UNIVERSIDADE FEDERAL DO CEARÁ

***CAMPUS* SOBRAL**

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

PABLO EDUARDO ESPINOZA LARA

AUTOMATIC MULTICHANNEL VOLCANO-SEISMIC CLASSIFICATION USING MACHINE LEARNING AND EMD

SOBRAL

2019

PABLO EDUARDO ESPINOZA LARA

AUTOMATIC MULTICHANNEL VOLCANO-SEISMIC CLASSIFICATION USING MACHINE LEARNING AND EMD

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em Engenharia Elétrica e de Computação do *Campus* Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia Elétrica e de Computação. Área de Concentração: Sistemas de Informação

Orientador: Prof. Dr. Carlos Alexandre Rolim Fernandes

Coorientador: Dr. Adolfo Inza Callupe

SOBRAL

2019

PABLO EDUARDO ESPINOZA LARA

AUTOMATIC MULTICHANNEL VOLCANO-SEISMIC CLASSIFICATION USING MACHINE LEARNING AND EMD

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em Engenharia Elétrica e de Computação do *Campus* Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia Elétrica e de Computação. Área de Concentração: Sistemas de Informação

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Carlos Alexandre Rolim
Fernandes (Orientador)
Universidade Federal do Ceará (UFC)

Dr. Adolfo Inza Callupe (Coorientador)
Instituto Geofísico del Perú (IGP)

Prof. Dr. Jarbas Joaci de Mesquita Sá Junior
Universidade Federal do Ceará (UFC)

Prof. Dr. Jérôme Igor Mars
Institut Polytechnique de Grenoble (Grenoble-INP)

To all those who sacrifice family time to study and research outside the country of origin, but know with pride, that studies and the intellect gained, is the best inheritance of our parents, and this cannot be stolen.

ACKNOWLEDGEMENTS

To Prof. Dr. Carlos Alexandre Rolim Fernandes for believing in me from the beginning and guiding me in the master's course.

To Dr. Adolfo Inza Callupe of the Geophysical Institute of Perú, for guiding me in the seismic and scientific field.

To Dr. Jérôme Mars from GIPSA-lab / France, and to Dr. Jean-Philippe Métaxian from IRD / France, for guiding me in the methodology in the present investigation.

To National Volcanological Center of the Geophysical Institute of Peru for providing the catalog with classified seismic events of the Ubinas volcano.

To my coordinator and also professor of the master's degree Prof. Dr. Iális Cavalcante for all the help and advice during my stay in Brazil.

To my professors of the master's degree, Dr. Jarbas Joaci, Dr. Rui Vigelis, Dr. Marcio Amora, for the teachings and willingness to solve all my doubts and questions.

To the Organization of American States (OAS), the Coimbra Group of Brazilian Universities (GCUB), and the Higher Education Personnel Improvement Coordination (CAPES), for funding this research through a scholarship.

To my mother Mercedes Lara, to my grandparents Marina and Leocadio, and to my aunt Lilian Lara, for the unconditional support throughout my life.

“The important thing is not to stop questioning. Curiosity has its own reason for existence. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries merely to comprehend a little of this mystery each day.”

(Albert Einstein)

RESUMO

Esta pesquisa propõe o design de um classificador automático usando a decomposição do modo empírico (EMD), juntamente com técnicas de aprendizado de máquinas para identificar os cinco tipos de eventos mais importantes do vulcão Ubinas, o vulcão mais ativo no Peru. O método proposto utiliza atributos dos domínios temporal, espectral e cepstral, extraídos da EMD dos sinais, bem como um conjunto de técnicas de pré-processamento e correção de instrumentos. Devido ao fato de atualmente sensores multicanais estarem sendo instalados em redes sísmicas em todo o mundo, a abordagem proposta utiliza sensores multicanal para realizar a classificação, ao contrário da abordagem usual da literatura de usar um único canal. O método apresentado é escalável para usar dados de várias estações com um ou mais canais. O método de análise de componentes principais (PCA) é aplicado para reduzir a dimensionalidade do vetor de características e a classificação supervisionada é realizada por meio de vários algoritmos de aprendizado de máquinas, sendo que a Máquina de Vetores de Suporte (SVM) fornece os melhores resultados. A investigação apresentada foi testada com um grande banco de dados que possui um número considerável de eventos de explosão, medidos no vulcão Ubinas, localizado em Arequipa, Peru. O sistema de classificação proposto alcançou uma taxa de sucesso superior a 90%.

Palavras-chave: Vulcão. Sinais sísmicos. Domínio cepstral. Domínio espectral. Domínio temporal. Decomposição do modo empírico. Aprendizado de Máquina.

ABSTRACT

This research proposes the design of an automatic classifier using the empirical mode decomposition (EMD) along with machine learning techniques for identifying the five most important types of events of the Ubinas volcano, the most active volcano in Peru. The proposed method uses attributes from temporal, spectral and cepstral domains, extracted from the EMD of the signals, as well as a set of pre-processing and instrument correction techniques. Due to the fact that multichannel sensors are currently being installed in seismic networks worldwide, the proposed approach uses a multichannel sensor to perform the classification, contrary to the usual approach of the literature of using a single channel. The presented method is scalable to use data from multiple stations with one or more channels. The principal component analysis (PCA) method is applied to reduce the dimensionality of the feature vector and the supervised classification is carried out by means of several machine learning algorithms, the support vector machine (SVM) providing the best results. The presented investigation was tested with a large database that has a considerable number of explosion events, measured at the Ubinas volcano, located in Arequipa, Peru. The proposed classification system achieved a success rate of more than 90%.

Keywords: Volcano. Seismic signals. Cepstral domain. Spectral domain. Temporal domain. Empirical mode decomposition. Machine Learning.

LIST OF FIGURES

Figure 1 – Structure of a vertical inertial sensor. 20

Figure 2 – Electromagnetic seismometer. 20

Figure 3 – PDF of the seismic noise. 21

Figure 4 – Seismic signal recorded by short period, long period, and broadband seismometer. 22

Figure 5 – Example of the IMFs of a signal HB. 24

Figure 6 – Flowchart of the EMD. 25

Figure 7 – Architecture of a MLP with 2 hidden layers. 27

Figure 8 – a) Hyperplane separation between the two classes. b) Support vectors and separation margin in SVM. 29

Figure 9 – SVM with soft margin. 31

Figure 10 – Satellite image of the Ubinas volcano on May 15, 2014, the image shows a pale ash plume generated by the volcano. 33

Figure 11 – Volcanological Seismic Network of Perú 33

Figure 12 – Telemetry network of Ubinas volcano. 34

Figure 13 – Simplified data acquisition system from the Ubinas volcano signals to the central base located in Lima. 34

Figure 14 – Example of the Ubinas seismic signal displayed in Lima (vertical component of the sensor). 35

Figure 15 – Time series samples of the five classes with their respective PSD. 36

Figure 16 – Map of the Ubinas volcano with the UB1, UB2, UB3 and UB4 permanent stations. 38

Figure 17 – Simplified block diagram of the proposed classification system. 39

Figure 18 – Feature extraction block diagram. 39

Figure 19 – Data transmission from the Ubinas volcano to the Arequipa observatory. . . 40

Figure 20 – Block diagram of signal conditioning. 40

Figure 21 – a) Time series of a VT signal with a sampling rate of 100 Hz and its spectrum. b) The same signal with a sampling rate of 41

Figure 22 – a) Time series of a EX signal with offset, and its spectrum. b) The same signal without offset, and its spectrum. 42

Figure 23 – Bode diagram of the CMG40T sensor transfer function. 43

Figure 24 – Simplified block scheme of the data acquisition system from the ground motion to the reception system. 44
Figure 25 – a) Time series of a EX signal. b) The same signal with instrumental correction. 45
Figure 26 – Time series of a LP and EX signal with instrumental correction. 46
Figure 27 – Example of Variance Contribution Ratio of the three selected IMFs of a signal VT. 47
Figure 28 – A sample of an LP signal, its three selected IMFs and their PSD. 48
Figure 29 – Timeseries of an example VT signal. 49
Figure 30 – Energy signal of the 3 selected IMFs of the example signal. 50
Figure 31 – Envelope of the 3 selected IMFs of the example signal. 50
Figure 32 – Timeseries, Periodogram, and Welch's PSD of the 3 selected IMFs of the example signal. 51
Figure 33 – 13 first MFCC of the 3 selected IMFs of the example signal. 52
Figure 34 – Principal component analysis of the feature vector. 54
Figure 35 – Learning curves of the machine learning techniques a) LDA, b) MLP, c) RF and d) SVM. 57
Figure 36 – Number of predictions per second of the machine learning models. 60

LIST OF TABLES

Table 1 – Calibration sheet for Guralp CMG40T 0.033 Hz - 50 Hz 43

Table 2 – Attributes extracted from the IMFs 53

Table 3 – Success rate for several classification techniques - single-channel and multi-channel cases 56

Table 4 – Confusion matrix with the true and predicted classes, using multiple channels and SVM - without EMD and with EMD (in parentheses) 58

Table 5 – Confusion matrix with true classes and predicted classes, using multiple channels and SVM - without instrument correction. 59

Table 6 – Success rate using attributes from different domains. 59

CONTENTS

1 INTRODUCTION . . . 13

1.1 State of the art . . . 14

1.2 Contributions . . . 16

1.3 Work Organization . . . 18

2 THEORETICAL BACKGROUND . . . 19

2.1 Sensors . . . 19

2.2 PSD . . . 22

2.3 EMD . . . 23

2.4 Classifiers . . . 26

2.4.1 Multilayer Perceptron . . . 26

2.4.2 Linear Discriminant Analysis . . . 26

2.4.3 Random Forest . . . 27

2.4.4 Support Vector Machine . . . 28

3 UBINAS DATABASE AND ACQUISITION SYSTEM . . . 32

3.1 Ubinas Volcano . . . 32

3.2 Ubinas Seismic Network . . . 32

3.3 Description of the Volcano Classes . . . 35

3.4 Database . . . 36

4 CLASSIFICATION SYSTEM . . . 39

4.1 Signal Conditioning . . . 39

4.1.1 Re-sampling . . . 40

4.1.2 Offset removal . . . 41

4.1.3 Instrumental correction . . . 42

4.2 Feature Extraction . . . 45

4.2.1 Choice of the Number of IMFs . . . 46

4.2.2 Calculation of Attributes . . . 48

4.2.2.1 Time domain . . . *49*

4.2.2.2 Spectral domain . . . *50*

4.2.2.3 Cepstral domain . . . *50*

4.2.2.4 List of Attributes . . . *51*

4.3 Principal Component Analisys (PCA) . . . 52

4.4 Classification 53

5 SIMULATION RESULTS 55

5.1 Multi-channel vs single-channel, and Machine Learning model choice . 55

5.2 EMD performance 56

5.3 Performance of instrumental correction 58

5.4 Performance of the temporal, spectral and cepstral domains 58

5.5 Latency in predicting new data 59

6 CONCLUSIONS AND PERSPECTIVES 61

REFERENCES 62

APPENDICES 67

APPENDIX A – Attributes 67

1 INTRODUCTION

Volcanoes are a latent threat to society and the occurrence of an eruption can be very early, with no indicator of date and size. Moreover many cities and towns are in areas of impact and high risk. History shows that an eruption is a continuous threat that can lead to a lot of human losses.

The recent eruption of the *Volcan de Fuego* volcano (June 2018, Guatemala) showed the catastrophic effects of a small volcanic eruption. Cataloged with an index of 3 on the Volcanic Explosive Index (VEI 3) scale, this eruption destroyed a large amount of infrastructure and killed more than 300 people. Volcanic activities have been a latent threat to humans since the existence of humanity. Indeed, many people live in areas of high risk, such as the city of Arequipa and the valleys of the volcanic chain in southern Peru, and the city of Yogyakarta, Indonesia, close to the Merapi volcano (RAHMAN *et al.*, 2016).

The magma interacts with the surrounding environment during its way to the crater in a system of ducts, causing disturbances when it is near the surface and generating seismic activity that can be observed by the seismic sensors. When the volcanic seismicity increases, the probability of eruption gets high. Although it can be just a mild activity, it can also be a catastrophic eruption. This question can be elucidated by analyzing seismic time series, through the classification of volcano-seismic patterns. The volcano-seismic signals can be categorized into 5 main classes (MCNUTT, 2005; WASSERMANN, 2012, Chapter 13): Long Period (LP), Tremors (TR), Explosion (EX), Volcano-Tectonic (VT) and Hybrid (HB).

Seismicity is an important parameter to distinguish the active volcano manifestation. Depending on its activity, seismicity can be used as an important indicator for the prediction of volcanic events, as in (CHOUET, 1996), which uses the LP-type signals to make a prediction of an eruption.

Thanks to the advance of technology, there are currently more and more volcanoes monitored with seismic networks. A large amount of seismic data is observed worldwide and the analysis of these time series can be used to predict or detect the eruptive state of volcanoes. However, in many places, this data is still classified manually, which can lead to errors or delays in event detection.

The present research is developed with data collected in the Ubinas volcano, located in Arequipa city, in Perú, whose catalog is prepared by experts of the National Volcanological Center of the Geophysical Institute of Peru (IGP). The IGP is continuously monitoring the seismic activity of the Ubinas volcano.

Just to get an idea of the amount of volcano-seismic events generated by the Ubinas volcano, the daily average of events for the months of January, February, March and April was 44.5, 164.5, 382.8, 462.0, respectively. This means that the experts in charge of generating reports and the writing catalogs had to carry out 462 analyses of the seismic signals per day in the month of April 2014, which requires a great demand for analysis and personnel, which can lead to errors in the classification, besides not being an optimal method.

The present investigation proposes a solution to this problem using signal decomposition techniques, as well as artificial intelligence to perform an automatic classification, in order to generate an automatic catalog.

1.1 State of the art

The analysis of seismic signals is a very important issue, because the study of these types of signals includes earthquakes, eruptions, nuclear tests, which involve people's lives. In addition, due to the study of seismic signals we have a better idea of the internal Earth structure and composition. One of these applications is the use of seismic reflection for the study of the subsurface in order to find oil, as in (ZHANG *et al.*, 2006).

