Deep Neural Networks Approach to Automatic Recognition Systems for Volcano-Seismic Events

Manuel Titos*, Angel Bueno*, Luz García*, Carmen Benítez*
*Department of Signal Theory, Telematic and Communications, University of Granada, Spain

***Abstract*—Deep neural networks could help to identify the internal sources of volcano-seismic events. However, direct applications of deep neural networks are challenging, given the multiple seismic sources and the small size of available datasets. In this paper, we propose a novel approach in the field of volcano seismology to classify volcano-seismic events based on fully-connected Deep Neural Networks (DNNs). Two DNN architectures with different weights scheme initialization are studied: stacked Denoising Autoencoders (sDA) and Deep Belief Networks (DBN). Using a combined feature vector of Linear Prediction Coefficients (LPC) and statistical properties, we evaluate classification performance on seven different classes of isolated seismic events. These proposed architectures are compared to Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Random Forest (RF). Experimental results show that DNNs can efficiently capture complex relationships of volcano-seismic data and achieve better classification performance with faster convergence when compared to classical models.**

I. INTRODUCTION

Volcanic eruptions are natural spectacles due to their physical power. But these spectacles might be lethal for nearby populations, as volcanoes liberate hazardous gases and produce intense earthquakes. The seismic anomalies associated to volcanic eruptions are the result of an energy exchange between volcanoes and the environment. This energy exchange is produced by stress and relaxation processes, pressure changes or fluid movements [1]. It generates complex seismic signals with unique characteristics: when magma is ascending towards the surface, resonance effects produce seismic signals known as long-period events [2]. If gases are accelerated within the volcano edifice, heavy explosions can be registered and volcanic debris processes may occur at surface level. Volcanic seismology is the most reliable approach to understand active volcanoes, and to characterize their eruptive behavior. Signal processing and machine learning algorithms provide an appropriate framework to analyze these seismic signals and assess the hazardous impact of an eruption. Monitoring volcanoes based on robust automatic recognition systems will help to refine our knowledge of underlying seismic sources, and to improve human understanding of internal volcano dynamics. In addition, new seismic networks provide vast amounts of high-quality seismic data, opening a new horizon to build robust forecasting systems by analyzing that data through the lens of artificial intelligence and machine learning.

Inspired by neuroscience, the idea of artificial neural networks (ANNs) composed by multiple processing layers to learn representations of data with high level of abstraction was introduced by [3] and [4]. Known as Deep Neural Networks (DNN), these models are defined as an ANN with multiple hidden layers between the input and output. DNNs can discover intrinsic patterns within large datasets, and fine-tune their internal parameters in each layer using back-propagation algorithm. One of main pillars for success of DNNs algorithms is the increased availability of data in certain domains, including speech recognition or computer vision [3]. In volcanic seismology, obtaining data from volcanoes around the world is a challenging problem by itself, as careful geophysical interpretation is needed in order to determine seismic sources [2].

The complexity of a DNN is related with the number of hidden layers. Some authors [4] consider a model as "deep" when more than one layer of hidden units is present between its input and output layers. Others, [5] when the number of hidden layers is high. Based on this affirmation, and the aforementioned literature in the field of volcano-seismic recognition, we assume MLP as a neural network with only one layer of hidden units, in order to differentiate it from deep networks.

This research is concerned with the practical applications of deep neural networks in the context of volcano seismic monitoring, where data is scarce, hard to obtain and expensive to label. Hence, the aim of this paper is to study how generalization capabilities of deep neural networks can exploit volcanic-seismic data patterns to classify isolated events. We explore initialization schemes and provide new empirical results for the proposed models, extending the study to the effects of unsupervised pre-training with respect to data-set size and hidden layers depth. Our dataset is composed by 9332 labelled seismic signals recorded at *"Volcan de Fuego"* in Colima, Mexico. The network of sensors, located at different parts of the volcano topography, acquired over time seven different types of seismic events associated to distinct eruptive periods of *"Volcan de Fuego"*. This dataset of isolated events, with no temporal relations across eruptive periods, is very suitable to be modelled by the classification capabilities of DNNs. Concretely, given our dataset size, we focus on stacked denoising autoencoders (sDAs) and deep belief networks (DBNs), as well as network initialization effect to classify isolated events. In order to develop these systems, we needed to take into account two aspects: (i) the input layer of both DNNs must have the same length regardless of the duration of the event; and (ii) the performance of the systems depend on how good the statistical modelling is, and how accurately the parameters of the model can be estimated using the available training data. To address these considerations the well-known Linear Prediction Coding

(LPC) is applied to extract features from the signals [6]. Thus, all events are represented by the same number of LPC coefficients, regardless of their duration. With this setting, we aim to address the robustness of DNN as classifiers for multi-class seismic events and assess the capabilities of DNN to extract hierarchical representations of seismic data.

The rest of the paper is organized as follows: section 2 introduces the related research in the field of volcanic seismology. Section 3 provides a theoretical framework of deep neural networks architectures, and how they can be used for recognition of volcano-seismic events. Section 4 describes, from a geophysical point of view, the seismic signals registered at *"Volcan de Fuego"*. Section 5 describes the experimental setup. Section 6 presents the results and discussions, and section 7 concludes the study. An Appendix with confusion matrices for the best architectures obtained is included.

II. Related Research

Machine Learning algorithms have been widely applied for the classification of volcano-seismic events [7]. Support Vector Machine (SVM) with Gaussian Kernels were used by [8] to discriminate volcanic tremors, landslide and explosions at Stromboli volcanoes. Research by [9], [10], [11], [12], uses hidden Markov models (HMM) to analyze temporal sequences of seismic data in real time. ANNs have been applied as automatic classifiers for volcano-seismic signals. The Multilayer perceptron (MLP) with one hidden layer was first introduced by [13]: trained with spectral features, this MLP was used as a binary classifier to discriminate noise and volcano-tectonic earthquakes from Stromboli volcano. Further research by [14] employs one hidden layer neural network to detect underwater explosions at Stromboli volcanoes. Similar approaches by [15] and [16] use multilayer perceptrons to classify observed wave seismograms, using extracted features and parametrized attributes. These architectures are well suited to solve simple or structured problems, but show limitations when applied to real world data [3]. In volcano seismology, these classifiers are constrained by the data acquisition process: volcanic signals are hard to obtain and easily corrupted by external factors, such as terrain composition or environmental noise [17]. In addition, volcanic data often requires a careful engineering process to design meaningful representations of the data in order to train pattern recognition systems.

DNNs define the state-of-the-art frameworks in many scientific disciplines, such as speech recognition [4], [18] and computer vision [19]. These architectures have been introduced in the field of remote sensing for hyper-spectral image classification [20], [21], [22], weather forecasting [23], and cyclones forecasting [24], although others works include super-resolution, semantic segmentation and object detection [25], [26], [27]. These computational models address the limitations of *shallow-structured* architectures with one level of non-linear feature transformations, and are able to learn intrinsic patterns from vast amounts of labelled data.

To our knowledge, DNNs have not been extensively applied to volcano-seismic data. DNNs, as stated by [3], started its uprising with the discovery that greedy layer-wise unsupervised pre-training can be used to find a good initialization for a learning procedure over all layers, leading to an efficient training of fully connected architectures. The first deep networks trained with speech data used Mel Frequency Cepstral Coefficients (MFCC) to perform acoustic modelling, on Switchboard dataset [4]. These computational models were based on an unsupervised pre-training stage proposed by [5], [28], [29], [30] in order to provide a good initialization of the internal parameters of the network.

III. Deep Neural Networks

Deep Neural Networks (*DNNs*) addresses the limitations of *shallow-structured* architectures, with one layer of non-linearity, and allows computational models to learn rich representations of data with multiple levels of abstraction [3]. Whilst shallow architectures are enough to solve simple well constrained problems, they lack the power to build internal representations of data.

