~~How can~~ Could seismo-volcanic ~~catalogues~~ catalogus be improved or created using ~~robust neural networks through~~ weakly supervised approaches with pre-trained systems?

Manuel Titos[1,2], Carmen Benítez[1,2], Luca D'Aria[3], Milad Kowsari[4], and Jesús M. Ibáñez[5,6]

[1]Department of Signal processing, Telematics and Communications, University of Granada, Granada, 18014, Spain
[2]Research Center on Information and Communication Technologies of the University of Granada (CITIC-UGR)
[3]Volcanological Institute of the Canary Islands, Tenerife, 38400, Spain
[4]University of Iceland, Faculty of Civil and Environmental Engineering, Reykjavík, 102, Iceland
[5]Instituto Andaluz de Geofísica, University of Granada, Granada, 18071, Spain
[6]Department of Theoretical Physics and the Cosmos, University of Granada, Granada, 18071, Spain

Abstract. Real-time monitoring of volcano-seismic signals is complex. Typically, automatic systems are built by learning from large seismic ~~catalogues~~catalogs, where each instance has a label indicating its source mechanism. However, building complete ~~catalogues~~ catalogs is difficult owing to the high cost of data-labelling. Current machine learning techniques have achieved great success in constructing predictive monitoring tools; however, ~~catalogue-based~~ catalog-based learning can introduce bias into the system. Here, we show that while monitoring systems ~~recognize almost~~ trained on annotated data from seismic catalogs achieve performance of up to 90% ~~of events annotated in seismic catalogues~~in event recognition, other information describing volcanic behavior is not considered or either discarded. We found that weakly supervised learning approaches have the remarkable capability of simultaneously identifying unannotated seismic traces in the ~~catalogue and correcting mis-annotated~~ catalog and correcting misannotated seismic traces. When a system trained with a master dataset and ~~catalogue~~ catalog is used as a pseudo-labeller within the framework of weakly supervised learning, information related to volcanic dynamics can be revealed and updated. Our results offer the potential for developing more sophisticated semi-supervised models to increase the reliability of monitoring tools. For example, the use of more sophisticated pseudo-labelling techniques involving data from several ~~catalogues~~ catalogs could be tested. Ultimately, there is potential to develop universal monitoring tools able to consider unforeseen temporal changes in monitored signals at any volcano.

1 Introduction

Understanding the dynamics of active volcanoes ~~, and~~and, even more so, carrying out Early Warning protocols for volcanic eruptions require multiparametric observations focused on ~~accomplish~~ accomplishing accurate and effective monitoring (Sparks, 2003). The objective of identifying precursors that warn of a possible volcanic eruption involves the analysis of long temporal series of data, characterizing and relating them with source models associated with the internal dynamics of the volcano (Witze, 2019; Palmer, 2020). Currently, the availability of multiparametric long-time data series, such as seismology, deformation, measurements of volcanic gases and fluids, space imaging, and other processes, is limited to a few volcanoes around the world. For this reason, volcanic seismology continues being the backbone of the analysis, both in real time and using data from previous eruptive episodes (Chouet, 2003; McNutt, 2015). This is because the installation and acquisition of seismic data continues to be the most efficient procedure of volcanic monitoring, and because the existence of numerous open access repositories allows the scientific community reviewing consolidated databases to understand what occurred in the past for modelling future eruptions.

In volcanic seismology, the presence of various seismic signals—such as volcano-tectonic earthquakes (VT), long-period events (LP), ultra-long-period (~~VLP~~ULP) events, hybrid (HY) events, explosions (EXP), and volcanic tremors (TR)—indicates the existence of multiple seismic sources, which can sometimes operate simultaneously and must be considered. Thus, models of brittle rock fracturing, conduit resonance, pressure transients in fluids, bubbles, cracking in viscoelastic mediums, elastic energy transfer by fluid flow, debris flows, and many others are used (Ibáñez et al., 2000; McNutt and Roman, 2015; Minakami, 1974)) ~~. Hence, this complex system~~ (Table 1 summarizes the source models and classifications for different authors). The complexity of seismic sources ~~results in~~ leads to varying interpretations of ~~volcano dynamicsbased on~~ volcanic dynamics, influenced by the predominant signal type and its spatio-temporal evolution. ~~Recently, Rey-Devesa et al. (2023, 2023) developed a series of algorithms based on the processing of seismic signals that allow efficient early warnings of volcanic eruptions. However, comprehending~~ Comprehending the underlying physics behind the eruptions, and thus understanding why ~~eruptions~~ they occur, cannot be solely explained through such signal processing. It requires knowledge of the frequency and types of seismic events that ~~occurred.~~take place. This understanding is primarily gained by constructing seismic catalogs, which are then analyzed to infer volcanic dynamics in future crises. However, building complete catalogs presents significant challenges due to factors such as noisy signals, human error, intense seismic activity, and overlapping signals, all of which complicate the identification and classification of seismic events.

Historically, seismic catalogs have been manually created by experts, with the classification of seismic signals based on time-frequency characteristics and wave-field properties. The process relies heavily on expert knowledge, which, while essential, can introduces potential biases. These biases may arise from various factors, such as the prevailing scientific understanding at the time of labeling, or the occurrence of intense seismic activity where, due to time constraints, only the most energetic events are highlighted, or even when the energies are not high enough, overlapping signals are classified as a single event, leading to the combination of different types of signals under a single label. This issue was notably observed during the 2011 eruption on the island of El Hierro, where continuous VT events resulted in a high-frequency signal resembling volcanic tremor due to the overlap of hundreds of VTs per hour (Ibáñez et al., 2012; Díaz-Moreno et al., 2015). Despite the efforts made, such challenges remain widespread across seismic databases worldwide, highlighting the need for improved methods of signal classification and event labeling.

The introduction of automatic recognition procedures for earthquake-volcanic signals almost two decades ago (e.g. Ohrnberger 2001, Scarpetta et al., 2005; Alasonati et al., 2006; Benítez et al., 2006; Ibáñez et al., 2009, Curilem et al., 2009~~) has made~~

Ibáñez, J.M et al. (2000)	McNutt, S. and Roman, D. (2015)	Minakami, T. (1974)	Frequency [Hz]	Example source models
Volcano Tectonic Earthquakes Tectonic Short Period Earthq.	High Frequency (HF)	A-Type	>5	Shear failure or slip along faults, usually as swarms within the volcanic edifice
Long Period Event Volcanic Long Coda Event Tornillo	Low Frequency (LF)	B-Type	1-5	Fluid driven cracks, pressurization processes (bubbles), and attenuated waves
Hybrid Event Medium Frequency	Mixed Frequency (MX)	-	1-12	Mixture of processes (e.g., cracks and fluids, frictional melting)
Explosion Volcanic Explosion	Explosion Quake (EXP)	Explosion Quake	>10	Accelerated emissions of gas and debris to the atmosphere
Volcanic Tremor Harmonic Tremor	Volcanic Tremor (TRE)	Volcanic Tremor	1-12	Pressure disturbance, gas emissions, debris processes, and pyroclastic flows

Table 1. Representative volcano-seismic scientific labels and associated source models proposed by Ibáñez, J.M. et al. (2000). Other labels and associated source models proposed by different authors have been included for comparison.

~~the process of identifying and characterizing signals more efficient, faster, and comprehensive. The use of procedures based on the application of classic Machine Learning (ML) models as Support Vector Machine (SVM) or, Hidden Markov Models (HMM), and different types of Artificial Neural Networks as Baeysian Neural Networks (BNN), Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN) among others (e.g.~~ , Bhatti et al. 2016; Canario et al., 2020 ; Cortés et al., 2021; Bueno et al. (2021, 2022); Martínez et al. 2021; Titos et al. (2017, 2018, 2019), Bicego et al., 2022~~) allowed progress in a higher quality of reconnaissance procedures (unsupervised, semi-supervised and supervised) and their implementation in real time for the~~ , etc) has made the process of identifying and characterizing signals more efficient, faster and comprehensive, allowing progress in both building robust catalogs and real-time monitoring of active volcanoes. ~~Although there are currently numerous and highly advanced works showcasing successes in this field~~However, the results obtained have begun to reveal potential ~~new problems.~~ problems: *monitoring systems loss effectiveness when recognizing events over time, which biases the construction of seismic catalogs and, in turn, affects experts' ability to analyze and understand volcanic dynamics.*

These outcomes raise open questions that should be efficiently addressed to adequately comprehend ~~the derived source models. Some of these questions include: (a) Is~~ and solve such problems: a) Why do monitoring systems lose effectiveness? Could it be because volcanoes do not behave uniformly over time, displaying different unrest patterns from eruption to eruption and from one volcano to another? (b) Could it be that automatic monitoring systems show weakness due to seismic catalog-induced bias in their development? That is, is the database used ~~as a reference properly labelled? That is, are~~ during the development process properly labeled? Are the signal names or labels ~~adequately~~ accurately identified? (~~b) How~~ c) Finally, how do seismic attenuation processes or source radiation patterns influence changes in the appearance of a signal, thus confounding the associated source models? ~~(c)~~ How could background seismic noise affect the identification of seismic events?~~For the second~~

For the last open question, ~~there are potential solutions. In~~ it is well-know that seismic waves carry information not only on volcanic activity but also on the intricate internal structure of the volcanic edifice, which influences the seismic wave-field and complicates its interpretation (Titos et al. (2018)~~, it was demonstrated that a signal interpreted as a VT at short distances appeared as an LP at another seismic station only a few kilometres away from the reference station due to the high seismic attenuation in volcanic environments. As potential solution to these ambiguities,~~). At many volcanoes, rugged and pronounced topography introduces additional complexities, such as wave interference, high attenuation, and path alterations for direct seismic waves. Consequently, even for the same volcano and the same originating seismic source, recordings vary in shape and wave-field characteristics depending on seismometer placement. Furthermore, even at the same seismic station, similar sources may produce different signal patterns due to variations in the source's energy radiation. These effects are broadly categorized into path-related (attenuation) and source-related (energy and radiation pattern) influences (Titos et al. (~~2023)~~2018)). As a potential solution, experts propose using a ~~system~~ network of multiple seismic stations for signal recognition and ~~"establishing"~~ defining rules or conditions ~~for identifying~~ to identify signals simultaneously.~~Specifically, perform a second recognition procedure and assign the 'type of signal' based on the station with the highest high-frequency component.~~ The first and ~~third~~ second open questions may potentially be more difficult to resolve. ~~An incomplete or mislabelled seismic catalogue can result from various~~ Volcanic behavior is highly variable, exhibiting different signs of unrest between eruptions and between volcanoes. Environmental and geological factors, such as ~~noisy signals, human error, intense and overlapping seismic activity, among others. Most available seismic catalogues were manually developed based on time-frequency characteristics as well as wave-field properties obtained from the seismic signals. Human factors, including bias in manual labelling or influence from~~ geology, magma composition, and the ~~prevailing scientific understanding at the time of labelling, may also contribute to incomplete or inaccurate catalogues. The occurrence of intense seismic activity can significantly impact the completeness of catalogues, with only the most energetic events being highlighted, or even combining several types of signals in the same label when the energies are not high enough and time pressures require quick responses. (e.g., labelling a superposition of signals as volcanic tremor).In some eruptive processes, such as the eruption of the island of El Hierro in 2011, the continuous occurrence of VT events led to a signal resembling volcanic tremor, but at a high frequency, due to~~ volcanic edifice, influence how seismic signals propagate and are recognized. This variability poses a challenge for automatic recognition systems, which are typically built by learning from large seismic catalogs, where each instance has a label indicating its source mechanism. The more diverse the data, the better the system's adaptability. However, as stated before, constructing complete catalogs is challenging because of the high cost of data labeling, which often leads to inaccuracies or mislabeling in seismic catalogs. Such inaccurate or mislabeled seismic catalogs could bias the effectiveness of the ~~overlapping of hundreds of VTs per hour (Ibáñez et al., 2012; Díaz-Moreno et al., 2015) . The use of denoising techniques can help reduce seismic noise levels in certain frequency~~

~~bands, but in general, they will also affect the frequencies of~~ systems, meaning that their performance may be influenced not only by changes in volcanic dynamics, but also by inadequate modeling of those dynamics.

In this work, we propose a comprehensive analysis of seismic catalog-induced bias when developing automatic recognition systems. We evaluated the ability of several monitoring systems trained using a master seismic catalog from Deceptio Island volcano (referred to as the ’Master database’) to adapt to new different volcanic environments from Popocatépetl (Mexico) and Tajogaite (Canary Island, Spain) volcanoes. We hypothesize that, often, automatic recognition systems are not capable of modeling the spatial-temporal evolution of seismic events. Instead, they learn to recognize the probabilistic pattern-matching observed in their training data. In other words, rather than simply learning to characterize volcanic dynamics by describing the latent physical model, catalog-induced learning biases the ~~signals. Therefore, in many cases, they may not be efficient for improving labelling task. Another potential solution is to manually relabel the databases using new expertise and state-of-the-art techniques, but this is a highly time-consuming process in practice. Nevertheless, this issue is prevalent in many seismic databases worldwide. Furthermore, having a database with potential biases will interfere with the implementation of automatic recognition methodologies based on ML. Biased labels will propagate in the generation of models and feature vectors used for comparison and the implementation of these technologies.~~ system’s performance as it learns the description of the data annotated in the catalog, potentially discarding useful data that describes volcanic dynamics. Therefore, we conclude that using systems trained with a master database (complete and large) as pseudo-labeler, could help create less biased catalogs from which the systems can be retrained and adapted to different volcanic environments.

~~In this work, we propose an alternative solution that avoids the need for manual relabelling and the use of denoising-based filtering techniques, while also addressing the correction of potential biases in databases and completeness in seismic catalogues.Based on our experience with transfer learning approaches (e.g. , Bueno et al. , (2021, 2021); López-Pérez et al. , 2020; Titos et al. , (2018, 2019, 2023)), we suggest the utilization of a high-quality, meticulously labelled volcano database (referred to as the ’Master database’)~~ To test our hypothesis, we conduct three independent experiments with three different automatic monitoring systems. In the first experiment, aimed at demonstrating that any state-of-the-art machine learning model can effectively learn the information contained in a seismic catalog, we will build monitoring systems within the Transfer Learning framework. In this approach, systems that have previously been trained on Deception Island volcano, will be re-trained using a seismic catalog from the Popocatépetl volcano. Once trained, the models will be evaluated in terms of performance and analyzed in detail. The outcomes reveal a key issue: when the catalog is not meticulously constructed, and events are not accurately annotated—where multiple events are combined as a single label—the systems fail to recognize each individual event, leading to the loss of valuable data that describes volcanic dynamics. In the second experiment, instead of re-training the pre-trained systems using a given catalog, we use the pre-trained systems as a foundational seed ~~for retraining another seismic catalogue. We suggest employing ML approaches to assess the completeness and accuracy of the tested database. In the literature, as mentioned earlier, there are several successful methodologies that could satisfactorily address this transfer learning approach. Building on the successful applications demonstrated using new-generation Artificial Neural Networks (ANN) in sequence modelling tasks, this work will compare the performance of three architectures in achieving this goal. We will utilize a Recurrent Neural Networks with Long Short-Term Memory cells (RNN-LSTM) and a Temporal Convolutional~~

~~Networks (TCN) previously developed by Titos et al. (2018, 2022), as well as an improved version of the RNN-LSTM based on multi-resolution dilated recurrent skip connections known as Dilated Recurrent Neural Networks (DRNN) (Chang et al. 2017; Titos et al. 2024). We aim to demonstrate that when properly fine-tuned, any of these neural networks will yield successful recognition results. However, the critical point lies in selecting the Master database for performing this tuning.~~(pseudo-labeler) for labeling the new database and construct a new catalogs. Using these new catalogs as training knowledge, we will re-train the systems. Afterwards, we will compare and analyze the results obtained from both approaches. The outcomes reveal that a significantly higher number of events, compared to those annotated in the original catalog, are recognized. This finding could offer a new potential perspective on the volcanic dynamics. Finally, to prove the robustness of our hypothesis, we will conduct a new experiment with data from the eruption of Tajogaite volcano in 2021, for which only an earthquake catalog is available, demonstrating that the application of automatic seismo-volcanic monitoring systems based on weakly supervised techniques can offer an effective alternative for both building and revising seismic catalogs.