In the field of earthquakes signal analysis, there are many investigations in the literature, such as (ALLEN, 1978), which uses the famous and traditional earthquake detection method STA/LTA (Short-term average/Long-term average), based on the algorithm that compares the ratio between STA and LTA of the seismic signal with a threshold value. Moreover, (VAEZI; BAAN, 2015) used the power spectral density for the detection of microseismic events, while (BAILLARD *et al.*, 2014) presented a method for the automatic detection of Primary (P) and Secondary (S) seismic waves using kurtosis-derived characteristic functions with three component multichannel seismic signals.

Science about seismic waves advanced so much that we can estimate the damage caused by earthquakes, estimating seismic intensities, as in (HOSOKAWA *et al.*, 2009), which uses synthetic-aperture radar (SAR) images, or the methodology that the United States Geological Survey (USGS) uses, based on the estimation of intensities using the modified Mercalli scale, as in (WALD *et al.*, 2006).

Decomposing the seismic signal is an approach that has been very successful in seismology, such as the compression of seismic data. It is known that a seismic database is very large, since currently the seismic sensors are generally configured to record data from 20 Hz to 200 Hz. In this sense, decomposing the signal into significant minor functions provides a solution, as in (NUHA; SUWASTIKA, 2015), which uses the PCA with Fractional Fourier Transform to achieve the compression.

In the literature, there are also research about optimizations of the seismic resolution through wavelets decomposition, as in (DU; ZHANG, 2010a), which proposes a methodology to optimize the resolution with seismic data in volcanic rock, using the signal decomposition by wavelets, obtaining a better resolution.

Decomposition applications are also found using EMD and Local Mean Decomposition (LMD), as in (YU; ZHANG, 2017), which uses EMD and LMD to denoise the sensible earthquake signal. Even the use of EMD has been used to locate nuclear experiments, as in (CHILO *et al.*, 2008), and to develop a monitoring and precaution system for pipelines security, as in (WEN; SUN, 2010).

The artificial intelligence (AI) applied to the geophysics has given excellent results, both for regression and classification problems. Problems involving seismic vulnerability in buildings have been solved using regressions with Machine Learning (ML) models, such as (PANAGIOTA *et al.*, 2012), which uses support vector regression (SVR) to estimate building vulnerability. The use of SVR for reconstructing seismic data from under-sampled or missing traces is also found in the literature, as in (JIA; MA, 2017).

In classification, many works proposed systems for classifying seismic signals, such as (BENBRAHIM *et al.*, 2007), which uses time-frequency representations to classify local earthquakes, far earthquakes and chemical explosions, and (GROSS; RITTER, 2009), that uses power spectral density (PSD) spectrograms to classify urban seismic noise. Moreover, (YILDIRIM *et al.*, 2011) used neural networks to classify earthquakes and quarry blasts, while (KISLOV; GRAVIROV, 2017) presented a method for automatic identification of noisy seismic events. Also, new techniques are found in the literature for classifying seismic signals with machine learning models, such as the use of the cepstral domain with SVM in (ZHOU *et al.*, 2012) or with Hidden Markov Model (HMM) in (PENG *et al.*, 2019). Besides, a three-channel seismic signal decomposition using wavelets with kernel ridge regression is presented in (RAMIREZ; MEYER, 2011).

In the case of volcano-seismic signals, research on volcanic seismology is the most complex topic that seismologists investigate. This is because seismic sources in volcanoes include movements of liquids, solids and gases. Moreover, the propagation paths are extremely heterogeneous and anisotropic, for this reason, scientists must conduct a thorough study and design many experiments to understand the physics of the volcano, as mentioned in (AKI, 1992).

In this type of signals, machine learning classification is also a promising method. For instance, the works (SCARPETTA *et al.*, 2005; CURILEM *et al.*, 2009; BUENO *et al.*, 2018) use time-frequency features with neural networks, while (OHRNBERGER, 2001; BENITEZ *et al.*, 2006; GUTIÉRREZ *et al.*, 2009) use HMM for classifying volcano-seismic signals. In addition, the work (DU; ZHANG, 2010b) uses wavelet decomposition as part of the classifier and (MALFANTE *et al.*, 2018) uses attributes in the temporal, spectral and cepstral domains for the extraction of attributes, along with the SVM classifier.

1.2 Contributions

The present work presents the design of a classifier for identifying the aforementioned five most important types of events of a volcano, with a methodology that can be easily implemented in monitoring centers in real time. The objective is to automatically classify volcano-seismic signals and generate a catalog of time series of this type of signals, with the aim of finding a seismic pattern associated with the magma behavior.

The empirical mode decomposition (EMD) (HUANG *et al.*, 1998) is used to include more physical contrast in the machine learning algorithm, as the EMD is a natural adaptive decomposition method that is well-suited for non-stationary signals, as in the case of the seismic signals. The basic idea is to decompose multichannel seismic signals into components that occupy different frequency bands, called intrinsic mode functions (IMFs). In the present dissertation the natural characteristics of the IMFs are exploited to generate the feature vector using attributes in temporal, spectral and cepstral domains, in order to obtain a better representation of the different types of signals to be classified.

Contrary to the usual approach of the literature of using only a single channel, the proposed approach makes use of a multichannel triaxial sensor to perform the classification. Indeed, apart from the vertical channel, the east and north channels are also considered. This kind of triaxial sensors is currently being installed in seismic networks worldwide. The presented methodology is scalable to use data from multiple stations with one or more components, since our database includes data collected simultaneously in more than one seismic station. Logically, with more information added by multiples sensors and channels, the efficiency of the result is improved. The presented classification system also proposes to perform a set of processing and instrument correction operations in the original seismic signals captured by the sensors (signal conditioning), in order to transform the signals of each channel in meters per seconds (HAVSKOV; ALGUACIL, 2004, Chapter 6).

The proposed automatic classification system can be summarized in the following steps. Firstly, a signal conditioning is performed on the seismic signals, including offset elimination, instrumental correction, among other operations. The EMD is then calculated, with the 3 most significant IMFs being selected. Next, the extraction of the attributes in temporal, spectral and cepstral domains is performed. After, the principal component analysis (PCA) method is applied to reduce the dimensionality of the feature vector. Finally, the supervised classification is carried out by means of several machine learning techniques.

As mentioned in previous lines, this research is developed with a large database collected from two stations of the Ubinas volcano, located 70 km northeast of the city of Arequipa, in Peru. The data catalog was made by experts of the National Volcanological Center of the Geophysical Institute of Peru (IGP). The catalog of the Ubinas volcano showed, in the last years, a high number of volcano-seismic events, half a thousand events per day, which represents a difficult job for the volcanologist experts. A relevant characteristic of the database is its relatively high number of explosion events, when compared with databases of other works.

The main original contributions of this dissertation can be summarized as follows:

- The inclusion of a multichannel sensor and data from two seismic stations to model the behavior of the volcano, contrary to previous works that use only a single channel.
- The use of instrumental correction in order to make the volcanic classifier independent of the type of sensors used and to give the energy of the signals a physical sense.
- The use of the EMD along with machine learning to classify the events of a volcano.
- The use of database with a high number of explosions events, which can be considered the most important event to be detected.
- The presentation of simulation results showing an excellent performance of the proposed classifier when compared with other approaches. Indeed, the proposed classification system achieved a success rate higher than 90% when the SVM technique is used.

This dissertation has originated a publication at the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, with the title: "Automatic Multichannel Volcano-Seismic Classification Using Machine Learning and EMD".

1.3 Work Organization

The rest of the work is organized as follows.

Chapter 2 details the seismic data acquisition system of the Ubinas volcano, detailing the transmission-reception system and describing the positioning of the sensors. Moreover, the description of the volcano-seismic classes to be classified is presented and the database used in this investigation is detailed. A brief history of seismic instrumentation is also mentioned.

Chapter 3 presents the design of the classification system, in other words, it describes the proposed methodology, starting with the conditioning of the multi-channel seismic signal, followed by the generation of the feature vector. In this part, the use of the EMD in order to extract attributes from each decomposed signal is explained as part of the feature extraction block. The extraction of attributes in the temporal, spectral and cepstral domain is also described, as well as the use of the PCA dimensionality reduction method. Finally, several machine learning models used to perform automatic classification are reported.

Chapter 4 presents the numerical results of the investigation, the confusion matrix of the proposed model being the main figure of merit used. Many simulation scenarios were tested in order to obtain the best results.

Finally, Chapter 5 details the main conclusions and perspectives that are available for future work.

2 THEORETICAL BACKGROUND

This chapter presents a survey of the main methods, models and techniques used in this work for developing the proposed volcano-seismic classification system. In Section 2.1, a brief history of seismic sensors is described. In Section 2.2, the estimation of the PSD by Welch's method is detailed. In Section 2.3, some works related to EMD and its computation are presented. Finally, Section 2.4 describes the supervised machine learning classifiers used in this investigation.

2.1 Sensors

This section presents a brief history of seismic sensors until reaching triaxial broadband seismic sensors, which are used in this investigation. It also describes the importance of three component sensors in a seismic classification. As mentioned in the research methodology in Section 1.2, the present investigation uses multichannel seismic signals to perform the classification, contrary to the usual approach of using single channel sensors.

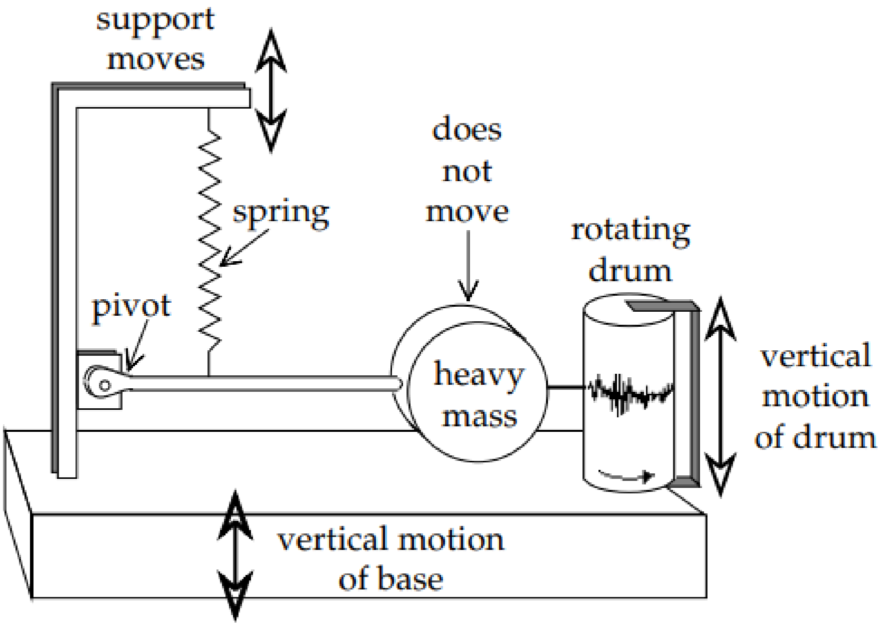
Seismic sensors are transducers that convert ground motion into an electrical velocity signal, but it was not always so. History shows that, from the 1950s until the 1970s, sensors were used to capture the ground motion through springs attached to a heavy mass, as mentioned in (OKUBO *et al.*, 2014). As the ground moves, the spring exerts a force that moves the heavy mass which is attached to a pen, recording the signal in a rotating drum, as shown in Figure 1, which shows the structure of a vertical inertial sensor. Note that this figure only captures the upward and downward movement of the ground, i. e. the vertical signal.

This type of sensors are based on the principle of inertia, since stationary objects, such as heavy mass, will remain unmoved unless an external force makes them move. This type of sensors are commonly called inertial seismometers.

In the late 1970s, the first magnetic sensors began to be used, which consisted of joining the spring with a coil and, through the force exerted by the spring, the coil moves generating a magnetic flux. The variations of the magnetic flux generate a potential difference in the electrical circuit. Figure 2 shows a simplified scheme of a electromagnetic seismometer. Note that this voltage signal is still analog signal. A recorder or digitizer is then necessary to perform the digital signal processing.

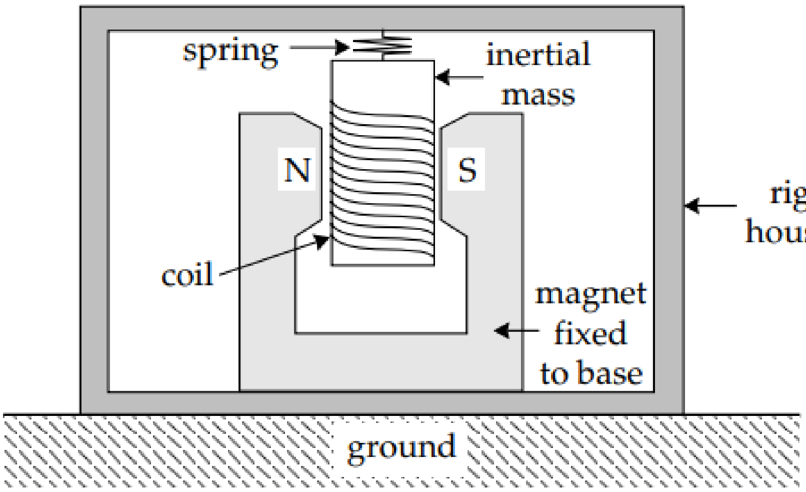
On the other hand, the frequency range of a seismic signal that a seismometer can

Figure 1 – Structure of a vertical inertial sensor.



Source: Adapted from (LOWRIE, 2007).

Figure 2 – Electromagnetic seismometer.



Source: Adapted from (LOWRIE, 2007).

capture plays a very important role in the analysis of the seismic signals.

