Neural Computing and Applications
https://doi.org/10.1007/s00521-023-08994-z

ORIGINAL ARTICLE

Deep learning and multi-station classification of volcano-seismic events of the Nevados del Chillán volcanic complex (Chile)

Alejandro Ferreira[1] · Millaray Curilem[2] · Walter Gomez[1] · Ricardo Rios[3]

Received: 27 June 2023 / Accepted: 22 August 2023

Abstract
This paper presents a methodology for developing a volcano-seismic event classification system using a multi-station deep learning approach to support monitoring the Nevados del Chillán Volcanic Complex, which has been active since 2017. A convolutional network of multiple inputs processes the information from an event recorded up to five seismic stations. Each record is represented by its normalized spectrogram; thus, the network may receive from one to five spectrograms as input. The design includes entering additional information into the network, like the stations configuration and the event duration, information not provided by the spectrograms. Finally, this work includes the design and implementation of a relational database to access the continuous traces of events, showing different subsets of data quickly and efficiently. The results show that the classification of an event recorded up to five stations is substantially more effective than a single-station strategy. However, incorporating additional information of the signal does not significantly improve the classification performance.

Keywords Volcano monitoring · Deep learning · Data science · Classification of events

1 Introduction

Chile is a natural laboratory for studying volcanoes as it has more than 2000 volcanoes, of which around 90 are considered dangerous. The Observatorio Vulcanológico de los Andes Sur (OVDAS) is the technical entity of the National Geology and Mining Service (SERNAGEOMIN), charged with monitoring the Chilean most active volcanoes. Each volcano has multiple sensors to monitor variables such as the chemical content of the gases, the temperature, and the deformation. However, it is the volcanic seismicity that stands out in importance, as it provides information on the internal activity of a volcano. The structure of volcanoes is composed of chambers and pipelines through which the magma and gases pass, exerting continual pressure on the walls. Occasionally, this constant pressure releases energy that generates ground movements, called volcano-seismic events, which are recorded by seismic stations. Different processes can produce energy release or exchange within the volcanic structure, generating different patterns of the volcano-seismic events.

1. Long-period (LP) events are caused by an abrupt transit of magmatic and hydrothermal fluids within the volcanic conduits.
2. Tremors (TR) begin with a sustained pressure disturbance on the magmatic and hydrothermal fluid, which can be continuous or a sequence of transitory signals similar to those generated by the LP.

Millaray Curilem, Walter Gomez and Ricardo Rios have contributed equally to this work.

✉ Alejandro Ferreira
alejandro.ferreira@ufrontera.cl

Millaray Curilem
millaray.curilem@ufrontera.cl

Walter Gomez
walter.gomez@ufrontera.cl

Ricardo Rios
ricardoar@ufba.br

[1] Departamento de Ingeniería Matemática, Universidad de La Frontera, Temuco, Chile

[2] Departamento de Ingeniería Eléctrica, Universidad de La Frontera, Temuco, Chile

[3] Institute of Computing, Federal University of Bahia, Salvador, Bahia, Brazil

Published online: 13 September 2023

3. Explosions (EX) are episodic explosive outgassing that can occur due to many mechanisms. They have a surface expression and generate sound recordings.
4. Volcano-tectonic events (VT) are associated with the fracture of rocks in the conduits inside the volcanic structures.
5. Very long-period events (VLP) belong to a sub-group of LP with very low frequency, but they have the same origin.
6. The breakage of ice causes IceQuakes (IC), a phenomenon due to the dynamics of glaciers.
7. (AV) are produced by avalanches.
8. Finally, the (ZZ) class mainly contains noise caused by artifacts or conditions external to the volcano, such as storms, wind, and rain.

Classifying the seismic signals of volcanoes consists of labeling each seismic event with respect to the process that gave rise to it. This stage is generally carried out directly by analysts since, as each volcano has its own internal structure, the seismic patterns are different from one volcano to another; thus, the classification is complex despite being essential for monitoring. In addition, classification is affected by deformations in seismic patterns due to the stations' location. Seismic stations are placed at variable distances in relation to the crater. When a seismic event occurs, it travels in all directions, reaching the monitoring stations at different times. A volcanic seism can be recorded at one or several stations based on its energy, but it suffers different deformations, which depend on many factors such as the different soil types in the path from the source to the station, the quality of the soil where the station is installed (site effect), and the type of sensor. The seismic patterns recorded by the stations can be heavily deformed, and analyzing them separately may produce false conclusions. Therefore, the multi-station approach becomes necessary to improve the seismic analysis.

Thus, volcano monitoring consists of continuously processing the data from all the sensors installed in the volcano monitoring stations. This task becomes more intense as the number of volcanoes monitored, the number of stations, and volcanic activity increase. Consequently, volcano observatories generate an enormous amount of information that requires reliable and opportune processing. This need has driven the progressive development of automatic processing systems to support the different monitoring activities, in particular classification of the seismic events. As part of this continuous research effort to develop automatic monitoring and classification systems, deep neural networks (DNN) stand out for their excellent performance in pattern recognition. The literature contains many methodologies to address the problem by applying deep learning, and they differ in aspects such as the network architectures, input data, representation of the signals, and number of considered classes but there is still no consensus on the suitable design of classifiers for volcanic seismicity. In addition, due to each volcano having its own behavior, implementing universal classifiers is complex.

In this paper, we propose a novel multi-station approach that merges the information of many stations for the classification of volcano-seismic events, addressing the problem of individual station distortions. A database of events collected at the Nevados de Chillán Volcanic Complex over 7 years and labeled by OVDAS' experts was implemented. Eight classes (see Table 3) were included according to the classification given by OVDAS; see also [1]. However, due to the nature of the phenomena occurring in the volcano and its surroundings, the database is extremely unbalanced, requiring a data balancing technique, thus a process for the training, validation, and test sets definition was necessary to prepare the data.

Each seismic event was represented by its spectrogram, so it can be processed as an image. The proposed architecture uses a convolutional neural network with 5 inputs or processing tunnels, which make it possible to process the spectrograms of events recorded from one up to five monitoring stations simultaneously. Depthwise separable convolution is used as the main operation. Additional information related to the stations' location and the duration of the events was also integrated into the network. Three neural models were designed: the first two combine the multi-station information, while the third is a single-station model. Four experiments were performed to evaluate different weight update strategies and the additional input information, and to compare the multi- and single-station approaches.

The paper is divided into the following sections. State-of-the-art literature on the topic is given in next Sect. 2. Section 3 is devoted to the material. The methodologies used are presented in Sect. 4. The main results are shown in Sect. 5, and finally, a discussion on the results of the paper is provided in the last section.

2 Literature review

Next, the state-of-the-art is studied with respect to the recent research on the classification of volcanic events, using deep learning models.

In [1], the authors use convolutional neural networks [2] (CNN) to classify spectrograms that represent the volcano-seismic signal of events belonging to the Llaima volcano. Representing a signal as a spectrogram is considered a signal processing step. These were recorded between 2010

and 2016 and corresponded to the LP, TR, VT, and tectonic event (TC) classes, generating a data set of 3592 events.

In [3], they use recurrent neural networks to classify spectrograms belonging to events in classes LP, TR, VT, and hybrid events belonging to the Deception Island volcano. The data corresponded to volcano-seismic signals collected between 1994 and 1996 and 2001 and 2002. Yields of 94% were obtained regarding the detection and correct classification in training, whereas in a validation set, the accuracy reaches the value of 0.7522. All the work was done considering only 2193 events for the training, validation, and test set.

In [4], the authors present a classification system that combines 5 classification models: Naive Bayes, SVM, *K*-nearest neighbors (KNN), random forest, and an MLP network. The system was validated on 587 LP events and 81 VT events, recorded in the Cotopaxi volcano in Ecuador.

In [5], an extensive comparison of DNN architectures was carried out to classify volcano-seismic events belonging to the Llaima volcano. Spectrograms from the LP, TR, VT, and TC classes were used, making a total of 3592 events. The study concluded that a CNN yields the best result in classifying those events, compared to architectures such as Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM).

In [6], an automatic seismogram monitoring system is proposed and applied to the Nevados de Chillán Volcano. The monitoring is done in two steps: detection of the discrete events in the raw data and classification. A VGG-16 model was used for the classifier, pre-trained in the ImageNet dataset [7]. Four types of volcano-seismic events are classified: LP, TR, VT, and Others events (ZZ). The network was retrained with 15,557 events, reaching 75% accuracy.

A different exercise applied to seismic signals, for smaller databases, is done in [8], where they propose an Active Learning procedure based on temporal CNNs, to systematically select the most informative seismic samples to maximize the performance. It is mentioned that this type of learning has not been applied widely in volcano monitoring. Considering four types of seismic events, an event recognition performance of 83% is obtained.

A similar approach is taken to classify volcano-seismic events in [9], where diversity-based active learning is combined with an existing CNN classifier for data from the Nevado del Ruiz volcano in Colombia and the Llaima Volcano in Chile. The total number of data items for each volcano was 5614 and 3592, respectively. In addition, they consider LP, TR, VT, Hybrid, and Tectonic events. With active learning, they obtain an increase in the yield for the Nevado del Ruiz volcano, but not for the Llaima volcano.

Similar work for databases and deep learning was done in [10], where they propose the use of Gaussian processes (GP) and deep Gaussian processes (DGP) for the classification of volcano-seismic events using data from the Volcán de Fuego in Mexico, contrasting the performance of these methods with the traditional methods to train deep neural networks (DNN). It is demonstrated that for small data sets, the GP surpass the shallow DNN classifiers and that the DGP surpass traditional DNN approaches.