~~Our Master catalogue was created using data from Deception Island volcano~~ The rest of this paper is organized as follows. Section II describes the seismic dataset and signals used in this study. . Section III provides the experimental framework, and describes how weakly supervised techniques can be used for developing automatic volano-seismic recognition systems. Section IV and V presents the results and discussions. Section VI concludes this paper.

2 Seismic data and catalogs

As previously stated, in this study, we will use three datasets from three volcanoes of different nature: Deception Island (Antarctica)~~and has already been successfully applied in different Deep Learning architectures such as BNNs, RNNs, DNNs, and TCNs, among others. The tested database belongs to~~ , Popocatépetl ~~volcano~~ (Mexico) and ~~was created during a seismic experiment conducted in 2002. This database was selected not because there were doubts about its quality; quite the contrary, it is considered a carefully prepared database due to the quality of the human team that created it and the importance of the volcano. Therefore, controlling the possibility of improving it represents a challenge to our suggestion of how to further enhance, if possible, existing databases. Furthermore, the eruptive dynamics of both volcanoes are entirely different, so we also seek evidence of transferability capability – that is, if a high-quality master database is available, it is possible to export its knowledge to other volcanic systems.In this way, this work will not only be useful for potentially improving existing databases, but also for creating seismo-volcanic catalogues using ML in volcanoes where no catalogue has ever been previously made.Therefore, ensuring appropriate seismic catalogues and support for developing monitoring tools should be a priority, as significant as applying new and more effective artificial intelligence (AI)techniques. It is important to note that the tested catalogue cannot be considered the official database used by the surveillance institution working on Popocatepetl volcano; therefore, we are not correcting any past or future work performed in this area. Our aim is to demonstrate how any seismic catalogue can be improved or even created with success and low time-consuming effort.~~

3 ~~Seismic data and catalogues~~

~~In this work~~Tajogaite (Canary Island, Spain). Due to the extensive expertise and in-depth knowledge that our research group has on Deception Island volcano, providing a comprehensive understanding of its structure and dynamics through numerous campaigns conducted since 1994 (Ibáñez et al., 2000; Martínez-Arévalo et al., 2003; Zandomeneghi et al., 2009; Carmona et al., 2012; Ibáñez et al., 2017), we will consider the dataset associated with this volcano as the reference or "master" dataset, thus granting it a high level of reliability and robustness. Therefore, to corroborate our hypotheses, we will use ~~a labelled database obtained from Deception Island volcano in Antarctica, which we consider to be reliable and homogeneous.~~the Popocatépetl and Tajogaite databases as benchmarks.

Deception Island (~~62º59‘S, 60º41’~~62°59’S, 60°41’W) is a horseshoe-shaped volcanic island that emerged during ~~Quaternary period and placed in~~ the Quaternary period. It is located within a marginal basin-spreading ~~centre~~ center of the Bransfield Strait, where the South Shetland Islands and the Antarctic Peninsula are ~~being separated~~ separating (Smellie, 1988; Martí et al., 2011; Carmona et al., 2012). ~~Deception Island volcano has been extensively studied by our research group from various perspectives, providing a comprehensive understanding of its structure and dynamics (Ibáñez et al., 2000 ; Martínez-Arévalo et al., 2003 ; Zandomeneghi et al. , 2009; Carmona et al. , 2012;~~ The Deception Island dataset (hereafter referred to as MASTER-DEC) was created using seismic data collected during the 1994-1995 campaign organized by the Andalusian Institute of Geophysics (IAG) with a short-period array of 8 channels. The array consisted of a three-component Mark L4C seismometer with a lower frequency band of 1 Hz and 5 Mark L25 sensors with a vertical component frequency of 4.5 Hz, electronically extended to 1 Hz. After analyzing the 8 channels, the one with the highest Signal-to-Noise Ratio (SNR) was selected (Ibáñez et al., ~~2017). Similarly, seismic data collected through numerous campaigns conducted on this volcanic island since 1994 have allowed us to create a highly reliable database. This database serves as the cornerstone for all our advancements in the field of recognizing and segmenting seismo-volcanic signals. It has served as the foundation for studies on the application of HMM models, various neural networks, parameter reduction algorithms, and many more. By combining~~ 2000). The data were sampled at a frequency of 100 Hz. Since this sampling frequency allows for the analysis of frequencies up to 50 Hz and our parameterization workflow primarily operates within the 1-20 Hz range, the data were filtered within this range. This filtering minimizes the influence of the sensorization used for signal recording and ensuring the comparability of the data recorded by different sensors over various time periods or at different volcanoes. By integrating our understanding of the structural, source, and dynamic models of Deception Island volcano with ~~our~~ advancements in signal processing and ~~ML~~Machine Learning (ML), MASTER-DEC has played a crucial role in the development of seismo-volcanic signal segmentation and classification. It has also served as the foundation for studies involving hidden Markov models, artificial neural networks, parameter reduction algorithms, and more (e.g., Bueno et al., 2021; López-Pérez et al., 2020; Titos et al., 2018, 2019, 2023; Cortés et al., 2021). Therefore, we can confidently assert that this database is both highly reliable and ~~ideal~~ ideally suited for our intended purpose~~of~~: serving as a reference seed ~~for building any other seismic catalogues. Therefore, we designate this database, named MASTER-DEC, as the 'Master database' to be utilized as a control and testing process for the Popocatepetl volcano database~~(pseudo-label) for constructing other seismic catalogs or improving existing ones, particularly those designed for early warning systems for volcanic eruptions.

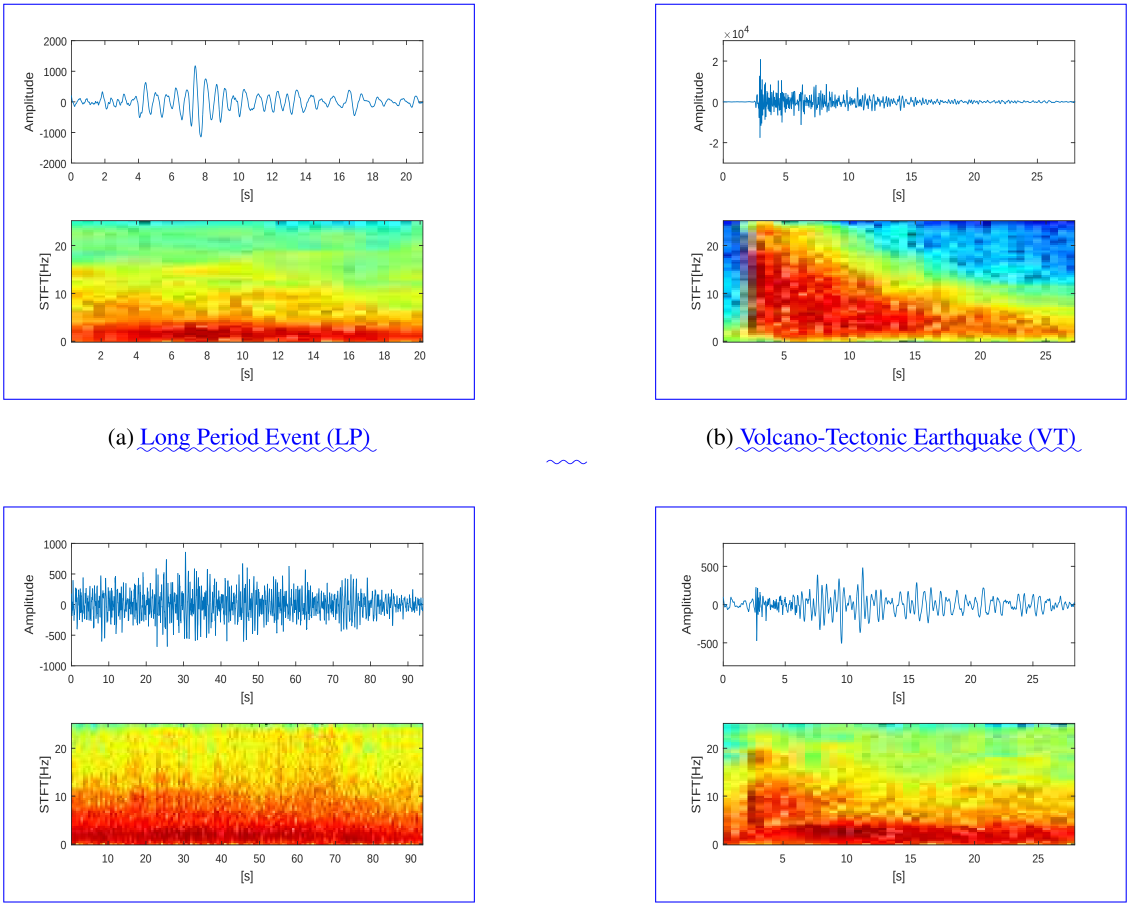
While it is true that not all types of signals are ~~present in this 'Master database'~~represented in MASTER-DEC—especially those associated with ongoing eruptive processes—its primary objective aligns with our ML application, which ~~aims to understand~~ focuses on understanding pre-eruptive processes.~~Hence, this database is ideal for preparing new ones or improving existing ones focused on serving as Early Warning tools for volcanic eruptions.~~

For the current study, we extracted a subset of ~~data considered the most reliable~~ reliable data, consisting of 2,193 seismic events. ~~The data is categorized into 5 classes(aligning~~ These data are categorized into five classes, which align with the volcano-seismic scientific labels and the accompanying source models ~~put forward by Ibañez~~ proposed by Ibáñez et al. (2000~~, see Table 1) , with the duration of the events varying: Background seismic noise (BGN)-1,222 events , Volcanic Tremor (TRE) -77 events, Long Period events (LPE)-765 elements, Volcano-Tectonic earthquakes (VTE)-75 elements, and Hybrid events (HYB)-54 events. It is important to note that although the number of selected events for some classes may seem limited for training ML procedures requiring a considerable number of elements, these events were chosen to be the most representative and of the highest quality from our Deception Island database. The selected database for testing proposes belongs to Popocatepetl volcano, Mexico~~) (Table 1 summarizes the source models and classifications). Table 2 presents a detailed summary of the seismic events and their distribution. Figure 1 depicts an example of each type of event corresponding to the prototypes in the database. Figure 2 illustrates the UMAP (Uniform Manifold Approximation and Projection) projection, showing the distribution of the five MASTER-DEC event types within the feature representation space. This visualization highlights how different seismic events occupy unique but sometimes overlapping regions, revealing potential challenges in distinguishing between event categories. The projection provides an intuitive view of the clustering tendencies and the proximity of events with shared characteristics, underscoring the inherent variability and possible misclassification risk in automatic seismic event recognition systems even in thoroughly analyzed and refined datasets.

Class	nEvents	min(sec)	mean(sec)	max(sec)	total(sec)	std(sec)
BGN	1222	0.3	15.4	128.2	18835.2	11.8
TRE	77	10.4	93.3	150.0	7184.2	43.63
HYB	54	7.8	29.4	136.8	1587.1	18.9
VTE	75	5.4	19.1	89.9	1434.5	12.88
LPE	765	2.4	9.8	30.7	7469.8	3.81

Table 2. MASTER-DEC summary. The table reflects statistics on the duration of the signals and the number of events for each class. Seismic categories: Background Seismic Noise (BGN), Volcanic Tremor (TRE), Long Period Events (LPE), Volcano-Tectonic Earthquakes (VTE), and Hybrid Events (HYB). Duration) is in seconds (sec).

Popocatépetl Volcano (19°1'N, 98°37'W) ~~. This volcano~~ is placed within a different geodynamic framework and exhibits a different eruptive style compared to Deception Island; a subduction region in confront to a rift area. Popocatepetl is a large dacitic–andesitic stratovolcano covering > 500 km2 of the eastern Trans-Mexican volcanic belt (Alaniz-Álvarez et al., 2007; Siebe et al., 2017). It is surrounded by a densely populated area with ~~a population of~~ around 25 million ~~people~~ inhabitants (Arango-Galván et al., 2020). The volcano is highly active, with the current active period beginning in December 1994 (Arango-



(a) Long Period Event (LP)

(b) Volcano-Tectonic Earthquake (VT)

(c) Tremor (TRE)

(d) Hybrid Event (HYB)

Figure 1. Amplitude and spectrograms of the main four prototypes of volcano-seismic events recorded at *Deception Island* volcano, during three seismic surveys: 1994-1995, 1995-1996, and 2001-2002.

Galván et al., 2020). The dataset used in this study (hereinafter called POPO2002) was collected during a seismic experiment conducted between November and December 2002, using short-period seismic stations. There is no detailed information regarding the type or specifications of the sensors used to record the seismic signals. Data labelling was manually performed by a group of geophysicists with extensive knowledge and experience of the volcano's dynamics. ~~The Popocatepetl 2002 catalogue (named Popo2002)~~ It consists of 4,883 events, divided into similar classes as the MASTER-DEC ~~catalogue~~ catalog (again aligning with the volcano-seismic scientific labels and accompanying source models proposed by Ibañez et al. 2000)~~, but further subdivided based on spectral content. This includes three classes for TRE (T1, T2, T3)-273 events, six classes for LPE (LPA1, LPA2, LPB1, LPB2, LPC1, LPC2)-1155 events, and VTE distinguished between local and regional earthquakes (VTE, REG)-371 events~~. Additionally, the ~~catalogue~~ catalog includes noisy events (labelled as GAR)-2739 events, and due to

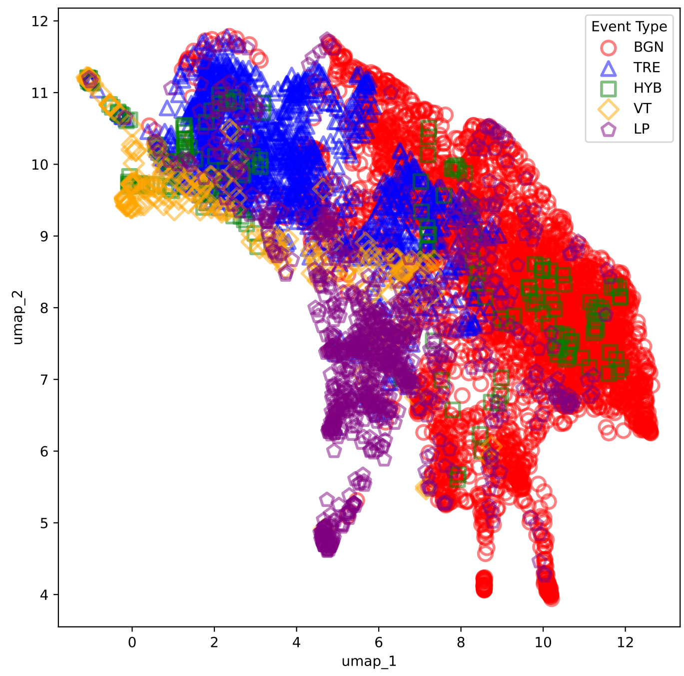


Figure 2. UMAP (Uniform Manifold Approximation and Projection) projection obtained for the input vector forming the original data of MASTER-DEC dataset. Different seismic categories may have elements located in overlapping areas of the representation space, where they share similar projected features.

Popocatepetl's activity, there is a ~~class~~ category for explosions (EXP)~~-4 events~~. Along with the event ~~catalogue~~catalog, we have continuous seismograms from this period that will be used for segmentation and identification processes. Table 3 summarizes the POPO2002 catalog. With the aim of minimizing the influence of the sensors used for signal recording and ensure data comparability, the signals were first filtered to match the frequency range of MASTER-DEC, followed by a subsampling process to adjust the sampling frequency accordingly.