DNN are defined as several fully connected layers, stacked on top of each other, let the information flow sequentially, being the output of the previous layer the input for the next one. However, as the model becomes deeper, limitations appear due to the diffusion of gradients. The many distinct local minima that the optimization process can find in the highly non-convex objective function describing the model parameter space [31].

Extensive work by [28] introduced a type of neural network known as *Deep Belief Networks* (DBNs). These networks can be efficiently trained using a procedure known as *greedy layer-wise* pre-training [32]. Similar architectures, but based on a different pre-training procedure, were developed by [33]. In both cases, these pre-training stages revealed an efficient approach to further train deep architectures. In [5] and [34], it is suggested how weights initialization render the optimization process more effective, and unsupervised pre-training has been reported to achieve faster convergence by initializing the network parameters near a convergence region. Glorot initialization proposes to initialize the network weights based on a gaussian (or uniform) distribution [35]. Random weights initialization has been used in the field, with poor results [34]. Based on [5], [28], [29], [30], [31], [34], it is possible to summarize that the complete process of training a fully connected DNN with unsupervised pre-training comprises two phases: the greedy layer-wise unsupervised pre-training and the fine-tuning.

A. Greedy layer-wise unsupervised pre-training.

In this section we briefly introduce the main concepts behind the greedy layer-wise unsupervised pre-training procedure, the two main architectures derived of such strategy and how they can be applied for volcano-seismic signals.

Regardless of the number of layers, pre-training stage is supposed to be a first step, before applying any learning algorithm to fine-tune all the layers of the model. As a greedy

layer-wise algorithm, each layer is initialized via unsupervised pre-training, and the output of the previous layer can be used as the input for the next one. This implies that each layer uses a single-layer algorithm to learn latent representations from the previous layer, with higher relation to discriminative parameters. The two most used architectures during the unsupervised pre-training phase are *Restricted Boltzmann Machine* (RBM) [28], [36], [37] and *Denoising Autoencoder* (DA) [5], [30]. Regardless which unsupervised algorithm is chosen, the training procedure is the same in all cases. As seen in Figure 1, by stacking and pre-training RBMs, we build Deep Belief Networks (DBNs); by stacking and pre-training DA, we build Stacked Denoising Autoencoders (sDA).

- *Deep Belief Networks (DBN)*: *DBNs* are generative models in which each layer has been pretrained as a RBM. The input layer (or visible) is real-valued, whereas the hidden units are binary-valued [28]. Training one layer at a time, each hidden layer is obtained independently from the hidden states of a RBM via unsupervised learning, in a bottom-up procedure [28]. This unsupervised pre-training proves useful, as each hidden layer learns a meaningful relationship from the units in the lower layer, whereas the higher layer representation become more complex. Each RBM is trained in using the contrastive divergence algorithm *CD-k*, proposed by [36], [37], [38]; forcing each layer to improve the variational lower bound of the training data distribution. To build a deep multilayer generative model, we must take the inverse direction for each *kth* hidden layer, which is given by the transpose of its weights matrix w_k^T. On top, we add a *softmax* layer to normalize per-class output probabilities (2):

$$p(Y = i|x) = \frac{e^{a_i(x)}}{\sum_j e^{a_j(x)}} \tag{1}$$

  where a_i is the output from the previous hidden layer, linked to the weights matrix i, and $\sum_j e^{a_j(x)}$ is the sum of all the hidden units outputs from the previous layer. When unsupervised pre-training is done, we have a deep belief network composed of one visible layer and many hidden layers, that can be fine-tuned for a specific task.
- *Stacked Denoising Autoencoders (sDA)*: (*sDA*) is a multilayer generative model in which each layer has been pre-trained using an autoencoder with noise corruption criteria. This autoencoder is trained to minimize the reconstruction error by the explicit corruption of the input feature vector during the training stage [30]. The main idea behind noise corruption is that obtained representation remains stable even under corruptions of the input, being the denoising task essential to extract characteristic patterns from the input distribution. Thus, pre-training with noisy inputs will enhance generalization for a supervised learning task. Once the first *auto-encoder* has been trained to minimize the reconstruction error, its hidden representation is used as the input for the upper layer. This pre-training procedure is repeated layer-wise for $k - th$ layers, obtaining a set of hidden layers, that can be stacked together in multilayer generative model. Similarly to the DBN, a *softmax* layer can be added on top to compute per-class probabilities.

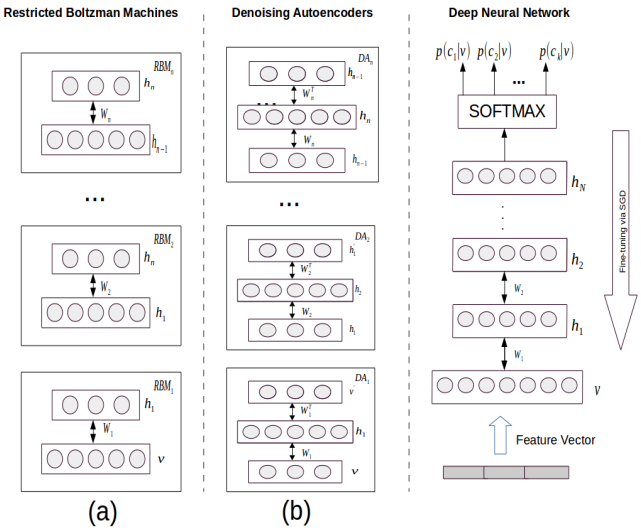


Figure 1. General overview of deep neural network building procedure with pre-training initialization. If we use the hidden states of a restricted Boltzmann machine (RBM) (a), then we build a Deep Belief Network. If we use the hidden states of an autoencoder trained with noise corruption criteria (b), then we build a Stacked Denoising Autoencoder (sDA)

For both models, a normal *fine-tuning* operation is done by applying *Stochastic Gradient Descent (SGD)* algorithm and minimizing a cost function [3]. The entire deep system can be fine-tuned to perform classification of seismic events.

B. Applying deep networks to volcano-seismic data.

One of the main attributes of deep learning is to replace the feature extraction process with deep-network features learned from the data. This contrasts with the criteria followed in this work, where an algorithm for feature extraction is used. The main success of deep learning is based on computational models, trained with vast amounts of labelled data. This opens a new challenge when applying deep learning to volcano-seismic data, given the high dimensionality of the input data and relatively small amount of available labelled data. Nature limits the data registered during an eruption, with multiple incoming signals and noise in the background, that require a careful interpretation by geophysicist. These limitations translates into datasets that geophysicists carefully label, known as ”snapshots”, in which magmatic processes are intensively studied. Given the typical size of a snapshot, a direct application of state-of-the-art deep learning architectures is a challenge, having failed to reach satisfactory results when they have been applied to Colima dataset. Thus, our main objective is to find an application space between classical and deep models, in which by using deep learning ideas, we can explore if this approach improves current state of the art. Following previous research by [34], [31] and [39], our study

focus on the effect of greedy layer-wise unsupervised pre-training as network pre-conditioner.

Seismic events have different temporal duration, limiting the application of temporal models. As an example, volcanic tremors can last from minutes to days, which lead to high-dimensional signals with many long-range dependencies that are hard to model. The classification of isolated events with such temporal variation requires an input vector which efficiently exploits the signal information. Furthermore, sDA and DBN based on fully connected architectures require input vectors of equal lengths. These associated constraints forbid the windowing of each signal: the training of a fully-connected architecture will reveal complicated due to the high number of windowed parts from long-lasting signals. To alleviate this, we transform the raw data into a domain where each event has the same length (see subsection IV-B), encoding each signal using Linear Prediction Coefficients (LPC). With input vectors of same length, DNNs can be trained to classify volcanic-seismic data.