In [11] the authors propose a classification approach of LP and TC events based on combining mathematical morphology and techniques of similarity criteria, exploring the seismic signal domain to calculate a new feature space based on the edges map of the seismic events pattern represented as spectrogram images in grayscale. This transformation feeds a set of classifiers based on similarity, defined as the Euclidean distance between two vectors in the new feature space. They conclude that the execution of the proposed method is faster and the complexity of the algorithm is less than the cutting-edge methods.

In [12], a classification method for earthquakes and volcanic tremors is developed based on the analysis of the correlation of seismic records. Ten-minute time windows with data from the Sakarajima volcano in Japan were continuously analyzed. With the method, they correctly classify 80% of volcanic tremors and 75% of earthquakes. They conclude that the applied technique can be used as an automatic tool for detecting and locating earthquakes and volcanic tremors.

Something different is proposed in [13], where the authors study the frequency bands of the main volcano-seismic events, including LP, TR, VT, IC, Hybrid and Tectonic events. The values observed in practice are compared with those published in the literature. They conclude that establishing the frequency bands that characterize the typical events of each volcano makes it possible to know them better. In addition, knowing the bands improves the spectrogram generation processes, or features for correctly classifying the events in automatic classification systems.

In [14], they systematically test almost 100 groups of characteristics in 4 automatic learning-based classifiers. The groups of features include variables associated with time, characteristics related to the representation of the signal using short-term Fourier transform (STFT), Wavelet transform, and shape, intensity, and texture statistics of the spectrograms from the signal. The data set is distributed in 587 LP events type and 81 VT events from the Cotopaxi volcano in Ecuador. The models were trained representing the events in vectors of 99 features, including the variables associated with the previously mentioned groups of characteristics. They emphasize the use of classifiers based on neural networks of backpropagation and the groups of

characteristics associated with shape and texture (from the spectrogram) in classifying those types of volcano-seismic events.

There are few recently published papers aiming to use the information of several stations to build classifiers of volcanic events.

In [15], two different strategies were considered to classify a database of events from the Llaima volcano and recorded by 3 monitoring stations. In the first strategy, 5 characteristics per station were combined, generating a total of 15 characteristics with which a classifier was trained. The second strategy consisted of training one classifier per station, considering 5 characteristics of the signals. Then, the classifiers were combined by simple voting to obtain a single classification of the event. They used a total of 2307 signals from the volcano corresponding to the classes LP, TR, VT and OT. They report a significant improvement in the classification performance with an approach combining information from the stations, obtaining a better performance with the first proposed strategy.

The paper [16] considers also a multi-station approach to build a warning system based on the classification of patterns of the volcanic tremor by using Self-Organizing Maps (SOM) and fuzzy clustering. The classifier in the paper aims to detect patterns typical for volcanic unrest. The system was tested at Mt. Etna (Italy) using the records of 11 permanent seismic stations located on two rings around the crater. Triggering parameters for unrest were defined for each single station and a voting system with station weights was built to provide three levels of alert. The multi-station warning system implemented proved to have good detection accuracy.

In [17] a seismic-event classification system for monitoring activity at the Piton de la Fournaise volcano observatory (OVPF, La Réunion Island) was implemented. The classifier used base on the Random Forest algorithm and consider eight classes of seismic signals. A multi-station approach is followed to select the best features for each station and combination of stations. The best performance is observed by using three-station combination.

In [18] a multi-station automatic classification system for volcanic seismic signatures is trained, that is close to our approach in this paper. The model for classification uses Transfer Learning on base of the network AlexNet. The data base used comprises 6145 events of the Lascar volcano labeled into five categories (HY, LP, TC, TR, VT). The classes are highly unbalanced and the authors conduct different experiments to include data augmentations and study the effects on the quality of the classifiers. The event are recorded in several stations, but by constructing the final database only one station for each event is selected by analysts on base of the corresponding signals and spectrograms of the event. Consequently by training the classifier the information of only one station for each event is provided to the network. Our data base of events and records is larger, but more important, our multi-station approach is also quite different, since our network (by training and classifying) gets simultaneously as input the spectrograms of several signals recorded in different stations for the same event. In particular a selection of station for each event by analysts is not required.

The summary of papers reviewed shows that approaches based on deep learning are the most used. However, it is also noted that most papers concentrate on identifying and classifying VT and LP events and use databases with few examples, so there is little variability in information for automatic learning systems, even more for deep learning systems. In addition, only recently only few papers aim to use a multi-station classification approach. In fact, those adopting a multi-station approach to classify do not use information additional to the signal or its spectrograms. Furthermore, in none of the reviewed works is the signal recorded in multiple stations combined train a classifier or to classify a single particular event.

3 Materials

The experimental setup was design using the following libraries: Python 3.9, Pandas 1.4.4, Numpy 1.20.3, Obspy 1.4.0, Scipy 1.9.1, TensorFlow 2.10, and Keras. The models were trained on the Khipu Server with 256 GB of RAM and distributed in 4 GPU's Nvidia RTX A4000. The data were made available by OVDAS through Project FONDEF ID19I10397. The stations included in this project are described in Table 1.

To manage the data, this work needed to implement a database structure that is presented in Fig. 1. The main databases are as follows: a database of a list of discrete events (DB1), and one with the continuous signal (DB2). Both contain seismic information recorded at the Nevados

Table 1 Description of the permanent seismic stations of the Nevados de Chillán Volcanic Complex

Code	Long W	Lat S	Alt. [m]	Dist. [km]	Type	Period
FRE	71.39°	36.87°	2630	1.58	Trillium	120
PLA	71.45°	36.83°	2080	8.15	Reftek	30
SHG	71.38°	36.88°	2640	2	Trillium	40
FU2	71.34°	36.90°	2599	5.5	Trillium	40
LBN	71.38°	6.85°	2658	1.32	Güralp	30
CHS	71.34°	36.87°	2466	2.68	Trillium	120

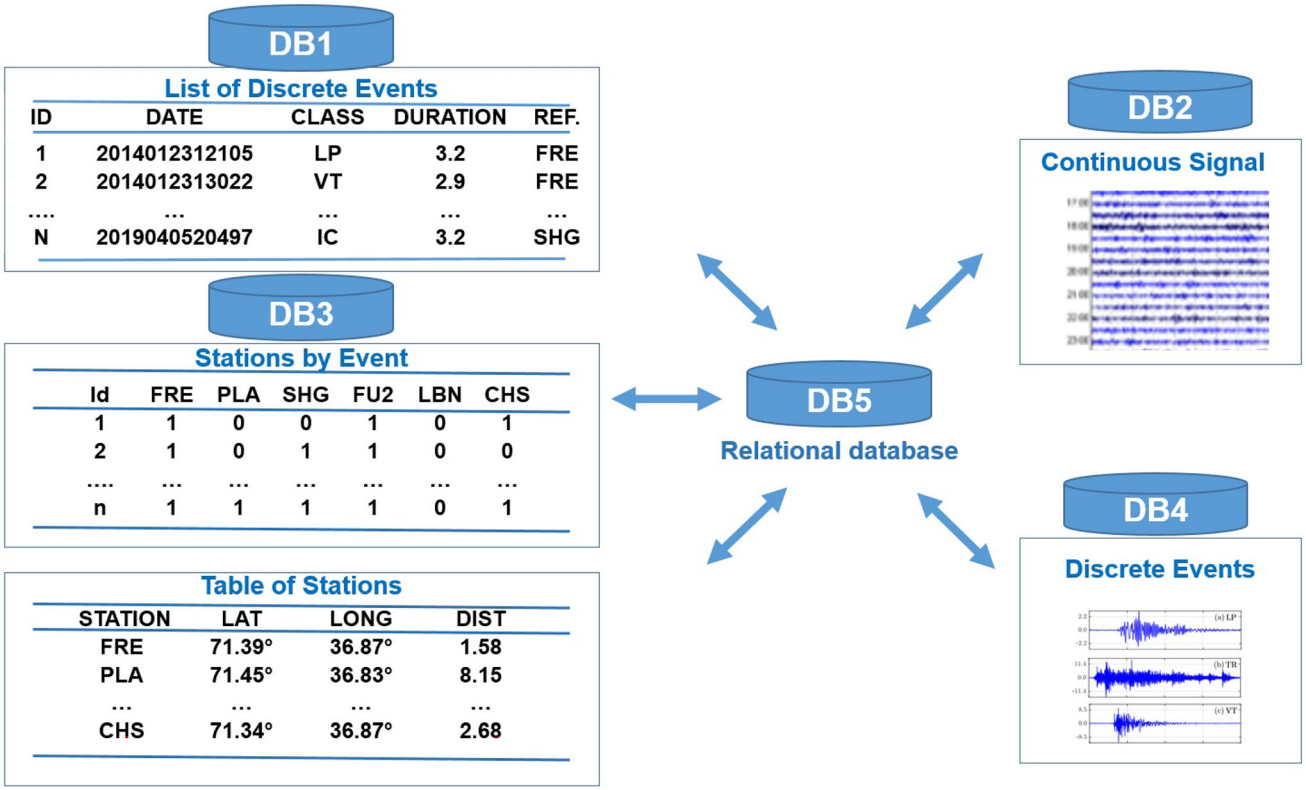

Fig. 1 Database structure. The structure is formed by descriptive lists of events as well as continuous and discrete traces of the seismic events. The descriptive lists are related to events (DB1), the stations (Table of Stations) and both (DB3)

de Chillán Volcanic Complex from January 1st, 2014 to July 31st, 2020.