3 ~~Methodology~~

~~The well-established procedure for creating a volcano-seismic database from scratch using supervised ML techniques involves selecting and segmenting a large, reliable set of well-labelled seismic events that cover the maximum possible range of events occurring in the studied volcanic area. As is widely known, this initial labelling process relies on expert/human decisions,~~

height~~Ibáñez, J. M et al.(2000)~~ **Class**	**nEvents**	**min(sec)**	**max(sec)**	
~~Volcano Tectonic Earthquakes Tectonic Short Period Earthq.~~BGN	~~High Frequency (HF)~~ 340	~~A-Type~~ 0.63	~~>5~~ 5048.09	~~She~~
~~Tornillo~~TRE	~~Low Frequency (LF)~~ 273	~~B-Type~~ 10.14	~~1-5~~ 357.17	
~~Medium Frequency~~HYB	~~Mixed Frequency (MX)~~ 1	~~-~~ 32.63	~~1-12~~ 32.63	
~~Volcanic Explosion~~VTE	~~Explosion Quake (EXP)~~ 371	~~Explosion Quake~~ 6.33	~~>10~~ 1202.7	
~~Harmonic Tremor~~LPE	~~Volcanic Tremor (TRE)~~ 1155	~~Volcanic Tremor~~ 8.95	~~1-12~~ 1227.99	
~~processes, and pyroclastic~~EXP	4	76.82	240.59	
~~flows~~GAR	2739	0.78	14228.95	

Table 3. ~~Representative volcano-seismic scientific labels and associated source models proposed by Ibáñez, J~~POPO2002 summary. ~~M~~The table reflects statistics on the duration of the signals and the number of events for each class. ~~et al~~The table reflects statistics on the duration of the signals and the number of events for each class. Seismic categories: Explosions (~~2000~~EXP), Garbaje (GAR), Hybrids (HYB), Long Periods (LP), Volcano-Tectonic Earthquakes (VT), Background Seismic Noise (BGN), Volcanic Tremor (TRE). ~~Other labels and associated source models proposed by different authors have been included for comparison~~Duration is in seconds (sec).

~~which can be extremely time-consuming. These events serve as the initial seed for the training procedure. Once our initial selection of high-quality, differentiated types of event classes is completed, encompassing as many events as possible, various techniques can be employed for the training process. Several data-driven approaches have been utilized to uncover descriptive patterns that characterize the diverse types of events appearing in seismic signals. Ultimately, these advancements have resulted in the development of automatic early warning systems (Malfante et al., 2018; Lara et al.~~ Tajogaite volcano (28º40’N, 17º52’E) is located on the island of La Palma in the Canary Islands, Spain. The eruptive activity started in September 19, 2021~~).In the literature, seismic event recognition (detection and classification) has been classically explored from two distinct perspectives: (i) the classification of isolated events, where the classifier receives individual events (previously isolated by expert geophysicists) and assigns them to one of the available seismic categories (Hibert et al., 2017; Titos et al., 2018; Bueno et al., 2021; Titos et al., 2019; Canario et al., 2020; Bicego et al., 2022), and (ii) sequence modelling, where the model processes continuous sequences of seismic signals to detect, and classify the events present (Alasonati et al., 2006; Benítez et al., 2006; Köhler et al., 2010; Bhatti et al., 2016; Titos et al., 2018; Bueno et al., 2021; Titos~~ , following a period of seismic activity, marked by several VT swarms and then carried by continuous volcanic tremor, becoming the first eruption on La Palma since 1971. The eruption started with the opening of a fracture in the southwest part of the island, and the emission of material persisted for nearly three months, generated extensive lava flows and pyroclastic deposits (D’Auria et al. ~~,~~ (2022). ~~Considering real-time monitoring of seismo-volcanic signals as a sequence modelling problem, ANNs, including both recurrent and temporal convolutional approaches, have played an important role in recent years due to their temporal modelling capabilities. These approaches have yielded successful results in improving seismic catalogues and advancing in both the procedure of real time seismic monitoring and early warning protocols. However, as mentioned above, several databases can be biased by factors such as seismic attenuation, site effects, radiation patterns, and human decisions and experience. Furthermore, starting this procedure from scratch, regardless of the time it consumes, implies that creating a~~

~~database in new scenarios could be discouraging. Moreover, in the case of new volcanic processes, old databases may not be suitable or adequate for seismic monitoring. We will assess the efficacy of three ANNs to demonstrate the transferability of knowledge from one volcanic scenario to another, aiming to : a) verify the reliability of the existing database and ; b) create a new databasewithout the initial time-consuming human supervision. Specifically, we will employ a RNN-LSTM, a Dilated-RNN, and a TCN.RNN–LSTM is a RNN with Long Short-Term Memory (LSTM) cells (Hochreiter, S., Schmidhuber, J., 1997). LSTM cells are a special kind of RNN units that are capable of learning long-term dependencies and mitigating the vanishing/exploding gradient problems, which are common in traditional RNNs (Schmidhuber, J., 2015). LSTM cells improve their performance through a complex architecture that includes a gates mechanism (input, forget, and output gates) which regulate the flow of information, allowing the network to maintain and update a cell state over extended periods of time (Schmidhuber, J., 2015~~). ~~This makes LSTMs particularly effective for tasks that involve long-term sequential data, such as speech recognition, language translation,~~ This event significantly affected the surrounding environment, infrastructure, and ~~time series prediction within emerging research geosciences fields such as climate change or remote sensing (Racic et al., 2020; Yan et al., 2020). This framework is especially useful for monitoring temporal evolution of volcanic systems in quasi real time.Dilated Recurrent Neural Network (Dilated-RNN) is a type of RNN designed to handle long-term dependencies more efficiently by using "dilations" in the sequence processing steps (Chang et al. 2017).Analogous to the dilated CNN proposed in (Aa¨ron van den Oord et al. 2016, Fisher et al., 2017), this architecture modifies the standard RNN structure by skipping certain time steps allowing the network to introduce multi-resolution dilated recurrent skip connections between layers, therefore enhancing efficiency and reduced training parameters (Chang et al. 2017). The dilation technique allows for the selection and analysis of non-correlative temporal segments in each network layer during the learning process. This enables an advanced trainable selection of individual features based on their relevance to the recognition task. The flexibility to select non-correlative temporal segments offers the advantage of focusing on the search and detection of seismic events with a broader perspective of the previous temporal steps. Additionally, it demonstrates comparable state-of-the-art performance in sequence modelling tasks involving very long-term dependencies, while helping to mitigate issues of vanishing and exploding gradients. Regarding TCNs (Lea et al., 2017), unlike traditional RNNs which process sequences sequentially, they leverage dilated 1D convolutional layers and residual connections to process entire temporal sequences simultaneously. The dilated convolutions (similar to the multi-resolution dilated, recurrent skip connections between layers in the Dilated-RNN) introduce gaps between the input elements of the convolutional layers, enabling the network to cover a larger temporal context without a significant increase in computational complexity. Residual connections (Fisher et al., 2027) help mitigate the vanishing gradient problem and facilitate the training of deeper networks by allowing gradients to flow smoothly. This approach efficiently captures long-range dependencies and temporal patterns, offering several advantages over traditional RNNs, such as better handling of the exploding/vanishing gradient problem, improved memory retention, and enhanced parallelization during training (Lea et al.~~ regional air traffic. The volcanic process yielded comprehensive seismic and geochemical data, providing valuable insights into volcanic behavior in the Canary Islands and serving as a key reference for improvements in volcanic monitoring and hazard assessment. The seismic catalog for this volcano (from this point forward referred to as LAPALMA2021) differs from previous seismic catalogs since it only includes annotations of the occurrence of VT-type events. That is, the catalog consists solely of a series of entries describing the date of the event's occurrence, along with its magnitude and depth. There is no detailed information regarding the type or specifications of the sensors used to record the seismic signals. Given the nature of this catalog and database, ~~2017)~~we believe that the inclusion of this use case could be of interest for evaluating the capability of the proposed approach to improve a catalog from scratch. Once again, to further minimize the impact of sensor differences and ensure data comparability, the signals were first filtered to match MASTER-DEC's frequency range, then adjusted to the same sampling frequency.

3 Methodology and experimental framework

This section details the methodolgy and experiments conducted to test our hypothesis that, beyond the changing dynamics of volcanoes between eruptive periods, and intrinsic factors like attenuation effects and source characteristics that alter the shape and spectrum of seismic signals, the effectiveness of automatic seismic monitoring systems is further compromised by the *incompleteness* of the seismic catalogs on which they rely. To accomplish this task, the proposed algorithm will first be described, and then, once its functioning is understood, the three experiments conducted will be detailed. The results of each of these experiments will be detailed in the results section.

~~Building on these~~

3.1 Methodology

Building on the architectural strengths and ~~by integrating the ML's advanced temporal modelling capabilities~~ integrating the advanced temporal modeling capabilities of machine learning techniques, this work proposes using a weakly supervised transfer learning ~~(TL) algorithm to improve the reliability of existing databases and create new databases~~ algorithm to create new seismic catalogs from which the systems can be retrained with minimal initial human supervision.

3.2 ~~Weakly supervised learning as a pseudo-labelling approach~~

In other words, instead of retraining the pre-trained systems with a given catalog, this approach proposes using the pre-trained systems as a foundational seed (pseudo-labeler) to weakly label the new database and construct new catalogs. These new catalogs will then serve as the training knowledge for retraining the systems to the new volcanic environment.

Weakly supervised learning is a branch of ~~ML~~ machine learning covering the construction of predictive models ~~by learning with weak supervision (Zhou , 2018)~~with minimal or indirect supervision (Zhou, 2018). Such techniques focus on learning with incomplete, inexact, and/or inaccurate information derived from noisy, limited, or imprecise supervision processes. The objective is to automatically provide supervision for ~~labelling~~ labeling large amounts of data using ~~labelling~~ labeling functions derived from domain knowledge. This approach replaces the costly and impractical ~~hand-labelled~~ hand-labeled process with inexpensive weak labels, understanding that although imperfect, they can be used to create a strong predictive model. ~~In this study~~In this framework, the source domain (denoted as D_s) is the MASTER-DEC dataset (based on refined physical models and a strong revision process). The target domain (denoted as D_t) is ~~the Popo2002~~ a new given dataset (whose available seismic ~~catalogue~~ catalog will not be considered). The goal ~~of this study~~ is to address a domain adaptation task ~~(Kouw , 2019; Farahani et al. , 2021)~~ (Kouw and Loog, 2019; Farahani et al., 2021) to reduce the cost of developing a reliable seismic ~~catalogue and database~~ catalog and database for a new given dataset with minimal initial human supervision. That is, automatically provide supervision for labelling large amounts of data from D_t using labelling functions derived from domain knowledge D_s.

In a domain adaptation framework, typically D_s and D_t have the same feature space but different distributions. However, in this study, for the ~~pseudo-labelling~~ pseudo-labeling task we assumed that:

- The marginal distributions of D_s and D_t are the same: $P_s(X_s) = P_t(X_t)$, where X_s and X_t are the input feature vectors associated with different seismic windows or frames in both domains. As such, the ~~pseudo-labelled~~ pseudo-labeled samples do not need to contain any domain information, and the occurrence of different seismic events is equally likely in both domains.

- The conditional distributions of D_s and D_t are the same: $Q_s(Y_s|X_s) = Q_t(Y_t|X_t)$. As such, the ~~pseudo-labelled~~ pseudo-labeled samples are valid in both domains.

Such assumptions have important implications since in the target domain, while the marginal distributions of D_s and D_t are the same $[P_s(X_s) = P_t(X_t)]$, the conditional distributions could be different $[Q_s(Y_s|X_s) \neq Q_t(Y_t|X_t)]$. This shows how similar feature vectors taken as the input could output different probabilistic event detection matrices. That is, the description or ~~characterisation~~ characterization of seismic categories could change between domains, or D_t could contain seismic categories unforeseen in D_s.

Therefore, ~~by~~ leveraging the probabilistic detection matrices output by the system trained in D_s, we can apply a weakly supervised learning technique as a pseudo-labeller in D_t to construct a new ~~dataset~~ catalog from which to train a new system in a supervised way. Those ~~parts~~ subset of the unlabelled dataset with high per-class probability, and then high confidence, are added to the new ~~training set~~catalog. Although imperfect, this method guarantees that~~at least~~ , at least, events showing characteristics similar to those annotated in the master ~~catalogue~~ catalog will be included in the new training dataset. As a result, after the re-training phase, the target ~~catalogue is~~ catalog could be enlarged and updated. ~~Our proposed weakly supervised transfer learning (TL) algorithm is outlined as follows and depicted in Fig. 1.~~ It is important to ~~highlight that our methodology is just one possible approach; many other weakly supervised methods could also be applied to achieve similar results.~~note that this experiment does not aim to correct the catalog created by our colleagues with utmost dedication and effort; it simply seeks to highlight that a pseudo-labeler can be a valuable tool in constructing and reviewing it with success and low time-consuming effort.

Taking these factors into account, our proposed approach is outlined as follows and depicted in Figure 3:

1. **~~Subset Analysis~~Recognition:** ~~A subset of the Popo2002 dataset (20%-40% of the total) is analyzed using an ML~~ According to Figure 3 a, the recognition block analyzes a subset of data from the new dataset using a pre-trained system

(RNN-LSTM, Dilated-RNN, TCN) ~~trained on~~ and gets a probabilistic event detection matrix with per-class membership outputs. The data stream illustrates continuous or streaming analysis (allowing near real-time processing). To carry out the recognition step using the network seed (trained with the MASTER-DEC dataset), streaming or continuous signals are filtered between 1 and 20 Hz and split into frames or windows; the same algorithm of feature extraction used the MASTER-DEC is applied. For each window, a feature engineering pipeline based on a logarithmic scale filter bank is applied. This pipeline reduces the dimensionality of the input vector associated with each analysis window (compared to raw signals), which facilitates the training and convergence of the systems, as it increases the separability of the data based on well-studied features in the literature (review Titos et al. 2024 for a detailed understanding of the parameterization pipeline).

2. **Event Detection and Confidence Analysis (Concept drift detection):** Ignoring the information contained within the available ~~Popo2002 seismic catalogue, we analyze~~ seismic catalog, the concept drift detection block analyzes the confidence of each detected event using ~~a~~ the previously obtained probabilistic event detection matrix with per-class membership output~~from the softmax layer~~. This step allows us to quantify the severity of drift between datasets (usually knows as "concept drift~~detection~~") ~~(Lu et al. , 2018)~~(Lu et al. 2018). High or extremely high per-class recognition probabilities for each event type indicate that the systems are well-fitted to the master database. Low per-class probabilities indicate a change in the description of the analyzed information. Accurate and robust dissimilarity measurement and statistical hypothesis evaluation are not strictly necessary ~~,~~ given the well-known dissimilarity between ~~the~~ volcanic environments.

3. **Concept Drift Adaptation~~Mechanism~~:** An adaptive threshold mechanism ~~is adopted. Detected events~~ where a probability threshold is defined to select the events that will be included in the new database is employed. Events with an average per-class probability ~~above a certain~~ exceeding this threshold are selected and ~~included~~ incorporated as training instances in the training set.

4. **Re-training process:** Finally, the ~~pre-trained ML systems~~ ML systems trained with the MASTER-DEC used in step 1 are re-trained using the selected instances and labels obtained in step 3.

5. **Iterative Refinement:** Repeat steps 2 to 4 iteratively until the desired result is achieved.