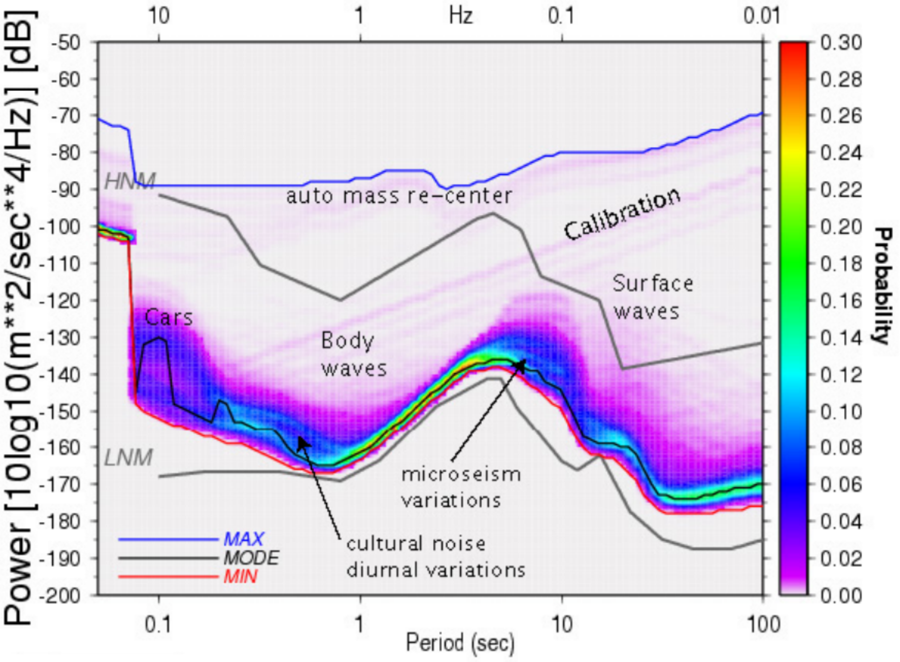
In this sense, the first sensors were short-period seismometer and long-period seismometer. The short-period seismometer works for periods in the range of 0.1 - 1 s (or its equivalent in frequency 1 - 10 Hz), while the long-period seismometer works for periods in the range of 10 - 100 s (frequencies in the range of 0.01 - 0.1 Hz).

The problem with this type of sensors is when recording data in the frequency range of 0.1 - 1 Hz. This frequency range is usually caused by microseisms, as shown in Figure 3, which is generated by a nearly continuous succession of small ground movements, as said in

(LOWRIE, 2007). Figure 3 shows the computation of many PSDs of a seismic noise, obtaining ranges where noise is caused by different sources. This type of graph is called Power Density Function (PDF), which is generated by a histogram of PSDs.

Microsisms noise can be caused by vehicular traffic, animals, factories, etc., this is called cultural noise. It can also be caused by storms, tides, among others. For this reason, collection of microsisms is essential. Figure 3 shows the computation of many PSDs of a seismic noise, obtaining ranges where noise is caused by different sources. This type of graph is called Power Density Function (PDF), which is generated by a histogram of PSDs.

Figure 3 – PDF of the seismic noise.

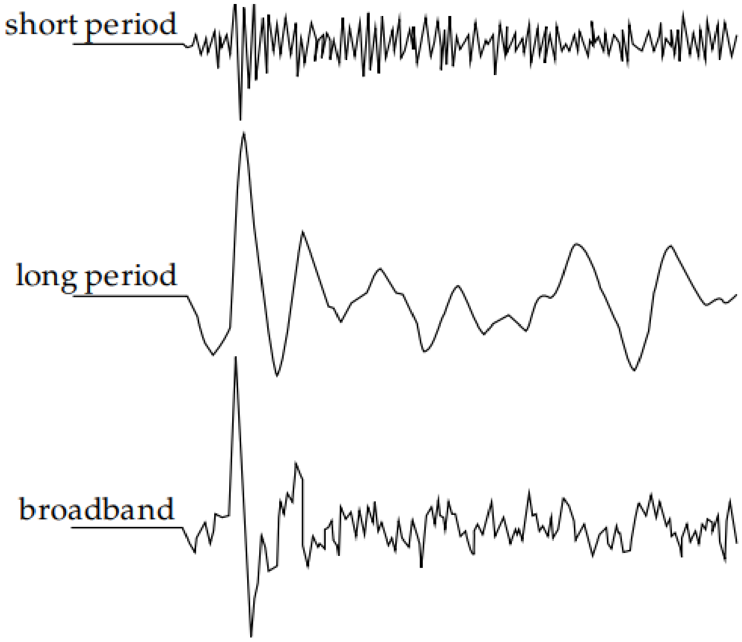


Source: Adapted from (BORMANN; WIELANDT, 2012).

Short-period sensors are dominated by high frequency signals, while long-period sensors tend to smooth noise without being able to detect it, as shown in Figure 4. For this reason, broadband sensors were created because they have high sensitivity in a very wide frequency spectrum. Figure 4 shows a seismic signal collected by a short-period, long-period, and a broadband sensor.

In order to capture the seismic waves in all directions, the current seismometers contain embedded sensors oriented in three orthogonal axes, the east (BHE), north (BHN) and vertical axis (BHZ). To clarify the importance of using multichannel sensors, let’s imagine that we recorded the seismic signal in a 1-channel vertical sensor. The signal that will be recorded in, is in fact, the projection of the seismic wave in the vertical channel. If we want to obtain all the

Figure 4 – Seismic signal recorded by short period, long period, and broadband seismometer.



Source: Adapted from (LOWRIE, 2007).

possible information, we need the projection of the seismic wave in the other 2 axes.

In this investigation, 3-component broadband sensors are used and it will be shown, by mean of simulation results, that the use of more than 1 channel improves the result in the automatic classification.

2.2 PSD

The PSD is defined by the Wiener-Khintchine theorem, as the Fourier transform of the autocorrelation of the signal $x[n]$, and is given by the following equation:

$$S(\omega) = \sum_{k=-\infty}^{+\infty} \phi_{xx}[k] e^{(-j\omega k)}, \tag{2.1}$$

where $S(\omega)$ is the power spectral density and $\phi_{xx}[k]$ is the autocorrelation of the signal $x[n]$. The following equation is used to compute the autocorrelation:

$$\phi_{xx}[k] = \frac{1}{N} \sum_{n=-\infty}^{+\infty} x[n]x[n+k], \tag{2.2}$$

where N is the number of samples of $x[n]$.

Using (2.2) and (2.1), the following expression is obtained to estimate the PSD, called the Periodogram (ROWELL, 2008).

$$S(\omega) = \frac{1}{N} |X(\omega)|^2, \tag{2.3}$$

where $X(\omega)$ is the Fourier transform of $x[n]$.

To enhance the magnitudes of the frequencies of interest through the PSD, the Welch's method is used in this work to compute the PSD. This method consists of the following steps, as mentioned in (PARHI; AYINALA, 2014):

(a) The input signal $I_i[n]$ of length N, is divided into L segments of length M, overlapping in $M-S$ points. For example, if $S=M/2$, then $(M-S)/M$ would be 0.5, which means a 50% of overlapping. In this investigation, an overlap of 75% is used. The segmented signals are labeled as: $p_d[n]$, where $d \in [1, L]$.

(b) Multiply the segmented signals by a window function $w[n]$, in order to avoid discontinuities at the beginning and end of the segmented signals. In this research, the Hanning window is used:

$$s_d[n] = p_d[n].w[n]. \tag{2.4}$$

(c) Compute the periodogram of each windowed signal $s_d[n]$:

$$P_d(\omega) = \frac{|s_d(\omega)|^2}{W}, \tag{2.5}$$

where W is the power of the window $w[n]$:

$$W = \sum_{n=0}^{M-1} |w[n]|^2. \tag{2.6}$$

(d) Repeat the steps from (b) to (c) for each segment, to obtain all the periodograms and average them:

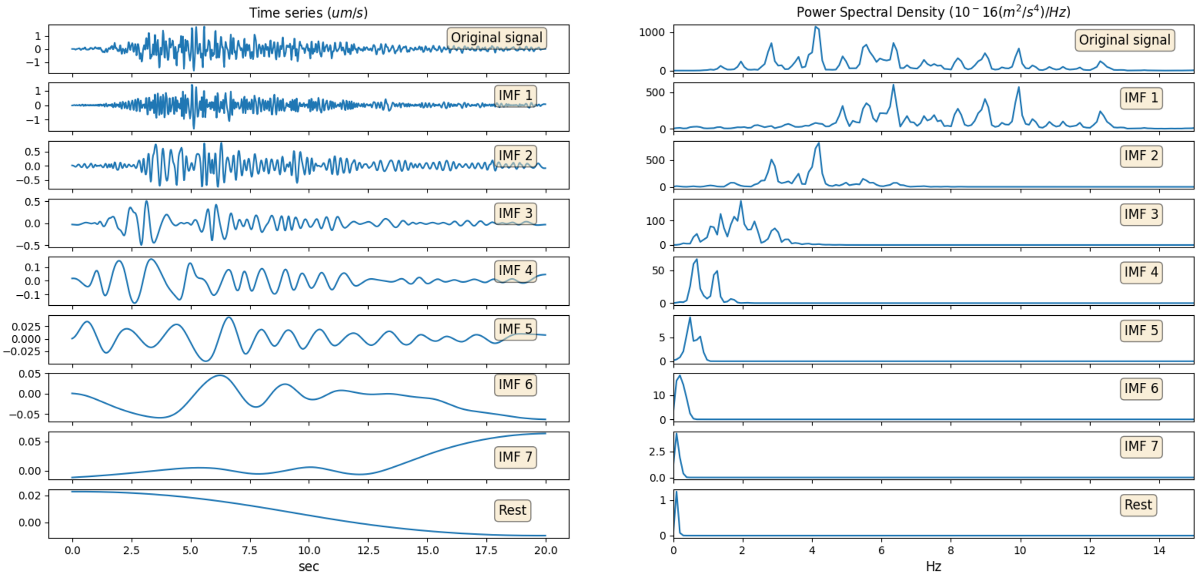
$$PSD = \frac{\sum_{l=1}^{L} P_d(\omega)}{L}. \tag{2.7}$$

2.3 EMD

The EMD is a self-adaptive filter developed by Huang in 1998 for analysis of nonlinear and non-stationary signals (HUANG *et al.*, 1998). This method has been applied in the study of gravitational waves (CAMP *et al.*, 2007), noise analysis (KOMATY *et al.*, 2012), acoustic signals (GRULIER *et al.*, 2008), image processing (NUNES; DELÉCHELLE, 2009), etc. The use of EMD method is relatively new in seismology. For instance, the works (INZA, 2013), (HAN *et al.*, 2018), (SAEED, 2011), (HAN, 2014) recently applied the EMD for this type of signal. Up to now, to the best of our knowledge, the use of EMD for classification of volcano-seismic signals using machine learning was not found in the literature.

The principle of the EMD is to decompose a signal, through a sifting process, into different modes called IMFs. The term IMF is due to the fact that the EMD does not use a fixed type of basis function to compute the decomposition, as it usually happens in signal transforms, like, for instance, the Wavelet and Fourier transforms. On the contrary, the EMD performs the decomposition based on natural or intrinsic characteristics of the signal. The IMFs occupies different frequency bands, the first IMFs containing roughly high frequencies and the last ones containing roughly low frequencies. In the present work, the natural characteristics of the IMFs are exploited to generate the feature vector, in order to obtain a better representation of the different types of signals to be classified. An example of the use of EMD in a Hybrid (HB) signal is shown in Figure 5, which shows a HB signal, its first seven IMFs and the residual signal, with their respective PSDs.

Figure 5 – Example of the IMFs of a signal HB.



Source: Own author.

The steps of the EMD of a discrete-time signal $x[n]$ are illustrated in Figure 6, with the following remarks:

- In this work, we used cubic spline to interpolate the upper envelope and lower envelopes.
- The conditions for $h_1[n]$ to be an IMF are the following:

  a) the number of extrema and the number of zero-crossing points must be equal or differ at most by one;

  b) $m_k[n]$ must be 0 at some point.

- The standard deviation (SD) is defined as:

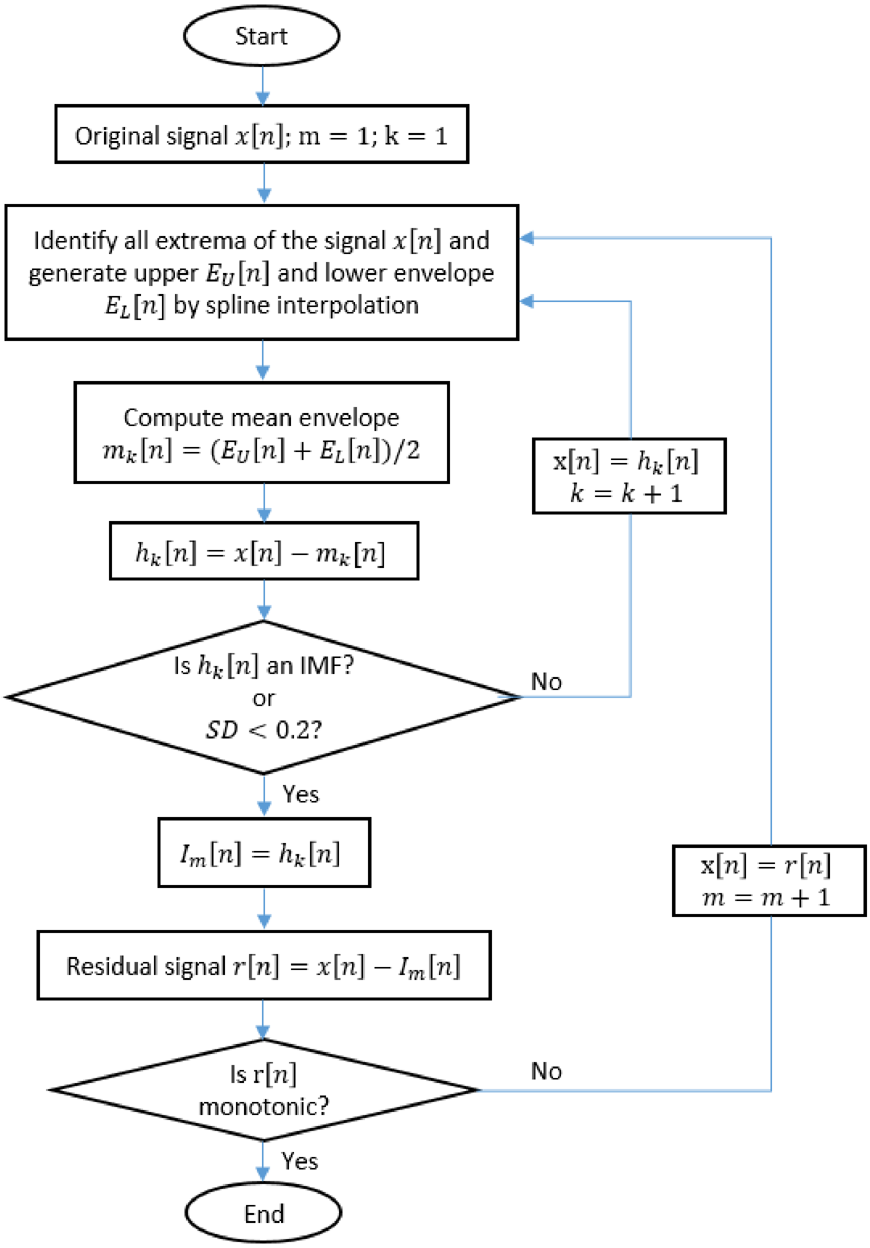
$$SD = \sum_{k=1}^{N} \frac{|h_{k-1}[n] - h_k[n]|^2}{h_{k-1}^2[n]}. \tag{2.8}$$

- $I_m[n]$ is the m^{th} IMF of $x[n]$.
- The original signal $x[n]$ can be represented as follows:

$$x[n] = \sum_{m=1}^{M} I_m[n] + r[n], \tag{2.9}$$

where M is the number of IMFs and $r[n]$ is the final residual signal.