DB1 is a table that contains the historical list of 504,772 events. Each event was identified and classified by OVDAS' experts and is described according to its Id, Date (beginning), Class, Duration (in seconds), and station of Reference at the moment the event occurred. The DB1 was debugged, eliminating events with null or duplicated values.

DB2 is the database (on-premise) that contains the historical continuous traces of the monitoring stations.

Using the criteria of the experts, 6 stations were selected (out of 11) according to their location and online time. This reduced the DB1 examples to 313,951, obtaining the distribution of events presented in Table 2. It should be noted that each event of DB1 is recorded by at least one station, that none of the 6 stations recorded all the BD1 events and that the maximum number of stations that recorded a particular event is 5. For this reason, the classification model proposed in this work incorporates the signal of up to 5 stations as input. The signals can come from any of the 6 stations considered.

The Table of Stations by event (DB3) was implemented, which indicates the stations where each event of debugged DB1 is recorded. Every row in this table contains the id of the event and 6 columns (one for each station) that indicate in a binary manner if it was recorded (1) or not (0) by the corresponding station.

Table 2 Number of events recorded by the selected stations. The same event can be recorded at several stations

Station	Nr. of recorded events
FRE	294,446
PLA	134
SHG	301,299
FU2	247,912
LBN	53,296
CHS	241,503

In addition, a **Table of stations** was implemented, which indicates the location of the stations. Every row represents a station and the following 3 columns indicate the latitude, the longitude and the distance to the crater (in km). This information of the stations will be included in the classification architecture, as explained in section 4. Using the name of the station as the key, this table is related to DB3.

The **Trace of Discrete events DB (DB4)** was implemented and stored locally. It is formed by the traces extracted from DB2 that contain the volcano-seismic events listed in DB1 and according to DB3 that gives the information of which of the 6 stations considered in this study recorded the event (see Table 2). Each trace of an event was stored in pickle format in directories labeled with the name of the class to which the event belongs. In the name of the pickle file, the information is included that can identify an event, followed by the station where the trace of the event was recorded.

The implementation of the DB4 database, was performed by an algorithm that executes two processes: 1) it requests (by SQL) the events in the on-premise server where DB2 is stored, identifying the stations and the corresponding segments; 2) it downloads the trace of the events and stores them in the corresponding local directories. In this way, one event has from 1 up to 5 segmented records stored in DB4. Since the stations register each event with a delay, events are segmented considering one second before and one second after the beginning and the end defined by the experts.

Finally, a **Relational Database (DB5)** was implemented to make it easier to read and process the data. DB5 relates the events and the stations to segment the events stored locally in the Continuous Trace **(DB4)**, using as the primary key the id of the event of **(DB1)** and the Table of Stations by event **(DB3)**. In this way, different subsamples of the total events can be easily generated. Moreover, only the event traces of the subsamples need to be loaded into memory, and can be discarded and reloaded when they are required again; either for training or testing the models in the different experiments. Using this procedure a full loading of all the data into memory can be avoided.

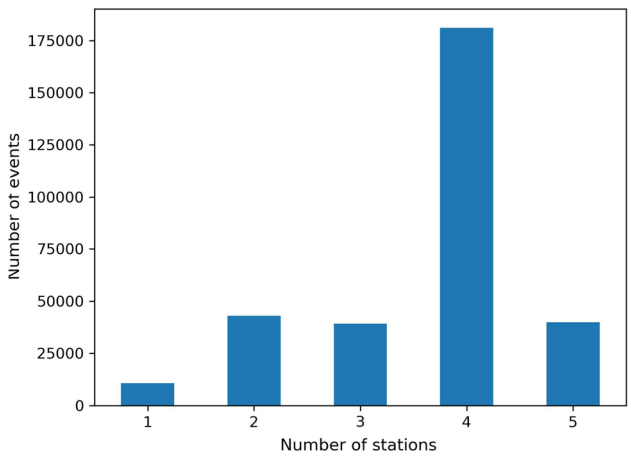

Fig. 2 Distribution of the number of events with record in number of stations. Most events were recorded in 4 stations, whereas few were recorded in only one station

Next, in Fig. 2, the distribution of the number of events with the record in 1, 2, and up to 5 monitoring stations is shown in the chart.

The distribution of the events by classes is shown in Table 3.

4 Methodology

The methodology proposed in this work for the classification of seismic-volcanic events uses a deep convolutional network composed of 5 inputs that processes 5 different records of the same event obtained from up to 5 stations. The experiments allow to evaluate different architectures and additional information that can feed the network. Also, in one of the experiments (Experiment 1) the traditional backpropagation algorithm is simplified, in order to reduce the time required to adjust the network.

The CRISP-DM methodology [19] was used, including in some steps the good practices for adjusting hyperparameters and model assessment, although the deployment phase was not implemented. For model training, in the modeling phase, 5-fold cross-validation was used [20]. Due to the imbalance of the amount of data in the classes (see Table 3), a method was proposed to define the training, validation, and test sets. This procedure is described next.

4.1 Method to generate the training, validation and test sets

A random sub-sample was taken from 10% of the available data for the test set. The training set was implemented from the 90% remaining data. It is worth noting that the data in DB1 is significantly unbalanced regarding the stations (see Fig. 2 and the classes (see Table 3). Thus, these distributions are replicated in the test and training sets. However,

Table 3 Distribution of the number of events per class

Class	Number of events	Proportion (%)
LP	194,288	61.87
TR	64,367	20.50
EX	33,703	10.73
VT	13,570	4.32
LV	3086	0.98
IC	1916	0.61
AV	1850	0.59
ZZ	1171	0.38
TOTAL	**313,951**	**100**

Total numbers indicate the bold letters

for training it is important to balance the classes, so the following procedure was applied:

Procedure 1: Creation of the training set.

1. All the events of the classes with the fewest number of examples (ZZ, LV, IC, and AV) were included in the training set.
2. A random sub-sample of the data from the classes with the highest number of events (LP, TR, EX and VT) was performed, forming a data set where these classes reach a number not superior to 3 times the ZZ class, which is the one with the fewest number of events.

A description of the distribution of the training data, according to the stations and the number of events per class, is shown in Fig. 3 and Table 5, respectively.

The models were adjusted using 5-fold cross-validation. Then, considering the imbalance of the data, the following procedure was applied to define the folds.

Procedure 2: Definition of the folds.

1. Separate the training set by class (see Table 5), obtaining 8 groups of data.
2. Subdivide each group into five folds with an equal number of events.
3. For $K = 1$:
   (a) The validation fold is formed by the first fold of each group.
   (b) The training set is formed by the 4 remaining folds of each group. For the ZZ class, replicate each event 3 times and include all them in the training set. For the intermediate classes (LV, IC, and AV), replicate each event twice, get a random sample of the duplicated events and include them in the training set until obtaining sets of up to 3 times the minority class. Finally, include the entire groups that contain the events of the majority classes (LP, TR, EX, and VT).
4. For $K = 2, 3, 4, 5$:
   1. For the validation set, take the *kth* fold of each group.
   2. For the training set, take the 4 remaining folds of each groups and repeat the procedure in step 3 letter b.

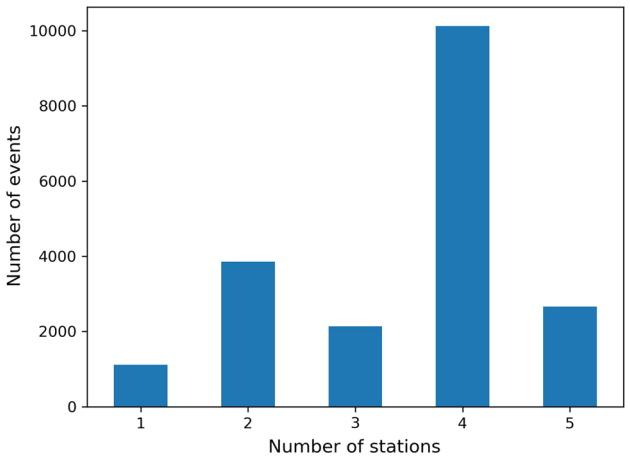

Fig. 3 Distribution of the number of events recorded in a number of stations for the training set. It is observed that most of the events were recorded in 4 stations, whereas few events were recorded in only one station, the same as the original set

4.2 Resulting training, validation and test sets

The test set contains 31389 events. The distribution of the events in this set, in terms of the number of stations, remains the same as the full set of events (see Fig. 2). The distribution of the events per class in the test set is presented in Table 4.

The distribution of the number of stations that recorded an event in the training set is shown in Fig. 3. The distribution of the classes obtained applying **Procedures 1 and 2** can be seen in Table 5.

So the models were trained and validated with 19,890 different events from the Nevados de Chillán volcanic complex (see Table 5). In addition, after applying the data augmentation method described in **Procedure 2**, the training folds contain approximately 20,224 balanced events (see Table 6 for $K = 1$) while the validation folds are of approximately 3978 events (see Table 7 for $K = 1$).