4 ~~Result~~

~~This section presents the results supporting the objectives outlined in the previous section: a)verifying the reliability of the existing database and~~

3.1 Experimental framework

While the literature offers a variety of accurate machine learning architectures used to uncover descriptive patterns in seismic signals (Malfante et al., 2018; Lara et al., 2021; Hibert et al., 2017; Titos et al., 2018; Bueno et al., 2021; Titos et al., 2019;

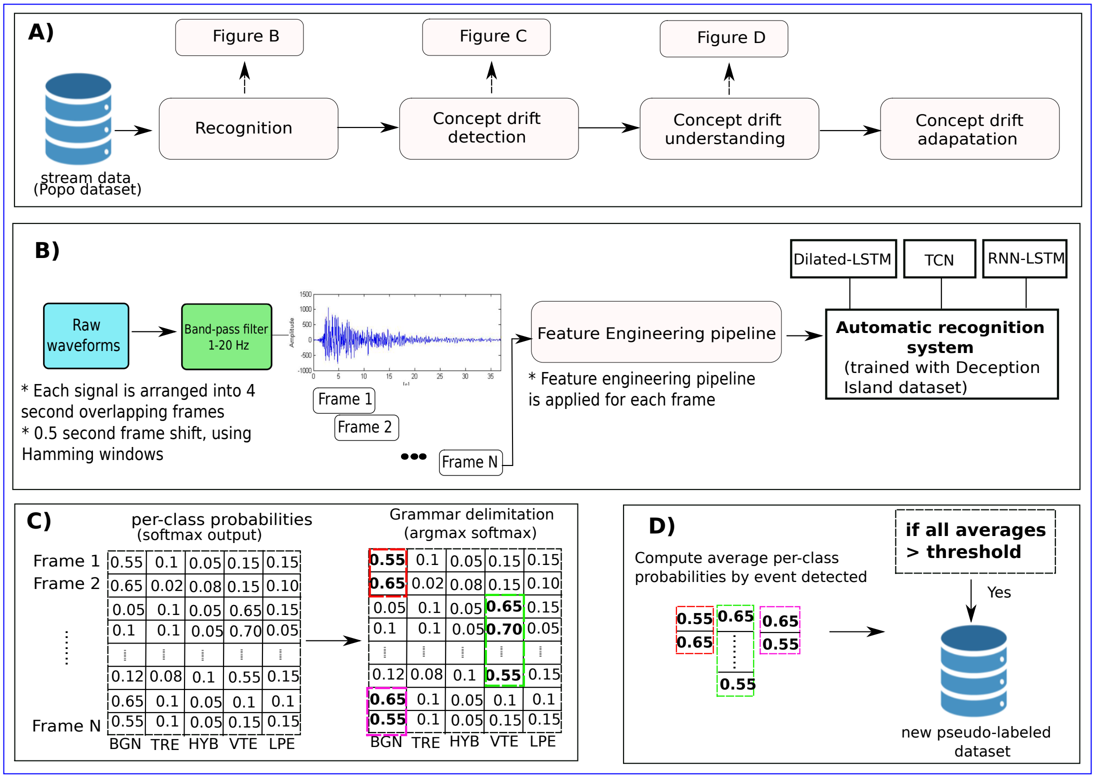


Figure 3. a) ~~Weakly~~ Overview of the weakly supervised event selection algorithm ~~applied to Popo2002 dataset~~developed. A subset of the dataset (in our case 40% of the total) is used as a training set by the ~~previously trained ML models~~reference pre-trained systems. The rest of the data is used as a test set. Only high per-class probability recognized events are selected as new training instances. b) Workflow structure and the specific preprocessing steps employed~~in this article. The specific preprocessing elucidates the feature engineering process~~, which relies on frequency analysis within the logarithmic filter bank domain (~~ref~~Titos et al. 2022). This processed information serves as the input for the different volcano-seismic recognition systems. c) For each detected event, the confidence of the detection is analysed using a probabilistic event detection matrix with per-class probabilities output by the ~~softmax layer~~systems. d) Drift adaptation mechanism based on an adaptive threshold ~~was~~ is then adopted. Those events whose average number of per-class probability ~~was~~ is greater than a given threshold ~~were~~ are selected and included as new training instances.

Canario et al., 2020; Bicego et al., 2022; Alasonati et al., 2006; Benítez et al., 2006; Köhler et al., 2010; Bhatti et al., 2016; Titos et al., 2018; Bueno et al., 2021; Titos et al., 2022), some of these methods may not be as effective for the specific challenges posed by continuous or streaming data (such as POPO2002 and LAPALMA2021). Given the inherent variability and complexity of these data (consisting of seismic signal sequences containing multiple events, where the goal is to detect and classify each individual event), specialized approaches capable of adapting to these conditions are required. More specifically, we will base our experimental framework on the pre-trained systems previously published in Titos et al. (2018, ~~b)creating a new database without the initial time-consuming human supervision.The reliability of the dataset will be verified through the~~

~~self-consistency results of the Popo2002 database. Although self-consistency could be achieved through two~~ 2022 and 2024). These systems correspond to the Recurrent and Dilated Recurrent Neural Networks (Hochreiter, S., Schmidhuber, J., 1997; (Schmidhuber, J., 2015); (Chang et al. 2017)), both with LSTM cells, along with Temporal Convolutional Networks (Lea et al., 2017) (henceforth referred to as RNN-LSTM, Dilated-LSTM, and TCN, respectively).

3.1.1 Developing automatic recognition systems from available catalogs

The standard procedure for developing an automatic volcano-seismic monitoring system from scratch using supervised machine learning techniques involves having a sufficiently representative seismic catalog (selecting and segmenting a large, reliable set of well-labeled seismic events that cover the maximum possible range of events occurring in the studied volcanic area). These events serve as the initial seed for the training procedure. This training can be carried out using different approaches, ranging from training the system from scratch ~~or applying transfer learning, in this work we focus on transfer learning approaches since they offer significant advantages(they reduce training time and resource usage by leveraging~~ to using transfer learning techniques (Weiss et al. (2016)). Transfer Learning offers significant advantages, especially when the available data for a particular volcano is limited. This technique allows for reusing a model pre-trained ~~models and they enhance performance with limited data by utilizing knowledge gained from diverse datasets~~on data from another volcanic region (for example, a previously studied volcano) and adapting it to new data with considerably less computational and labeling effort. By transferring knowledge acquired from one volcano to another, the system's ability to recognize seismic patterns and adapt to different volcanic characteristics could be enhanced, leading to ~~better generalization and mitigating overfitting).Therefore, the self-consistency results correspond to the automatic recognition performance obtained by conducting a classical supervised Transfer Learning where a model previously trained in one volcanic environment (MASTER-DEC)is used as a starting point to build a new system in a different volcanic environment (Popo2002)using as training the available original catalogue. Finally, to verify the feasibility of creating a new database without human supervision, we present the recognition resultsobtained using our weakly supervised learning proposal. Checking the self-consistency of Popo2002 dataset: Traditional transfer of knowledge from~~ improved accuracy and generalization.

Thus, in the first experiment, to demonstrate that state-of-the-art machine learning model can effectively learn the information contained in a seismic catalog (assuming catalog-induced learning biases), a recognition system based on transfer learning approaches will be developed from scratch utilizing three different architectures. To achieve this, three systems pre-trained on MASTER-DEC ~~to Popo2002The most efficient procedure to check the self-consistency of a given database is to tune an automatic system encompassing all types of available events. As mentioned before; to accelerate development and achieve high-quality resultscompared to training from scratch, we will employ a classical Transfer Learning approach. If the events classified as elements of the same class are not homogeneous, the model's performance will be low (usually below 70The self-consistency of~~ will serve as the foundation for adapting recognition systems to the ~~Popo2002 database is checked using the three proposed ML architectures (~~Popocatépetl volcano. Specifically, these systems will be re-trained with the available data and catalog from POPO2002 dataset. Given that the POPO2002 catalog contains 7 seismic categories, a recognition system based on transfer learning can be constructed in different ways. One approach is to consider only the categories present in MASTER-DEC, while another includes all the categories (i.e., incorporating Explosions and Garbage events in the training set, thus expanding the number of seismic categories by two). From a machine learning perspective, these two approaches have no major implications. In the first case, where only 5 seismic categories are used, the systems would undergo retraining with the new catalog. In the second case, when using 7 categories, systems are adjusted to accommodate 2 additional categories, leveraging the pre-existing parameters while updating only the output layer. After that, the models are trained as usual.

3.1.2 Developing automatic recognition systems with weakly supervised pseudo-labeling

To test our initial hypothesis and following the workflow outlined in Figure 3, this second experiment highlights the use of weakly supervised approaches to enhance seismic-volcanic catalogs. By leveraging an existing unbiased master catalog, we can incorporate prior knowledge into the new dataset under review. This process involves using each of the three reference systems (RNN-LSTM, ~~Dilated-RNN and TCN)~~ Dilated-LSTM, TCN), considered well-trained, to reassess and label the seismic categories in the new dataset, then retraining themselves based on these pseudo-labels. Therefore, Each system will analyze a subset of the total POPO2002 database to create a training set for the retraining process. Once retrained, the systems will generate a new seismic catalog, which will then be compared and analyzed against the original POPO2002 catalog to assess the results.~~We apply a Leave One Out~~

Since MASTER-DEC is composed of five seismic categories, and the weakly supervised approach relies on pre-trained models, the experiments presented here are based solely on these five categories. This limitation is a consequence of the methodology and must be properly understood in order to ensure a correct interpretation and discussion of the results, as it directly influences the way the data is analyzed and compared with the original catalog. An important consideration in this experiment is that the recognition percentage obtained by the systems before and after retraining, using the original catalog annotations as a reference, can provide valuable insights into the algorithm's behavior. Therefore, both results will be taken into account in this experiment, with the aim of analyzing in detail how the retraining process with the new pseudo-catalog affects the system's performance.

3.2 Building a new catalog during an eruptive crisis: The Tajogaite volcano use case, 2021

The third and final experiment aims to analyze the robustness of the proposed methodology by building a seismic catalog from scratch in a highly demanding use case, such as during an eruptive crisis. Since we have not had the opportunity to test it in an actual eruptive scenario, we propose using data from the Tajogaite volcano during the 2021 eruptive episode. We also suggest abstracting this offline test to simulate a real-time episode, as if data were being analyzed in real-time, since the functionality would be exactly the same. As previously mentioned, the selected pre-trained systems are capable of operate in near real-time, with particularly short latency times, analyzing (not re-training) 24 hours of data in few seconds.

Therefore, for this experiment, a pair of 24-hour seismic records from the PPMA and PLPI seismic stations, corresponding to September 12, 2021, just a few days before the eruption began when an increase in activity was detected, have been used. To conduct an analysis and comparison of the results, we have a seismic catalog created by geophysical experts from that volcano during the eruption crisis itself. Given the large number of recorded events, the significance, and the urgency of the moment, we believe that this catalog meets the human requirements of the time. Again, just as we argued in the case of the POPO2002 catalog, this experiment does not aim to correct the catalog created by our colleagues with utmost dedication and effort; it simply seeks to highlight that a pseudo-labeler can be a valuable tool in constructing and reviewing it.

4 Results

This section presents the results supporting the experiments outlined in the previous section. For each experiment, tables describing the system performances in terms of accuracy, along with detailed confusion matrices are presented. For Experiments 1 and 2, accuracy (%) metric evaluates the capability of the developed systems to accurately recognize (detect and classify) the events annotated in the POPO2002 seismic catalog. The normalised confusion matrices provide a breakdown of true positives, false positives, false negatives, and true negatives, allowing a thorough analysis of each system's robustness in recognizing each event type. All results were obtained using a Leave-One-Out cross-validation ~~method~~ process with 4 random partitions. Each time, we select T% of the entire database as training set, and the remaining (100-T) % as test set to check the performance of the systems. ~~Given that using a~~ This analysis helps to identify specific areas where the model may struggle, such as mis-classification between event types with similar features. Finally, in experiment 3, where only partial knowledge of the earthquakes recorded during the crisis is available, results evaluate the model's ability to generate a more comprehensive and reliable catalog. This catalog will serve as a basis for inferring potential volcanic dynamics, with confusion matrices helping to assess how well the model distinguishes between known and newly identified event patterns, which is critical in real-world eruptive crisis scenarios.

The optimal RNN-LSTM configuration consists of a single hidden layer with 210 units and no dilations. For the Dilated-LSTM model, the configuration that yielded the best performance included three hidden layers, each with 50 units and 2–4 dilated recurrent skip connections per layer. The TCN model, achieved optimal performance with 50 filters, a kernel size of 2, and dilation values of 8, 16, and 32. Only one residual block was used, as additional blocks are more suitable for longer sequences, such as waveforms with extensive time samples. Data normalization was performed using standard deviation normalization, where each feature was normalized by subtracting its mean and dividing by its standard deviation, calculated from the training set. The model was optimized using Stochastic Gradient Descent (SGD) with a fixed learning rate, ranging from 0.004 to 0.01, with the optimal learning rate found to be 0.001. To prevent overfitting, early stopping and L2 regularization were applied during training.

4.1 Developing automatic recognition systems from available catalogs

Table 4 presents the recognition results obtained by the pre-trained ~~model~~ systems after being trained on POPO2002 catalog. Since using a transfer learning approach allows for more efficient ~~utilization~~ use of computational resources~~and that~~ , and the fine-tuning phase ~~generally~~ typically requires fewer resources than training a ~~model~~ system from scratch, ~~we conducted two experiments with~~ two experiments were conducted. These experiments considered 5 and 7 seismic categories, each using 20% and 40% of the total data for the training set (T = 20 and T = ~~40.~~ 40). This means that the results were obtained using 80~~Table 2 presents the results of the preliminary self-consistency test for the three proposed architectures. As indicated, the confidence matrix indices are high, suggesting good self-consistency. Table 3 shows~~ % and 60% of the data in the test partition, respectively. Table 5 summarizes the averaged normalised confusion matrices ~~of the Leave One Out cross validation process for the Popo2002 dataset. The results of these tests indicate that inside of each seismic class there is a coherence of the elements selected to belong to them. However, as noted by Titos et al. (2018), propagation and source effects can influence the characterization of seismic events. For example, according to the Table 3 and the original Popo2002 catalogue, VTE events are not properly identified, as the confusion level is over 60% in some cases, meaning only~~ belonging to the test using 5 seismic categories and 40% of the ~~VT events are correctly classified. In all cases, the highest levels of confusion are observed with the LPE class. It could be explained because LPE events could share similar characteristics than high attenuated VTs, producing potential biases in the allocation of events in the right category. Consequently, different seismic categories may have elements located in overlapping areas of the representation space, where they share similar projected parameters. These events, despite being assigned to a specific cluster, could easily transition to another. Therefore, a high performance on the self-similarity test does not guarantee an unbiased classification. While this self-consistency suggest uniformity in the training procedure, it's important to acknowledge that potential biases in the labelling process may not be evident in this measure of goodness.~~total data for the training set.

	5 seismic categories		7 seismic categories	
	Training percentage		Training percentage	
	20%	40%	20%	40%
RNN-LSTM	77.38	**88.99**	84.01	**84.39**
Dilated-LSTM	82.88	**84.70**	84.05	**85.21**
TCN	82.46	**88.30**	**85.77**	83.27

Table 4. Self-consistency results using 5 and 7 seismic categories, with 20% and 40% of the data for training and 80% and 60% for testing, respectively. The results correspond to the average accuracy over the four partitions.

4.2 ~~Weakly~~ Developing automatic recognition systems with weakly supervised ~~Transfer Learning~~pseudo-labeling

~~The direct application of classic TL could be considered a kind of blind test because, the automatic system acquires the pre-existing knowledge in the catalogue, since they used the original labels during training, thereby propagating the existing bias. According to section 3.1, we propose a solution to improve the TL approach through a weakly supervised method. By leveraging an existing master database, we can impose prior knowledge on the database being tested. The seismic categories of the testing database will be re-reviewed under the supervision of the system trained with the 'unbiased' dataset. In this~~

	RNN-LSTM					Dilated-LSTM		TCN						
	BGN	TRE	HYB	VTE	LPE	BGN	TRE	HYB	VTE	LPE	BGN	TRE	HYB	VTE	LPE
BGN	**0.97**	0.02	0	0	0.01	**0.96**	0.02	0	0	0.02	**0.98**	0.01	0	0	0.01
TRE	0.06	**0.78**	0	0.05	0.11	0.13	**0.69**	0	0	0.18	0.11	**0.68**	0	0.09	0.12
VTE	0.08	0.13	0	**0.51**	0.28	0.12	0.17	0	**0.31**	0.4	0.14	0.09	0	**0.59**	0.18
LPE	0.05	0.07	0	0.03	**0.85**	0.04	0.18	0	0	**0.78**	0.05	0.05	0	0.04	**0.86**

Table 5. Averaged normalized confusion matrices associated with the Leave One Out cross validation process for the Popo2002 dataset. These results belong to the test using 5 seismic categories.

~~way, the potential bias of the tested database will not be propagated to either domain, as the Master catalogue imposes its knowledge.Table 4 displays the self-consistence of the Popo2002 dataset when the~~ Table 6 presents the recognition accuracy achieved by the pre-trained systems, which were retrained using the proposed weakly supervised approach with the training partition set to 40% of the total POPO2002 dataset. As previously stated, since MASTER-DEC ~~knowledge is imposed. We observe a noticeable decrease in accuracy, attributable to the aforementioned bias in event classification~~consists of five seismic categories and the weakly supervised approach builds on pre-trained models, the results presented here include only these 5 seismic categories. The first column of ~~Table 4 shows the recognition results obtained when directly applying~~ the ~~models trained on the MASTER-DEC dataset to recognize events in Popo2002 compared to the original catalogue~~table represents the results obtained by directly applying recognition with the pre-trained models. This column shows the degree of similarity between the original POPO2002 catalog and the pseudo-catalog constructed using the pre-trained systems as pseudo-labelers. The second column ~~presents the results obtained after applying our weakly supervised TL approximation, comparing the new Popo2002 catalogue with the original one. These results highlight how the new POPO data base has been modified in comparison with the original.~~reflects recognition results compared to the original POPO2002 catalog after the systems have been retrained using the previously constructed pseudo-catalog. Table 7 summarizes the averaged normalized confusion matrices of the new systems based on the weakly supervised approach, with the POPO2002 catalog as the reference. The results are over the whole test set using 40% of the whole set for training and five seismic categories. The y-axis corresponds to the real label or ground-truth and the x-axis corresponds to predicted labels. Finally, Table 8 summarizes the comparison between the events initially annotated in the POPO2002 catalog and the events detected by the new automatic systems following the weakly supervised approach.

