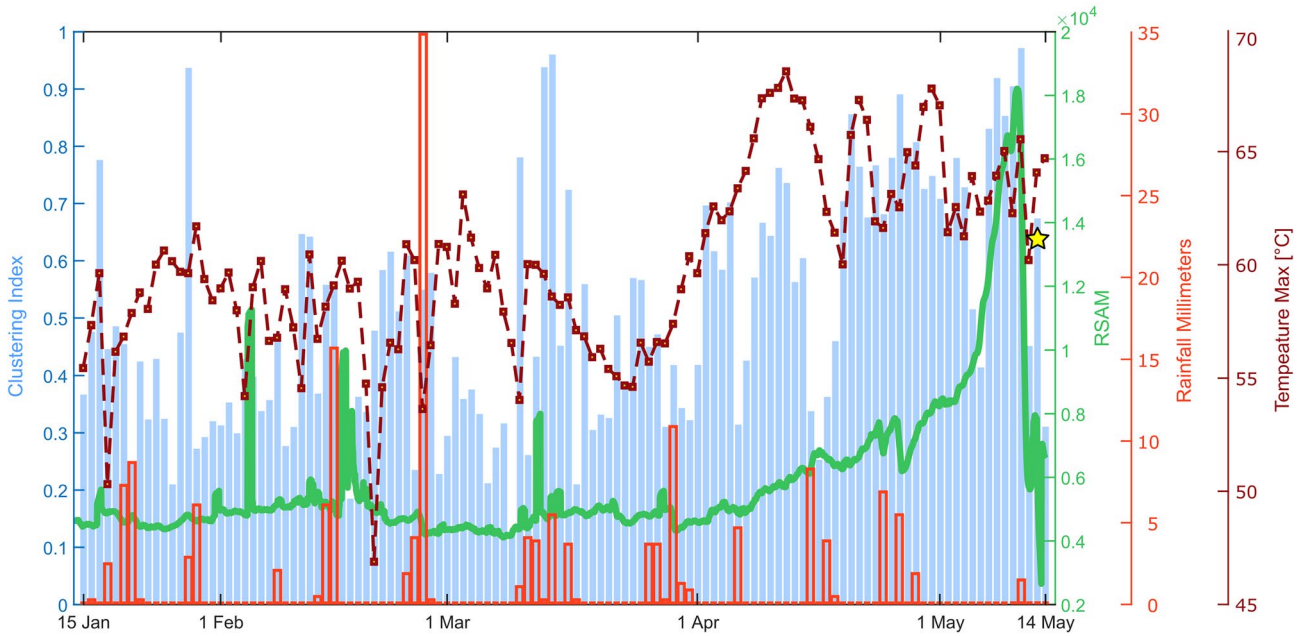

Fig. 10. Variation in the clustering index and RSAM from CPIS station seismic data (Jan–May 2025). Daily values of the clustering index (blue bars) and RSAM (green line) were computed from CPIS station data after projection onto the SOM map, which was trained using the V0102 signals. Also shown are daily fumarole temperature (brick-red line) and daily rainfall (orange bars). The clustering index shows a persistent increase from April to mid-May 2025 that follows the RSAM trend, which reflects rising hydrothermal tremor activity. This pattern suggests an intensification of degassing and seismic signal complexity during this period. On May 13, 2025, the Md 4.4 event, marked by the star, followed a sharp decrease in both RSAM and clustering index.

indicates an intensification of degassing and greater complexity of seismic signals during this timespan. Finally, it is worth noting that the implemented algorithm was able to collect the Md 4.4 event that occurred on 2025-05-13 at 10:07:44.91 in the event cluster. The event was located near the port of Pozzuoli, approximately 3 km from the CPIS station, and occurred right after a sharp decrease of the RSAM value toward the background value.