4.3 Signal preprocessing

The following preprocessing was applied to each event of the training, validation, and test sets, to obtain the spectrograms:

Table 4 Distribution of the number of events per class for the test set

Class	Number of events	Proportion (%)
LP	19,314	61.53
TR	6517	20.76
EX	3423	10.90
VT	1366	4.35
LV	291	0.93
IC	185	0.59
AV	175	0.56
ZZ	118	0.37
TOTAL	**31,389**	**100**

Total numbers indicate the bold letters

Table 5 Distribution of the number of events per class for the training set, after the class balancing procedures

Class	Number of events	Proportion (%)
LP	3159	15.88
TR	3159	15.88
EX	3159	15.88
VT	3159	15.88
LV	2795	14.05
IC	1741	8.75
AV	1665	8.37
ZZ	1053	5.3
TOTAL	**19,890**	**100**

Total numbers indicate the bold letters

Table 6 Distribution of events per class after applying procedure 2. The table shows the training set for the 4 remaining folds ($K = 1$)

Class	Number of events	Proportion (%)
LP	2528	12.5
TR	2528	12.5
EX	2528	12.5
VT	2528	12.5
LV	2528	12.5
IC	2528	12.5
AV	2528	12.5
ZZ	2528	12.5
TOTAL	**20,224**	**100**

Total numbers indicate the bold letters

Table 7 Distribution of events per class after applying procedure 2. The table shows the validation set for the first fold ($K = 1$)

Class	Number of events	Proportion (%)
LP	632	15.88
TR	632	15.88
EX	632	15.88
VT	632	15.88
LV	559	14.05
IC	348	8.75
AV	333	8.37
ZZ	210	5.3
TOTAL	**3978**	**100**

Total numbers indicate the bold letters

1. 5th-order Butterworth passband filter, with a cutoff frequency between 0.5 and 10Hz to reduce the noise in the signal.
2. Calculation of the spectrogram, with a window of 100 samples, 90% overlap, and resolution of 1024 frequency samples.
3. Normalization of the spectrograms and adjustment of the color map to emphasize the shape of the spectrogram of each event, applying the preprocessing method suggested in [1].
4. Resize the spectrogram to fit the network input of $112 \times 112 \times 3$, where each of the 3 channels represents an RGB component. This implies a normalization in time since regardless of its duration, time is represented in 112 pixels.

4.4 Description of classification models

For this work, three architectures or models based on deep learning were proposed. The first two are the Base and the Intermediate classifier models. In addition to being multi-station, these models incorporate additional information about the signal in their intermediate layers. On the other hand, the third model (Single input model) is not multi-station.

Both, the base model and the intermediate classifier model are neural networks with 5 processing tunnels or inputs. Each tunnel processes a spectrogram. The tunnels are fed at the same time with the spectrograms of the same event recorded at different stations, so the network has a multi-station view of each event. The order of the stations at the input is not relevant, since there is no tunnel-station specialization. Some tunnels can receive the null vector as input for those events recorded in less than five stations.

The output of the 3 models is an 8-dimensional vector representing the probability that the spectrograms of an event is from some of the 8 volcano-seismic event classes considered here. For this, the models were trained by minimizing the categorical cross-entropy loss function [21] and using Adam's optimization method [22] in the 3 cases. A learning rate with a value equal to 0.01 was chosen, which was the best value noted in the range proposed in [23], using the values 0.1, 0.01, and 0.001. Next, the three models are described.

4.4.1 Base model (M1)

A diagram of this network model is shown in Fig. 5. A convolutional neural network was used, with 5 input tunnels, each of one is formed by 4 blocks composed of a separable Depthwise convolution layer [24] with ReLU activation followed by a max-pooling layer. The size of the kernel used in the depthwise convolution was 3×3 and stride equal to 1, and the size of the pool in the max-pooling operations was 2×2. In addition, 16, 32, 64, and

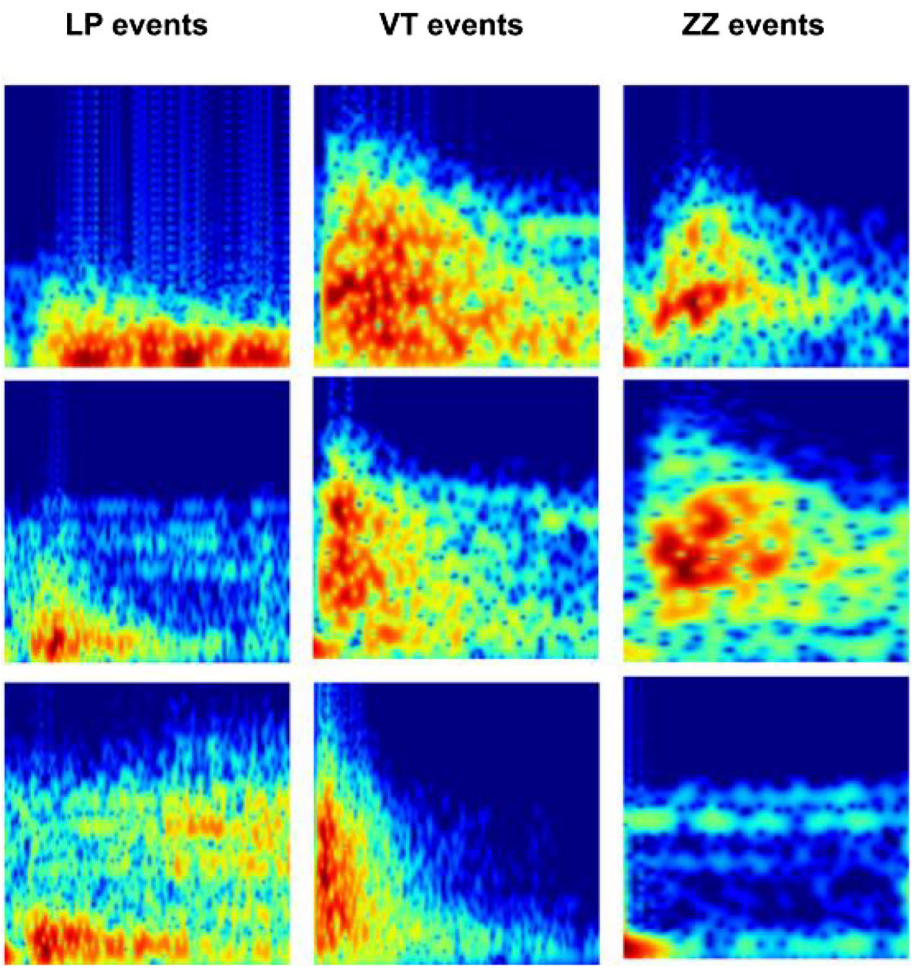

Fig. 4 Spectrograms generated from the signal preprocessing steps mentioned above. From left to right, 3 events are observed for the LP, VT and ZZ classes

128 feature maps were generated in each convolution layer. After these 4 blocks, a flatten operation was added to reconstruct the resulting tensor $7 \times 7 \times 128$ in size in a vector 6272 in size. This operation was followed by two fully connected layers with 200 nodes and sigmoid activation, to then add a concatenation layer to group or combine the information that comes from the 5 inputs. It should be noted that, in this layer, the station-associated additional information is added. This station-related information contains the latitude, the longitude and the distance to the crater (in km) and is recovered from the Table of stations (see Fig. 5). Finally, 2 fully connected layers with 400 nodes and sigmoid activation are added, where the additional event-related time information is added at the output of the second layer (see Fig. 5). The time information added is the duration of the event in seconds and is recovered from table DB1. The last layer of the network is a softmax classifier with 8 nodes. Additionally, for the training of the network, a dropout operation (0.4) is added to each of the completely connected layers. This way, the total number of parameters of this model is 7,103,583.

4.4.2 Intermediate classifier model (M2)

A diagram of this network is presented in Fig. 6. This model comprises an architecture similar to the previous one but includes an intermediate classifier before the concatenation layer. Thus, this model classifies each of the 5 input spectrograms, considering 8 classes, and then combines the classifications to obtain the final classification (single) associated with the event. The additional station-related information is entered at the intermediate classifier that reduces the information from the signal to 8-dimensional probability vectors (see Fig. 6). The duration of the event is entered into the network after the second dense layer (see Fig. 6).

This model is proposed under the hypothesis that obtaining an intermediate classification from each spectrogram, that is, having a classifier for each tunnel, may improve the class estimation.