~~This procedure could be applied iteratively until a reliable catalogue is achieved. In our case, after three iterations, the improvements only reached 2% compared to the first iteration. However, it could change when using a different test dataset. In the next section, we will demonstrate how this new catalogue dramatically improves the original one, confirming that the new Popo database could be better adapted.~~

5 ~~Discussion~~

	Five seismic categories blind test	'Weakly supervised TL' using five seismic categories TL
RNN-LSTM	55.95	64.89
Dilated-RNN	50.13	55.72
TCN	58.27	66.16

Table 6. Classification accuracy (acc. %) on the test set ~~obtained when directly applying the models trained on~~ achieved by the ~~MASTER-DEC dataset to recognize events in Popo2002 compared to~~ pre-trained systems, which were retrained using the ~~original catalogue and after applying our~~ proposed weakly supervised ~~TL approximation keeping only five seismic categories and using~~ approach with the training partition set to 40% of the total POPO2002 dataset ~~as the training set~~and only 5 seismic categories.

	RNN-LSTM					Dilated-LSTM		TCN							
	BGN	TRE	HYB	VTE	LPE	BGN	TRE	HYB	VTE	LPE	BGN	TRE	HYB	VTE	LPE
BGN	**0.88**	0.09	0	0	0.03	**0.67**	0.32	0	0	0.01	**0.8**	0.19	0	0	0.01
TRE	0.29	**0.36**	0.03	0.02	0.03	0.29	**0.5**	0	0	0.21	0.19	**0.7**	0	0	0.11
VTE	0.27	0.41	0.08	**0.03**	0.21	0.46	0.28	0	**0.03**	0.23	0.36	0.46	0.03	**0.06**	0.09
LPE	0.36	0.19	0.06	0.06	**0.33**	0.47	0.18	0	0.01	**0.34**	0.41	0.33	0.01	0.01	**0.24**

Table 7. ~~Normalised~~ Normalized confusion matrices ~~related to~~ for the new retrained system using a weakly supervised approach~~implemented for the three architectures~~, ~~using~~ with the ~~Popo2002 catalogue~~ POPO2002 catalog as reference. The results are over the whole test set using 40% of the whole set for training and five seismic categories. The y-axis corresponds to the real label or ground-truth and the x-axis corresponds to predicted labels.

~~This work emphasizes in the use of Weakly supervised approaches to improve seismic-volcanic catalogues. We have verified that~~

	Popo2002 catalog	
BGN	340	>20,~~when effectively use, these procedures could significantly enhance the detection and identification capabilities of va~~
TRE	273	
VTE	371	
LPE	1,155	

Table 8. Comparison between the events initially annotated in the catalog and the events detected by the new automatic systems following the implementation of a weakly supervised approach.

4.1 Building a new catalog during an eruptive crisis: The Tajogaite volcano use case, 2021

Table 9 shows the recognition results obtained in this experiment using 24-hour seismic traces from the PLPI and PPMA stations on 9/12/2021 at Tajogaite volcano. The number of events manually annotated by experts during the volcanic crisis for the analyzed day, serving as a guide for the subsequent analysis is 247 earthquakes, both tectonic and volcanic in origin. As mentioned earlier, it is important to highlight that these results correspond to an experiment where only a tentative earthquake catalog (constructed during the eruptive crisis under the urgency and challenges that such situations entail) is available. For this reason, to conduct a rigorous comparative analysis, we have included the recognition results from a widely-used tool like PhaseNet (Zhu and Beroza (2019)). PhaseNet is a neural network-based system designed for automatic phase picking of seismic events. It detects and labels seismic phases and estimates the probability of each phase type (P and S) across the trace. After analyzing the two seismic stations, PLPI and PPMA, for September 12, 2021, 1173 P-phases and 1518 S-phases were obtained for PLPI, and 390 P-phases and 522 S-phases for PPMA.

	RNN-LSTM		Dilated-LSTM		TCN	
	PLPI	PPMA	PLPI	PPMA	PLPI	PPMA
BGN	4344	4641	1800	3005	6409	8642
TRE	109	64	229	241	152	139
HYB	12	14	5	8	-	-
VTE	187	131	194	161	333	403
LPE	1008	1032	564	711	516	761

Table 9. Recognition results obtained by the pre-trained reference models on the seismic traces recorded on 12/9/2021 at the PLPI and PPMA stations. Results are without considering re-training process.

5 Discussion

5.1 Developing automatic recognition systems from available catalogs

The classical way to assess the robustness of an automatic recognition system is by evaluating its recognition accuracy across all events included in the catalog. Typically, a system with an average performance below 75% is considered unreliable. However, this lack of reliability is often not due to the system's ability to learn to distinguish between different events, but rather results from how the catalog is constructed. Specifically, if the seismic categories are not homogeneous and events of different natures are assigned to the ~~recognition performances for Popo2002 dataset are high, with accuracy levels up to 85%. This result is often used in many works as the sole criterion for validating the quality of the database.However, a closer inspection of Fig. 1 (Uniform Manifold Approximation and Projection for dimension reduction (UMAP)(McInnes et al.~~ same type, the system's performance will drop. If events classified as part of the same category are not consistent, the system will struggle to make accurate predictions, as the inherent variability within each type undermines the learning process. In such cases, ~~2018) visualization of~~ the recognition accuracy typically falls below 70%. Therefore, Tables 4 and 5 not only provide information about the reliability of the developed systems but also about the consistency of the catalog itself.

According to such results, the three proposed systems are shown to achieve a high degree of recognition in both experiments (including 5 and 7 seismic categories), allowing us to conclude that the systems effectively learn to recognize the events annotated in the catalog. It is worth noting, however, that in the second experiment, when the number of seismic categories increases from 5 to 7, the recognition rate of the 3 systems is slightly affected. This result is clearly influenced by the imbalance in the dataset. The seismic category Explosion (EXP), with only 4 events, has no impact on the outcome. In contrast, the inclusion of the Garbage (GAR), with 2,739 events of varying durations, significantly changes system performance. Firstly, because it is the predominant category in terms of both number and duration, performing an analysis by windows results in a considerable increase in labels of this type, biasing system learning. Secondly, the spectral characteristics describing GAR events are very similar to those of BGN events. The former represents a set of events without a clear definition, while the latter represents seismic noise. Therefore, including both in the training process leads the systems to confuse the two, with GAR emerging as the more dominant qualitative category due to its imbalance.

Regarding the confusion matrices across the 3 systems, the analysis suggests that, the POPO2002 catalog is consistent, within each seismic category, there is coherence among the elements classified within the same category. However, propagation and source effects can influence seismic event characterization. For instance, VTE events are not well-identified, with confusion rates exceeding 60% in some cases, meaning only 40% of VT events are accurately classified. The highest confusion levels are observed between the VTE and LPE categories, possibly due to shared characteristics, as LPE events may resemble highly attenuated VTs, causing potential biases in event categorization. This overlap suggests that some seismic categories have elements positioned in overlapping areas of the representation space (mathematical space where data points are mapped according to learned features), where they share similar projected features, and events, despite being assigned to a specific cluster, could transition between categories (similar to MASTER-DEC ~~) as well as the study by Bueno et al. (2021)for the Bezymianny volcano indicate that the identification of classes of seismic events is not free of uncertainties. Different seismic events share many parameters in the representation space, and as a result, the definition of the event type depends on the training process, which could be influenced by human factors. Therefore, while systems' performances range from approximately~~ and described in Figure 3). Thus, although system performances range between 85% ~~to~~ and 90%, ~~it~~ this does not always reflect ~~that the seismic catalogue is~~ a complete or unbiased ~~, but rather that the training process has been self-consistent. More than learning to characterise the volcano dynamics describing the latent~~ seismic catalog. Rather than solely learning to characterize volcano dynamics based on an underlying physical model, the systems ~~are~~ may be learning the information contained ~~in the catalogue. For instance, in a tableware classification process, if a trained class includes both spoons and forks, the result of the accuracy matrix will be optimal, although the bias is evident. That is, catalogue-induced learning could bias the generalisation capability of a system, with information or dynamics relevant to improving background knowledge about the volcano remaining hidden. Our proposal to address this trend and potentially improve the system's generalization capabilities as well as the completeness of the seismic dataset is based on weakly supervised procedures. By leveraging an existing unbiased database, we can impose prior knowledge on the database being tested. This involves using an automatic system, which we consider well-trained, to review the seismic categories of the new dataset and retrain itself using such pseudo-labels~~within the catalog itself. Consequently, catalog-induced learning could limit a system's ability to generalize, potentially obscuring information relevant to advancing our understanding of volcanic behavior.

5.2 Developing automatic recognition systems with weakly supervised pseudo-labeling

Once the construction of catalogs through transfer learning has been discussed, we are now ready to discuss the use of weakly supervised pseudo-labeling approaches. Results demonstrate that, when applied effectively, these methods can significantly improve the detection and identification of diverse earthquake-volcanic signals. According to Table ~~4, when using this approach, the performance decreased substantially compared with the classical TL approaches (Table 2~~6, using pre-trained systems as pseudo-labelers results in a substantial decrease in overall performance compared to building automatic monitoring systems from available catalogs (Table 4). However, a closer inspection of the ~~results~~ Table 8 shows other aspects of the performance being very encouraging.

First, the ~~models detected~~ new systems recognized events that were originally not ~~recognised by expert geophysicists during data-labelling. Table 6 summarizes the comparison between the events initially~~ annotated in the ~~catalogue and the events detected by the three automatic systems~~preliminary catalog during data-labeling. The vast majority of ~~undetected events, which were not annotated in the preliminary catalogue,~~ such recognized events, were discovered within ~~segments labelled as GAR (Fig. 3a)) . Second,~~ long segments labeled as GAR or TRE. An example of this behavior can be seen in Fig. 4, which shows LP events (red boxes) that were not initially annotated during labeling within a trace labeled as TRE, along with the correction of an event originally labeled as LP, now relabeled by the ~~data-labelling process carried out by geophysical experts located the ends and beginnings of some events in positions very different from those provided by the models, which decreased per-frame recognition . This can clearly be seen during earthquake recognition (Fig. 3b)), where an averageof just~~ system as VT. This scenario occurs many times throughout the dataset, and these additional labels reduce overall recognition accuracy relative to the original labeling, although they do not necessarily represent errors.

Second, among the seismological community, there is a marked interest in associating different types of seismo-volcanic signals with models of seismic sources in order to better understand the physics of the underlying processes. At present, there are two main complementary lines of research within volcano seismology: a) the detection and identification of different types of volcanic events and b) the investigation of physical source models that explain the origin of these signals. As scientific knowledge has advanced, a paradoxical situation has developed: there is a lack of uniformity in the naming of observed seismic signals. Therefore, the subjectivity of human operators during the labeling process can lead to discrepancies in catalog construction. As a result, catalogs and automatic recognition outcomes often vary across different volcanoes and researchers, which ultimately reduces the system's ability to be universally applied and impacts its performance. A clear example of this discrepancy can be seen in Table 7. According to such table, on average, only 5% of ~~frames were correctly recognised (see Table 5 for detail) :~~ the analysis windows labeled as VTE in the original catalog were recognized by the retrained systems. On initial inspection, these results might suggest poor systems recognition for this seismic category, but interestingly, it is one of the most distinctive events due to its high-frequency content and exponential energy decay. So, what accounts for the low recognition rate? A detailed analysis shows that it is mainly due to labeling discrepancies between the MASTER-DEC event prototypes and POPO2002 catalog annotations. On the one hand, the start and end points of some events are often marked in positions that differ significantly from those annotated by the automatic systems. Instead of recognizing entire seismic traces such as volcano-tectonic earthquakes (VTE) as annotated in the original catalog, the systems detect background noise (BGN) segments before and after the earthquakes. While segments with high spectral content were ~~correctly~~ detected and classified as

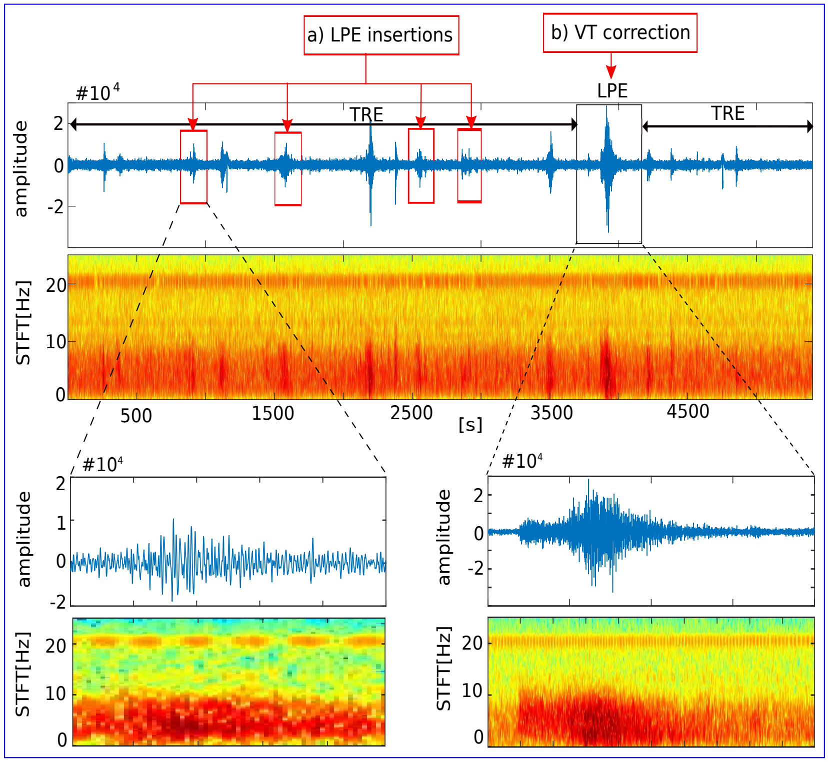


Figure 4. Insertion-based errors when retraining systems using a weakly supervised approach. Detection of LP events (red boxes) that were initially overlooked during the labeling process within a trace labeled as TRE–LPE–TRE. Correction of an event originally labeled as LP, which the system now re-labels as VT. This scenario occurs frequently throughout the dataset, and these additions reduce per-frame recognition accuracy compared to the original labeling; however, they do not always indicate errors.

VTE, those with low spectral content were classified as BGN or TRE. ~~However, after a posterior supervision by a geophysical expert, we consider that the outputs obtained by the models were better, as the earthquakes were generally well recognised. Attenuation and sources effects also affected the final recognition~~ These additional detections reduce per-frame recognition accuracy. This can be clearly seen in Fig. 5 during earthquakes recognition.

On the other hand, the VTE prototype events used in MASTER-DEC have very specific characteristics. However, some of the VTE events labeled in POPO2002 do not reliably share these characteristics. This may be due to the fact that catalogs are often constructed using data from multiple seismic stations, with strong attenuation and source effects, while imposing rules or conditions for identifying signals. Therefore, the original labeling of an event does not always align with the waveform and spectral content of the analyzed signal, as it may vary depending on the station being analyzed. As a result, if the signal being analyzed does not align with the characteristics of the prototype event used to construct the system, such signal will be labeled or associated with the event prototype that most probabilistically resembles it. This behavior reduces the recognition rate for this seismic category. Figure 6 illustrates this behavior, showing two examples of events annotated as VTE in the POPO2002

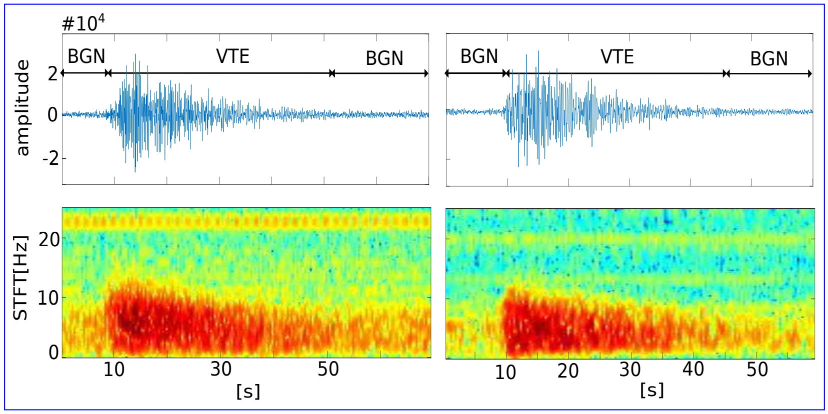


Figure 5. Insertion-based errors when retraining systems using a weakly supervised approach. Event delimitation: examples of the labeling process for the systems. Instead of recognizing entire seismic traces such as volcano-tectonic earthquakes (VTE) as annotated in the original catalog, the systems detect background noise (BGN) segments before and after the earthquakes. These additional detections reduce per-frame recognition accuracy; however, after a posterior revision, they should not be considered errors. The current colormap in the spectrogram represents the energy levels. The blue color corresponds to the minimum energy, while the red color corresponds to the maximum energy.

catalog that are recognized as TRE by the systems. The Power Spectral Density (PSD) of both events shows a clear content in low and intermediate frequencies (1-12 Hz), perfectly aligning with the source model proposed by Ibañez et al. (2000) in Table 1, which is also followed by the MASTER-DEC. Similar to the previous analysis, this behavior is repeated throughout the database, not only with TRE but also with LPE events, which explains the high degree of confusion addressed. A potential solution to this situation would be to apply the algorithm to different stations.