Figure 6 – Flowchart of the EMD.


Start
Original signal $x[n]$; m = 1; k = 1
Identify all extrema of the signal $x[n]$ and generate upper $E_U[n]$ and lower envelope $E_L[n]$ by spline interpolation
Compute mean envelope $m_k[n] = (E_U[n] + E_L[n])/2$
$h_k[n] = x[n] - m_k[n]$
Is $h_k[n]$ an IMF? or $SD < 0.2$?
No
$x[n] = h_k[n]$ $k = k + 1$
Yes
$I_m[n] = h_k[n]$
Residual signal $r[n] = x[n] - I_m[n]$
Is r$[n]$ monotonic?
No
$x[n] = r[n]$ $m = m + 1$
Yes
End

Source: Own author.

The problem now is to choose how many IMFs will be used. Since the EMD method is a natural decomposition, the number of IMFs will depend on each signal, so it is essential to fix a number of IMFs to perform the extraction of attributes of each IMF. This issue will be addressed in subsection 4.2.1.

2.4 Classifiers

This section details the supervised machine learning classifiers used in this investigation.

2.4.1 Multilayer Perceptron

Multi-Layer Perceptron (MLP) is the most used Artificial Neural Network (ANN), generally using back-propagation training (LINS; LUDERMIR, 2005). Although the architecture of a MLP has given good results in the literature, the disadvantage in this model is that the convergence tends to be slow, leading to a high computational cost in the training stage. This fact is a significant problem in a real-time classification system if it is required to retrain this model for each new seismic event in the volcano, since many times there is a large number of events per day. This model may also be subject to the possibility of being trapped in a minimum unwanted location, causing problems in the classification.

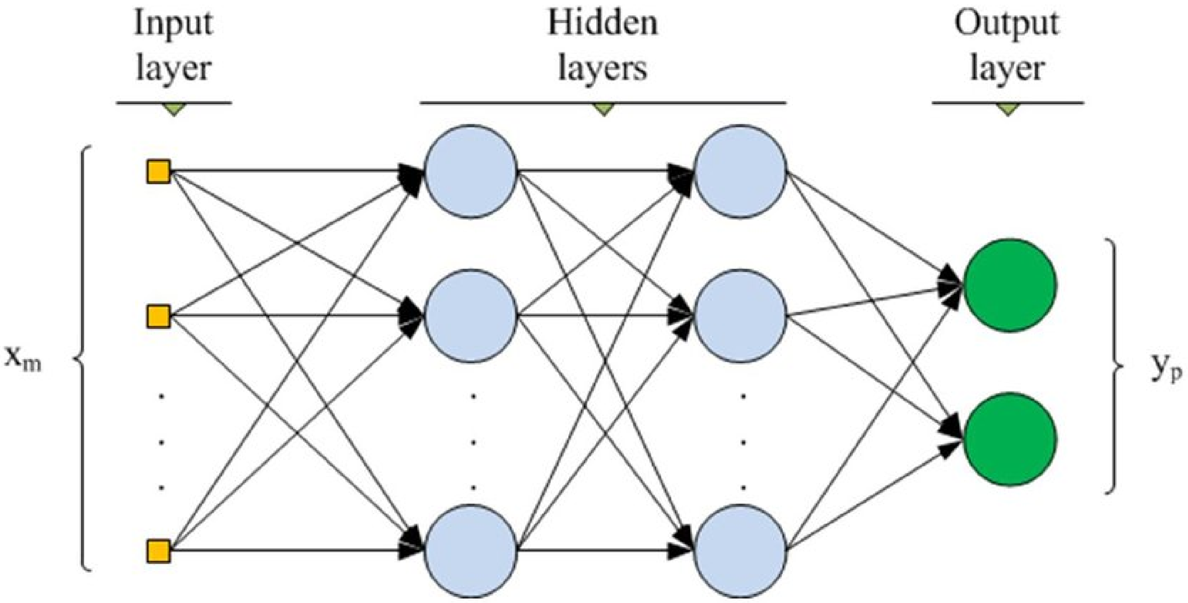
A MLP model is divided into three parts: input layer, hidden layers, and output layer. The input layer receives the feature vector, in this investigation, the input layer is fed by the components generated by PCA. The hidden layers are responsible for performing the non-linear processing of the feature vector, and the output layer performs the final processing of the data classification.

To achieve a significant result, parameters such as connection weights play an important role in this model. A large number of connections do not guarantee an optimal result since it can lead to overfitting results, while a low number of neurons can lead to a model that is unable to solve the classification problem.

2.4.2 Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) classifier, which is a generalization of the Fisher classifier (1936), assumes that all classes have the same covariance matrix. In that sense,

Figure 7 – Architecture of a MLP with 2 hidden layers.



Source: Adapted from (SUZUKI, 2013).

starting from the equation of the posteriori probability $P(c_i|\vec{x})$ that the vector $\vec{x}$ belongs to the class c_i, using Bayes theorem, as in (BACKES; SÁ JUNIOR, 2016), the following expression can be obtained:

$$P(c_i|\vec{x}) = -\frac{1}{2}ln|\Sigma_i| - \frac{1}{2}(\vec{x}-\vec{\mu}_i)^T(\Sigma_i)^{-1}(\vec{x}-\vec{\mu}_i) + ln(P(c_i)), \tag{2.10}$$

where μ_i is the vector mean of class c_i, $|\Sigma_i|$ its covariance matrix and $P(c_i)$ is a priori probability of the class c_i.

As all classes have the same covariance matrix, then $|\Sigma_i| = |\Sigma|$, and the first term of (2.10) can be neglected, obtaining a linear separator given by the following equation:

$$P(c_i|\vec{x}) = -\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T(\Sigma)^{-1}(\vec{x}-\vec{\mu}_i) + ln(P(c_i)). \tag{2.11}$$

2.4.3 Random Forest

Random Forest (RF) is a classifier created by Breiman (BREIMAN, 2001), that can be used as a classifier or as a regression method. As a classifier, RF makes a prediction using a combination of decision trees. In this model, each tree depends on an independent random vector whose distribution is the same for each tree. To know which class is the winner, a vote is taken among all the trees and the class with the highest number of votes is the winner.

In summary, the RF procedure can be summarized as follows:

- From the training dataset, a group of random samples are selected using replacement sampling to create different data groups of the same size (forest). This method of creating groups through samples chosen from the dataset with replacement is called bagging.
- A decision tree is created with each group of data, obtaining different trees, since each group contains different data.
- When creating the trees, random features are chosen in each node of the tree, allowing the tree to grow in depth, that is without pruning.
- The classification of the trees is carried out and the votes obtained in each tree are calculated, winning the one with the majority vote.

The use of bagging in the RF algorithm improves variance by reducing correlation between trees and it also improves the accuracy of decision tree algorithm, but the model interpretation is lost, since RF is seen as a black box.

2.4.4 Support Vector Machine

Support Vector Machine (SVM) is a binary supervised machine learning classifier created by Boser, Guyon and Vapnik (BOSER *et al.*, 1992). The SVM classifier performs the separation of the classes through hyperplanes that are optimized for generating the greatest possible distance between the classes.

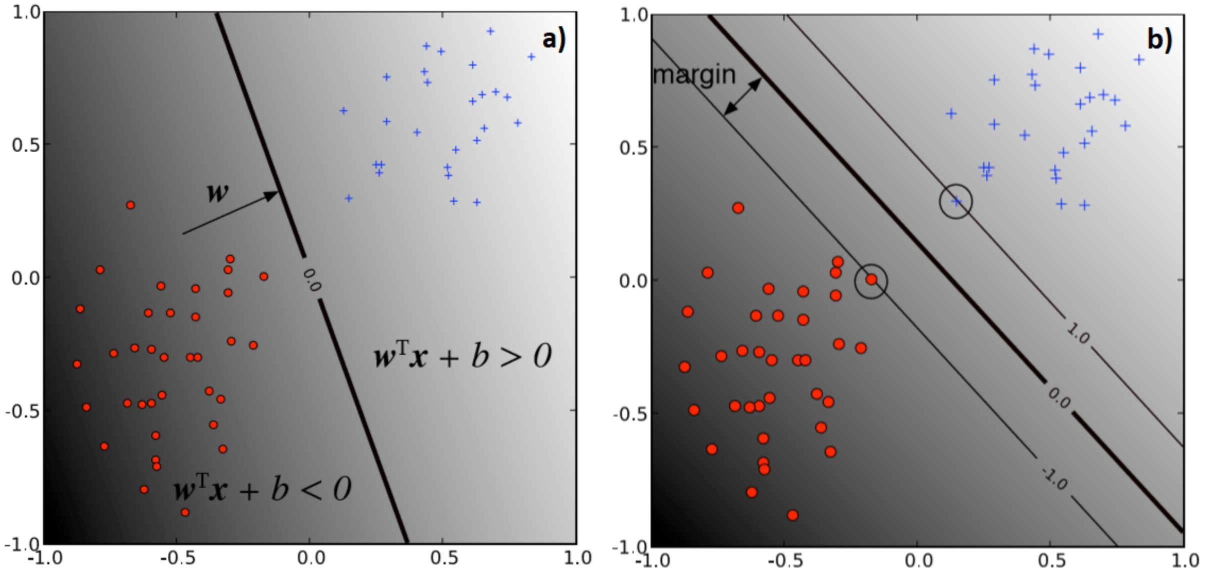
Let us consider two linearly separable classes as shown in Figure 8a, which shows one of the separation hyperplanes between the two classes, labeled $y = +1$ for blue circles and $y = -1$ for red circles. The equation for a separating hyperplane is given by the following equation:

$$f(x) = \mathbf{w}^T\mathbf{x} + b = 0, \tag{2.12}$$

where $\mathbf{w} = [w_0, ..., w_{n-1}]^T$ is the vector of weights and b is the bias. Then for $\mathbf{w}^T\mathbf{x} + b > 0$, the circles will be blue, and for $\mathbf{w}^T\mathbf{x} + b < 0$ the circles will be red.

However there are infinite hyperplanes that can separate the classes. The objective of SVM is to find the optimal hyperplane, that is, find the hyperplane in which the distance of the hyperplane to both classes is as large as possible. To this end, the concept of separation margin is introduced, defined as the distance between the separating hyperplane towards the closest sample to each class, as shown in Figure 8b. These points are called the support vectors and they verify the following conditions:

Figure 8 – a) Hyperplane separation between the two classes. b) Support vectors and separation margin in SVM.



Source: Adapted from (BEN-HUR; WESTON, 2010).

- For the support vectors of class $y = +1$ (blue):

$$\mathbf{w}^T \mathbf{x}_{sup+} + b = 1, \tag{2.13}$$

- For the support vectors of class $y = -1$ (red):

$$\mathbf{w}^T \mathbf{x}_{sup-} + b = -1, \tag{2.14}$$

where $\mathbf{x}_{sup+}$ represents the support vector of the class $y = +1$, and $\mathbf{x}_{sup-}$ the support vector of the class $y = -1$.

We can express $\mathbf{x}_{sup+}$ as a linear combination of $\mathbf{x}_{sup-}$ and $\mathbf{w}$, as follows:

$$\mathbf{x}_{sup+} = \mathbf{x}_{sup-} + c\mathbf{w}, \tag{2.15}$$

where c is a constant and the separation margin τ is:

$$\tau = ||c\mathbf{w}||, \tag{2.16}$$

where $||.||$ the norm vector.

Replacing (2.15) in (2.13):

$$\mathbf{w}^T (\mathbf{x}_{sup-} + c\mathbf{w}) + b = 1, \tag{2.17}$$

simplifying:

$$c||\mathbf{w}||^2 + \mathbf{w}^T \mathbf{x}_{sup-} + b = 1, \tag{2.18}$$

using (2.14) in (2.18):

$$c = \frac{2}{||\mathbf{w}||^2}. \tag{2.19}$$

Finally, the separation margin is obtained by replacing (2.16) in (2.19) and its value is:

$$\tau = \frac{2}{||\mathbf{w}||}. \tag{2.20}$$

To get the optimal hyperplane, τ should be as large as possible so that the distance between class $y = +1$ and $y = -1$ is as far apart as possible. In that sense, we must maximize $\frac{2}{||\mathbf{w}||}$ which is equivalent to minimize $\frac{||\mathbf{w}||^2}{2}$.