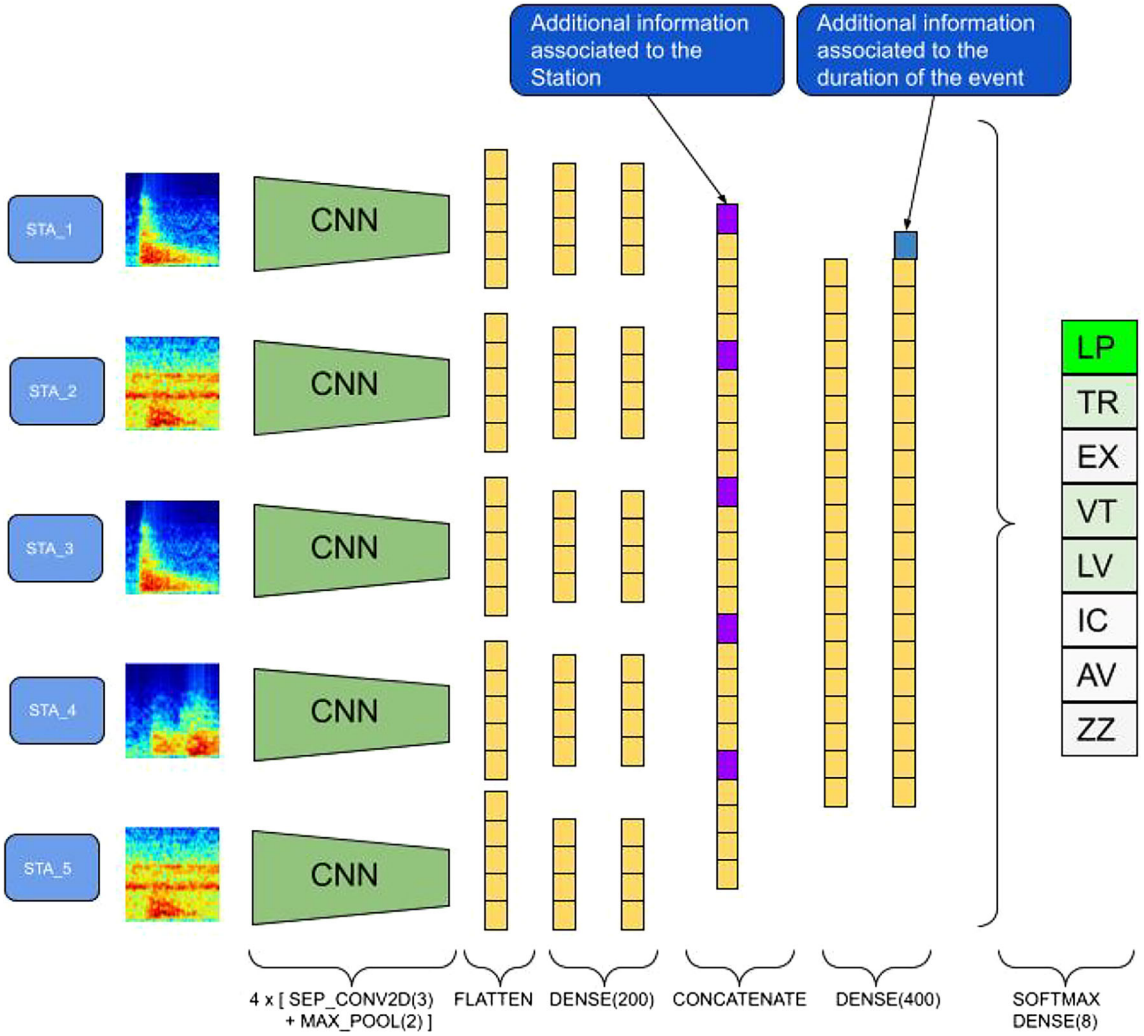

Fig. 5 Multi-station Base Model (M1). Five spectrogram inputs for the multi-station classification are considered. The M1 architecture includes additional information to the spectrograms in the deep layers of the network: information of the station and information on the duration of the event

4.4.3 Single input model (M3)

This model uses a unique CNN structure (single tunnel) and does not add additional information from the stations nor the duration of the event.

As shown in Fig. 7, the M3 model classifies a volcano-seismic event considering its record in a single monitoring station. This model is a traditional classification approach, without considering a multi-station approach, and is used for comparative purposes.

4.5 Description of the experiments

Four experiments were performed using the models defined in the previous sub-section. The M1 and M2 models were used in experiments 1 and 2 whereas only the M1 model was used in experiment 3 and the M3 model in experiment 4.

4.5.1 Experiment 1

In this experiment, a simplified training for models M1 and M2 is proposed: in the backpropagation step [25], instead of updating all the model weights, the weights of the CNNs structures of the STA_2, STA_3, STA_4 and STA_5 inputs are frozen, including the blocks SEP_CONV2D(3)-MAX_POOL(2), FLATTEN and the DENSE(200) layers (see Figs. 5 and 6), and only the weights of the STA_1 CNN input are updated, as well as the weights associated with the fully connected layers (MLP) that follow the concatenation layer (CONCATENATE in Figs. 5 and 6). In this way, the total of parameters (or weights) that are updated at each training step with the backpropagation

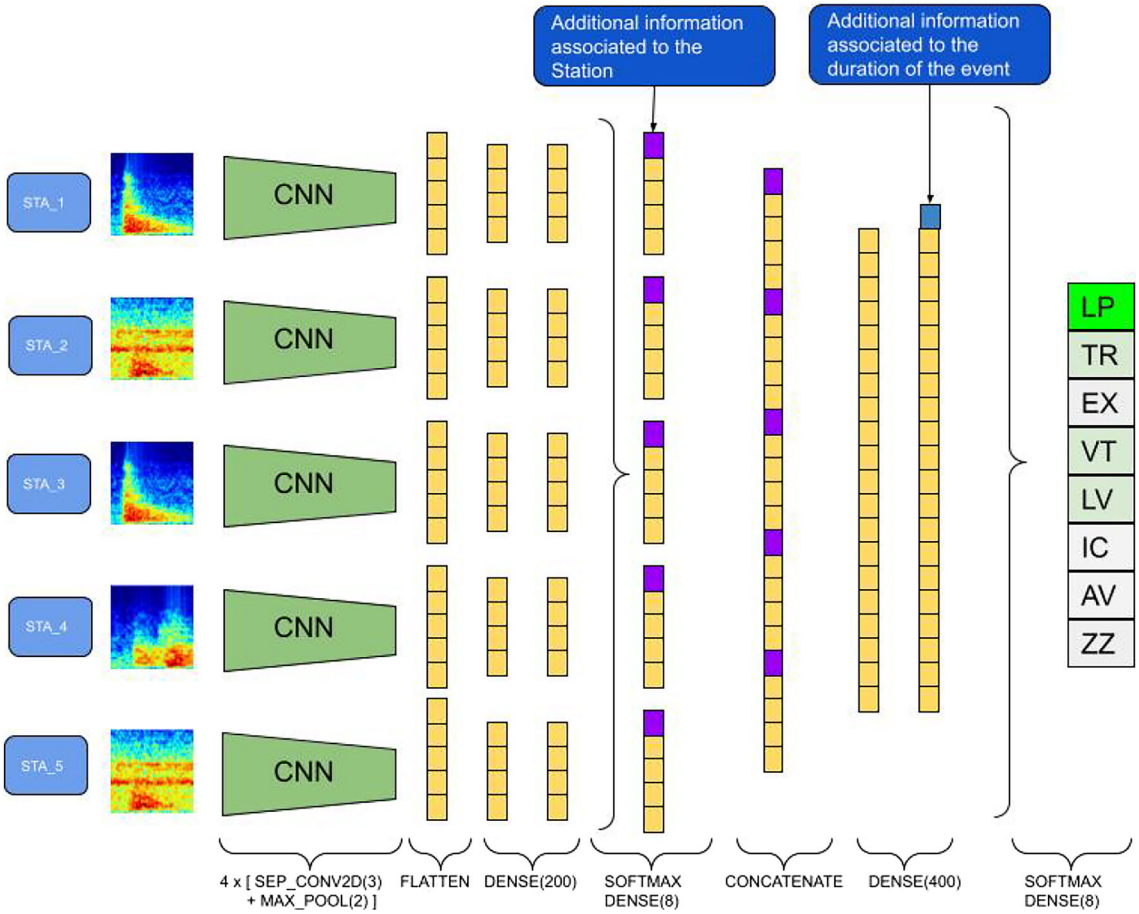

Fig. 6 Multi-station Intermediate Model (M2). An intermediate classifier is included in the SOFTMAX-DENSE(8) layer. After reducing the inputs to the 8-dimensional probability vectors (intermediate classification), the station-related information is added. Then, the information from the 5 inputs is concatenated and a single classification per event is obtained. The event-related information (duration) is entered into the dense layer, before the final classification

algorithm is reduced to 1,877,043, representing 26.4% of the original total of parameters. A schematic summarizing the training steps of this proposed experiment is shown in Fig. 8.

As shown in Fig. 8, in the forward step (Step 1), the information flows through all the input structures (the 5 CNN tunnels are considered), then this information is condensed in the concatenation layer (intermediate layer, yellow and brown vector) and prediction is obtained through MLP. At the end of step 1, the loss $L(\hat{y}, y)$ is calculated by comparing the model prediction, $\hat{y}$, with the expected output, y. Then, in Step 2, the weights associated with the STA_2 to STA_5 CNN structures are frozen (Step 2.1) and the loss is backpropagated only through the weights of the MLP and the STA_1 CNN structure (Step 2.2), obtaining MLP* and CNN*, the updated weights structures. In Step 3, the weights of the remaining 4 CNN structures are updated, copying the same weights obtained for CNN* during Step 2. Then, in Step 4, all CNN structures are identical to CNN*, to finally, in Step 5, repeat Steps 1 to 4, until completing one training epoch.

The model is trained in a batch of N events. As the events have at least one station record, during the training it is prioritized that the recorded signals are entered in the first processing tunnels (STA_1, STA_2, etc.), thus there is no tunnel-station specialization. For events recorded in less than 5 stations, the missing records are replaced by adding the null vector (tensor of dimension $112 \times 112 \times 3$ with components equal to 0) in the corresponding entries (see Fig. 8). In this sense, adding the null vectors to complete the 5 stations records has the effect of "turning off" the CNNs structures in which there is no record, obtaining a more robust model that learns to classify seismic-volcanic events recorded from 1 to 5 stations.