However, to use deep classifiers, we need to adapt the model to the uniqueness of seismic data. First, the pre-training procedure of the DBN architecture uses a Gaussian-Bernoulli RBM (as seismic-volcanic data is real-valued) as first layer. The hidden states of the first layer will be used as input data for pre-training the upper RBMs (binary hidden states, modelled with a Bernoulli-Bernoulli distribution) [37]. For the sDA, each input vector has been 10% corrupted with additive Gaussian noise, a natural choice for real valued inputs. Minimization error can be optimized by minimizing the cross-entropy error between the denoised output, and the uncorrupted input [33]. Both architectures, sDA and DBN, have a softmax layer with seven probabilistic outputs, corresponding to the target labels from the dataset (see section IV). The cost function is given by the negative log-likelihood $NLL(x, \theta)$, defined as:

$$NLL(x, \theta) = -\sum_{n=0}^{n=N} log(p(c_k|x_n, \theta)) \quad (2)$$

Being N the total number of instances for training, x_n a training instance, c_k the class-label assigned to input x_n and θ, the network weights. Afterwards, Equation 2 is minimized using *Stochastic Gradient Descent (SGD)*, with batch training, performing a weights update for every batch of n training instances. [40].

One of the biggest problems that arises when designing DNN models is the selection of the best model with the optimal number of layers, hidden units per layer and learning rate. This problem is known in the literature as hyper-parameters optimization [41]. Traditional algorithms to solve this problem are based on grid search, random search or manual setting. Due to the amount of time required to train deep neural networks, in this work, the procedure used to select the best set of parameters is similar to [42]. The hyper-parameter search is based on fixing a neural network architecture and search for a good optimization of hyper-parameters within a constrained grid of best results. In addition, to prevent over-fitting, models are trained with dropout ([43]), a regularization technique which randomly drop weights during training time to avoid weights to learn redundant representations, and reduce model complexity. Early stopping criteria is also used to prevent overfitting: at the end of each epoch, performance on the test set is evaluated, and if results outperform the previous best model, a copy of the model is saved. Otherwise, training continues for a determined number of epochs (known as patience interval) and testing performance is evaluated. If there is not further improvement of the model, training is stopped [40].

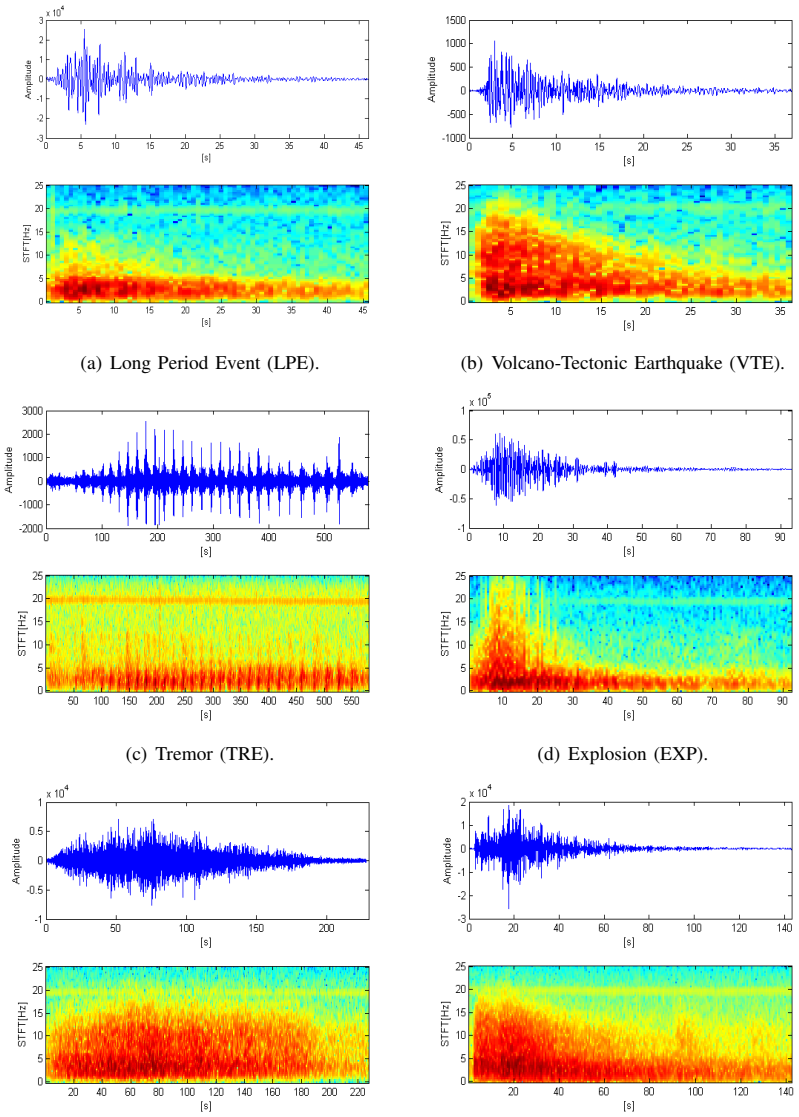
IV. VOLCANO-SEISMIC DATA

DNNs architectures presented in the previous section have been trained as discriminative models, using 9332 seismic signals registered at *"Volcán de Fuego"*. The dataset was collected using two monitoring seismic stations plus one broadband station, at different soil locations, during the eruptive periods of 1998, 2004, 2005 and 2006. Data labelling was performed by geophysicists based on their professional criteria, knowledge and experience of the volcano.

A. Description of the volcano-seismic events.

Located within the East of the trans-Mexican volcanic belt, the *"Volcán de Fuego"* in Colima is a very active andesitic stratovolcano, with a maximum height of 3860 m. In [44], a detailed analysis of the geophysical activity at *"Volcán de Fuego"* is described. Figure III-B shows representative waveforms and spectrograms of each type of volcano-seismic event, recorded at Colima. Following the guidelines proposed by [2] [17], volcano-seismic signals of *"Volcán de Fuego"* can be classified according to their waveform and spectral content, and additionally associated to potential sources as:

1) *Long period events (LPE)* (Figure 2(a)): These seismic signals remain in duration to small VTE earthquakes but wit different frequency contain showing a clear harmonic signature [45]. In general they are quasi-monochromatic signals with a narrow frequency band centered, in the majority of the cases, between 1 to 6 Hz. Their source models are associated to volumetric modes of deformation of the propagation medium. In general the proposed models are related to resonance of the medium as a consequence of fluid displacement inside of the volcanic edifice or the generation of pressure transients in fluids. We can mention as example, a crack in which a resonance occurs when the fluids (magma, gas or water) are ascending towards the surface or the existence of pressure transients within the fluid-gas mixture inside of the volcanic edifice, causing also resonance phenomena [45]. All proposed models are able to explain the behavior of the observed features in time and spectral domains. They are usually located in particular areas of the volcanic structure where fluids generate disturbances. They have been used as short-term precursors of volcanic eruptions. In many cases, they appear in time forming the so call "seismic swarms": thousands of LPE events in a short time period, often overlapped.



(a) Long Period Event (LPE).

(b) Volcano-Tectonic Earthquake (VTE).

(c) Tremor (TRE).

(d) Explosion (EXP).

(e) Lava Flow (COL).

(f) Regional Earthquake (REG).

Figure 2. Spectrogram of volcano-seismic signals registered at *"Volcán de Fuego"*, Colima (Mexico).