Third, intra-category variability can also affect the overall recognition of the systems. The new dataset contains high ~~intra-class~~ variability in some categories (~~those composed of different subcategories as LPE or TRE~~ categories composed of distinct events with shared characteristics are grouped into a single category, such as various LPs, TRE events, or regional and volcano-tectonic earthquakes all labeled collectively as earthquakes). Again, the nature of the seismic data played an essential role. Within the feature space, the representation of events belonging to a given subcategory in the new domain (~~Popo2002~~POPO2002) was closely related to the representation of events belonging to a different category in the source domain (MASTER-DEC). For example, similar to what occurs with some events in Fig. ~~1~~3, the representations of some ~~LPE subcategories in Popo2002 were~~ LPEs in POPO2002 are very close to the representation of TRE in MASTER-DEC (Fig. ~~4a)~~7)a). As such, the ~~weakly supervised selection and labelling algorithm assigned~~ algorithm assigns the TRE label during the training phase. This decreased the overall systems performance since many frames (33%, 19%, and 18% for TCN, RNN–LSTM, and Dilated–LSTM, respectively) were detected as TRE~~, resulting in an ‘error’ that was not really an error~~. The same issue arose for some attenuated earthquakes, which were labelled as LPE in the original seismic ~~catalogue but ‘mis-classified’~~ catalog but classified as VTE or TRE ~~in the new one. After a posterior supervision by a geophysical expert, we consider the output obtained by the models to be correct~~ since, even when attenuated, they align with the feature space

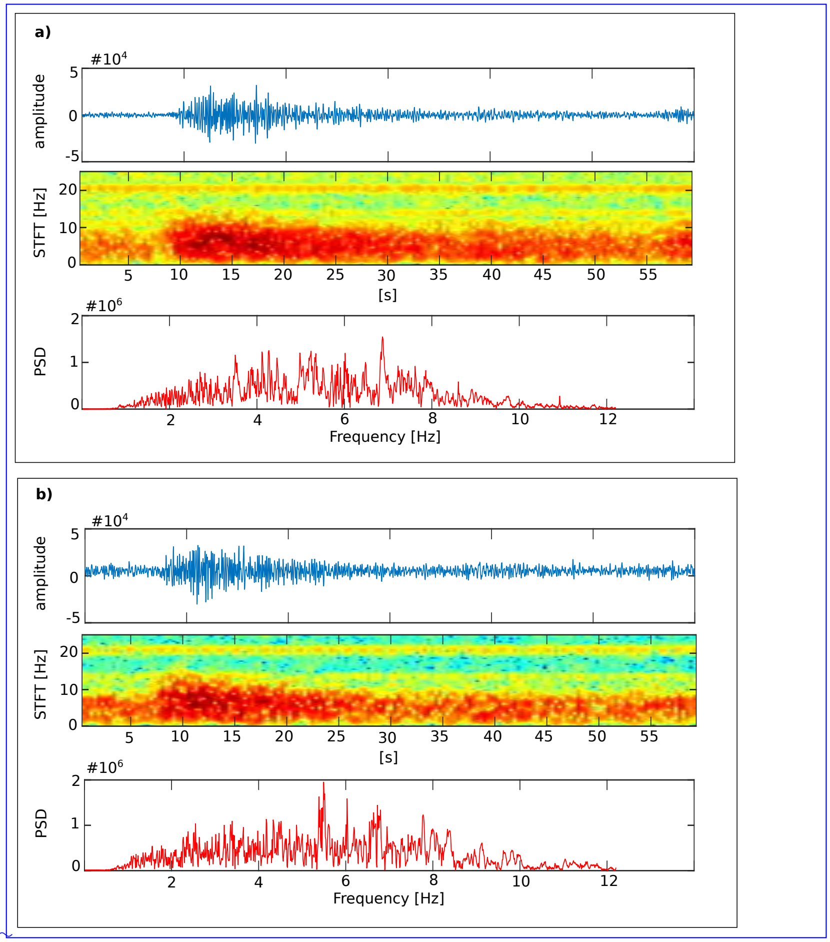


Figure 6. Two examples of event annotated as VTE in the POPO2002 catalogue being recognized as TRE for the systems. The current colormap in the spectrogram represents the energy levels. The blue color corresponds to the minimum energy, while the red color corresponds to the maximum energy. The PSD reflects the distribution of a signal's energy among the frequencies.

representation of an earthquake event in MASTER-DEC (Fig. ~~4b)since the seismic traces correspond to earthquakes with attenuated high frequencies~~7b). Finally, low-energy TRE events were clearly mis-classified as BGN because the peak-to-peak amplitude degradation of the signals was related to attenuation effects. This complex scenario was widely discussed by Titos et al. (2018); therefore, to correctly deal with these errors, further information from several seismic stations is needed. ~~Additional illustrative examples can be found in the Weakly Supervised Learning section of the supplementary material (refer to Fig. F1, F2, and F3).~~ The results suggest that ~~per-frame~~ overall recognition can be strongly biased by the intrinsic limitations addressed when developing the seismic ~~catalogue~~ catalog and from which the comparative metrics were obtained. Therefore, if labelling criteria between datasets differ, per-frame recognition results will vary widely. ~~Until now~~Hitherto, the development of new monitoring systems has focused primarily on improving existing recognition rates. However, our findings confirm that

~~weakly supervised learning approaches~~ by leveraging an existing unbiased master catalog, we can incorporate prior knowledge into the new dataset under review. Using automatic pseudo-labelers have the remarkable capability of simultaneously identifying unannotated seismic traces in the ~~catalogue~~ catalog and help to correct the labels of mis-annotated seismic traces. ~~When a weakly supervised approach based on a master seismic catalogue guides the re-training phase, although~~ Although the general performance of the system ~~decreases~~seems to decrease relative to the original catalog, previously hidden information that can improve knowledge of the volcanic dynamic background can be obtained. ~~**Popo2002 catalogue RNN-LSTM Dilated-LSTM TCN**~~

5.3 Building a new catalog during an eruptive crisis: The Tajogaite volcano use case, 2021

To conduct a detailed analysis of the results obtained in this experiment, it is essential to know the reference data. As mentioned earlier, this experiment considered the seismic traces from two stations, PLPI and PPMA, for September 12, 2021, a few days before the eruption of Tajogaite volcano began. On this day, given the volcanic activity and monitoring conditions only 247 earthquakes, both tectonic and volcanic, were annotated in the catalog.

~~340 >20, 000 >20, 000 17,~~ Considering this information, we now proceed to discuss the results. For the sake of the comparison, we will start analyzing the outcomes obtained by PhaseNet. PhaseNet detected several hundreds of P and S phases, with the number of S phases being higher at both stations. It is due to the greater energy associated with these waves. However, as it can be seen in Figure 8a, when fixing a phase score threshold highlighting the reliability of the detections, the number of detections decreases rapidly with high values. For example, for values close to 80%, only approximately 722 P-phases–503 S-phases at PLPI and 282 P-phases–216 S-phases at PPMA are detected. This significantly reduces the number of potential events that could be included in the catalog. Fig. 8b shows the match between detections and the cataloged events. Of these 247 annotated events, Phasenet detects 206 P-phases and 199 S-phases at PLPI; and 157 P-phases and 28 S-phases at PPMA, all without applying any probability threshold. Again, when setting the phase score threshold greater than or equal to 80%, the detections decrease to 163 P-phases and 164 S-phases at PLPI, and 116 P-phases and 21 S-phases at PPMA. This behavior underscores the complexity of constructing seismic catalogs, as even when focusing solely on seismic phase detections, there is no consistent criterion between a human operator and advanced automatic systems for choosing events. More importantly, even when considering the inclusion of these potential events, extensive human supervision would be required to validate and categorize them.

Looking at the recognition results obtained by the pre-trained reference systems (see Table 9), it can be observed that a big amount of events are being detected. However, similar to Phasenet, some of such events should be discarded because the reliability of the recognitions. Figure 9 depicts such reliability based on the belonging probabilities outputted by the systems. To dive into these results: 1) we will analyze how the number of detections changes as the reliability changes (we focus on more specific or sensitive systems); 2) we will examine how the systems perform using as reference the 247-events annotated in the catalog; and 3) we will assess the reliability of the remaining detected events in order to evaluate the reliability of the new pseudo-catalogs.

~~273 3, 291 2, 538 3, 204 371 1, 741 1, 032 94 1, 155 2, 230 2, 250 2, 159 Comparison between the events initially annotated~~

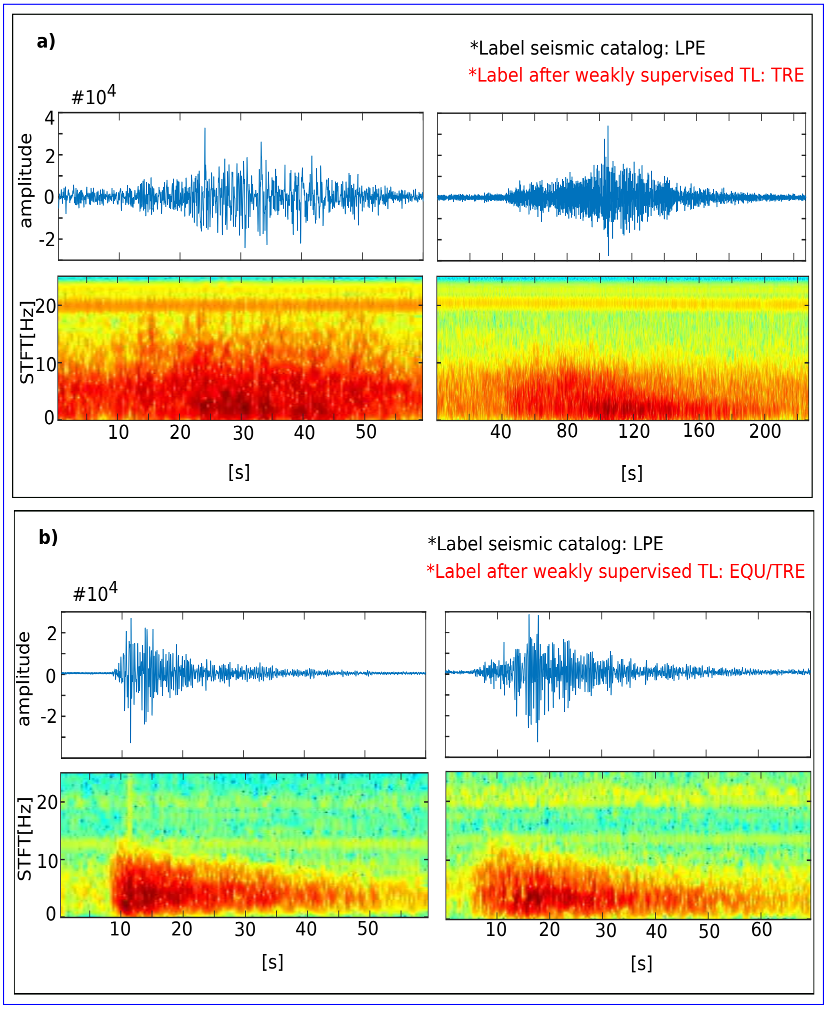
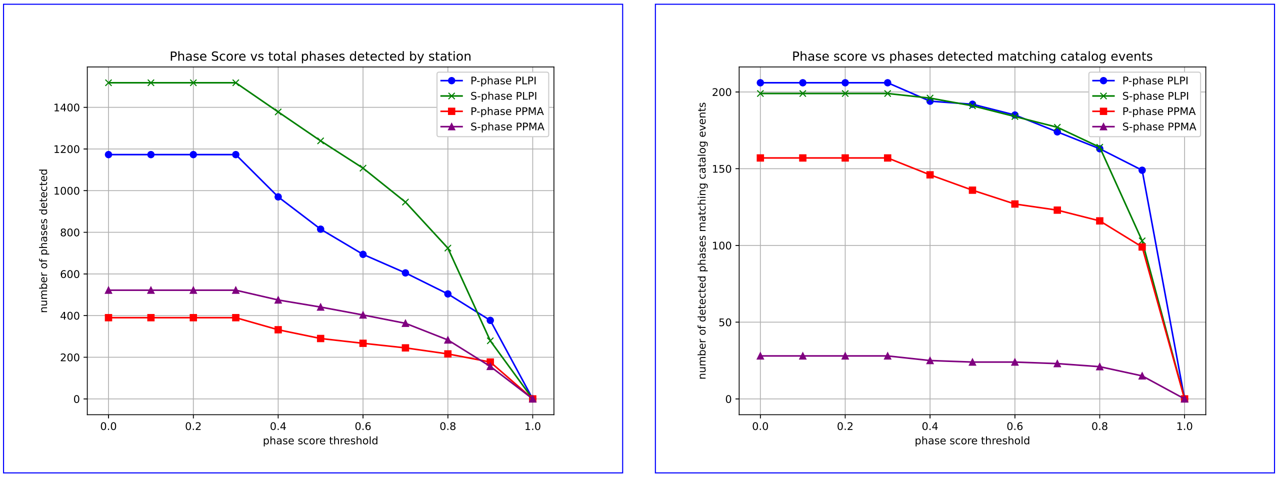


Figure 7. Detailed analysis of intra-class variability and attenuation-based errors when applying a weakly supervised approach. (a) Intra-class variability-based errors: some long period event (LPE) subcategories in POPO2002 are very close to the representation of tremor (TRE) in MASTER-DEC. (b) Two attenuated earthquakes labelled as LPE in the seismic catalog, but classified as volcano-tectonic earthquake (VTE) or TRE. The current colormap in the spectrogram represents the energy levels. The blue color corresponds to the minimum energy, while the red color corresponds to the maximum energy.

~~in the catalogue and the events detected by the three distinct automatic systems following the implementation of a weakly supervised learning TL approach.~~ Across all systems and at both stations, the number of detected events decreases significantly as the probability threshold increases, particularly for values above 80%. At higher thresholds, the detections are predominantly limited to events closely correlated to the prototype events on which the systems were trained. Figure 9c shows that for thresholds above 80%, the number of detected earthquakes by both RNN-LSTM and Dilated-LSTM averages between 120 and 150 events at both stations. For TCN, the number of detected earthquakes is significantly higher, highlighting that its specificity



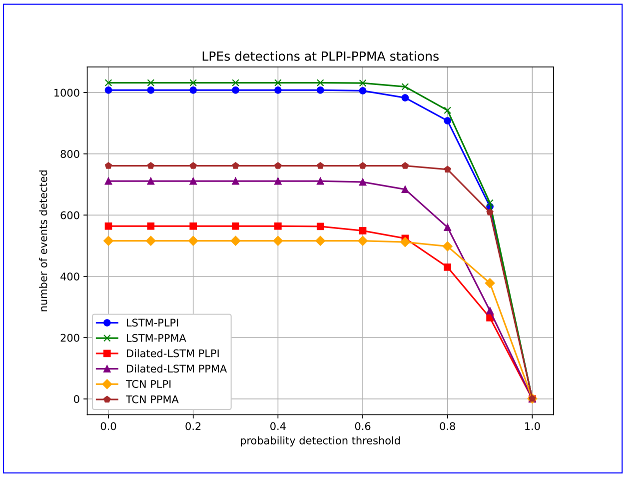
(a) (b)

Figure 8. Evolution of the number of detected phases at the seismic stations as the phase score threshold varies. A) Total number of phases detected at both stations. B) Number of phases matching the 247 events recorded in the catalog on 12/9/2021.