Conclusions

This study aims to assess the efficacy of employing the Self-Organizing Map (SOM), an unsupervised machine learning algorithm, and distinct data encoding methodologies for the analysis of continuous, large, and complex seismic datasets, such as those from the Campi Flegrei volcanic complex. Unsupervised machine learning techniques can be instrumental in identifying and classifying significant seismic transients, distinguishing them from background noise efficiently and reliably. This approach allows for a simplified visualization of large datasets and effective clustering that preserves the topological relationships of the original dataset. Furthermore, another advantage of the method lies in its single-station applicability, which makes it particularly valuable in settings where the installation of a seismic network is not feasible. By employing various encoding features, including Linear Predictive Coding (LPC) coefficients, STA/LTA ratio, and Multiscale Entropy (MSE) values, this study evaluates the use of the SOM for seismic dataset exploration and characterization. The application of the SOM algorithm on waveforms encoded using LPC coefficients revealed hidden anomalies in the dataset (Jun 9–Nov 12), underscoring the utility and usefulness of this technique in detecting significant variations within seismic data. The joint application of STA/LTA and Multiscale Entropy for data preprocessing on the restricted dataset (Jul 20–Nov 12) has resulted in the clustering of catalogued events and the subsequent recognition of uncatalogued seismic events, thereby demonstrating the efficacy of such protocol in seismic monitoring and analysis in sensitive areas. By using this protocol, seismic waveforms that require further investigation can be isolated and analyzed, providing valuable insight for local seismicity characterization. Moreover, the study of the daily clustering variations indicates the presence of analogous peaks in values for clustering index, CO_2 emissions, and precipitation in the Campi Flegrei caldera, suggesting that environmental factors can influence fumarolic seismic activity. Multiscale Entropy was demonstrated to be effective in capturing changes in seismic waveforms during fumarolic tremor episodes, opening the way for further investigation to confirm the potential of such analysis on seismic waveforms. Originally developed for physiological signal processing, such as electroencephalogram (EEG) and electrocardiogram (ECG) data, MSE's ability to detect waveform changes in a complex setting like the Campi Flegrei caldera highlights its versatility and suggests promising applicability in other volcanic contexts. The same protocol was applied to seismic datasets recorded in a similar period from different stations, which demonstrated that the proximity of the station to the fumarole-mud pool system is a major factor to consider in order to have a proper and efficient clustering of the fumarolic tremor. The final experiment involved the projection of data recorded by the CPIS station between January and May 2025 onto a map that had been trained on the V0102 dataset from July to November 2023. This process necessitated the downsampling of the signals. In this case, no significant association with rainfall was observed. However, a continuous and steady increase in the clustering index between April and May 2025 was noted, which was temporally aligned with a steady increase in RSAM over the aforementioned period. This finding indicates the activation of clustering behavior of the map in the occurrence of hydrothermal degassing along with temperature increase recorded in the Pisciarelli area.

Methods

Dataset

The primary dataset used in this study includes continuous recordings from June 9th to November 12th, 2023 acquired at the V0102 temporary seismic station operated by the Istituto Nazionale di Geofisica e Vulcanologia, Osservatorio Vesuviano, (OV-INGV). The station is equipped with a TELLUS Lunitek short-period sensor and a 6-channel ATLAS C Lunitek datalogger, with a sampling frequency of 200 Hz. We also utilized recordings of two close stand-alone seismic stations, PESG and RENG, that are equipped with a Guralp CMG-40T broadband sensor and a Marslite datalogger, with a sampling frequency of 125 Hz. These two stations are located in the Agnano-Pisciarelli area, about 500 m away from the fumarole-mud pool system, as well as recordings from January to May 2025 of the CPIS seismic station, equipped with a GURALP CMG-40T-60S broadband, with a sampling rate of 100Hz, located about 20 meters away from V0102 old location. The analysis was limited to the vertical component of the recorded signals, as the tremor near the Pisciarelli fumarole is predominantly polarized along this component[8]. To validate the events identified through the analysis of 1-min windows using the technique proposed in this work, we utilized the INGV seismic catalogue. The catalogue covers the study period and includes 2,028 earthquakes with duration magnitude ranging from -0.76 to 4.2, depths between 0.18 km and 4.5 km, and locations within latitudes $[40.797 - 40.840]$ and longitudes $[14.083 - 14.170]$. The average, minimum, and maximum daily temperature of the Caldera, the daily CO_2 emissions measured at the station FLXOV8[24] (Pisciarelli), and the daily rainfall catalogue produced by the Meteorological Observatory of San Marcellino are made available. It should be noted that although the Meteorological Observatory of San Marcellino is located 10 km E from the seismic station V0102, the rainfall data can be taken as valid for comparison[21].