Then, the following equation should be minimized.

$$minimize_{\mathbf{w}} \ \frac{||\mathbf{w}||^2}{2}, \tag{2.21}$$

subject to the condition that there are no samples between the hyperplanes $y = +1$ and $y = -1$, that is:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \tag{2.22}$$

where $\mathbf{x}_i$ is the vector of samples and y_i its label.

Problems of this type can be solved through Lagrange multipliers, whose function to minimize is:

$$\mathscr{L}(\mathbf{w}, b, \alpha) = \frac{||\mathbf{w}||^2}{2} - \sum_{i=1}^{n} \alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1), \tag{2.23}$$

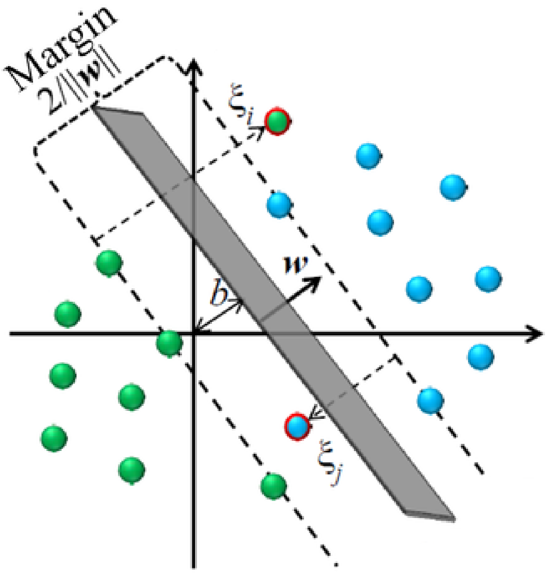
where α_i are the Lagrange multipliers.

To minimize this equation, it is necessary to use the primal or dual problem subject to the Karush-Kuhn-Tucker conditions. If we use a kernel function, we must use the dual problem. The LIBSVM library (CHAN; LIN, 2011) facilitates the calculation of the optimal hyperplane.

In the case of two classes that are not linearly separable, as in the case of Figure 9, which shows margin violation and misclassified in some samples, slacks variables ξ are introduced to model this relaxation SVM. In this sense, the cost function of the SVM with the relaxing variables ξ can be rewritten as:

$$minimize \ \frac{||\mathbf{w}||^2}{2} + C\sum_{i=1}^{n} \xi_i, \tag{2.24}$$

Figure 9 – SVM with soft margin.



Source: Adapted from (KHAN *et al.*, 2016).

subject to the conditions:

$$\begin{aligned} y_i(\mathbf{w}^T\mathbf{x}_i+b) &\geq 1-\xi_i, \\ \xi_i &\geq 0. \end{aligned} \tag{2.25}$$

The term C is called Penalty Parameter and indicates how much an erroneous classification of the training data should be avoided, that is, for large values of C the margin will be smaller, while for small values of C the margin will be larger, even if there is poorly classified data. Equations (2.24) and (2.25) are solved using Lagrange multipliers.

Since SVM is a binary classifier, strategies for multi-class classification such as One vs. One and One vs. All are used, as in (MILGRAM *et al.*, 2006). In addition, to improve the linear separation between classes, kernel functions are used, which help the separation by representing the data in a higher dimension. In this research, different kernel functions were tested, such as the linear kernel, Radial Basis Function (RBF), polynomial and sigmoid, and the kernel that got the best performance was the RBF, whose function is:

$$K_{x,x'} = e^{(-\gamma||x-x'||^2)}, \tag{2.26}$$

where γ is a free parameter.

3 UBINAS DATABASE AND ACQUISITION SYSTEM

In this chapter, the database and the acquisition system of the Ubinas volcano, in the city of Arequipa, Peru, are presented. The seismic database was built by experts of the National Volcanological Center of the IGP.

3.1 Ubinas Volcano

The Ubinas volcano (16 22 'S, 70 54' W, altitude 5672 m) began to erupt on March 25, 2006, after almost 40 years of inactivity. Located in the Central Volcanic Zone (CVZ, south of Peru), the Ubinas volcano is an active andesitic stratovolcano truncated in the upper part by a caldera with a diameter of 600m. The caldera floor is a flat area approximately 5,100 m above sea level. The active crater is located in the southern section and the bottom is 300 m below the floor of the caldera. Ubinas is considered the most active Peruvian volcano during the last 500 years, threatening 3,500 people living on the edge of the Ubinas Valley. The city of Arequipa, located 60 km away from this volcano, has been affected several times since 2006 due to the ash emissions (INZA, 2013). Figure 10 shows a satellite image of the Ubinas volcano captured on May 15, 2014.

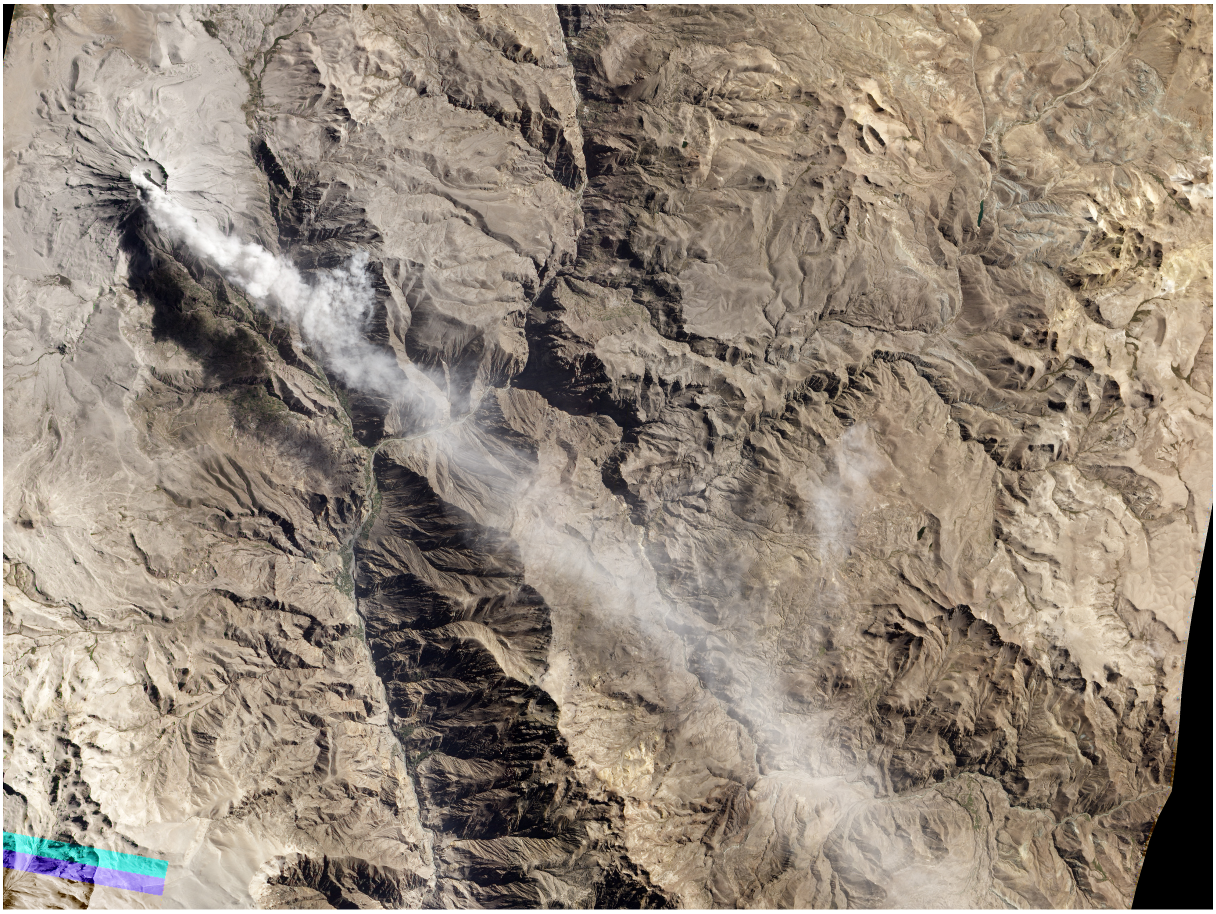
3.2 Ubinas Seismic Network

In order to monitor the seismic activity of volcanoes in Peru, the IGP has a Volcanological Seismic Network, as shown in Figure 11. This network is made up of sensors, digitizers and routers, which capture the ground motion and send the collected information to the Central Control Unit located in Lima. In the present investigation, we work with signals from the Ubinas volcano, UBI in the Figure 11. The numbers 1, 2, 3 and 4 represent the four existing sensors in Ubinas.

As the volcano sensors are positioned in inhospitable environments, they are subject to the IP68 standard (International Protection Marking in extreme environments). According to the standard, the instruments must withstand environments with extreme dust, water, and temperatures.

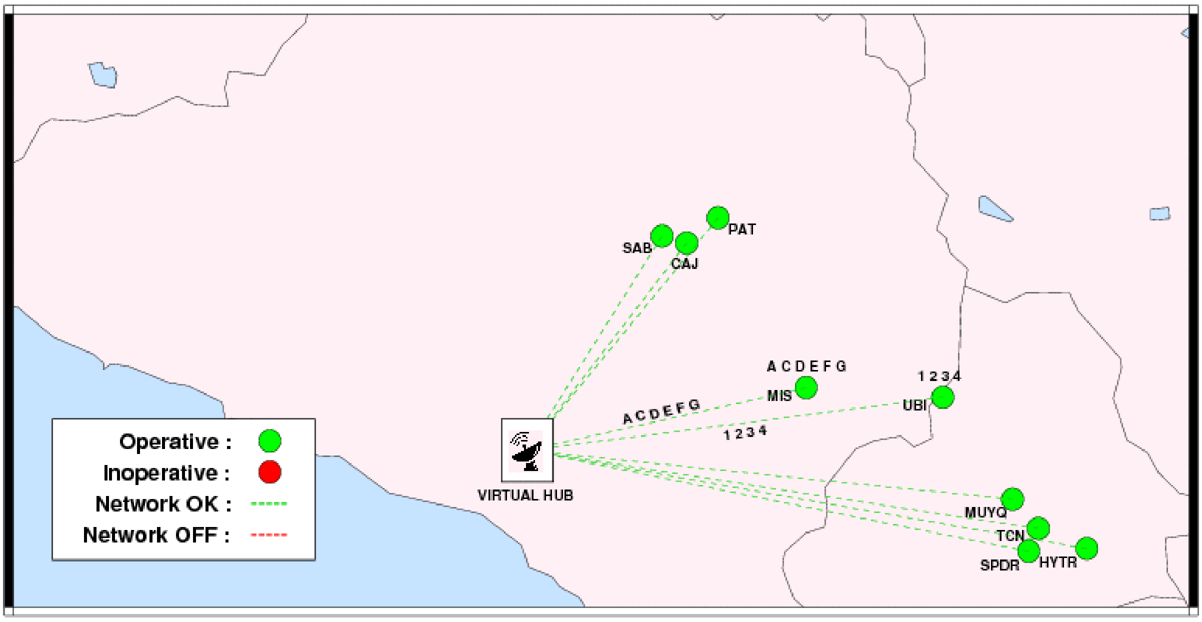
These sensors should not stop transmitting as they are of vital importance for seismic monitoring. However, due to the extreme environment of the volcano, the transmission through the internet is limited. Because of this, the data collected in real time is transmitted first to the

Figure 10 – Satellite image of the Ubinas volcano on May 15, 2014, the image shows a pale ash plume generated by the volcano.



Source: Adapted from the website of the Earth Observatory of NASA.

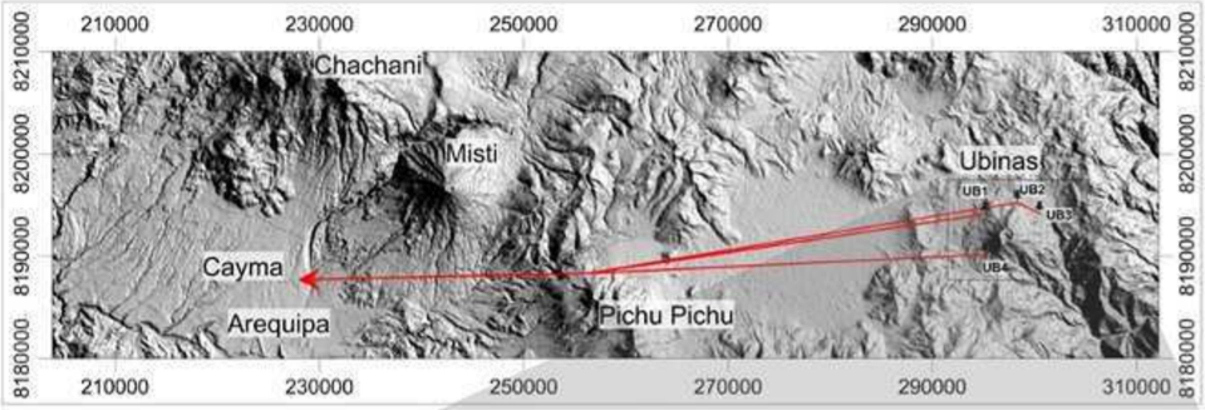
Figure 11 – Volcanological Seismic Network of Perú



Source: Own author.

Arequipa Observatory via radio link. Figure 12 shows a map with the telemetry network of Ubinas volcano.

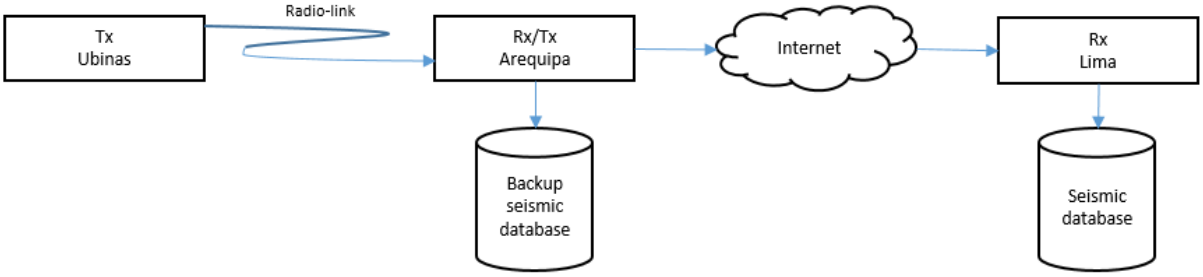
Figure 12 – Telemetry network of Ubinas volcano.