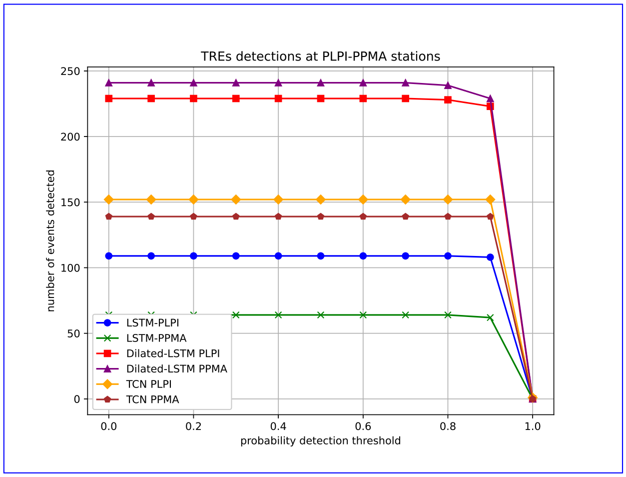
could be set at slightly higher thresholds, around 85-90%. The main reason for the non-detection of certain catalog-annotated events was their differing spectral content compared to the average spectral content of the earthquakes annotated in the catalog. Specifically, by comparing the spectral content of the undetected events with the average spectral content of all the annotated events, a clear attenuation of energy is observed at higher frequencies (>15 Hz). This characteristic is crucial, as the systems were trained with prototype events that had a clear energy component at high frequencies. Figure 11 illustrates a couple of examples of this behavior. The first row corresponds to the seismogram of the event being analyzed (annotated in the catalog but not detected by any of the systems). The second row corresponds to their spectrograms. The third and fourth rows show the average power spectral density (PSD) of all events annotated in the catalog for that day and the PSD of the event under analysis. The fourth row of both Figure 10a and 10b show a clear attenuation of energy at high frequencies and a higher level of energy at lower and intermediate frequencies, respectively. In general, these events reflect belonging probabilities ranging between 50% and 80%. It highlights the importance of adjusting the specificity or sensitivity threshold when creating new pseudo-catalogs.

Regarding the detection of events identified by the systems but not annotated in the catalog, on average, RNN-LSTM and Dilated-LSTM detected approximately 60 earthquake-type events, while TCN identified over 150. Figure 11 presents a couple of examples of such earthquakes. The PSDs reveals that they share characteristics consistent with those of earthquakes. However, as indicated by the probabilities shown at the top of the figure, their partial similarity in spectral content prevented them from being classified with higher confidence.

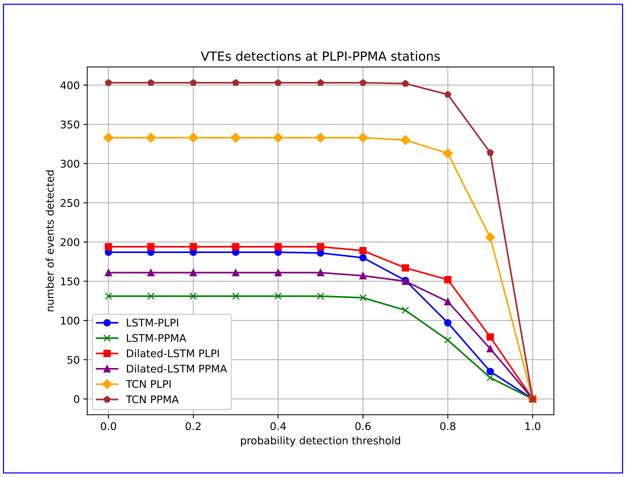
Finally, it is important to discuss the recognition of events different to earthquakes, for which there is no available information to contrast the results. Figures 9a and 9b show the number of LPE and TRE events recognized by the systems, along with their



(a)



(b)

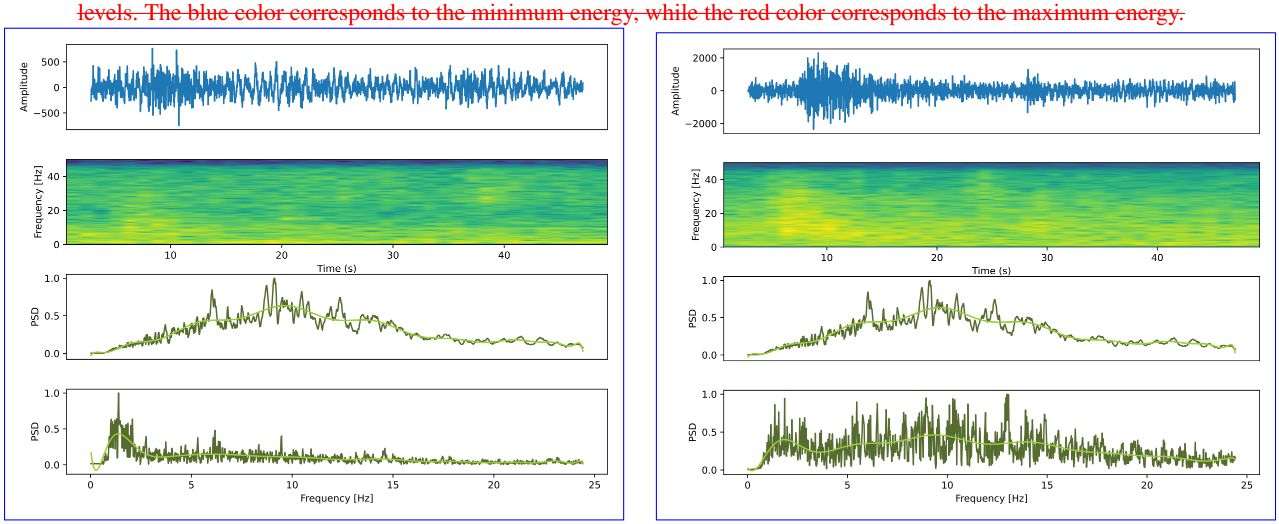


(c)

Figure 9. ~~Uniform Manifold Approximation and Projection (UMAP) obtained for~~ Evolution of the ~~input vector forming~~ number of detected event at the ~~original data~~ seismic stations as the belonging probabilities threshold varies. A) Total number of ~~MASTER-DEC dataset~~LPEs detected at both stations. ~~Different seismic categories may have elements located in overlapping areas~~ B) Total number of ~~the representation space, where they share similar projected parameters~~TREs detected at both stations. C) Total number of VTEss detected at both stations.

corresponding membership probabilities. From these figures, it can be concluded that the number of detected events is high for both categories, and the assigned membership probabilities are also relatively high, ranging from 80% to 95%. Unlike earthquakes, where high-frequency energy from external factors can lead to errors, TRE and LPE events are highly distinctive and well-defined at low frequencies. Since the systems were trained using parameter vectors based on logarithmic scale filter banks, which provide higher resolution at low frequencies than at high frequencies, the analysis of energy distribution across

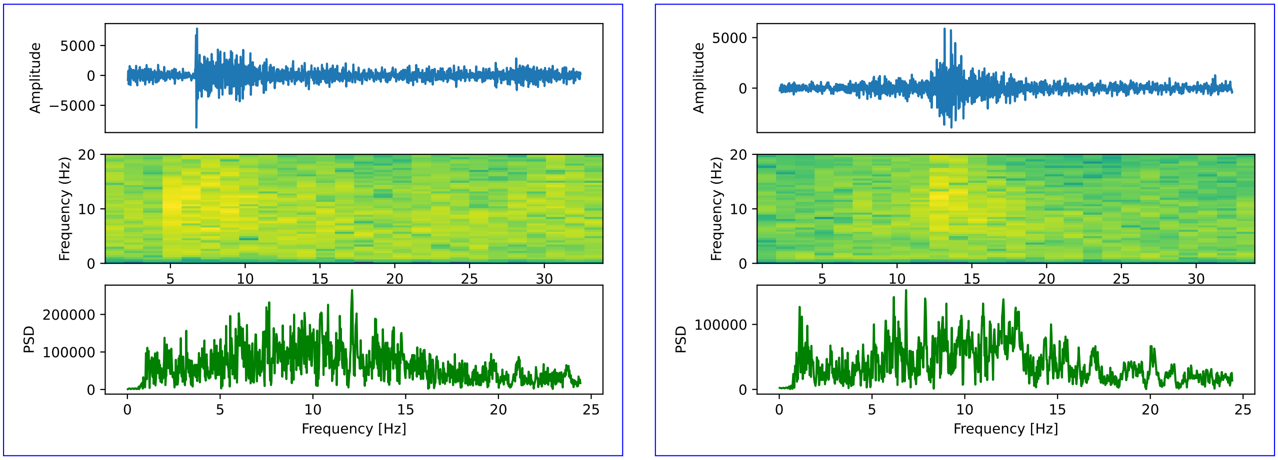
~~Insertion-based errors when applying a weakly supervised transfer learning (TL) approach. (a) Detection of events previously not labelled: long period events (LPE; red boxes) that were not originally taken into account during the labelling process have been added to the seismic trace labelled as TRE (tremor)–LPE–TRE. This scenario occurs many times throughout the dataset and these insertions decrease per-frame recognition with respect to the original labelling; however, they do not always correspond to errors. (b) Event delimitation: examples of the labelling process obtained for the models. Instead of recognising entire seismic traces, such as volcano-tectonic earthquakes (VTE), background noise (BGN) segments are detected before and after the earthquakes. Again, these insertions reduce per-frame recognition; however, after a posterior revision, they should not be considered as an error. The current colormap in the spectrogram represents the energy levels. The blue color corresponds to the minimum energy, while the red color corresponds to the maximum energy.~~



(a) (b)

Figure 10. ~~Detailed analysis~~ Example of ~~intra-class variability and attenuation-based errors when applying a weakly supervised transfer learning (TL) approach. (a) Intra-class variability-based errors: some long period event (LPE) subcategories~~ two earthquakes annotated in ~~Popo2002 are very close to~~ the ~~representation~~ catalog that were not detected by any of ~~tremor (TRE) in MASTER-DEC. Thus, the weakly supervised selection and labelling approach will assign the TRE label during~~ the ~~training phase; this decreases per-frame recognition~~3 reference systems. ~~(b~~a) ~~Two attenuated earthquakes labelled as LPE in the seismic catalogue, but 'miss-classified' as volcano-tectonic~~ Spectral analysis of an undetected earthquake~~(VTE) or TRE. After a posterior supervision~~, ~~we consider the model outputs to be correct since the seismic traces correspond to earthquakes with attenuated~~ where a clear attenuation of energy at high frequencies is observed. ~~The current colormap in the spectrogram represents the energy levels. The blue color corresponds to the minimum energy~~b) Spectral analysis of an undetected earthquake, ~~while the red color corresponds to the maximum~~ where an energy distribution in intermediate frequencies and attenuation at high frequencies are observed.

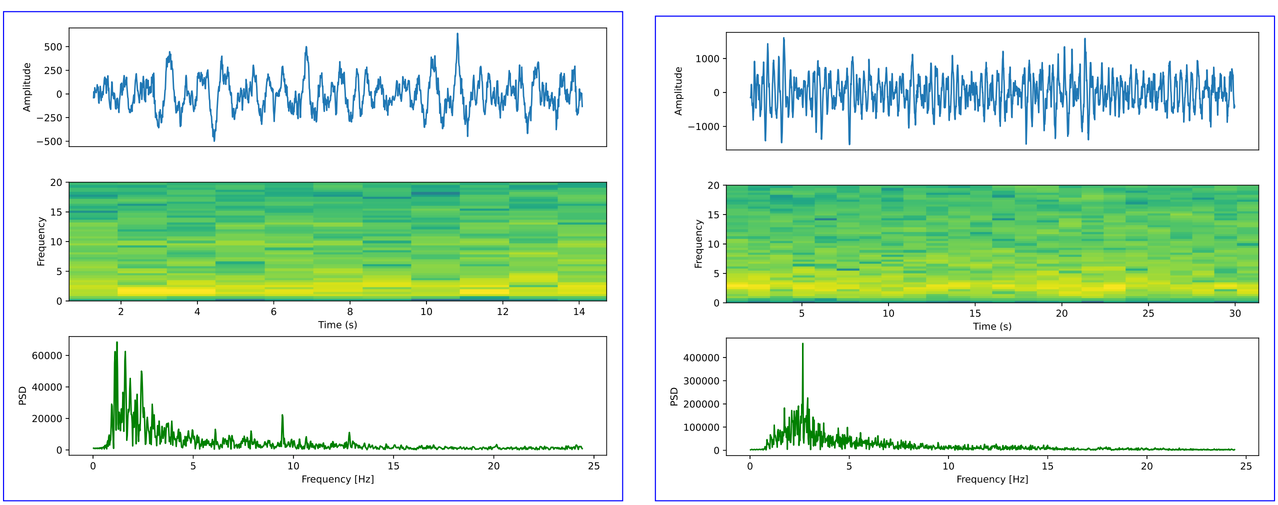
low frequencies is highly reliable. Figure 12 shows an example of the LPE and TRE detections. As shown, these events were recognized with very high probabilities. Analyzing their spectral content, waveform, and energy reveals a perfect correlation with the characteristics of the prototype events on which the systems were trained, as illustrated in Figure **??**. Therefore, we can conclude that a large percentage of the detected TRE and LPE events correspond to prototype events from MASTER-DEC,



(a) (b)

Figure 11. Example of two earthquakes not annotated in the catalog that were detected by the 3 reference systems with probabilities ranging from 63% to 78%.

which indicate the associated source mechanism of their label. It will be the responsibility of the volcano experts to analyze whether these detected events share the same source mechanism or whether they should be re-labeled before pre-training the systems to adjust to the volcanic environment under analysis.



(a) (b)

Figure 12. Example of a) LPE detected but not annotated and b) TRE detected but not annotated in the catalog.

5.4 Summary of Findings

The results presented in each experiment provide valuable insights into the development of automatic recognition systems with weakly supervised pseudo-labeling, highlighting both the strengths and limitations of the proposed methods. By synthesizing the outcomes, we aim to offer a comprehensive understanding of how leveraging an existing automatic pseudo-labeler based on a master catalog can incorporate prior knowledge into the new dataset under review, which can inform future research and applications in the field.

Among the main strengths identified, the systems's ability to recognize previously learned prototype events, even in scenarios quite different from those analyzed during the learning process. This feature enhances its usefulness in reducing biases when creating or improving catalogs. The results demonstrate that if systems would be trained across diverse volcanic environments with varied distributions of prototype events, recognition results could improve, suggesting good adaptability and, consequently, the construction of less biased catalogs in new scenarios and volcanic settings. However, the systems shows certain limitations, such as the detection of events that do not match any prototype, which could impact the final performance of the re-trained systems. This primarily occurs because the pre-trained reference systems from which the pseudo-catalogs are built must assign a category to each analyzed window. Therefore, the systems will always assign a seismic category, even when the prototype is far from the signal under analysis. Once again, this challenge can be addressed by creating more comprehensive training datasets that describe different event distributions. Finally, another major challenge identified is the decision of the membership threshold from which events are included in the new pseudo-catalogs, indicating a need for post-analysis to assess the confidence of the detections, which would help distinguish between very sensitive or very specific pseudo-catalogs. Adjusting low probability thresholds will allow the creation of highly sensitive catalogs, which may result in many false positives—events that do not match the prototype. Retraining the systems with these catalogs could drop the performance and detection skills. On the other hand, a high probability threshold might not be sufficient to adapt the systems to the new volcanic environment.