2) *Volcano tectonic earthquakes (VTE)* (Figure 2(b)): VTE events are *classical earthquakes* originated inside of volcanic environments. Their main characteristic is a signal with a broad frequency contains reaching up to 40 Hz with duration from a few to tens seconds. They are the result of a brittle response of the medium caused by seismic stress producing a shear failure of the volcanic edifice generating a broad set of seismic waves. This seismic stress could be produced by several causes, from local tectonic regime to fluid (water, gas or magma) displacement inside of the volcanic edifice. The consequence of this fracturing of the medium is the generation of two kinds of seismic waves (body waves) with different propagation velocity: P-waves (longitudinal displacement) associated to change of Pressure in the medium, and S-waves (transverse motion) associated to shear displacement of the elastic medium. They should appear spread in space and time inside of the volcanic edifice. Many times they have been used as long-term volcanic eruption precursory activity appearing from days, months or years before the eruption.
3) *Volcanic tremor (TRE)* (Figure 2(c)): These events are in general characterized by harmonic signals with sustained amplitude and highly variable duration, lasting from minutes to hours or even months. Their spectral characteristics resemble LPE events with quasi-monochromatic signature, but in some cases their peak of frequency could reach up 10 Hz of more. Sources of volcanic tremor are

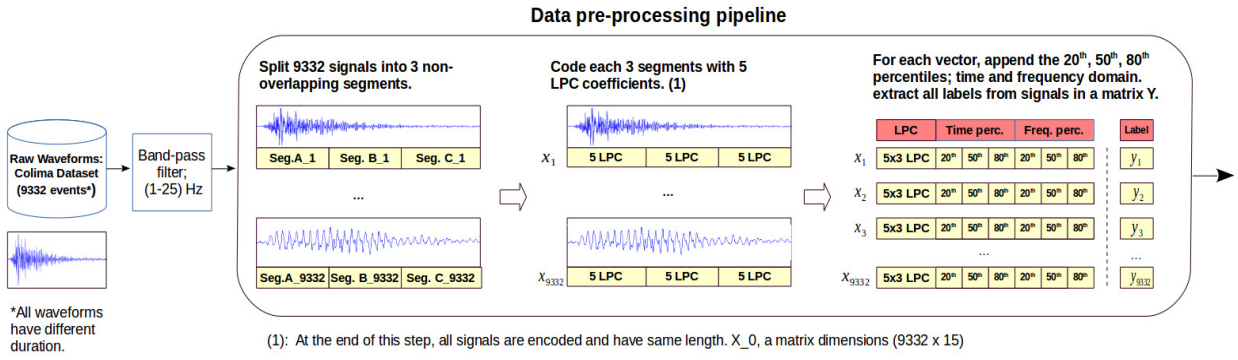


Figure 3. Overview of the data pre-processing pipeline. First, signals are band-pass filtered between 1 and 25 Hz. Each signal has been windowed into three non-overlapping segments. LPC coefficients, along with statistical properties, have been computed. All seismic events (9332) are represented by 21 dimensional vectors, regardless of their waveform duration. During the training phase, each of these vectors are batched and used as input for DNN models. The associated labels will be used as targets for our classification tasks.

diverse, from inner pressure disturbance to external gas emission, debris avalanches or pyroclastic flows, among others. Some theories suggest that, when the source of the tremor is located inside the volcanic edifice, its source is identical to those associated with LPE events, being the tremor the consequence of a non-lineal overlapping of multiple LPE events. This overlapping model is the consequence of the observation of the similar spectral characteristics, and that both TRE and LPE events share, in many cases, space inside of the volcanic edifice, appearing associated in the time.

4) *Explosions (EXP)* (Figure 2(d)): These signals are associated to the external activity of the volcano due to the sudden emission of gas and ash to the atmosphere (explosion). Since mostly of them can be visible and recorded by video, it is possible to associate the external effect to the signals recorded in the seismometers. They are characterized by an initial short duration LPE event followed by high frequency signals with a narrow energy peak with peaks located at different frequencies, from 4 up to 20 Hz.
5) *Lava Flow (COL)* (Figure 2(e)): As mentioned above, a class of tremor is originated by debris flow located at the volcano surface. Since they can be monitored using video record, it is possible to associate the surface lava movement with the generation of this type of volcanic tremor. These events provide very useful information on the final consequence of the internal activity in a volcano. These events exhibit frequency content between 5 to 10 Hz.
6) *Regional Earthquakes (REG) (Figure 2(f))*: Tectonic earthquakes might occur anywhere in the earth if there is enough elastic strain energy stored to drive the fracture propagation along a fault plane. They normally have a bigger duration than volcano-tectonic earthquakes, but similar spectral content.
7) *Seismic Noise (NOISE)*: Overlapped over any seismic signal, there is a type of signal, mainly of low amplitude, originated by multiple natural and artificial sources. This signal is named *seismic noise* and typically contaminates the registered seismic signals. As natural sources, we can mention wind, atmospheric pressure variation or rain. In case of artificial sources, this noise is known as *"cultural noise"* and is mainly introduced by nearby populations and human activity. In some cases, this noise could interfere the frequency range in which most of the volcanic spectral content is located.

A total of 9332 volcano-seismic events compose the dataset used in the experiments, with the following per-class distribution: 1738 VTE, 2699 LPE, 1170 TRE, 455 REG, 1406 COL, 278 EXP, and 1586 NOISE have been selected. Captured events are *isolated*: a recorded seismic signal with a specific duration, and its associated label.

B. Data processing and feature extraction.

As we above-mentioned in section III-B, a direct application of state-of-the-art architectures on raw volcano-seismic data has failed to reach satisfactory results. Therefore, data preparation and feature extraction is a crucial step to build reliable classifiers, as we aim to provide useful features for the discriminative stage. The feature extraction process has been automated in a *pipeline*: the input is the full dataset of 9332 seismic signals in the time domain, sampled at 50 Hz and band-pass filtered between 1 Hz and 25 Hz. Figure 3 shows two examples of the *pipeline* procedure for two events with different duration. Then, after the feature extraction stage, a data-set of 9332 features vectors is obtained, with their associated labels. This new dataset will be used as training data for the classifier. As suggested by [4], rich-features carry useful discriminative information for DNNs. In addition, experimental work by [13] has shown that Linear Prediction Coefficients (LPCs) are robust features for volcano-seismic classification, as LPCs can encode the values of any given signal in a linear combination of k. coefficients, regardless of its duration. An important advantage of LPC-based feature vectors is their computational simplicity, being this a key

factor when deploying this system in real time. Following the mentioned guidelines, our feature extraction *pipeline* can be summarized as follows:

1) To characterize temporal evolution of volcano-seismic events, each signal has been windowed into three non-overlapping segments of equal length. We decided to compute three segments in order to capture information at start, center and end of the signal. As suggested by [46], and similar to human phonemes [47], windowing the signal into three sections provides useful information about how the signal behaves temporally, and can be discriminitave enough for certain types of seismic events.
2) Aiming to provide a good representation of volcano-seismic signals, and encode the three non-overlapping segments into same dimensions, LPC with order $k = 5$ have been computed for a given signal and its segments. Following [13], and after exhaustive analysis of our data, we decided to use $k = 5$ LPC coefficients to avoid excessive redundancy and maintain trade off between dataset and architectures parameters.
3) For a given signal, LPC coefficients of each segment are complemented with statistical features proposed by [12]. These statistical features help to discriminate the impulsiveness of the signal in both, time and frequency domain. The $20th$, $50th$ and $80th$ percentiles of the cumulative sum of the signal amplitude in time and frequency domain are appended to the feature vector.

The final feature vector is generated by concatenating the parameters obtained for each frame. The size of each feature vector is $(3*k)+6$, being $k = 5$ the order of LPC coefficients, 3 the number of temporal frames, and 6 the number of features corresponding to the percentiles $20th$, $50th$ and $80th$, in both, time and frequency domains. This procedure yields into a set of $21 - dimensional$ input vectors, containing temporal and frequency features.

Given this set of features, we will train two DNNs architectures: Stacked Denoising Autoencoders (sDA) and Deep belief networks (DBN). We aim to determine the robustness of these architectures as classifiers of seven different types of seismic events, to understand how pre-training helps in this discriminative task, and to assess the capabilities of the model to extract meaningful features from the data.

V. EXPERIMENT DETAILS

A. Experimental setup.

Experiments to determine the power of deep neural networks architectures as classifiers of seismic events are conducted using the data from *"Volcán de Fuego"*, described in section IV-A. Data pre-processing is performed before training, as described in detail at subsection IV-B: At the end of the pipeline, all seismic events are represented by 21 dimensional vectors, regardless of their duration.