Source: Adapted from (INZA, 2013).

In the Arequipa Observatory, besides performing the retransmission of the signals from Ubinas volcano to Lima, the data is also saved as a backup, as a prevention method in case of external events. Once the data is collected, it is retransmitted via internet telemetry to the Central Control Unit located in Lima. Figure 13 shows a simplified block scheme of the data acquisition system from the Ubinas volcano to the central base located in Lima.

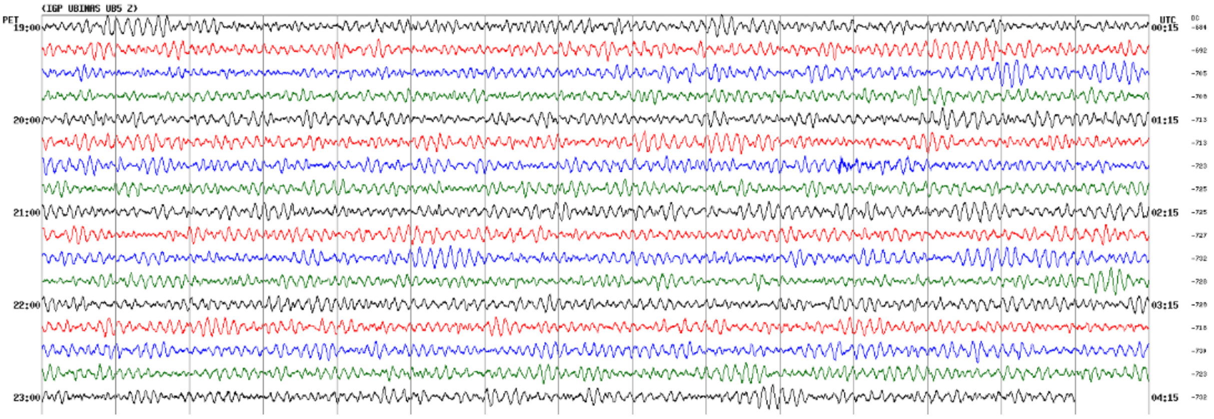
Figure 13 – Simplified data acquisition system from the Ubinas volcano signals to the central base located in Lima.



Source: Own author.

In Lima, this data is saved and monitored by experts in volcanology, who constantly perform the analysis of the signal received in real time, which often becomes a difficult task due to the large number of events generated by the volcano. Figure 14 shows an example of the Ubinas seismic signals displayed in Lima.

Figure 14 – Example of the Ubinas seismic signal displayed in Lima (vertical component of the sensor).



Source: Adapted from the website of the National Volcanological Center of the IGP.

3.3 Description of the Volcano Classes

Since the eruption of the Ubinas volcano in 2006, a large number and variety of types of waveforms have been generated, as presented in the literature (MACEDO *et al.*, 2009; INZA *et al.*, 2014; INZA *et al.*, 2011). These varieties of waveforms are associated with the behavior of magma, whose physical and chemical effects change depending on the trajectory and the environment it encounters on its route to the crater. The five main types of volcano-seismic events are the following:

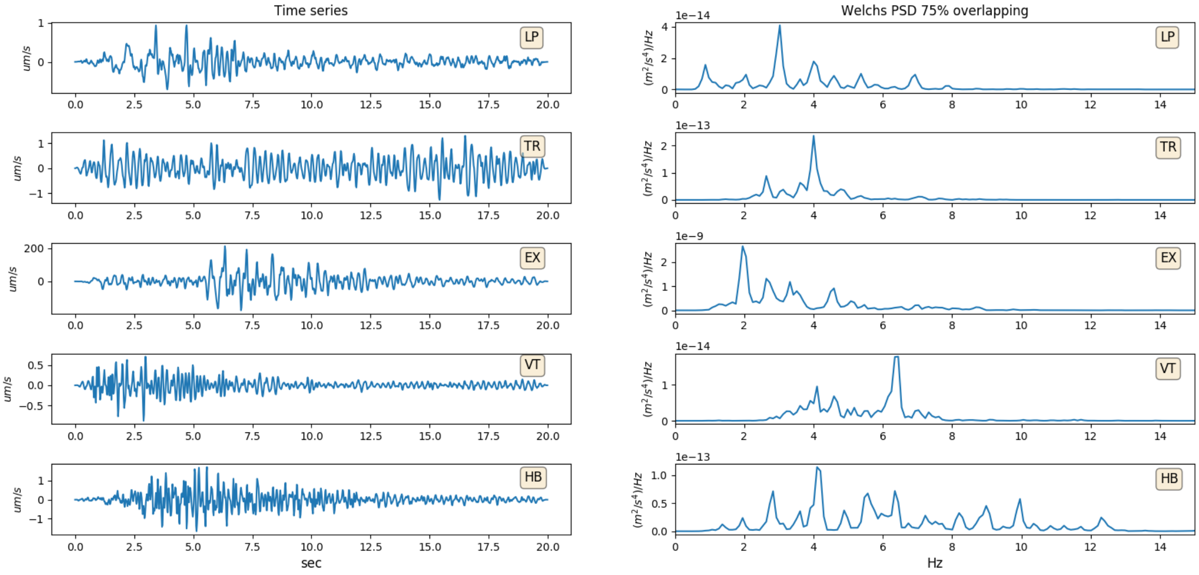
1. Volcano-Tectonic (VT): They are associated with the breakage of rocks due to the high pressure produced by the magma and can even activate internal faults in the volcanic building.
2. Long Period (LP): They correspond to the impact of the fluids moving in the volcanic system or interacting with the hydrothermal system.
3. Hybrids (HB): They are caused by fluids in the blocked ducts, which produce both VT and LP events at almost the same time.
4. Tremors (TR): These events are generated due to degassing or to the effect of resonance produced by the disturbance of the cavities of the duct systems under the crater. This type of signal may last from a few minutes to several days.
5. Explosions (EX): These events are originated due to the change of pressure and temperature of the magma, in conditions where volatile gases and bubbles explode.

Our initial simulations showed that a temporal analysis of the seismic signals can give us important characteristics of the classes. Indeed, many time domain features have proved to be useful for distinguishing one type of signal from the others. For instance, the EX events are characterized by high magnitudes, when compared to other classes, as illustrated in Figure 15, which shows one time series sample of each class, with their respective PSDs. In this example, it can be viewed that the maximum amplitude of the EX signal in the time domain is more than 200 $\mu m/s$, while the signals of the other classes reach less than 1 $\mu m/s$.

However, the simulations also showed that certain types of signals are better distinguished in the spectral domain, such as HB and VT signals. In Fig. 15, the HB and VT waveforms present a roughly similar behavior in the time domain. Nevertheless, it can be viewed that by their PSDs in Fig. 15 that the HB signal has a spectrum much broader than the VT signal, whose PSD is mainly concentrated around 6.5 Hz.

In the literature, the use of the cepstral domain has provided a relevant impact in the classification of seismic signals, e.g. (ZHOU *et al.*, 2012) and (PENG *et al.*, 2019). Due to this fact, attributes in the cepstral domain are also considered in the proposed method. Specifically, the mel frequency cepstral coefficients (MFCC) are used in this work, as they have given good results in the classification of volcanic signals, as in (MALFANTE *et al.*, 2018) and (BENITEZ *et al.*, 2007).

Figure 15 – Time series samples of the five classes with their respective PSD.



Source: Own author.

3.4 Database

The catalog with the seismic data used in our research was built by the IGP. In the Ubinas volcano, there is a permanent seismic monitoring with four seismic stations (with codes

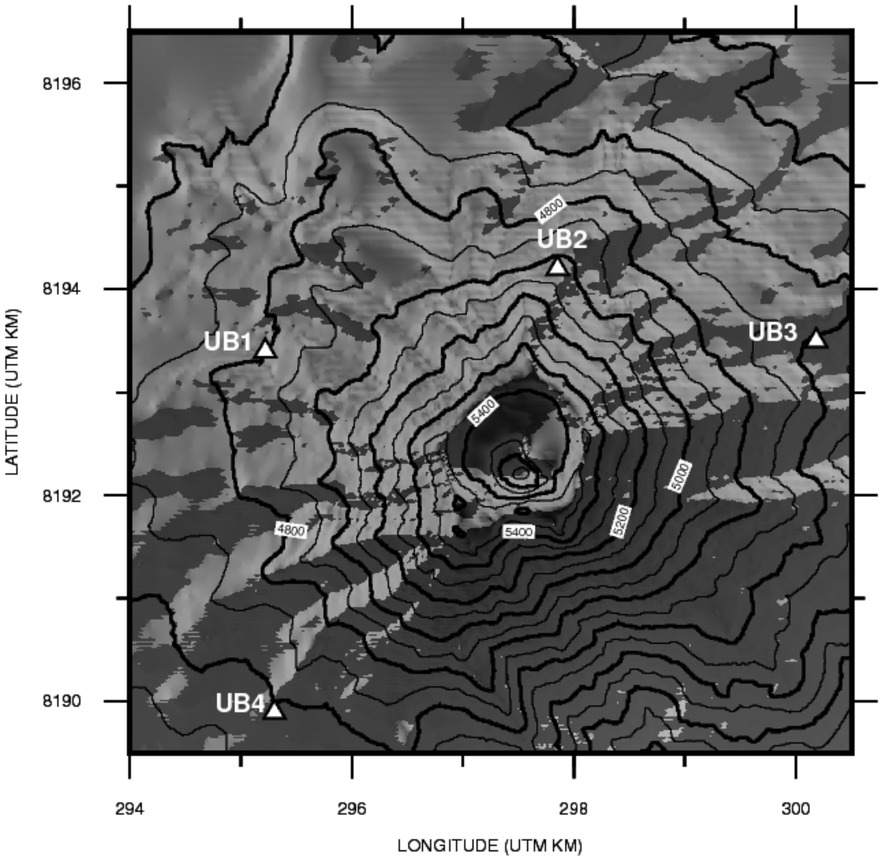
UB1, UB2, UB3 and UB4) distributed on the flanks of the volcano, as shown in Figure 16, which shows a map of the Ubinas volcano with the UB1, UB2, UB3 and UB4 permanent stations. From 2006 to 2007, the UB1, UB2, UB3 and UB4 stations were equipped with 1 Hz short period seismometers with an analog telemetry system for transmitting data to the observatory (IGP Arequipa). From 2008 to the present, these stations were progressively upgraded with broadband 3 component sensors and digital telemetry based on Guralp 40T and Reftek 130. The initial catalog was made mainly using the data recorded by these 4 permanent stations. However, the UB1 and UB2 stations have more stable instruments in terms of continuity of the data acquisition without gaps. Therefore, this work uses only data from the stations UB1 and UB2, located approximately 2 km to the west and north of the crater, respectively. The database of our research was collected in the year 2014 and, at that time, the station UB1 had a 3 component sensor, while UB2 had a single component sensor. This means each event is characterized by 4 simultaneous signals in the final database used in this dissertation.

The catalog used in this work consists of records of the five aforementioned main volcano-seismic events (VT, LP, HB, TR and EX) with the corresponding labels assigned by the experts of the IGP. There are other types of events exhibited in the literature, such as “Tornillo” (TOR), Very Long Period (VLP). However, there are not enough data found of these types of events. For this reason, they were not considered in this work.

The catalog is used to compare the responses of the automatic classifier with the ones of the experts, and to calculate the success rate of the classifier. Results of several temporary experiments, with other databases of seismic data collected by sensors located around the Ubinas volcano, were carried out in the years 2006, 2009, 2011, 2014 and 2015, with international participation, such as in the framework of the EU-VOLUME project (MACEDO *et al.*, 2009), a cooperation between the IGP and the *Institut de Recherche pour le Developpement* (IRD-France).

Due to the considerable number of volcanic events that occurred in 2014 (about 50,000 events), this database has a high number of events cataloged. In particular, it has a considerable number of explosion events, when compared with databases of other works, which can be considered one of the most important events to be detected. For instance, the database of (MALFANTE *et al.*, 2018) contains only 160 EX events, while our database has 592 EX events. Moreover, the database of the present work has a continuity in the data, that is, it has no gaps in the acquisition of the signals. The complete catalog has 28140, 11489, 8108, 1346 and 592 events of the classes LP, TR, HB, VT and EX, respectively. However, in order to maintain a balance among the number of samples in each class, we decided to use a smaller number of events, as it will be described in Chapter 5.

Figure 16 – Map of the Ubinas volcano with the UB1, UB2, UB3 and UB4 permanent stations.

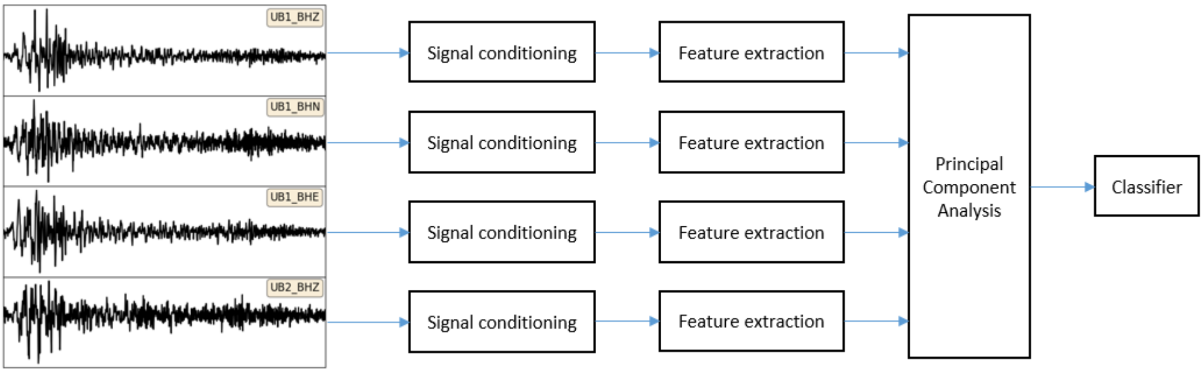


Source: Adapted from (INZA, 2013).