Data preparation

The continuous seismic recording acquired at the V0102 temporary seismic station was divided into 1-min windows, each containing 12,000 samples. Segmenting the data into 1-min windows offers a practical balance across the time scales of key seismic phenomena, such as fumarolic tremors or seismic events. For events with a shorter duration, the feature extraction process subdivides the waveform into smaller windows to ensure

adequate temporal resolution. Conversely, clustering metrics can capture persistent patterns that extend over an entire day of recordings for longer-lasting phenomena. Since the data from August 30th is missing, the total analysis period is 156 days. This approach yielded 224,640 seismic traces (156 days × 1440 traces per day), forming the dataset. Examples of two 1-minute windows are presented in Fig. S7. The same segmentation approach has been applied to generate the datasets for the PESG, RENG, and CPIS station recordings. The adopted approach can be summarized in the following steps: first, the continuous recording is segmented into one-minute waveforms; second, each waveform is converted into a vector of features that encodes its main characteristics (feature extraction); finally, the SOM algorithm is applied to the dataset with the features' vector as input to effectively distinguish different seismic transients within the dataset. The described workflow is shown in Fig. 2. The implementation of the proposed approach involves the application and testing of multiple feature extraction techniques. Feature extraction plays a crucial role in filtering out irrelevant information, reducing data dimensionality, and creating a compact, robust signal representation. To this aim, a tailored feature extraction methodology was implemented to emphasize key aspects of waveform complexity, spectral content, and temporal dynamics, ensuring robust clustering and meaningful insights into the fumarolic tremor characteristics. As features, we use the standard Linear Prediction Coding (LPC) coefficients to compactly encode the spectral content, the classic STA/LTA ratio to track amplitude fluctuations across consecutive moving time windows, and we leverage a promising feature, Multiscale Entropy, which evaluates the complexity of time series across different time scales. The following sections provide insights into each selected encoding feature.

Feature extraction

Linear prediction coding

The signal spectral content is described by the coefficients of the Linear Prediction Coding algorithm (LPC)[25,26]. Techniques based on the LPC algorithm are widely used in speech analysis to encode words[27,28]. Still, they are applied in several other scenarios, as a method for efficient seismic data encoding for neural network classification procedure at Mt. Vesuvius and Stromboli volcano[29,30] or for reducing the amount of data storage requirements for signals collected at ocean-bottom sismometers[31]. The idea is to estimate future values of a time series based on its past values, assuming that the current value in a temporal sequence can be approximated as a linear combination of the previous ones. The method aims to determine a set of coefficients that best fit the historical data, minimizing the least squares error. The inferred coefficients are then used to predict future data points by multiplying them by the corresponding past values and summing the results. Hence, the LPC algorithm coefficients can be used to encode the signal, providing information on its spectrum. The number of coefficients used to encode each signal was selected based on an analysis of the residual prediction error curves, which flatten progressively as the number of coefficients increases (see Fig. S8). The result suggests that adding more than 40 coefficients provides little gain in predictive accuracy. For this reason, 40 coefficients were chosen to encode each signal. Figure S9 shows the coefficients of the LPC algorithm for a trial seismic signal, as well as the comparison between the amplitude spectrum and its spectral envelope, computed using the LPC coefficients.