Once processed, the *"Volcan de Fuego"* dataset is divided into training (75%) and test (25%) sets. This yields a training set of 7000 training instances, and 2332 test instances. Moreover, we used 50% of the test set (1166 instances) as validation data. In addition, a balanced random shuffle of the data was done to avoid highly correlated batches during training stage. Cross-validation with four partitions over the test data has been used in order to test the model ability to generalize on unseen data.

Table I
BEST ARCHITECTURES FOR DBN AND SDA MODELS

	2 hidden layers	3 hidden layers
DBN	250-165	260-385-35
sDA	260-385	260-385-235

Hyper-parameters optimization and best model selection are based on grid-search, in a similar way to the approach used in [42]. Deep neural models are tested with a total number of 250100 different configurations of hyper-parameters. For both architectures, sDA and DBN, the number of hidden units at first, second and third layers are tested from 50 up to 1250 hidden units, with increments of 25 hidden units. In this case, the architecture is tested varying the number of hidden units of the last layer with increments and decrements of 5 hidden units, obtaining several new architectures. Learning rates have been tested within the range of 0.000001 to 0.01. Data has been normalized in mean and variance, and *sigmoid* is used as non-linearity function. Both models, sDA and DBN, have a softmax probability layer, defined by Equation 1 with seven target outputs, corresponding to each class of our dataset. These models have been trained with a batch size of 10 training instances. Dropout regularization technique was used with $p = 0.20$. To further alleviate over-fitting problem, early-stopping criterion with a patience interval of 10 epochs was used with the validation set.

Classification performance of deep architectures are compared to Glorot initialized MLP (with $tanh$ non linearity), Support Vector Machine (SVM) [48] and Random Forest (RF) [49]. Performance of the MLP architecture with one hidden layer is explored by varying the number of hidden units from 25 to 2000, obtaining the best performance with 500 hidden units. For SVM, radial and linear kernels are used. Similarly, for RF models, we explore a large range of estimators (up to 500), obtaining the best performance with 120 estimators. The best results obtained for DBN and sDA with 2 and 3 hidden layers are summarized in the Table I. Following the same procedure, an experimental study of unsupervised pre-training effects and layer depth with respect to the size of the dataset has been designed and compared against deep architectures with Glorot initialization.

B. Defining the metric of the architectures.

The reported metrics are based on $F1$ score, *Precision* (PR) and *Recall* (RC) [50]. The aim of using precision is to determine how good the model is at classifying specific classes. By using recall, we can assess how good the model is at selecting instances of a certain class from the dataset. $F1$ score is a trade-off measure between precision and recall. Precision is computed as:

$$PR = \frac{True\,Positives}{(True\,Positives + False\,Positives)} \tag{3}$$

and it shows which percentage of positive predictions were correct. Recall is computed as:

$$RC = \frac{True\,Positives}{(True\,Positives + False\,Negatives)} \quad (4)$$

and it measures which percentage of positive events were correctly predicted. Thus, precision is a measure of how good predictions are with regard to false positives, whereas recall measures how good the predictions are with regard to false negatives. $F1$ score can be computed as the weighted average of precision and recall:

$$F1_{score} = \frac{2 * (RC * PR)}{(RC + PR)} * 100\% \quad (5)$$

All deep neural models were implemented using *Theano* [51], a deep learning framework written in Python. In the case of SVM and RF, we used *scikit-learn* [52], an open-source Python framework for machine learning. Given the elevated number of computations required by deep models, training stage is done on two *Graphic Processor Unit* (GPU): NVIDIA K40c GPU, and NVIDIA GEFORCE GTX 1080 GPU.

VI. Results and discussions

In this section, we report the classification results for the DBN, sDA, MLP, SVM and RBF with data from *"Volcán de Fuego"*. Given the vast amount of experiments performed (about 250100), results are presented only for best configurations and best performances obtained in terms of *Precision*, *Recall* and F1 score (seet Table I). Initialization effect on dataset size for those models are compared with the same network topologies but with Glorot initialization. Moreover, given that we are working with real-world data, an experimental study using a threshold at the output probabilities of the softmax layer will help us to understand the confidence of the model when classifying volcano-seismic signals at real time. At the Appendix, confusion matrices for the DBN, sDA, SVM and RBF and the four test sets are included.

A. General performance of the system.

Table II shows per-class Precision and Recall of the implemented algorithms. There are several conclusions which can be drawn from these results. Compared to MLP, SVM and RF, DNNs attain higher recall, precision and F1 score in those seismic events that are hard to discriminate. Concretely, they work remarkably well on explosions (EXP), lava flows (COL) and volcanic tremors (TRE). During an eruption, explosions can be associated to lava flows and/or rock falls. This simultaneity in time is translated into less sensitivity for the SVM, RF and MLP, increasing the amount of false positives and inserting more errors across classes. Regarding table II, it is also important to notice that all models are able to classify seismic events that are very distinctive, such noise (NOISE), regional earthquakes (REG), volcano tectonic earthquakes (VTE) and long-period events (LPE). However, whilst SVM, RF and MLP attain good precision on these events, they do have smaller recall if compared with the sDA and DBN. This result suggests that the pre-training stage was indeed able to produce more useful feature detectors, and weights initialization via unsupervised pre-training does lead to better optimized models for classification of volcano-seismic signals.

In terms of precision and recall, we can see in table III that DNNs with 2 and 3 hidden layers can exploit information from seismic events and achieve good generalization. Therefore, the advantage of deep models is clear: whereas the SVM, RF and MLP can classify with high precision events in which the temporal and spectral contents are very characteristic, the sDA and DBN do the same but with higher precision and recall in their predictions. In addition, sDA and DBN tend to classify complex seismic events such explosions and tremors with higher recall and precision.

Table III shows per-class and average $F1$ scores. Additionally $Accuracy$ score has been calculated for the global dataset with a confidence interval of 95 %. In consistence with the analysis of table II, best results are obtained for deep architectures; being especially remarkable in the cases of explosions.

B. Effect of greedy layer-wise unsupervised pre-training

The experimental results at Table III show that sDA and DBN classify complex seismic events with higher recall, precision, and increased performance for the rest of classes when compared to classical models. This proves essential when monitoring volcanic environments, as explosions, earthquakes and long-period events can be seen as precursory of intensive volcanic activity, leading to potential eruptions. Thus, DNNs have proven useful to discriminate seven of the most important volcanic-seismic events in nature. However, the use of deep learning frameworks imposes the availability of large datasets to train computational models with millions of parameters. DNNs tends to overfit if the dataset size is small. As mentioned in subsection III-B, the direct application of DNNs as classifiers for volcano-seismic data is challenging given the number of available *snapshots*.

In spite of these drawbacks, Table III has shown that pre-training helps to initialize network weights, giving the model with a *a-priori* knowledge of the data distribution, and provides an optimization advantage which translates into fast convergence. This *a-priori* knowledge in both architectures can be explained by how hidden layers are pre-trained, as in both cases, a lower bound is being maximized [33], [37], which naturally leads to a minimization of the KL divergence between the true data distribution and model parameters [53]. Further evidence of this statement is supported by Table IV, which shows the effect of unsupervised pre-training in different dataset sizes, compared against a Glorot initialization scheme, in terms of global accuracy. DBN and sDA with Glorot initialization can be seen as deep neural models where pre-training phase is removed. The weights of the best architectures obtained for the sDA and DBN with 2 and 3 hidden layers (see Table I), are initialized with Glorot (no unsupervised pre-training) and compared against best unsupervised pre-trained architectures. Table IV shows that generative pre-trained models (DBN) with two hidden layers, behave slightly

Table II
PER-CLASS $Precision$ AND $Recall$ MEASURES OBTAINED FOR EACH OF THE LEARNING MODELS. VALUES ARE EXPRESSED IN %.