4 CLASSIFICATION SYSTEM

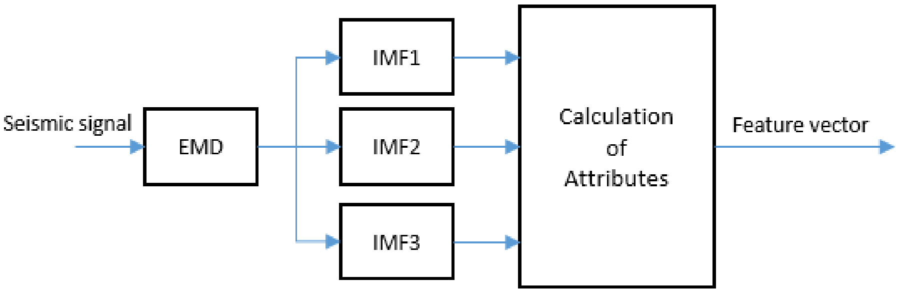
In this chapter, the proposed supervised classification method is presented. Figure 17 shows a simplified block diagram of the classification system, based on a multicomponent processing by using 4 seismic channels from 2 different stations. The classifier can be summarized in the following steps, which will be detailed below. Firstly, a signal conditioning that includes offset elimination and instrumental correction, among other operations, is performed. The extraction of the attributes in temporal, spectral and cepstral domains is then carried out, using the EMD. A simplified scheme of the feature extraction block diagram is shown in Figure 18. After this step, the PCA and the classification algorithm are applied.

Figure 17 – Simplified block diagram of the proposed classification system.



Source: Own author.

Figure 18 – Feature extraction block diagram.

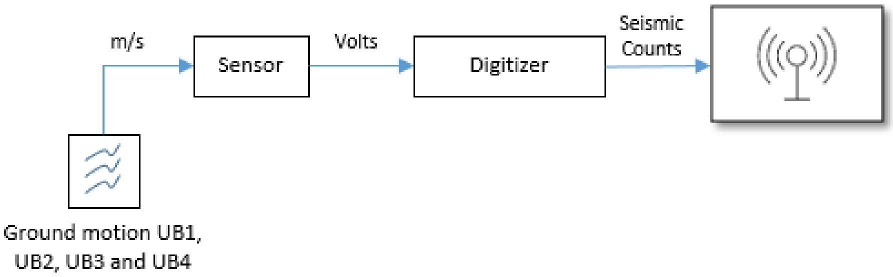


Source: Own author.

4.1 Signal Conditioning

In the seismic signal acquisition system, the sensors are responsible for capturing the ground motion, converting it into a voltage signal, as shown in Figure 19, which shows the simplified block diagram of the data recorded and transmitted from the Ubinas volcano to the Arequipa observatory. This signal is sampled and quantified through the digitizers, whose output is in the unit Seismic Counts. The digital signal (in Seismic Counts) is the one that travels through different telemetries, either by radio link, internet, satellite, etc. In our case, this signal travels from the Ubinas to the Arequipa observatory through a radio link and then from Arequipa to Lima through internet telemetry, where the signals are stored in a database, as mentioned in Chapter 3.

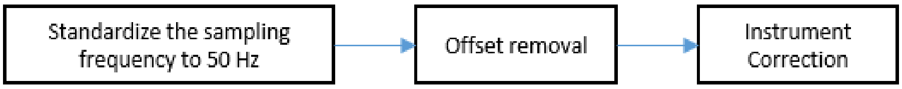
Figure 19 – Data transmission from the Ubinas volcano to the Arequipa observatory.



Source: Own author.

This database contains raw data from the volcano's signals. Certain methods are then necessary to condition the signal in order to extract attributes that represent and characterize the different types of volcano events. The signal conditioning consists in the following steps: unify sampling frequencies, eliminate the DC component of the signal, and perform the instrumental correction. Figure 20 shows the block diagram of the signal conditioning.

Figure 20 – Block diagram of signal conditioning.



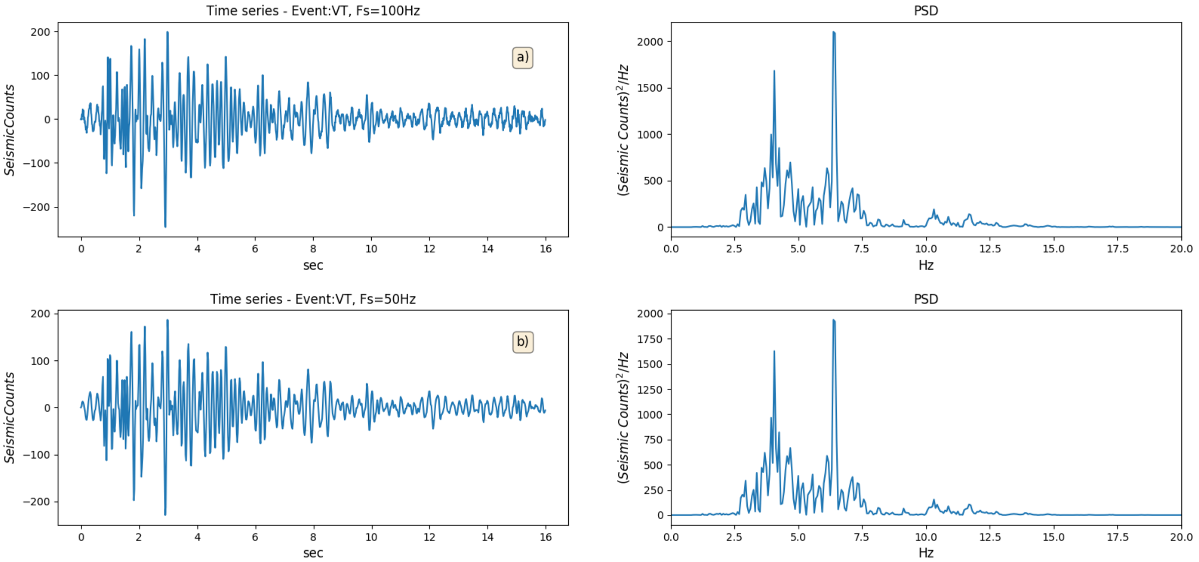
Source: Own author.

4.1.1 Re-sampling

The first step to condition the raw signal is to standardize the sampling frequency. Due to various experiments, different types of instrumentation were installed in the Ubinas volcano, mentioned in Section 3.4. The data collected in the database of this research has sampling frequencies of 50 Hz and 100 Hz. It is then essential to standardize all the sampling frequencies. By analyzing the PSD of the signals of the different classes, it was found that the spectral bandwidth of interest can be considered lower than 20 Hz. For this reason, the sampling rate of all the signals was set to 50 Hz, covering the entire spectrum of interest.

Figure 21 shows an example of re-sampling a VT signal. In this figure, the spectrum of the signal of the VT type at 100 Hz and 50 Hz are shown. In both cases, the fundamental frequency is 6.375 Hz, which is expected since a signal of the VT type has high frequencies (greater than 5 Hz).

Figure 21 – a) Time series of a VT signal with a sampling rate of 100 Hz and its spectrum. b) The same signal with a sampling rate of 50 Hz and its spectrum.


Time series - Event:VT, Fs=100Hz

PSD

Time series - Event:VT, Fs=50Hz

PSD

Source: Own author.

4.1.2 *Offset removal*

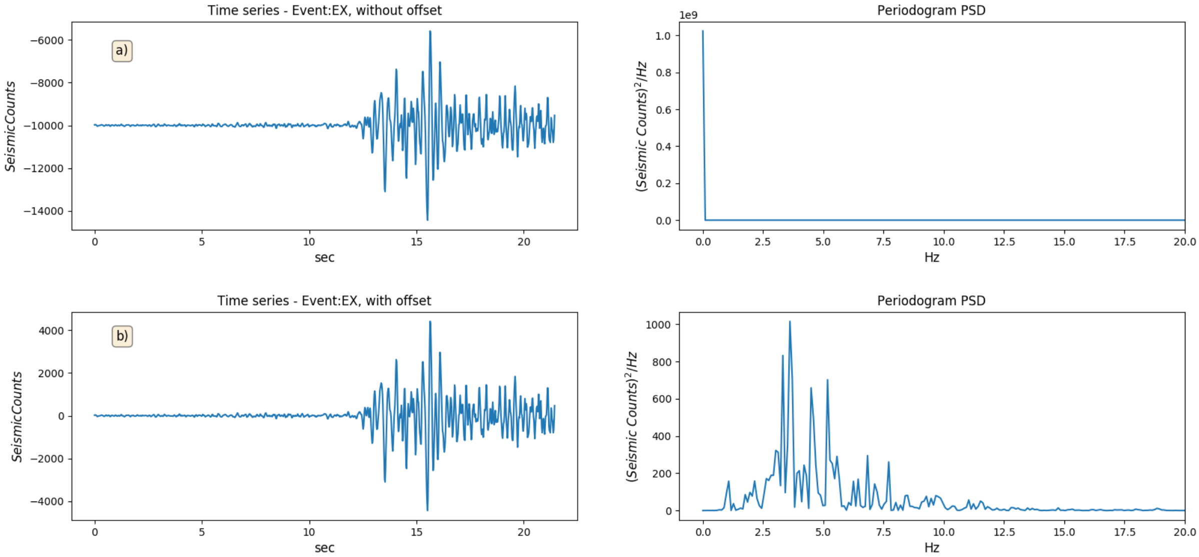
In this investigation, the PSD is used for the purpose of using physical measurements to the problem, since the amplitude of the PSD measures the energy density in the frequency space, which is a physical magnitude and gives a physical contrast to the problem.

In the literature, there are several methods to estimate the PSD, divided into two categories: parametric and non-parametric. The parametric methods assume that the signal is a stationary, while non-parametric methods do not assume any stochastic model (PARHI; AYINALA, 2014). Since the seismic signal is a non-stationary signal, the non-parametric approach will be used for estimating the PSD.

As seismic signals often have non-zero average, when the periodogram of a raw seismic signal is calculated, undesirable artifacts at very low frequencies may be generated. For this reason, the elimination of the DC component is essential in the seismic signal processing.

An example of this behavior is shown in Figure 22, which shows the computation of the periodogram of an EX raw signal centered in -10,000 Seismic Counts, and the periodogram of the same signal but eliminating the DC component. The results show that the frequency spectrum of the raw signal is governed by a very large magnitude at 0 Hz compared to the other frequencies (Figure 22a), obscuring the frequencies of interest. This effect is solved by eliminating the offset of the raw signal (Figure 22b), obtaining a much clear spectrum in the frequencies of interest.

Figure 22 – a) Time series of a EX signal with offset, and its spectrum. b) The same signal without offset, and its spectrum.



Source: Own author.

The magnitude of the PSD at 0 Hz of the signal in Figure 22, for $N = 512$, $Fs = 50$ Hz and $DC = -10,002.231$, is equal to $1.023x10^9$ (this is the giant magnitude shown in the PSD of the figure 22a), which obscures the frequencies of interest, whose magnitude corresponding to these frequencies are in the order of 1000 (Figure 22b).

4.1.3 Instrumental correction

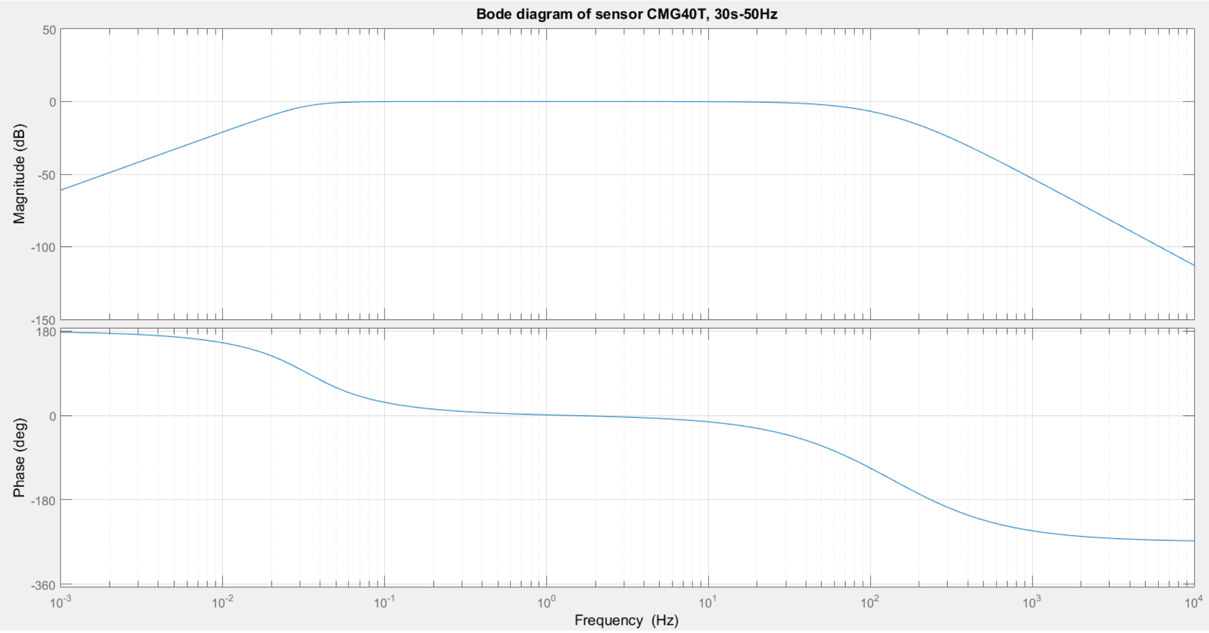
The signals recorded by a seismic sensor are the convolution of the ground velocity signal with the sensor transfer function, in this sense, the seismic amplitudes recorded by the sensor are proportional to the ground velocity in a frequency range given by the flat part of the Bode diagram of the transfer function, while the amplitude being attenuated for the other frequencies. A similar effect also occurs with the phase. The removal of this magnitude and phase effect caused by the sensors is called instrumental correction (HAVSKOV; ALGUACIL, 2004).