STA/LTA

Short-time Average over Long-time Average ratio (STA/LTA) is a well-assessed classical method to detect earthquakes in continuous seismic recordings[32]. The aim is to calculate the average values of the absolute amplitude of the seismic signal across two moving temporal windows with different lengths. Short-Time Average (STA) is sensitive to seismic events or sudden variations. At the same time, Long-Time Average (LTA) provides information on the evolution of the seismic amplitude trend over time. Following a testing phase, we chose a short-time window of 1s and a long-time window of 30s, which are reasonable values for local earthquake detection[33]. This choice also increases the chance of avoiding false triggers in an area characterized by high anthropic noise[32]. We obtained 60 values, one associated with each 1s-long window within each seismic trace, representing the evolution of the STA/LTA ratio. Additionally, the STA/LTA values are sorted in descending order to prevent the occurrence of unwanted information about the specific moment of the event's occurrence. The procedure is shown in Fig. S10.

Multiscale entropy

The multiscale entropy (MSE) method was originally introduced by Costa et al. (2002)[17] to analyze and characterize electrocardiogram (ECG) recordings[18,34]. Unlike other features that compress information related to waveform amplitude or frequency spectrum, multiscale entropy focuses on time series complexity. Time series complexity refers to significant, informative structures within the data, such as relevant correlations or dynamics[35]. Traditional analysis methods search for regularity, which is identified as repeated patterns within the series. However, methods like multiscale entropy (MSE) focus on comparing different complexity levels of the time series. As an example, the MSE algorithm assesses both completely deterministic and random signals by applying entropy measures across various time scales, revealing that neither type is complex. To perform a multiscale entropy analysis, for each discrete seismic signal $x_1, ..., x_i, ..., x_N$ we apply a coarse-graining approach to obtain a simplified, shorter time series $y(\tau)$ corresponding to a scale factor τ:

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, \quad 1 \le j \le \frac{N}{\tau} \tag{1}$$

Subsequently, we perform an entropy measure of the time series $y(\tau)$ for each time scale τ. The entropy measure is the Sample Entropy (SE)[36], and it is a measure of the likelihood that sequences similar within a certain tolerance remain similar when extended. More precisely, it is defined as:

$$S_E(m) = -\ln\left(\frac{U^{m+1}}{U^m}\right) \tag{2}$$

which is the negative natural logarithm of the conditional probability that two sequences, which are similar for m consecutive points, remain similar when one more point is added within a tolerance of r. As in the work of Costa et al.(2005)[18], here we use $m = 2$ and a tolerance of $r = 0.15$, where r is considered as a percentage of the original time series' standard deviation, and therefore remains constant for all scales. MSE values computed across multiple temporal scales are used to encode the original signals. Increasing the number of scale points in MSE captures signal complexity at progressively coarser temporal scales. In our analysis, Multiscale Entropy was computed over 20 scales, corresponding to a maximum coarse-graining window of 100 milliseconds (ms) for the V0102 recording, which was sampled at 200Hz. To assess the robustness of this choice, we retrained the SOM using different values of MSE scale points. We observed no critical dependence of the key results, such as the daily clustering index and event-to-node distributions, on this parameter. An example of an MSE vector is shown in Figure S11.