	NOISE		EXP		REG		COL		VTE		TRE		LPE	
	PR	RC	PR	RC	PR	RC	PR	RC	PR	RC	PR	RC	PR	RC
SVM-Rad	97.52	96.20	78.98	65.96	92.72	85.27	93.87	95.66	93.03	93.77	85.9	83.39	91.91	95.9
SVM-Lin	97.52	96.20	**87.07**	53.72	94.26	87.95	91.66	96.88	92.75	94.11	85.22	76.82	89.45	**96.29**
RF-120	97.62	95.47	86.57	61.70	92.09	88.39	93.53	96.07	93.32	94.9	86.90	85.95	92.27	96.06
MLP-H1	97.53	96.69	82.39	69.68	93.72	86.61	95.24	**97.69**	**93.68**	95.70	87.25	86.13	94.03	95.66
DBN-H2	97.20	**97.92**	82.39	69.68	92.27	**90.63**	97.03	97.56	92.96	95.70	89.63	88.32	94.66	95.03
sDA-H2	**97.78**	97.06	84.91	71.81	**94.31**	88.84	96.26	**97.69**	93.03	**96.72**	**89.64**	89.96	95.11	95.11
DBN-H3	97.55	97.43	82.82	71.81	91.47	86.16	**97.16**	97.42	93.37	95.70	87.97	89.42	94.35	94.79
sDA-H3	**97.78**	96.94	83.44	**72.34**	93.40	88.39	96.10	97.01	92.79	96.15	88.41	**90.51**	**95.56**	94.95

Table III
PER-CLASS F_1 SCORE OBTAINED USING $Precision$ AND $Recall$, GLOBAL ACCURY (ACC).

	NOISE	EXP	REG	COL	VTE	TRE	LPE		
	F1 (%)	F1 (%)	F1 (%)	F1 (%)	F1 (%)	F1 (%)	F1 (%)	F1 average (%)	Acc. Glob (%)
RF-120	96.53	71.88	90.13	94.78	94.10	86.41	94.13	89.71	92.80±0.61
SVM-Lin	96.85	66.39	91.03	94.19	93.42	80.80	92.75	87.92	91.55±0.80
SVM-Rad	96.86	71.86	88.81	94.75	93.39	84.63	93.87	89.17	92.32±0.76
MLP	97.11	75.33	90.03	96.45	94.68	86.67	94.85	90.73	93.57±0.70
DBN-H2	**97.56**	75.42	91.43	**97.29**	94.30	88.98	94.85	91.40	94.04±0.68
sDA-H2	97.41	77.78	**91.51**	96.97	94.84	**89.78**	**95.11**	91.92	**94.32±0.66**
DBN-H3	97.00	77.00	89.00	97.00	**95.00**	89.00	95.00	91.00	93.87±0.69
sDA-H3	97.00	**78.00**	91.00	97.00	94.00	89.00	95.00	**92.00**	94.10±0.68

worse than Glorot initalized models with two hidden layers without pre-training for critically scarce data (25%). When dataset increases, overall accuracy for all models increases, but the improvement is more substantial for pre-trained models. To our criteria, these substantial improvement can be explained on the underlying weights for the sDA and DBN. For the DBN, the RBM tries to learn a probability distribution over its input, by performing a sampling over data distribution with CD-k sampling. This sampling can approximate the probability distribution, even under data scarcity. Alternatively, the DA tries to minimize the expectation over the reconstruction error, maximizing the lower bound with respect to the true data distribution, and learning the most representative input data structure. Hence, when data is critically scarce the DA can not learn efficiently such approximation.

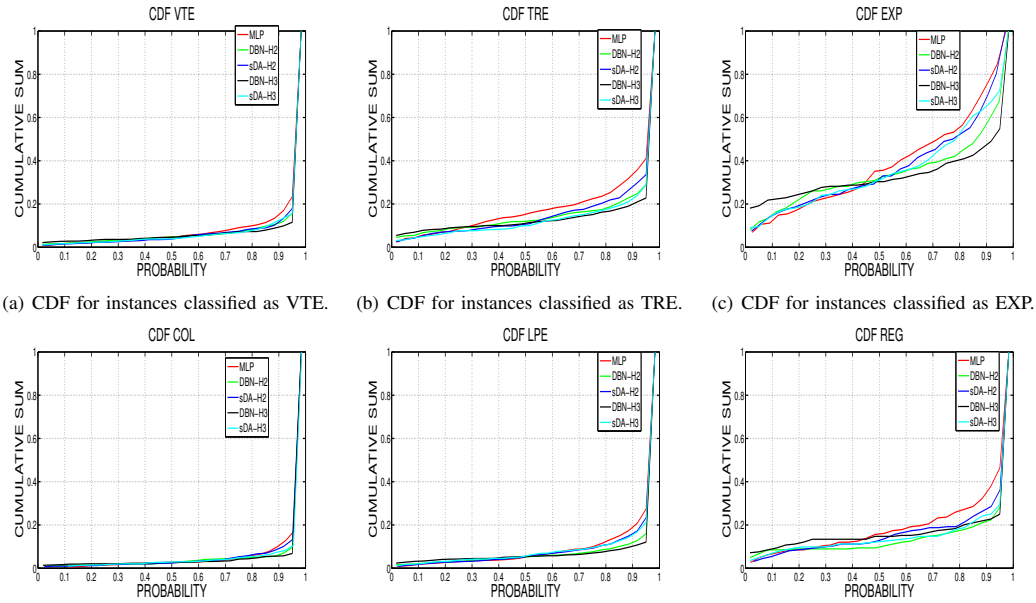
C. Analyzing the confidence of the classifications

Previously analyzed results suggest that classifiers emit high probabilities for overlapping classes. From a geophysical point of view, simultaneous seismic events would help to understand the evolution of the volcanoes. At seismic observatories, this problem is often solved by expert geophysicists. Analyzing the confidence of deep neural models when classifying seismic events would help to understand if the detection of unclear events can be enhanced

Previous results have shown that SVM and RF are outperformed by ANN models. Therefore, we will focus our analysis in neural network models. Using softmax probabilities at the output layer, we can compute the assigned class probabilities per event. Figure 4 depicts the cumulative distribution function (CDF) of class probabilities for the sDA, DBN and MLP. The CDF has been computed using the normalized cumulative sum of the histogram of output probabilities. By looking at the CDF of the predicted explosions (EXP), Figure 4 (c), notice that deep architectures assign higher class probabilities of belonging to explosion class for any given instance, whereas the MLP has less confidence when assigning class probabilities for same class. Moreover, from Figure 4, it is noteworthy that deep architectures always assign higher class-probability for COL and TRE classes than the MLP. Events such as VTE, COL and LPE, tend to have high class probabilities, close to one.

All CDFs have three distinctive probability regions: A low and high confidence one around 0.4 and 0.9 respectively, within an intermediate plateau of class-probabilities. The sparsity in this plateau for the MLP gives us information of how deep networks assign class probabilities. Thus, determining a certain value for the threshold leads to two types of model: For a small threshold, the system will be very sensitive, and all events with a probability above the threshold will be detected, and eventually, many of them classified in an erroneous manner. For high thresholds, the nature of earthquake-volcanic signals will make harder to characterize signals that are not very clear.

A study of the F1 score was performed by varying the threshold value from 0.4 to 0.8. The $F1$ score of sDA and DBN for 2 and 3 hidden layers outperforms the classic MLP. As seen from Figure 5, even for highly sensitive systems, $F1$ score for deep architectures is higher when compared to the MLP. When the threshold increases, the MLP is surpassed by deep architectures. These results confirm the previous hypothesis: Deep models can classify very characteristic events, such NOISE, VTE, COL and LPE. Complex events are classified with higher probabilites. Pre-training stage and increased depth (number of hidden layers) help the model to extract better features from seismic data, which translates in lesser sparse softmax probabilities.



(a) CDF for instances classified as VTE. (b) CDF for instances classified as TRE. (c) CDF for instances classified as EXP.