4.5.2 Experiment 2

In this experiment, the M1 and M2 models are trained by updating all the weights with the backpropagation method,

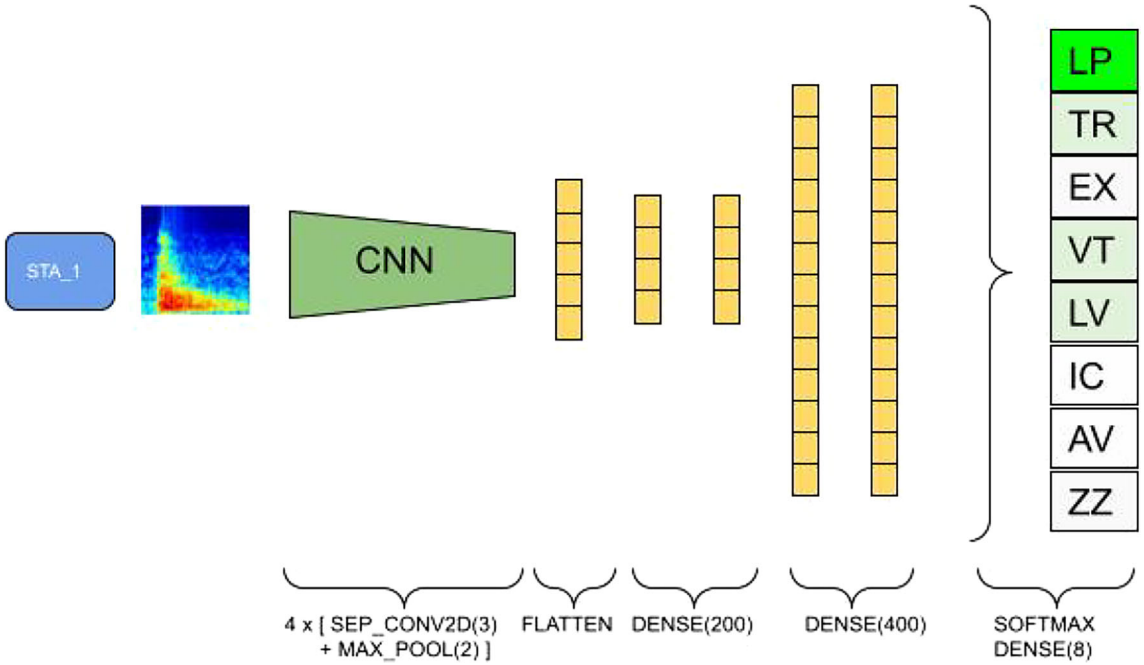

Fig. 7 Single-station Model (M3). This model is composed of a single input. The CNN bloc is identical to one of the CNN blocks of the base model, but only one spectrogram is processed and no additional information is provided

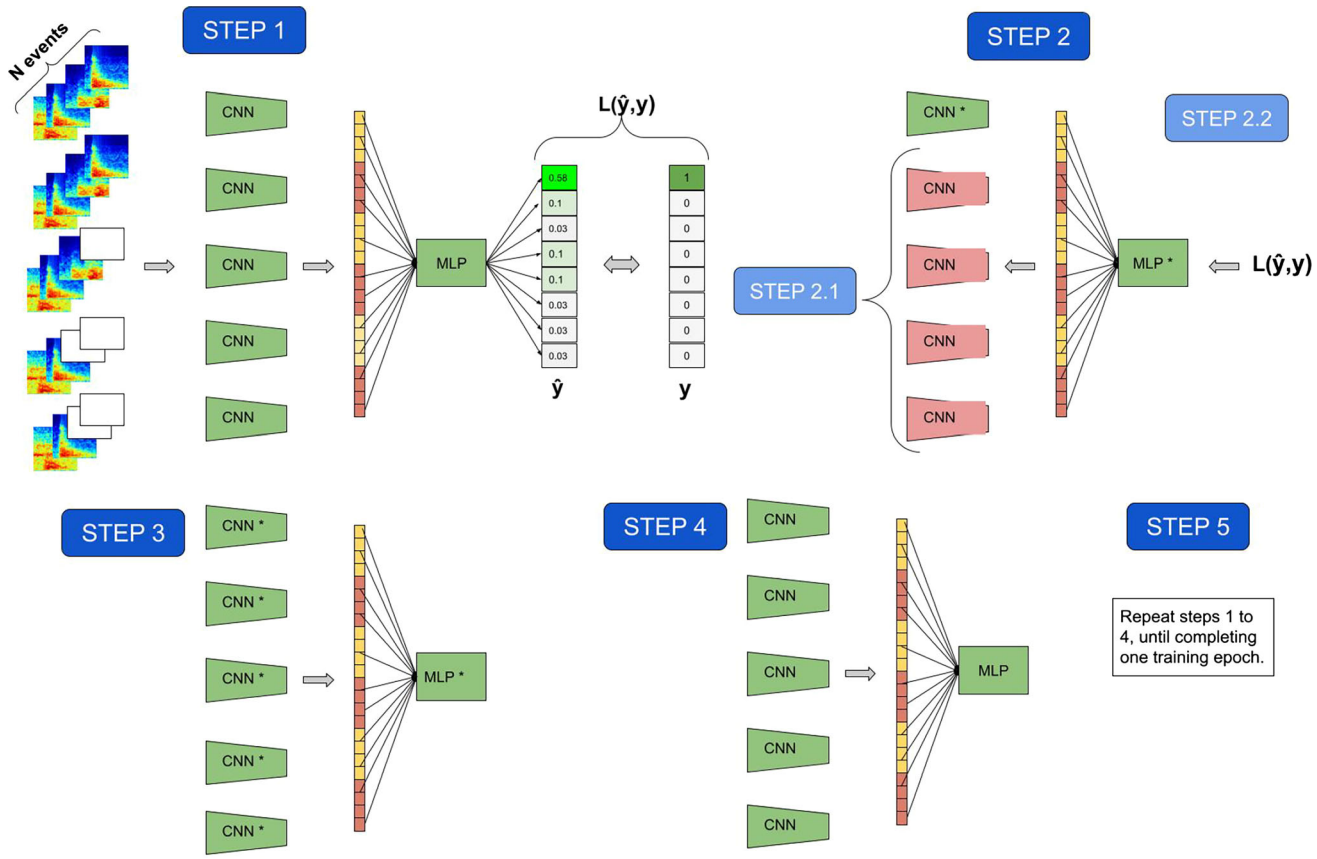

Fig. 8 Summary of steps considered for simplified training. The CNN and CNN* structures are identical in architecture, but have different weights (or parameters).The same happens with MLP and MLP*

Experiment	Models	Input	Additional information (station and duration)	Weights update strategy
1	M1 M2	Multi-station	Yes	STA_1 CNN and MLP update and duplicate the weights in all the stations
2	M1 M2	Multi-station	Yes	ALL
3	M1	Multi-station	**No**	STA_1 CNN and MLP update and duplicate the weights in all the stations
4	M3	**Single-station**	**No**	STA_1 CNN

Fig. 9 Summary of the experiments performed to evaluate the multi-stations models and the design strategies, as well as to compare with the traditional single-station model

from the CNNs structures to the MLP structure. In this sense, the training algorithm is the same than the one used to train any deep neural network [26].

4.5.3 Experiment 3

In this experiment, the additional information associated with the station and the duration of the event is removed from the M1 model, in the layers in which this information was entered into the network. The training algorithm is the same used in Experiment 1.

4.5.4 Experiment 4

In this experiment the model M3 is considered. Given the architecture of this model, the training algorithm is the same used to train any neural network. Regarding the training data, we considered the same events used to train the models proposed in the 3 previous experiments. Because events are recorded at up to 5 stations, only one of the station was considered for each event. In general the signals of the events registered in the FRE station were used, since the data belong to the period in which this was the reference station defined by the OVDAS. For the events that did not have a record in the FRE station (around 6% of the data), they were recorded in the SHG station, since after FRE, SGH was defined as the reference station. Finally, only 0.02% of the events had no record in either FRE or SHG. For these events, the recording was considered in the only station that recorded them, which were the LBN and FU2 stations.

Next Fig. 9 shows the main characteristics considered in each experiment to evaluate the multi-station approaches, the contribution of the additional information and the efficiency of the weights update strategies.

4.6 Performance metrics

To evaluate the different models, the following metrics are defined for binary classification:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} \tag{1}$$

$$\text{Sensibility} = \frac{\text{TP}}{\text{FN} + \text{TP}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

where TP is the number of true positive classifications, TN is the number of true negative classifications, FP is the number of false positive classifications and FN is the number false negative classifications of confusion matrix. The following identities can then be defined from the confusion matrix:

$$\text{AP} = \text{FN} + \text{TP}, \quad \text{AN} = \text{TN} + \text{FP} \tag{4}$$

$$\text{PP} = \text{TP} + \text{FP}, \quad \text{PN} = \text{FN} + \text{TN} \tag{5}$$

where AP are all the positive examples, AN are all the negative examples, PP are all the examples classified as positive and PN all the examples classified as negative. Then, the $F1$ score is defined as:

$$F1 = \frac{2 \times \text{TP}}{\text{AP} + \text{PP}} \tag{6}$$

All the metrics mentioned so far can be extended to multi-class classification problems, applying a one versus all strategy. Thus, to perform the metrics calculation for each class, the procedure is: (1) consider class c as the positive class and the sum of the events of the rest of the classes as the negative class; (2) calculate the Accuracy, Sensitivity, Specificity and F1-score, (3) perform this for all $k \in K$, where K is the number of classes considered. (4) finally, the values obtained for each class are averaged, in all the corresponding metrics, as performed in [1, 2, 5, 19].