Self-Organizing Map

The Self-Organizing Map (SOM)[37–39] is an unsupervised machine learning method well-suited for the visualization and exploration of large datasets. It clusters similar data and displays them on a low-dimensional grid while preserving their topological relationship. Unlike traditional clustering methods, such as k-means or hierarchical clustering, which require extensive pairwise comparisons and become computationally expensive as the dataset size grows, the Self-Organizing Map (SOM) scales linearly with the number of samples and requires minimal memory. This makes the SOM well-suited for exploring large datasets. The method has been applied in different fields, ranging from applications in data visualization[40] and telecommunication[41] to economics[42] and medicine[43]. A comprehensive collection of papers related to the Self-Organizing Map (SOM) is compiled in the work of Polla et al.(2009)[44]. In Geophysics, it has been widely applied to characterize the seismicity of volcanic areas, by detecting changes in the eruptive style of Stromboli volcano[45,46] or distinguishing the typical fumarolic tremor in a continuously recorded dataset at Campi Flegrei[20]. Each node of the SOM map is associated with a model or prototype, which is a vector computed as the mean of the nearest encoded signals, and the model is updated during the training process. For an efficient and meaningful application of the SOM algorithm, data preprocessing is required to extract the essential features for each trace, remove unnecessary or unwanted information, and reduce data size. Identifying the correct features to extract from the seismic signals is critical, as it affects the clustering we obtain from the SOM map. We have chosen to extract the three features discussed above from each 1-minute window. Each feature represents a different aspect of the seismic waveform: spectral content, amplitude variation, and complexity. Additionally, LPC and MSE values are standardized component-wise (to zero mean and unit variance) to ensure uniform weighting across all features during training. This standardization step is not applied to STA/LTA, as its information is encoded in a sequence of values sorted in descending order, and altering their relative scale would distort this structure. It is worth noting that all three features used in the input vectors share the same order of magnitude, so there is no concern regarding unbalanced weighting among them. The method was implemented in MATLAB, using the Deep Learning Toolbox for both training and evaluation of the SOM network. After training the SOM map, we show the number of waveforms that fall into each node. For each node, we also show the temporal distribution of when the traces assigned to that node were recorded during the registration period. It can be seen that some nodes have a temporal distribution with very low dispersion, highly concentrated within a relatively short time interval. Interestingly, low temporal dispersion also corresponds to low spatial dispersion on the map, highlighting the role of the SOM algorithm in identifying periods of anomalous seismic activity within a continuous dataset. Such daily spatial distribution is quantified by an index introduced in this work, the clustering index. The idea is to quantify the narrowness of the distribution of single-day hits on the map, around the node with the maximum number of single-day hits. In particular, we define the index I as follows:

$$I = \frac{hits_{max} + \frac{1}{2}hits_{nn} + c \cdot hits_{others}}{hits_{total}} \tag{3}$$

where $hits_{max}$, $hits_{nn}$ and $hits_{other}$ are the number of traces in the most densely populated node, the number of signals collected in the neighbourhood of such node and the remaining signals of the specific day, respectively, while $hits_{total}$ is the total number of recordings in a single day. The parameter c is the (negative) value that ensures that the index I is zero when the waveforms are uniformly distributed over the map. Due to edge effects, it depends on the location of the most densely populated node on the map. A clustering index of 1 indicates that the recordings for a single day are grouped into one node. If the recordings are uniformly distributed across all nodes of the map, however, the clustering index is 0.

Data availability

Seismic waveforms for station CPIS are available from the following seismic networks: Rete Sismica Nazionale (FDSN code: IV), operated by the Istituto Nazionale di Geofisica e Vulcanologia (INGV, 2005); Seismic waveforms for V0102, PESG and RENG stations are available from the corresponding author on reasonable request.

Received: 25 June 2025; Accepted: 29 January 2026
Published online: 07 February 2026