(d) CDF for instances classified as COL. (e) CDF for instances classified as LPE. (f) CDF for instances classified as REG.

Figure 4. Per class Cumulative Distribution Function (CDF) for different kind of events. The x-axis represents the class probabilities assigned by the models, whereas y-axis represents the normalized cumulative sum of events predicted within that class. For a high performance classifier, the *ideal* graph would tend towards probability one over x-axis, as the output probability vector would be less sparse.

Table IV
INITIALIZATION EFFECT ON DATASET SIZE FOR SAME DBN, SDA AND DNN BEST TOPOLOGIES.

	Acc. Global 25%	Acc. Global 50%	Acc. Global 75%	Acc. Global 100%
DNN-H2-Glorot	**91.07±0.39**	92.22±0.6	92.71±0.43	93.31±0.58
DBN-H2-preTra	90.69±1.4	**93.06±0.97**	**92.83±0.7**	**94.04±0.68**
DNN-H2-Glorot	**90.73±0.31**	**91.83±0.82**	92.77±0.72	93.17±0.66
sDA-H2-preTra	90.17±2	91.21±1.4	**92.83±1**	**94.32±0.66**
DNN-H3-Glorot	89.35±1.5	91.58±0.64	92.35±0.69	93.24±0.74
DBN-H3-preTra	**91.25±1.1**	**92.2±0.57**	**92.8±0.44**	**93.87±0.69**
DNN-H3-Glorot	90.25±1.4	91.67±0.52	92.63±0.63	93.09±0.55
sDA-H3-preTra	**90.77±0.88**	**92.09±0.76**	**92.97±0.53**	**94.1±0.68**

By choosing 0.8 as a safe threshold to avoid excessive errors, we show the relative improvement of DNNs over the MLP at Figure 6. Notice that explosions present a relative improvement (RI) of 13% for the DBN-H3 and 2.5% for the sDA-H3. On the other hand, there are significant improvements related to volcanic tremors, up to 4% in both architectures. Finally, it is remarkable the recognition improvement for REG events. They are easily confused with VTE and , in some cases, with moderate explosions (EXP) with very energetic arrival and quick decay. Increasing the class probability threshold leads to more confident models that can be deployed in real environments, and provide geophysicists with a tool to analyze most dubious signals.

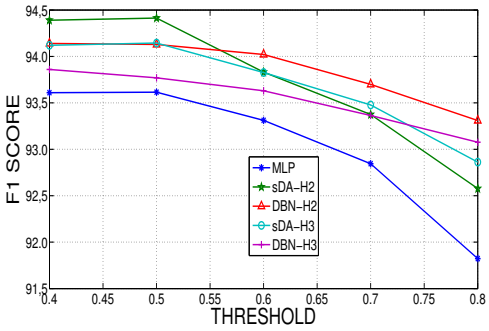


Figure 5. Evolution of F1 score for several class probability thresholds, ranging from 0.4 to 0.8.

VII. CONCLUSION AND FUTURE WORK

In this research, we proposed DNNs for automatic classification of volcano-seismic events based on pre-training initialization. Two different DNNs, DBN and sDA, are tested with seismic data recorded at *"Volcán de Fuego"*, Colima (Mexico). In our experiments, classification results show that deep architectures outperforms the SVM, MLP and RF. Pre-training initialization was found to be a particularly effective strategy to achieve this improvement. In addition, depth helps to increase the overall generalization capabilities of the system.

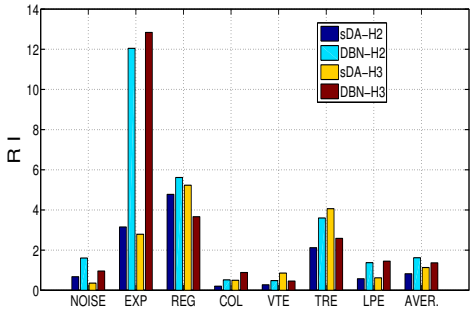


Figure 6. Per-event relative improvement of DNN architectures when compared to the MLP for a fixed threshold of 0.8.

We find that sDA and DBN can classify seismic events with higher precision and recall than classical architectures. Moreover, deep architectures are more sensitive to detect events that occur simultaneously in time, such as explosions and tremors. Classifiers based on deep neural networks can be deployed in real-environments to monitor the seismicity of restless volcanoes, and enhance current early warning systems. Finally, given the nature and size of volcanic snapshots (dataset), the use of raw volcanic events as training data results on non-useful representations, and therefore, a direct application of state-of-the-art deep learning architectures is still a challenge.

VIII. Acknowledgements

We would like to thank Prof. Hugo Larochelle and all the members of SMART Laboratory at University of Sherbrooke for their continuous support and advice when developing deep learning models in Theano. We thank *Instituto Andaluz de Geofísica* for providing us with Colima dataset and invaluable geophysical insight. This research was funded by TEC2015-68752 (MINECO/FEDER).

References

[1] B. Chouet. Volcano seismology. *Pure and Applied Geophysics*, 160(3-4):739–788, 2003.

[2] J.M Ibáñez, E. Del Pezzo, J. Almendros, M. La Rocca, G. Alguacil, R. Ortiz, and A. García. Seismovolcanic signals at deception island volcano, antarctica: Wave field analysis and source modeling. *Journal of Geophysical Research: Solid Earth*, 105(B6):13905–13931, 2000.

[3] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[4] G. Hinton, L. Deng, D. Yu, G-E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. Sainath. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[5] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks.

[6] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[7] C. Benítez, L. Garía, A. Alos, J. Prudencio, I. Alvarez, and A. de la Torre. A comparative study of classifiers based on hmm, gmm and svm for the vt, lp and noises discrimination task. In *EGU General Assembly Conference Abstracts*, volume 16, page 11783, 2014.

[8] M. Masotti, S. Falsaperla, H. Langer, S. Spampinato, and R. Campanini. Application of support vector machine to the classification of volcanic tremor at etna, italy. *Geophysical research letters*, 33(20), 2006.

[9] C. Benítez, J. Ramírez, J.C. Segura, A. Rubio, J.M. Ibáñez, J. Almendros, and A. García-Yeguas. Continuous hmm-based volcano monitoring at deception island, antarctica. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V, May 2006.

[10] J.M. Ibáñez, C. Benítez, L. Gutiérrez, G. Cortés, A. García-Yeguas, and G. Alguacil. 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):218–226, 2009.

[11] L. Gutiérrez, J.M Ibáñez, G. Cortés, J. Ramírez, C. Benítez, V. Tenorio, and A. Isaac. Volcano-seismic signal detection and classification processing using hidden markov models. application to san cristóbal volcano, nicaragua. In *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009*, volume 4, pages IV–522. IEEE, 2009.

[12] G. Cortés, R. Arámbula, L. Gutiérrez, C. Benítez, J. Ibánez, P. Lesage, I. Alvarez, and L. Garcia. Evaluating robustness of a hmm-based classification system of volcano-seismic events at colima and popocatepetl volcanoes. In *Geoscience and Remote Sensing Symposium,2009 IEEE International,IGARSS 2009*, volume 2, pages II–1012. IEEE, 2009.

[13] S. Scarpetta, F. Giudicepietro, E.C Ezin, S. Petrosino, E. Del Pezzo, M. Martini, and M. Marinaro. 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.

[14] E. Del Pezzo, A. Esposito, F. Giudicepietro, M. Marinaro, M. Martini, and S. Scarpetta. Discrimination of earthquakes and underwater explosions using neural networks. *Bulletin of the Seismological Society of America*, 93(1):215–223, 2003.

[15] S. Diersen, E. Lee, D. Spears, P. Chen, and L. Wang. Clasification of seismic windows using artificial neural networks. *Procedia Computer Science*, 00:1–10, 2001.

[16] M. Kuroda, A. Vidal, A. Maria, and A. De Carvalho. Interpretation of seismic multiattributes using a neural network. *Journal of Applied Geophysics*, 85:15–24, 2012.