6 Conclusions

This study provides the first comprehensive analysis of seismic ~~catalogue-induced~~ catalog-induced bias when developing automatic recognition systems. We evaluated the ability of several monitoring systems trained using a master seismic ~~catalogue~~ catalog from Deception Island volcano to adapt to a new seismic ~~catalogue~~ catalog from Popocatépetl volcano through our novel, proposed weakly supervised framework. Our results confirm the robustness of data-driven approaches as a basis for the construction of short-term early-warning systems. However, quantitative and qualitative analysis confirmed that the reliability of a system is strongly biased by the undetailed coverage of the seismic ~~catalogue~~catalog. While systems performance reached almost 90% per-frame recognition accuracy, intrinsic limitations when developing seismic ~~catalogues~~ catalogs led to extremely useful information describing the volcanic behaviour being ignored. Instead of simply learning to characterise volcanic dynamics by describing the latent physical model, ~~catalogue-induced~~ catalog-induced learning can bias the system by discarding useful data describing volcanic dynamics. However, when a weakly supervised learning approach based on a master seismic ~~catalogue~~ catalog is applied, an unknown amount of information related to volcano dynamics is revealed. This study raises important questions about the relevance of ~~catalogue-induced~~ catalog-induced learning when developing new monitoring systems. Our results demonstrate that systems based on iterative weakly supervised or even unsupervised learning techniques could offer a more successful approach than supervised techniques under crude seismic ~~catalogues~~catalogs. Therefore, we conclude that ensuring appropriate seismic ~~catalogues~~ catalogs and support for developing monitoring tools should be a priority to the same extent as applying new and more effective AI techniques. The use of more sophisticated pseudo-labelling techniques involving data from several ~~catalogues~~ catalogs could help to develop universal monitoring tools able to work accurately across different volcanic systems, even when faced with unforeseen temporal changes in monitored signals.

Code availability. Correspondence and requests for materials should be addressed to M.T.

Data availability. Correspondence and requests for materials should be addressed to M.T.

Code and data availability. Correspondence and requests for materials should be addressed to M.T.

7

6.1

Author contributions. All authors contributed to the conception of the study. M.T., C.B., conceived and conducted the experiments. All authors analysed and interpreted the results. All authors reviewed the manuscript.

Competing interests. The authors declare no competing interests

Disclaimer. TEXT

Acknowledgements. This work has been partially funded by the Spanish Project PID2022-143083NB-I00, "LEARNING", funded by MCIN/AEI /10.13039/501100011033 and by FEDER (EU), DIGIVOLCAN (PLEC2022-009271), SDAS (B-TIC-542-UGR20), and PROOF-FOREVER (EUR2022-134044). The authors would also like to acknowledge Instituto Andaluz de Geofísica, Instituto Volcanológico de Canarias

(INVOLCAN), Raúl Arámbula-Mendoza and CENAPRED for providing the data from Deception Island~~and Popo~~, Popo and Tajogaite volcanoes supporting this work. Any opinions, findings, conclusions, or recommendations expressed in this research are those of the authors and do not necessarily reflect the view of the funding agencies.

References

[1] ~~REFERENCE~~

[2] Sparks, R. S. J. Forecasting volcanic eruptions. Earth and Planetary Science Letters, 210(1-2), 1-15 (2003). Doi: 10.1016/S0012-821X(03)00124-9

[3] Witze, A. How ai and satellites could help predict volcanic eruptions. Nature 567, 156–158 (2019). doi: 10.1038/d41586-019-00752-3

[4] Palmer, J. The new science of volcanoes harnesses ai, satellites and gas sensors to forecast eruptions. Nature 581, 256–260 (2020), doi: 10.1038/d41586-020-01445-y

[5] Chouet, B. Volcano seismology. Pure applied geophysics 160, 739–788 (2003), doi: 10.1007/PL00012556

[6] McNutt, S. R., Roman, D. C. Volcanic seismicity. In The encyclopedia of volcanoes, 1011–1034 (2015), doi: 10.1016/B978-0-12-385938-9.00059-6

[7] Minakami, T. Prediction of volcanic eruptions. Dev. Solid Earth Geophys. 6, 313–333 (1974), doi: 10.1016/B978-0-444-41141-9.50020-6

[8] Rey-Devesa, P., Prudencio, J., Benítez, C. et al. Tracking volcanic explosions using Shannon entropy at Volcán de Colima. Sci Rep 13, 9807 (2023). doi:10.1038/s41598-023-36964-x

[9] Rey-Devesa, P., Benítez, C., Prudencio, J. et al. Volcanic early warning using Shannon entropy: Multiple cases of study. Journal of Geophysical Research: Solid Earth, 128, e2023JB026684 (2023). doi:10.1029/2023JB026684

[10] Ohrnberger, M. Continuous automatic classification of seismic signals of volcanic origin at Mt. Merapi, Java, Indonesia. Ph.D. thesis, Potsdam, Univ., Diss., 2001

[11] Scarpetta, S., Giudicepietro, F., Ezin, E. C., et al. Automatic classification of seismic signals at Mt. Vesuvius volcano, Italy, using neural networks. Bulletin of the Seismological Society of America, 95(1), 185-196 (2005). doi: 10.1785/0120030075.

[12] ~~REFERENCE 2~~

[13] Alasonati, P., Wassermann, J., Ohrnberger, M. Signal classification by wavelet-based hidden Markov models: application to seismic signals of volcanic origin. Statistics in Volcanology, H. M. Mader, S. G. Coles, C. B. Connor, L. J. Connor (2006). doi: 10.1144/IAVCEI001.13

[14] Benítez, M. C., Ramírez, J., Segura, J. C., et al. Continuous HMM-based seismic-event classification at Deception Island, Antarctica. IEEE Transactions on Geoscience and remote sensing, 45(1), 138-146 (2006). doi:10.1109/TGRS.2006.882264

[15] Ibáñez, J. M., Benítez, C., Gutiérrez, L. A., et al. The classification of seismo-volcanic signals using Hidden Markov Models as applied to the Stromboli and Etna volcanoes. Journal of Volcanology and Geothermal Research, 187(3-4), 218-226 (2009). doi:10.1016/j.jvolgeores.2009.09.002

[16] Curilem, G., Vergara, J., Fuentealba, G., et al. Classification of seismic signals at Villarrica volcano (Chile) using neural networks and genetic algorithms. Journal of volcanology and geothermal research, 180(1), 1-8 (2009). doi: 10.1016/j.jvolgeores.2008.12.002

[17] Bhatti, S. M., Khan, M. S., Wuth, J., et al. Automatic detection of volcano-seismic events by modeling state and event duration in hidden Markov models. Journal of Volcanology and Geothermal Research, 324, 134-143 (2016). doi: 10.1016/j.jvolgeores.2016.05.015

[18] Canario, J. P., Mello, R., Curilem, M., et al. In-depth comparison of deep artificial neural network architectures on seismic events classification. J. Volcanol. Geotherm. Res. 401, 106881 (2020). doi: 10.1016/j.jvolgeores.2020.106881

[19] Cortés, G., Carniel, R., Lesage, P., et al. Practical volcano-independent recognition of seismic events: Vulcan. ears project. Front. Earth Sci. 8, 616676 (2021). doi: 10.3389/feart.2020.616676

[20] Bicego, M., Rossetto, A., Olivieri, M., et al. Advanced knn approaches for explainable seismic-volcanic signal classification. Math. Geosci. 1–22 (2022). doi: 10.1007/s11004-022-10026-w

[21] Bueno, A., Benítez, C., Zuccarello, L., et al. Bayesian monitoring of seismo-volcanic dynamics. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-14 (2021). doi: 10.1109/TGRS.2021.3076012

[22] Bueno, A., Balestriero, R., De Angelis, S., et al. Recurrent scattering network detects metastable behavior in polyphonic seismo-volcanic signals for volcano eruption forecasting. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-23 (2021). doi:10.1109/TGRS.2021.3134198

[23] Martínez, V. L., Titos, M., Benítez, C., et al. Advanced signal recognition methods applied to seismo-volcanic events from Planchon Peteroa Volcanic Complex: Deep Neural Network classifier. Journal of South American Earth Sciences, 107, 103115 (2021). doi: 10.1016/j.jsames.2020.103115

[24] Titos, M., Bueno, A., Garcia, L., Benitez, C. A deep neural networks approach to automatic recognition systems for volcano-seismic events. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 11, 1533–1544 (2018). doi: 10.1109/JSTARS.2018.2803198

[25] Titos, M., Bueno, A., García, L., Benítez, C., Ibáñez, J. M. Detection and classification of continuous volcano-seismic signals with recurrent neural networks. IEEE Transactions on Geosci. Remote. Sens. 57, 1936–1940 (2019). doi: 10.1109/TGRS.

[26] Titos, M., García, L., Kowsari, M., Benítez, C. Toward knowledge extraction in classification of volcano-seismic events: Visualizing hidden states in recurrent neural networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, **15**, 2311–2325 (2022). doi: 10.1109/JSTARS.2022.3155967.

[27] Titos, M., Gutiérrez, L., Benítez, C., et al. Multi-station volcano tectonic earthquake monitoring based on transfer learning. *Frontiers in Earth Science*, **11**, 1204832 (2023). doi: 10.3389/feart.2023.1204832.

[28] Ibáñez, J. M., De Angelis, S., Díaz-Moreno, A., et al. Insights into the 2011–2012 submarine eruption off the coast of El Hierro (Canary Islands, Spain) from statistical analyses of earthquake activity. *Geophysical Journal International*, **191**(2), 659-670 (2012). doi: 10.1111/j.1365-246X.2012.05629.x.

[29] Díaz-Moreno, A., Ibáñez, J. M., De Angelis, S., et al. Seismic hydraulic fracture migration originated by successive deep magma pulses: The 2011–2013 seismic series associated to the volcanic activity of El Hierro Island. *Journal of Geophysical Research: Solid Earth*, **120**(11), 7749-7770 (2015). doi: 10.1002/2015JB012249.

[30] Chang, S., Zhang, Y., Han, W., et al. Dilated recurrent neural networks. *Advances in Neural Information Processing Systems*, **30** (2017).

[Titos et al.] Titos, M., Carthy, J., García, L., et al. Dilated-RNNs: A deep approach for continuous volcano-seismic events recognition. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* (to be published).

[32] Smellie, J. L. Recent observations on the volcanic history of Deception Island, South Shetland Islands. *Br. Antarctic Surv. Bull.*, **83-85** (1988).

[33] Carmona, E., Almendros, J., Serrano, I., et al. Results of seismic monitoring surveys of Deception Island volcano, Antarctica, from 1999–2011. *Antarctic Sci.*, **24**, 485–499 (2012). doi: 10.1017/S0954102012000314.

[34] Martí, J., Geyer, A., Aguirre-Díaz, G., et al. Deception Island (South Shetland Islands, Antarctica): An example of a tectonically induced collapse caldera. (2011). URI: http://hdl.handle.net/10261/162691.

[35] Ibañez, J. M., Pezzo, E. D., Almendros, J., et al. Seismovolcanic signals at Deception Island volcano, Antarctica: Wave field analysis and source modeling. *J. Geophys. Res. Solid Earth*, **105**, 13905–13931 (2000). doi: 10.1029/2000JB900013.

[36] Martínez-Arévalo, C., Bianco, F., Ibáñez, J. M., et al. Shallow seismic attenuation and shear-wave splitting in the short period range of Deception Island volcano (Antarctica). *Journal of Volcanology and Geothermal Research*, **128**(1-3), 89-113 (2003). doi: 10.1016/S0377-0273(03)00248-8.

[37] Zandomeneghi, D., Barclay, A., Almendros, J., et al. Crustal structure of Deception Island volcano from P wave seismic tomography: Tectonic and volcanic implications. *Journal of Geophysical Research: Solid Earth*, **114**(B6) (2009). doi: 10.1029/2008JB006119.

[38] Ibáñez, J. M., Díaz-Moreno, A., Prudencio, J., et al. Database of multi-parametric geophysical data from the TOMO-DEC experiment on Deception Island, Antarctica. *Scientific Data*, **4**(1), 1-18 (2017). doi: 10.1038/sdata.2017.128.

[39] Arango-Galván, C., Martin-Del Pozzo, A. L., Flores-Márquez, E., et al. Unraveling the complex structure of Popocatépetl volcano (Central Mexico): New evidence for collapse features and active faulting inferred from geophysical data. *J. Volcanol. Geotherm. Res.*, **407**, 107091 (2020). doi: 10.1016/j.jvolgeores.2020.107091.

[40] D'Auria, L., et al. (2022). Rapid magma ascent beneath La Palma revealed by seismic tomography. *Scientific Reports, 12*(1), 17654.

[41] Alaniz-Álvarez, S. A., Nieto-Samaniego, Á. F., Mexicana, S. G. *Geology of México: Celebrating the centenary of the Geological Society of México*, vol. 422 (Geological Society of America, 2007).

[42] Siebe, C., Salinas, S., Arana-Salinas, L., Macías, J. L., et al. The 23,500 y 14C BP white pumice plinian eruption and associated debris avalanche and Tochimilco lava flow of Popocatépetl volcano, Mexico. *J. Volcanol. Geotherm. Res.*, **333**, 66–95 (2017). doi: 10.1016/j.jvolgeores.2017.01.011.

Aquí está la referencia en el formato solicitado:

[43] Weiss, K., Khoshgoftaar, T. M., Wang, D. A survey of transfer learning. *Journal of Big Data*, **3**, 1–40 (2016). doi: 10.1186/s40537-016-0043-6.

[44] Zhu, W., and Beroza, G. C. PhaseNet: a deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, **216**(1), 261-273 (2019). doi: 10.1093/gji/ggy423.

[45] Malfante, M., Dalla Mura, M., et al. Machine learning for volcano-seismic signals: Challenges and perspectives. *IEEE Signal Process. Mag.*, **35**, 20–30 (2018). doi: 10.1109/MSP.2017.2779166.

[46] Lara, F., Lara-Cueva, R., Larco, J. C., et al. A deep learning approach for automatic recognition of seismo-volcanic events at the Cotopaxi volcano. *J. Volcanol. Geotherm. Res.*, **409**, 107142 (2021). doi: 10.1016/j.jvolgeores.2020.107142.

[47] Hibert, C., Provost, F., Malet, J. P., et al. Automatic identification of rockfalls and volcano-tectonic earthquakes at the Piton de la Fournaise volcano using a random forest algorithm. *J. Volcanol. Geotherm. Res.*, **340**, 130–142 (2017). doi: 10.1016/j.jvolgeores.2017.04.015.

[48] Köhler, A., Ohrnberger, M., Scherbaum, F. Unsupervised pattern recognition in continuous seismic wavefield records using self-organizing maps. *Geophys. J. Int.*, **182**, 1619–1630 (2010). doi: 10.1111/j.1365-246X.2010.04709.x.

[49] Hochreiter, S., Schmidhuber, J. Long short-term memory. *Neural Computation*, **9**, 1735–1780 (1997). doi: 10.1162/neco.1997.9.8.1735.

[50] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85–117 (2015). doi: 10.1016/j.neunet.2014.09.003.

[51] Lea, C., Flybbn, M. D., Vidal, R., et al. Temporal convolutional networks for action segmentation and detection. In *Proc. 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 1003–1012 (2017).

[52] Yan, J., Mu, L., Wang, L., et al. Temporal convolutional networks for the advance prediction of ENSO. *Nat. Sci. Reports*, **10** (2020). doi: 10.1038/s41598-020-65070-5.

[53] Racic, M., Oštir, K., Peressutti, D., et al. Application of temporal convolutional neural network for the classification of crops on Sentinel-2 time series. *ISPRS - Int. Arch. Photogramm. Remote. Sens. Spatial Inf. Sci.*, **XLIII-B2-2020**, 1337–1342 (2020). doi: 10.5194/isprs-archives-XLIII-B2-2020-1337-2020.

[54] Van den Oord, A., Heiga Z., Karen S., et al. A generative model for raw audio. *CoRR abs/1609.03499*, (2016). doi: 10.48550/arXiv.1609.03499.

[55] Fisher, Y., Vladlen, K., Thomas, F. Dilated residual networks. *arXiv preprint arXiv:1705.09914*, (2017). doi: 10.

[56] Zhou, Z. H. A brief introduction to weakly supervised learning. *Natl. Science Review*, **5**, 44–53 (2018). doi: 10.1093/nsr/nwx106.

[57] Kouw, W., M. Loog, M. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 766–785 (2019). doi: 10.1109/TPAMI.2019.2945942.

[58] Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H. R. A brief review of domain adaptation. *Adv. Data Science Information Engineering*, 877–894 (2021). doi: 10.1007/978-3-030-71704-965.

[59] Lu, J., Liu, A., Dong, F. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, **31**(12), 2346-2363 (2018). doi: 10.1109/TKDE.2018.2876857.

[60] McInnes, L., Healy, J., Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, (2018). doi: 10.48550/arXiv.1802.03426.