References

1. Giudicepietro, F. et al. Insight into Campi Flegrei caldera unrest through seismic tremor measurements at Pisciarelli fumarolic field. *Geochem. Geophys. Geosyst.* **20**. https://doi.org/10.1029/2019GC008610 (2019).
2. Calò, M. & Tramelli, A. Anatomy of the Campi Flegrei caldera using enhanced seismic tomography models. *Sci. Rep.* **8**. https://doi.org/10.1038/s41598-018-34456-x (2018).
3. D'Auria, L. et al. Repeated fluid-transfer episodes as a mechanism for the recent dynamics of Campi Flegrei Caldera (1989–2010). *J. Geophys. Res. Solid Earth* **116**, B04313. https://doi.org/10.1029/2010JB007837 (2011).
4. Chiodini, G. et al. Clues on the origin of post-2000 earthquakes at Campi Flegrei Caldera (Italy). *Sci. Rep.* **7**, 4472. https://doi.org/10.1038/s41598-017-04845-9 (2017).
5. Iannaccone, G. et al. Measurement of seafloor deformation in the marine sector of the Campi Flegrei Caldera (Italy). *J. Geophys. Res. Solid Earth* **123**, 66–83. https://doi.org/10.1002/2017JB014852 (2018).
6. Chiodini, G., Caliro, S., De Martino, P., Avino, R. & Gherardi, F. Early signals of new volcanic unrest at Campi Flegrei Caldera? Insights from geochemical data and physical simulations. *Geology* **40**, 943–946. https://doi.org/10.1130/G33251.1 (2012).
7. Tamburello, G. et al. Escalating CO2 degassing at the Pisciarelli fumarolic system, and implications for the ongoing Campi Flegrei unrest. *J. Volcanol. Geotherm. Res.* **384**, 151–157. https://doi.org/10.1016/j.jvolgeores.2019.07.005 (2019).
8. Chiodini, G. et al. Fumarolic tremor and geochemical signals during a volcanic unrest. *Geology* **45**, 1131–1134. https://doi.org/10.1130/G39447.1 (2017).
9. Giacomuzzi, G., Chiarabba, C., Bianco, F., De Gori, P. & Agostinetti, N. P. Tracking transient changes in the plumbing system at Campi Flegrei caldera. *Earth Planet. Sci. Lett.* **637**, 118744. https://doi.org/10.1016/j.epsl.2024.118744 (2024).
10. Ricciolino, P., Lo Bascio, D. & Esposito, R. GOSSIP—Database Sismologico Pubblico INGV-Osservatorio Vesuviano. https://doi.org/10.13127/gossip (2024).
11. Convertito, V. & Zollo, A. Assessment of pre-crisis and syn-crisis seismic hazard at Campi Flegrei and Mt. Vesuvius volcanoes, Campania, southern Italy. *Bull. Volcanol.* **73**, 767–783 https://doi.org/10.1007/s00445-011-0455-2 (2011).
12. Convertito, V., Giampaolo, F., Amoroso, O. & Piccialli, F. Deep learning forecasting of large induced earthquakes via precursory signals. *Sci. Rep.* **14**, 2964. https://doi.org/10.1038/s41598-024-52935-2 (2024).
13. Perol, T., Gharbi, M. & Denolle, M. Convolutional neural network for earthquake detection and location. *Sci. Adv.* **4**, e1700578. https://doi.org/10.1126/sciadv.1700578 (2018).
14. Weiqiang, Z. & Beroza, G. Phasenet: A deep-neural-network-based seismic arrival time picking method. *Geophys. J. Int.* **216**. https://doi.org/10.1093/gji/ggy423 (2018).
15. Wallet, B. C. & Hardisty, R. Unsupervised seismic facies using gaussian mixture models. *Interpretation* **7**, SE93–SE111. https://doi.org/10.1190/INT-2018-0119.1 (2019).
16. Meyer, S. G., Reading, A. M. & Bassom, A. P. The use of weighted self-organizing maps to interrogate large seismic data sets. *Geophys. J. Int.* **231**, 2156–2172. https://doi.org/10.1093/gji/ggac322 (2022).
17. Costa, M., Goldberger, A. L. & Peng, C.-K. Multiscale entropy analysis of complex physiologic time series. *Phys. Rev. Lett.* **89**, 068102. https://doi.org/10.1103/PhysRevLett.89.068102 (2002).
18. Costa, M., Goldberger, A. & Peng, C.-K. Multiscale entropy of biological signals. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **71**, 021906. https://doi.org/10.1103/PhysRevE.71.021906 (2005).
19. Vilardo, G., Guido, V., Bellucci Sessa, E. & Terranova, C. Morphometry of the Campi Flegrei Caldera (southern Italy). *J. Maps* **9**, 635–640. https://doi.org/10.1080/17445647.2013.842508 (2013).
20. Esposito, A., Cesare, W., Macedonio, G. & Giudicepietro, F. Efficient SOM's application to seismic fumarolic tremor for the detection of anomalous hydrothermal activity in Campi Flegrei volcano (Italy). *Appl. Sci.* **13**, 5505. https://doi.org/10.3390/app13095505 (2023).