[17] J. Wassermann. *IASPEI New manual of seismological observatory practice*, volume 1, chapter Chapter 13: Volcano seismology, page 42 pp. GeoForschungsZentrum Potsdam, 2002.

[18] L. Deng and Y. Dong. Deep learning: methods and applications. *Fundations and trends in Signal Processing*, 7:3-4:197–387, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] L. Zhang, L. Zhang, and B. Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, 2016.

[21] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 54(10):6232–6251, 2016.

[22] A. Romero, C. Gatta, and G. Camps-Valls. Unsupervised deep feature extraction for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1349–1362, 2016.

[23] M. Elhoseiny, S. Huang, and A. Elgammal. Weather classification with deep convolutional neural networks. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 3349–3353. IEEE, 2015.

[24] R. Kovordányi and C. Roy. Cyclone track forecasting based on satellite images using artificial neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(6):513–521, 2009.

[25] J. Tang, C .Deng, G. Huang, and B. Zhao. Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine. *IEEE Transactions on Geoscience and Remote Sensing*, 53(3):1174–1185, 2015.

[26] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3325–3337, 2015.

[27] F. Hu, G. Xia, J. Hu, and L. Zhang. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, 7(11):14680–14707, 2015.

[28] G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[29] M. Ranzato, S. Chopra, Y. LeCun, and F-J.Huang. Energy-based models in document recognition and computer vision. *Proc. International Conference on Document Analysis and Recognition*, 2007.

[30] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings*

of the 25th international conference on Machine learning, pages 1096–1103. ACM, 2008.

[31] D. Erhan, P.A Manzagol, Y. Bengio, S. Bengio, and P. Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. In *AISTATS*, volume 5, pages 153–160, 2009.

[32] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise trainning of deep networks. *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pages 153–160, 2007.

[33] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, December 2010.

[34] D. Erhan, Y. Bengio, A. Courville, P.A Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.

[35] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[36] A. Fischer and C. Igel. An introduction to restricted boltzmann machines. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 14–36. Springer, 2012.

[37] G. Hinton. A practical guide to training restricted boltzmann machines. *Momentum*, 9(1):926, 2010.

[38] M.A Carreira-Perpinan and G. Hinton. On contrastive divergence learning. In *Proceedings of the tenth international workshop on artificial intelligence and statistics*, pages 33–40. Society for Artificial Intelligence and Statistics NP, 2005.

[39] A.M Saxe, J.L McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *CoRR*, abs/1312.6120, 2013.

[40] S.Haykin. Neural networks:a comprehensive foundation. 1998.

[41] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[42] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 473–480, New York, NY, USA, 2007. ACM.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

[44] M. Palo, J.M Ibáñez, M. Cisneros, M. Bretón, E. Del Pezzo, E. Ocana, J. Orozco-Rojas, and A.M Posadas. Analysis of the seismic wavefield properties of volcanic explosions at volcan de colima, mexico: insights into the source mechanism. *Geophysical Journal International*, 177(3):1383–1398, 2009.

[45] B.A Chouet. Long-period volcano seismicity: its source and use in eruption forecasting. *Nature*, 380(6572):309, 1996.

[46] C. Valdés-González N. Varley G.R Dávila C. Navarro R. Arambula-Mendoza, P. Lesage. Seismic activity that accompanied the effusive and explosive eruption during the 2004-2005 period at explosive eruption during the 2004-2005 period at volcán de colima, méxico. *Journal on VOlcanology and Geothermal Research*, 205(1-2):30–46, 2011.

[47] L. Deng, J.Li, J. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, and J. Williams. Recent advances in deep learning for speech research at microsoft. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8604–8608. IEEE, 2013.

[48] F.Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on geoscience and remote sensing*, 42(8):1778–1790, 2004.

[49] M. Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26:217–222, 2005.

[50] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[51] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[53] D.M Blei, A. Kucukelbir, and J.D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted), 2017.

Appendix

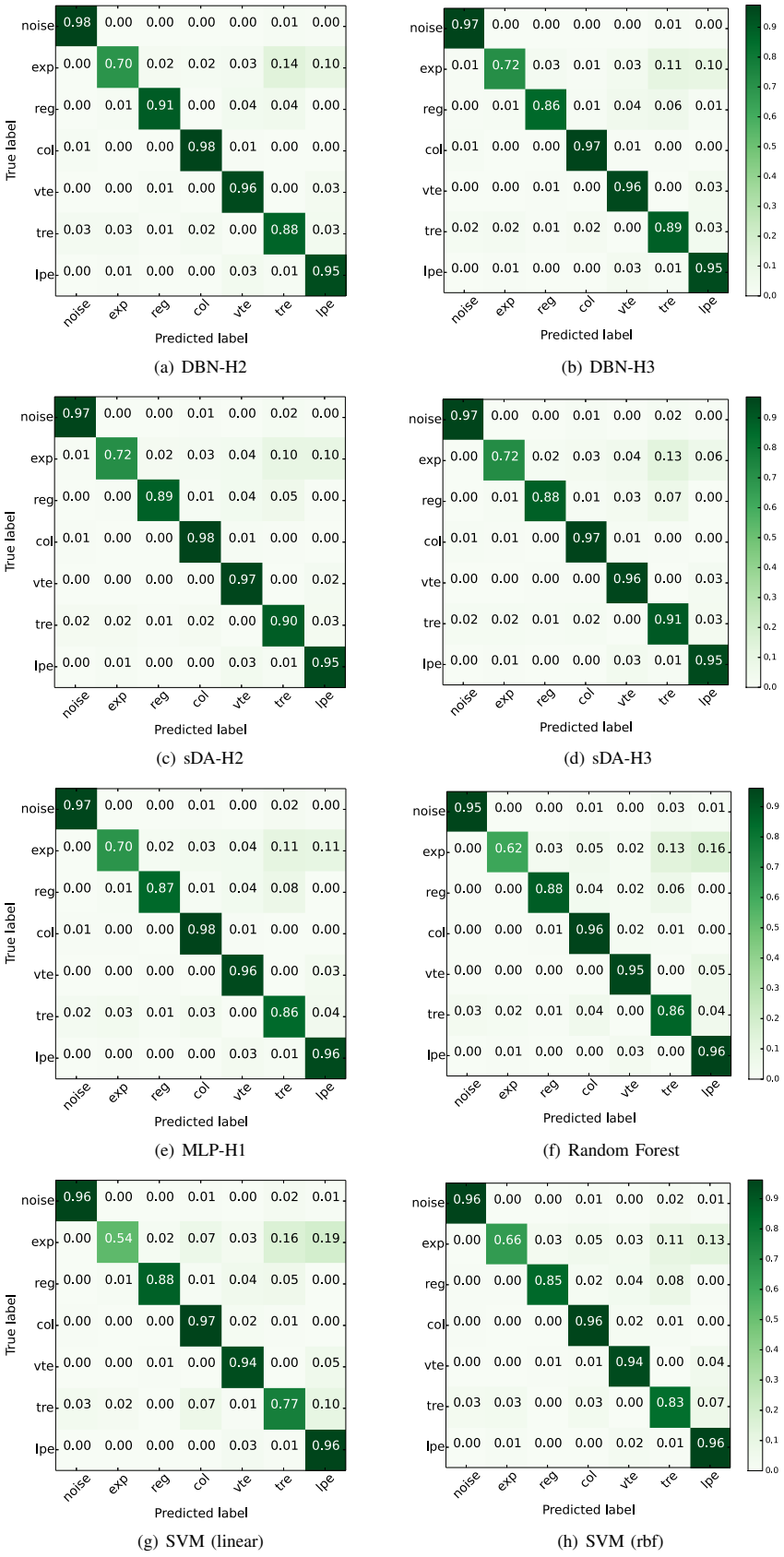


Figure 7. Normalized confusion matrices related to the implemented architectures. The results are over the whole set of tests.