Table 8 Performance metrics of the different models during the training process

Model	Accuracy	Sensibility	Specificity	F1	Kappa	MCC
M1-Exp1	**0.9386**	**0.7293**	**0.9646**	**0.7259**	**0.7156**	**0.7160**
	±0.001	±0.004	±0.0006	±0.004	±0.005	±0.005
M2-Exp1	0.9279	0.6922	0.9586	0.6829	0.6667	0.6688
	±0.01	±0.034	±0.006	±0.044	±0.05	±0.048
M1-Exp2	0.9377	0.7236	0.9641	0.7214	0.7113	0.7118
	±0.001	±0.006	±0.0009	±0.006	±0.007	±0.007
M2-Exp2	0.9331	0.7084	0.9615	0.704	0.6901	0.6906
	±0.002	±0.01	±0.0013	±0.01	±0.011	±0.011
M1-Exp3	0.9368	0.7223	0.9636	0.7174	0.7072	0.7077
	±0.001	±0.006	±0.0006	±0.006	±0.005	±0.005
M3-Exp4	0.9205	0.6414	0.9540	0.6464	0.6305	0.6316
	±0.001	±0.007	±0.0009	±0.006	±0.007	±0.007

Best performance for each metric indicates the bold letters

Table 9 Performance metrics of the different models in the test set

Model	Accuracy	Sensibility	Specificity	F1	Kappa	MCC
M1-Exp1	0.9221	**0.7367**	0.9519	0.4902	0.5356	0.5591
	±0.004	±0.004	±0.002	±0.014	±0.016	±0.012
M2-Exp1	0.9086	0.6998	0.9424	0.4599	0.476	0.4978
	±0.034	**±0.037**	**±0.018**	**±0.065**	**±0.136**	**±0.122**
M1-Exp2	**0.9274**	0.7326	**0.9531**	**0.4992**	**0.5539**	**0.570**
	±0.004	±0.004	±0.002	±0.008	±0.019	±0.014
M2-Exp2	0.9177	0.7176	0.9484	0.4736	0.5099	0.5322
	±0.004	±0.008	±0.0013	±0.013	±0.015	±0.01
M1-Exp3	0.919	0.7309	0.950	0.477	0.5198	0.544
	±0.003	±0.003	±0.0009	±0.013	±0.01	±0.007
M3-Exp4	0.9137	0.6501	0.9465	0.4838	0.4911	0.519
	±0.008	±0.014	±0.004	±0.015	±0.035	±0.029

Best performance for each metric indicates the bold letters

Now, as in [27], the Matthews Correlation Coefficient (*MCC*) [28] and the Cohen's Kappa (*K*) coefficient for multi-class classification are also calculated. They are obtained by the following equations:

$$\mathrm{MCC} = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2)(s^2 - \sum_k^K t_k^2)}} \tag{7}$$

$$K = \frac{c \times s - \sum_k^K p_k \times t_k}{s^2 - \sum_k^K p_k \times t_k} \tag{8}$$

where

$$c = \sum_{k \in K} C_{kk} \text{ is the total number of items correctly predicted.}$$
$$s = \sum_{i \in K} \sum_{j \in K} C_{ij} \text{ is the total number of elements.}$$
$$p_k = \sum_{i \in K} C_{ik} \text{ is the number of times class k was predicted.}$$
$$t_k = \sum_{i \in K} C_{ki} \text{ is the number of times class k actually occurred.} \tag{9}$$

With C the confusion matrix that in C_{ij} indicates the number of data belonging to class $i \in K$ and that were classified as class $j \in K$.

Table 10 Confusion matrix in the test set

	LP	TR	EX	VT	LV	IC	AV	ZZ
LP	**12,107**	2007	2317	1289	593	176	361	464
TR	365	**5073**	836	0	65	0	147	31
EX	294	619	**2494**	0	1	0	10	5
VT	59	1	0	**1156**	4	77	14	55
LV	6	5	0	1	**276**	0	2	1
IC	0	0	0	12	0	**157**	4	2
AV	3	10	2	9	6	13	**118**	24
ZZ	8	2	0	19	2	17	17	**53**

Number of correct predictions indicate the bold letters

5 Results

In this section, the first part presents the performance metrics of the models, for all the experiments, for both, training and testing. Then, a discussion and analysis is carried out focused on the best model obtained. This model is a multi-station classifier that incorporates additional information of the signal.

5.1 Results of the experiments

Table 8 shows the performance of the models on the training set, for all the experiments. The result of each metric in Table 8 is the average of the 5 cross-validation folds.

Table 9 shows the performance of the models obtained in all the experiments for the test set (described in Table 4). As for the training set, the metrics were calculated as the average of the performance of the 5 models, obtained in the 5-folds cross-validation process.

5.2 Discussion

Table 9 illustrates that the base model M1-Exp2 obtains the highest result in all the metrics, except Sensitivity, where M1-Exp1 (same architectures) obtains the highest value. Although M1-Exp1 comes second in the remaining metrics, the difference is not significant. In this context, it is worth noting that the models adjusted with Experiment 1 require half the time than those adjusted with Experiment 2, which implies a significant advantage in optimizing resources when training neural networks.

On the other hand, M1-Exp1 significantly surpasses M2-Exp1 and M2-Exp2. In addition, it was observed that M2-Exp1 is somewhat unstable for the classification of volcano-seismic events. This is noted in Table 9, since, for this model, a much higher standard deviation is obtained than that of the rest of the models, in every metric. This instability may be related to the design itself of M2, since this overly reduces the amount of information that comes from the 5 inputs before combining the information. Consequently, the MLP structure that follows the concatenation layer does not manage to learn the patterns necessary to classify this type of data systematically. In addition, in some cases, M2 suffered modal collapse during the training.

Tables 8 and 9 present a better overall performance of M1-Exp1 compared to M1-Exp3. However, this difference is not significant, so the additional information associated with stations and the duration of the event is not a great contribution to the classification.

Table 11 Performance metrics by class in the test set

Class	Accuracy	Sensibility	Specificity
LP	0.7469	0.6268	0.9391
TR	0.8997	0.7784	0.8936
EX	0.8698	0.7286	0.8871
VT	0.9509	0.8462	0.9557
LV	0.9781	**0.9484**	0.9784
IC	**0.9904**	0.8971	**0.9909**
AV	0.9802	0.6378	0.9822
ZZ	0.9793	0.4491	0.9813

Best performance for each metric indicates the bold letters

Finally, comparing the results obtained in M1-Exp1 and M3-Exp4, which uses the record from only one station (see Tables 8 and 9), a significant performance advantage in favor of M1-Exp1 is noted in all the metrics. This result clearly indicates that the multi-station strategy is advantageous over including the signal from only one station, even if this is the reference station and that the proposed design takes advantage of this improvement.

Considering this, the best classifier proposal of those discussed in this work is the multi-station one associated with the Base Model (M1) obtained from Experiment 1 (M1-Exp1). In addition to processing the signal of an event recorded in 5 monitoring stations, this model incorporates information in addition to the signal, such as the duration of the events and the coordinates of the stations.

Next, Table 10 presents the confusion matrix obtained by M1-Exp1 from the $K = 1$ cross-validation step of the test set.

It is observed that the M1-Exp1 model generally correctly labels 21,434 (diagonal of the matrix) of the 31,389 events. However, most errors in the labeling occur for the LP class, which is the most numerous class in the test set, with 61.53% of the data (see in Table 4).

Most of the works found in the literature and reviewed focus on the classes LP and VT. In this sense, achieving the separation of these classes is relevant, which is why we will concentrate on analyzing the confusion matrix for these two classes.

In particular, the model classifies 12,107 LP events well, out of the total of 19,314, representing 62.68% of the total events in the class. Of these, 2007 events (10.4%) are confused with TR events, 2317 events (11.9%) are confused with EX events and 1289 events (6.6%) are confused with VT events. In addition, 176 LP events (0.9%) are labeled as IC events, 361 (1.6%) as AV, and 464 (2.4%) as ZZ. By contrast, for the VT events, the model classifies 1156 out of 1366 well, representing 84.62% of the total events in the class. From this last amount, 59 events (4.3%)

are confused with LP events, and 1 event (0.07%) is confused with TR events. In addition, there are 4 VT events (0.3%) confused with LV events, 77 events (5.6%) confused with IC events, and 55 events (4%) confused with ZZ events. It is noted that there is an excellent performance in the classification of VT events; however, with respect to the LP class, a significant percentage of events are labeled as TR, EX and VT (near 10%). Experts at the OVDAS have corroborated that this confusion of classes is common in volcano-seismic event classifiers. In terms of future research, it remains to study a method that can improve the classification of LP events.

If these results are contrasted with those observed in the literature, a much higher accuracy value is observed (0.9207) than in [3], where they reached a value of 0.7522, also in a test set, with a single station approach. It should be indicated however that in [3] they only use 2193 volcano-seismic events to perform the training and test versus this work where 19,890 different events were used for the training, and 31389 events to test the model. In addition, in [3] only 4 classes are included compared to this work, which includes 8.

Additionally, Table 8 shows that a kappa value of 0.7156 $\pm$0.005 is obtained in the training of the model, whereas in [2] it is 0.662 $\pm$0.005. Therefore, it is deduced that this value is obtained in the training. Again, in [2], they only use 532 items of data for all the work and they only consider 4 classes.

Table 11 shows the metrics accuracy, sensitivity and specificity of the base model (M1-Exp1) for each class of the test set.

For the accuracy and specificity metrics, the highest values are obtained for the IC class with values of 0.9904 and 0.9909. For the sensitivity metric, the highest value is obtained for the LV class, indicating that the model manages to recognize 94.84% of LV events.