21. Scafetta, N. & Mazzarella, A. On the rainfall triggering of Phlegraean fields volcanic tremors. *Water* **13**, 154. https://doi.org/10.3390/w13020154 (2021).
22. Ambroise, C., Sèze, G., Badran, F. & Thiria, S. Hierarchical clustering of self-organizing maps for cloud classification. *Neurocomputing* **30**, 47–52. https://doi.org/10.1016/S0925-2312(99)00141-1 (2000).
23. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7 (1987).
24. Caliro, S. et al. Dataset of geochemical data acquired by the flxov8 multiparametric station in the period from October 2018 - May 2025 installed in Pisciarelli area at Campi Flegrei (southern Italy). https://doi.org/10.5281/zenodo.15704715 (2025).
25. Deng, L. & O'Shaughnessy, D. Speech processing: A dynamic and Optimization-Oriented Approach Marcel Dekkar Inc. (2003) ISBN: 0-8247-4040-8
26. Makhoul, J. Linear prediction: A tutorial review. *Proc. IEEE* **63**, 561–580. https://doi.org/10.1109/PROC.1975.9792 (1975).
27. Brown, M. K. & Rabiner, L. R. On the use of energy in lpc-based recognition of isolated words. *Bell Syst. Tech. J.* **61**, 2971–2987. https://doi.org/10.1002/j.1538-7305.1982.tb02287.x (1982).
28. Mada Sanjaya, W., Anggraeni, D. & Santika, I. P. Speech recognition using linear predictive coding (LPC) and adaptive neuro-fuzzy (ANFIS) to control 5 DOF arm robot. *J. Phys. Conf. Ser.* **1090**, 012046. https://doi.org/10.1088/1742-6596/1090/1/012046 (2018).
29. Esposito, A., D'Auria, L., Giudicepietro, F., Caputo, T. & Martini, M. Neural analysis of seismic data: Applications to the monitoring of Mt. Vesuvius. *Ann. Geophys.* **56**. https://doi.org/10.4401/ag-6452 (2013).
30. Esposito, A., Giudicepietro, F., Scarpetta, S. & Khilnani, S. *A Neural Approach for Hybrid Events Discrimination at Stromboli Volcano*. 11–21. https://doi.org/10.1007/978-3-319-56904-8_2 (Springer, 2018).
31. Bordley, T. Linear predictive coding of marine seismic data. *IEEE Trans. Acoust. Speech Signal Process.* **31**, 828–835. https://doi.org/10.1109/TASSP.1983.1164144 (1983).
32. Trnkoczy, A. Understanding and parameter setting of sta/lta trigger algorithm. In *New Manual of Seismological Observatory Practice (NMSOP-2)* (Bormann, P. Ed.). Chap. 8.1 (Deutsches GeoForschungsZentrum GFZ, 2009).
33. Carrera, E., Pérez, A. & Lara-Cueva, R. *Automated Systems for Detecting Volcano-Seismic Events Using Different Labeling Techniques*. 133–144. https://doi.org/10.1007/978-3-030-42520-3_11 (Springer, 2020).
34. Zhang, Y., Wei, S., Long, Y. & Liu, C. Performance analysis of multiscale entropy for the assessment of ECG signal quality. *J. Electr. Comput. Eng.* **2015**, 563915. https://doi.org/10.1155/2015/563915 (2015).
35. Eaton, W., Haindl, C. & Nissen-Meyer, T. Seismic scattering regimes from multiscale entropy and frequency correlations. *Geophys. J. Int.* **237**, 1109–1128. https://doi.org/10.1093/gji/ggae098 (2024).
36. Richman, J. & Moorman, J. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* **278**, H2039–49. https://doi.org/10.1152/ajpheart.2000.278.6.H2039 (2000).
37. Kohonen, T., Oja, E., Simula, O., Visa, A. & Kangas, J. Engineering applications of the self-organizing map. *Proc. IEEE* **84**, 1358–1384. https://doi.org/10.1109/5.537105 (1996).
38. Kohonen, T. *Self-Organizing Maps*. https://doi.org/10.1007/978-3-642-56927-2 (Springer, 2001).
39. Kohonen, T. Essentials of the self-organizing map. *Neural Netw.* **37**. https://doi.org/10.1016/j.neunet.2012.09.018 (2012).
40. Nakayama, H. et al. Comparative transcriptomics with self-organizing map reveals cryptic photosynthetic differences between two accessions of North American lake cress. *Sci. Rep.* **8**. https://doi.org/10.1038/s41598-018-21646-w (2018).
41. Tan, X., Wang, P., Hu, H., Cheng, R. & Bai, Y. Combining self-organizing map and Lipschitz condition for estimation in direction of arrival. *Open J. Appl. Sci.* **13**, 1012–1028. https://doi.org/10.4236/ojapps.2023.137081 (2023).

42. Länsiluoto, A., Eklund, T., Back, B., Vanharanta, H. & Visa, A. Industry-specific cycles and companies' financial performance comparison using self-organizing maps. *Benchmark. Int. J.* **11**, 267–286. https://doi.org/10.1108/14635770410538754 (2004).
43. Beckonert, O., Monnerjahn, J., Bonk, U. & Leibfritz, D. Visualizing metabolic changes in breast-cancer tissue using 1h-NMR spectroscopy and self-organizing maps. *NMR Biomed.* **16**, 1–11. https://doi.org/10.1002/nbm.797 (2003).
44. Pöllä, M., Honkela, T. & Kohonen, T. Bibliography of self-organizing map (SOM) papers: 2002-2005 addendum. In TKK Reports in Information and Computer Science, Helsinki University of Technology. Technical Report, Report TKK-ICS-R24 (2009).
45. Giudicepietro, F. et al. Changes in the eruptive style of Stromboli volcano before the 2019 paroxysmal phase discovered through SOM clustering of seismo-acoustic features compared with camera images and Gbinsar data. *Remote Sens.* **14**, 1287. https://doi.org/10.3390/rs14051287 (2022).
46. Romano, P. et al. Dynamic strain anomalies detection at stromboli before 2019 vulcanian explosions using machine learning. *Front. Earth Sci.* **10**. https://doi.org/10.3389/feart.2022.862086 (2022).

Acknowledgements

The daily rainfall catalogue produced by the Meteorological Observatory of San Marcellino are made available. The data are available from (https://www.meteo.unina.it,) accessed on 17 January 2025. The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

Author contributions

O.A., S.S. and V.C conceived the work, A.G., S.S. developed the software, A.G., G.M., O.A, S.S and F.N. analysed the results. A.G. O.A. S.S. V.C. F.N. and G.M. wrote the paper. D.G., L.N. and G.G. were responsible for the installation of the seismic stations and for providing the data in a processable format. P.C. provided part of the funding to conduct the research. All authors reviewed the manuscript.

Funding

The work has been partially supported by the following projects: TOGETHER - Sustainable geothermal energy for two Southern Italy regions: geophysical resource evaluation and public awareness financed by European Union–Next Generation EU Piano Nazionale di Ripresa (PNRR) e Resilienza Missione 4 "Istruzione e Ricerca"-Componente C2, Investimento 1.1, "Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)" CUP: D53D23022850001, project code: P2022JF5JE; D.I.R.E.C.T.I.O.N.S. - Deep learning aIded foReshock deteCTIOn Of iNduced mainShocks, project code: P20220KB4F, CUP: E53D23021910001, PRIN 2022 - Piano Nazionale di Ripresa e Resilienza (PNRR), Mission 4 "Istruzione e Ricerca" - Componente C2 Investimento 1.1, "Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)".

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-38257-5.

Correspondence and requests for materials should be addressed to O.A.