For the LP class, the value of the sensitivity metric tells us that the probability of detecting an LP event, given that the event is a true LP, is 0.6268 and the specificity metric tells us that the negative events for this class are predicted with 93.91% accuracy. The model behaves better for the remaining classes than for the LP class, except for the ZZ class. For this last class, the value of the sensitivity metric is 0.4491, which is why the model only manages to recognize 44.91% of the events in this class. This may be due to this class being less represented in the dataset used for this work and, being mainly noise, its spectra have great variability, which makes it difficult to learn the patterns, in addition to the small class set.

For the VT class, the model manages to correctly predict 84.62% of the events in this class, observing the sensitivity metric. Moreover, for the negative events in this class, the model manages to predict 95.57% of these correctly, now observing the specificity metric.

We consider that the good results of the proposal are due to the multi-station approach, where the information of many stations is merged and processed simultaneously, the same way that analysts do. As expected, the resulting model presented a better performance when compared to the single-station approach. However, the model is computationally expensive because there is an input for each station. Therefore, the model presents a limitation for the number of stations considered, as an increase in this number will increase the model complexity, reaching a limit that must be defined, despite the efficient training strategies that were proposed.

6 Conclusions

The main conclusions referring to the models and the experiments conducted in this work appear next.

The results obtained in this work confirmed that the multi-station method, which takes into account the signal of an event recorded up to five stations, is substantially more effective than a single-station strategy for classifying volcano-seismic events. In all the metrics considered here, the difference is significant. This is evidenced by comparing the results obtained by the base model M1 in Experiment 1, from simplified training, and the results obtained by the single input model M3 obtained from Experiment 4. Our result implies that the information about the event observed from a single station may be insufficient for automatic volcano monitoring systems to achieve truly high performances in the classification of volcano-seismic events. Although this result was obtained in previous works [15], it is maintained when applying deep neural models. Therefore, considering the combination of observed signals from a single event in different stations can lead to the required performance so that the automatic classification systems are more robust and reliable in the classification stage in volcano monitoring, one of the most important stages.

It is also concluded that incorporating additional information into the signal improves the result in the metrics, but not significantly. This is observed by comparing the results of the model M1 obtained in Experiment 1 and the same obtained in Experiment 3, with no additional information. This conclusion is consistent with the results in the same line reported in [14].

With respect to the training efforts, it is worth noting that the time in training the models M1 and M2 was reduced to half, comparing the training time of these models in Experiment 1, with simplified training, versus the training time in Experiment 2, where all the weights of

the models were adjusted with the original backpropagation. This difference represents a substantial improvement in the efficiency and optimization of the computer resources available for training these types of neural networks. Also, when comparing the results of these two experiments, particularly observing the number of times the models converge in the training, it can be concluded that the training algorithm suggested for Experiment 1 can reduce the overfitting of models without affecting performance of these (see Tables 8 and 9).

Finally, as to what was observed from the reviewed literature, this is a novel methodological proposal and could be the first to offer a multi-station classification approach based on a single neural network that combines the information of a "seen" event from 5 monitoring stations (or fewer). The results obtained from the metrics included here are similar (and in some cases superior) to those observed in the reviewed literature. However, in none of the reviewed works are 8 classes considered (most only consider 2 or 4 classes), nor are the models tested with such a large amount of data as in this work (31389 items of data in the test set).

Acknowledgements We thank OVDAS and the FONDEF ID19I10397 Project for having the data used in this work. In addition, thanks to the Department of Mathematical Engineering of the Universidad de La Frontera for having the Khipu Server to perform the computation and training of the models.

Data availability The data that support the results of this study and model M1-Exp1 are available on request from the repository https://drive.google.com/drive/folders/1Jjq0p4TZeT2vLSD1vsGOJlM5Pz7SMOr8?usp=sharing.

Declarations

Conflict of interest The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Curilem M, Canário JP, Franco L, Rios RA (2018) Using cnn to classify spectrograms of seismic events from llaima volcano (chile). In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. https://doi.org/10.1109/IJCNN.2018.8489285
2. Mou L, Jin Z (2018) Tree-based Convolutional Neural Networks: Principles and Applications. Springer, Singapore
3. Titos M, Bueno A, García L, Benítez MC, Ibañez J (2019) Detection and classification of continuous volcano-seismic signals with recurrent neural networks. IEEE Trans Geosci Remote Sens 57(4):1936–1948. https://doi.org/10.1109/TGRS.2018.2870202
4. Venegas P, Pèrez N, Benítez DS, Lara-Cueva R, Ruiz M (2019) Building machine learning models for long-period and volcano-tectonic event classification. In: 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), pp. 1–6. https://doi.org/10.1109/CHILECON47746.2019.8987505
5. Canário JP, Mello R, Curilem M, Huenupan F, Rios R (2020) In-depth comparison of deep artificial neural network architectures on seismic events classification. J Volcanol Geotherm Res 401:106881. https://doi.org/10.1016/j.jvolgeores.2020.106881
6. San Martín C, Fritz D, Ferreira A, Curilem M (2021) Continuous volcano seismic monitoring in two steps applied to the chillan volcano. In: 11th International Conference of Pattern Recognition Systems (ICPRS 2021), 2021, 211–216 . https://doi.org/10.1049/icp.2021.1453
7. Imagenet large scale visual recognition challenge (2015) Russakovsky, O., Deng, J., Su, H.e.a. Int J Comput Vis 115:211–252. https://doi.org/10.1007/s11263-015-0816-y
8. Bueno A, Titos M, Benítez C, Ibáñez JM (2022) Continuous active learning for seismo-volcanic monitoring. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2021.3121611
9. Manley GF, Mather TA, Pyle DM, Clifton DA, Rodgers M, Thompson G, Londoño JM (2022) A deep active learning approach to the automatic classification of volcano-seismic events. Front Earth Sci 10:807926
10. López-Pérez M, García L, Benítez C, Molina R (2021) A contribution to deep learning approaches for automatic classification of volcano-seismic events: Deep gaussian processes. IEEE Trans Geosci Remote Sens 59(5):3875–3890. https://doi.org/10.1109/TGRS.2020.3022995
11. Pérez N, Granda FS, Benítez D, Grijalva F, Lara R (2022) Toward real-time volcano seismic events' classification: A new approach using mathematical morphology and similarity criteria. IEEE Trans Geosci Remote Sens 60:1–13. https://doi.org/10.1109/TGRS.2020.3048107
12. Permana T, Nishimura T, Nakahara H, Shapiro N (2021) Classification of volcanic tremors and earthquakes based on seismic correlation: application at Sakurajima volcano Japan. Geophys J Int 229(2):1077–1097
13. Lara F, León R, Lara R, Tinoco A, Ruiz M (2021) A brief frequency analysis of various types of volcanic microearthquakes. In: 2021 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), 1–5. https://doi.org/10.1109/CHILECON54041.2021.9702950
14. Pérez N, Venegas P, Benitez D, Grijalva F, Lara R, Ruiz M (2022) Benchmarking seismic-based feature groups to classify the cotopaxi volcanic activity. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2020.3028193
15. Curilem M, Huenupan F, Beltrán D, San Martin C, Fuentealba G, Franco L, Cardona C, Acuña G, Chacón M, Khan MS, Becerra Yoma N (2016) Pattern recognition applied to seismic signals of llaima volcano (chile): An evaluation of station-dependent classifiers. J Volcanol Geotherm Res 315:15–27. https://doi.org/10.1016/j.jvolgeores.2016.02.006
16. Spampinato S, Langer H, Messina A, Falsaperla S (2019) Short-term detection of volcanic unrest at mt. etna by means of a multi-station warning system. Sci Rep 9:6506
17. Maggi A, Ferrazzini V, Hibert C, Beauducel F, Boissier P, Amemoutou A (2017) Implementation of a Multistation Approach for Automated Event Classification at Piton de la Fournaise Volcano. Seismol Res Lett 88(3):878–891. https://doi.org/10.1785/0220160189
18. Salazar P, Yupanqui F, Meneses C, Layana S, Yáñez G (2023) Multi-station automatic classification of seismic signatures from the lascar volcano database. Nat Hazards Earth Syst Sci 23(2):991–1006. https://doi.org/10.5194/nhess-23-991-2023

19. Schröer C, Kruse F, Gómez JM (2021) A systematic literature review on applying crisp-dm process model. Procedia Comput Sci 181:526–534. https://doi.org/10.1016/j.procs.2021.01.199
20. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'95, pp. 1137–1143. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
21. Zhang Z, Sabuncu MR (2018) Generalized cross entropy loss for training deep neural networks with noisy labels. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS'18, pp. 8792–8802. Curran Associates Inc., Red Hook, NY, USA
22. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. CoRR **abs/1412.6980**
23. Raschka S, Mirjalili V (2019) Python Machine Learning, 3rd edn. Packt Publishing, Birmingham, UK
24. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807. https://doi.org/10.1109/CVPR.2017.195
25. Rumelhart D, Hinton G, Williams R (1986) Learning representations by back-propagating errors. Nature 323:533–536. https://doi.org/10.1109/TPAMI.2013.50
26. Hecht-Nielsen (1989) Theory of the backpropagation neural network. In: International 1989 Joint Conference on Neural Networks, 593–6051 . https://doi.org/10.1109/IJCNN.1989.118638
27. Grandini M, Bagli E, Visani G (2020) Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756
28. Chicco D, Jurman G (2020) The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genom 21:6