Table 1 – Calibration sheet for Guralp CMG40T 0.033 Hz - 50 Hz

Sensor sensitivity	800 V/m/s
Normalization factor	5.71508E+08
Zeros in Hz Z = 2	0 0
Poles in Hz P = 5	-1.486000E-01 + j1.486000E-01 -1.486000E-01 - j1.486000E-01 -5.026500E+02 -1.005000E+03 -1.131000E+03

In this investigation, one of the sensors used is the Guralp CMG40T sensor, whose datasheet has a cutoff frequencies of 0.033 Hz (30 s in period units) and 50 Hz, and a sensitivity of 800 $\frac{V}{m/s}$. Figure 23 shows the Bode diagram of the CMG40T sensor. This diagram is made using the poles, zeros and the sensor normalization factor given in Table 1, where Z represents the number of zeros and P the number of poles. This information can also be found on the IRIS website (Incorporated Research Institutions for Seismology).

Figure 23 – Bode diagram of the CMG40T sensor transfer function.

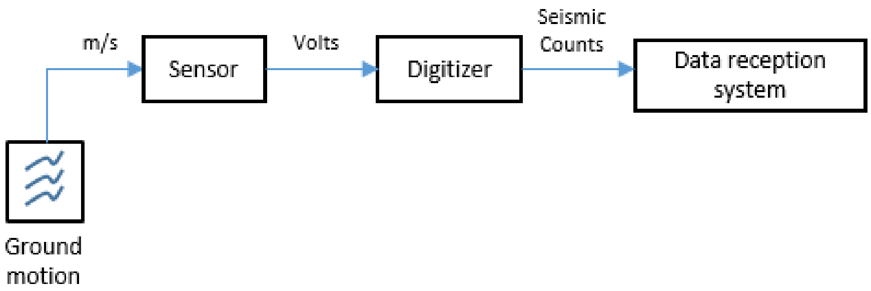


Source: Own author.

As we can see in the Bode diagram of magnitude in Figure 23, the magnitude is attenuated for both high and low frequencies. The instrumental correction corrects the frequency distortion caused by the sensor transfer function, increasing the available frequency range.

Figure 24 shows a simplified block scheme of the data acquisition system from the ground motion to the reception system. The velocity signals are recorded by the digitizers of the monitoring centers in units of Seismic Counts. The instrumental correction consists of computing the deconvolution associated with the transfer function of the data acquisition system (sensor transfer function multiplied by the digitizer sensitivity), bringing back the seismic signal to its original unity (m/s). This method for instrumental correction is detailed in (HAVSKOV; ALGUACIL, 2004).

Figure 24 – Simplified block scheme of the data acquisition system from the ground motion to the reception system.



Source: Own author.

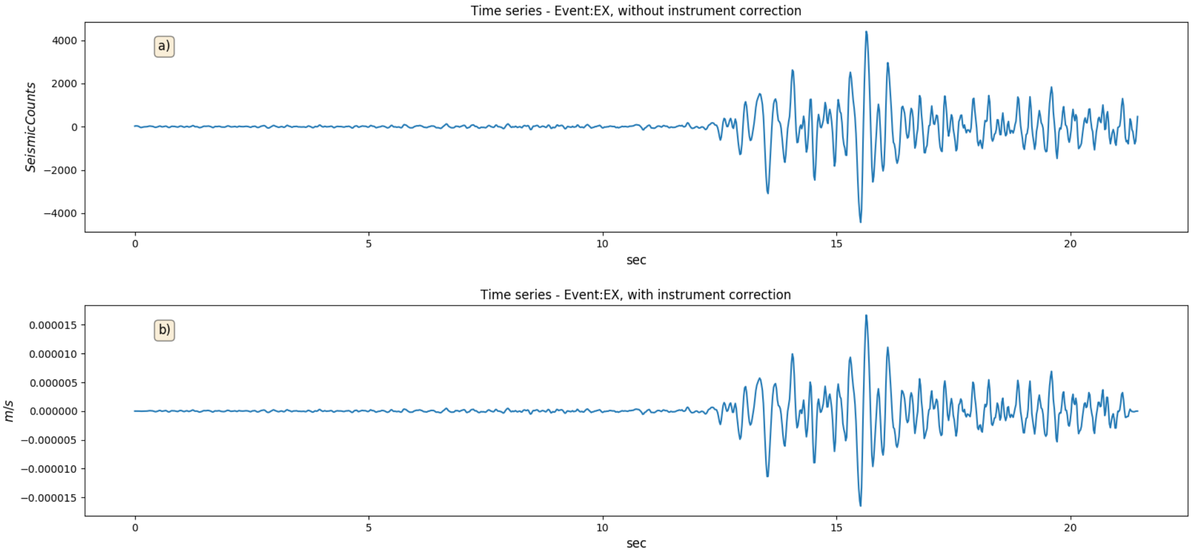
The use of the instrumental correction is due to a specific reason. By standardizing all the velocity signals to the unit m/s, the classifier becomes independent of the types of sensors used. Otherwise, we would be forced to normalize the signals, as seen in some cases in the literature (MALFANTE *et al.*, 2018), (CUEVA *et al.*, 2017). However, when a normalization is performed, valuable information of the physical energy of the signals is lost. This is particularly important in our study, as more than one type of sensor is being used simultaneously. That gives the energy of the signals a physical sense, allowing us to use the energy as an important attribute. In addition, this method give physical contrast to the investigation, because the signal to process will be the true seismic signal in m/s. The simulation results shows that the instrumental correction significantly improves the accuracy of the classifier.

As mentioned in Section 3.4, this investigation was carried out with Guralp CMG40T sensors and Reftek130 digitizers, the parameters of these instruments are in their respective manuals. Using these parameters we can perform the instrumental correction.

Currently, there are several programs that facilitate the use of instrumental correction, such as SAC (Seismic Analysis Code), Obspy (Python framework for processing seismological data), Matlab, etc.

Figure 25 shows an example of the instrumental correction of an EX type seismic signal. The signal in Seismic Counts is shown in Figure 25a, while Figure 25b shows the seismic signal in m/s after performing the instrumental correction.

Figure 25 – a) Time series of a EX signal. b) The same signal with instrumental correction.



Source: Own author.

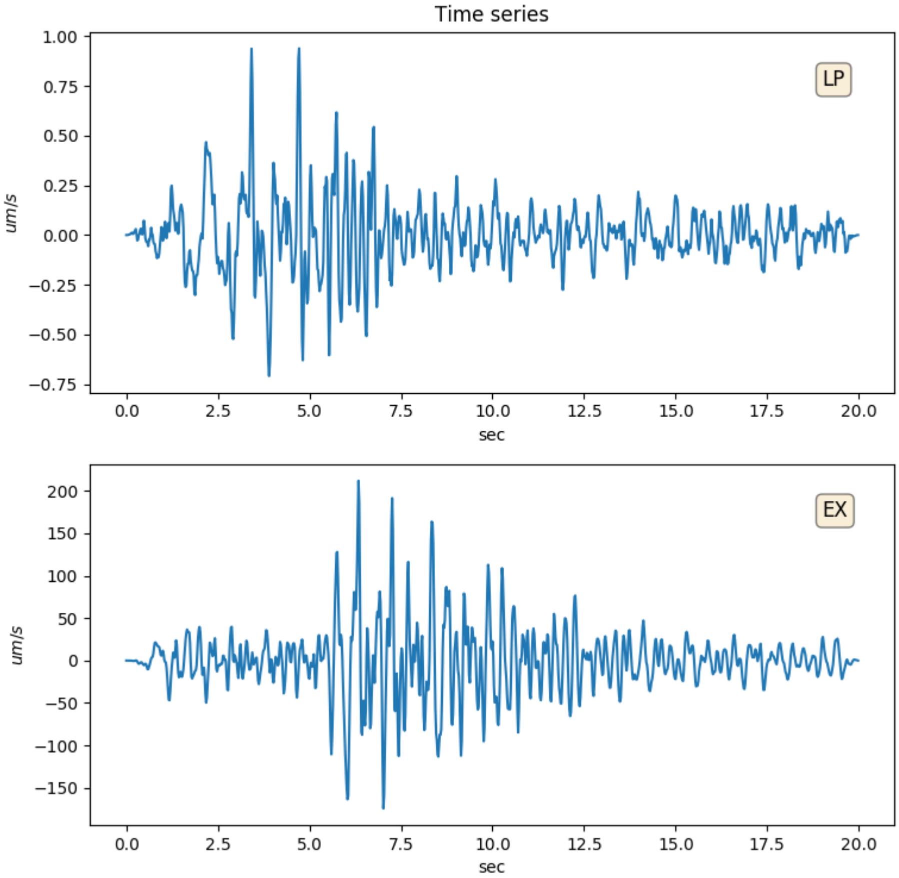
An example of the importance of the instrumental correction can be viewed by comparing the LP and EX signals in Figure 26. These signals have similar spectra, the main difference between these two signals being the high energy of the EX signal when compared to the LP. This characteristic can be reflected in attributes such as the energy or the maximum of the temporal signal. Indeed, in Figure 26, the maximum of the LP signal is 0.94 um/s, while the EX signal has a maximum of 211.7 um/s.

4.2 Feature Extraction

In this section, the feature extraction procedure is described. A simplified scheme of the feature extraction block was shown in Figure 18. As earlier explained, the proposed classifier is based on a multicomponent approach by using 4 seismic channels from 2 different stations, i.e. each volcano-seismic observation is represented by 4 simultaneous seismic signals: 3 signals observed by a station with a triaxial sensor (UB1) and 1 signal observed by a station with a single-channel sensor (UB2).

After the preprocessing steps described in Section 4.1, each of the 4 signals are decomposed with the use of EMD, the 3 most significant IMFs being selected and the others being neglected. A large number of attributes is then calculated for each IMF of each signal. These steps of the feature extraction block are detailed below.

Figure 26 – Time series of a LP and EX signal with instrumental correction.



Source: Own author.

4.2.1 Choice of the Number of IMFs

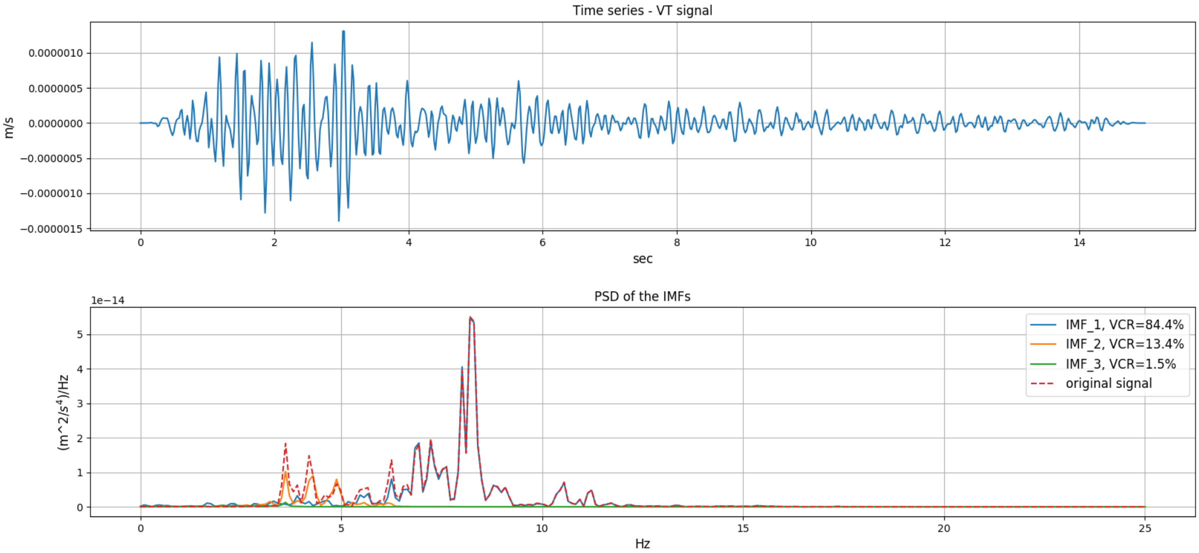
After computing the EMD, with the procedure described in Section 2.3, one must decide how many IMFs will be used for generating the attributes. Indeed, as the EMD method is a natural decomposition method, the number of IMFs is not fixed, it depends on each signal. It is then essential to fix a number of IMFs to perform the extraction of attributes. Otherwise, the number of attributes would be variable. In the present work, the Variance Contribution Ratio (VCR) is used for this purpose. The VCR represents the variance of each IMF with respect to the total variance, that is:

$$VCR_{I_m} = \frac{\text{var}(I_m[n])}{\sum_{m=1}^{M} \text{var}(I_m[n])}, \tag{4.1}$$

where $\text{var}(\cdot)$ is the variance operator. We have calculated all the VCR_{I_m} for a sample space of 2000 events (400 events per class), obtaining as a result that the first 3 IMFs from the 3 highest VCRs provide a VCR of at least 93%, i.e. $VCR_{I_1} + VCR_{I_2} + VCR_{I_3} \geq 0.93$. This means that the sum of the remaining IMFs represents, on average, less than 7% of the total energy. These remaining IMFs can then be considered noise, which means that the first 3 IMFs together account for most of the energy of the original signal. For this reason, we choose to use the first 3 IMFs that represent the highest amount of VCR. As a result, for each seismic event, we have 4 seismic signals, each one with 3 IMFs, leading to a total of 12 IMFs for each event.

To have a clearer idea of the choice of IMFs, a simulation was performed by calculating the VCR of the 3 selected IMFs of a VT signal, as shown in Figure 27, which shows the raw VT signal and its PSD, as well as the PSD of the 3 selected IMFs. A signal of the VT type is characterized by having high frequencies. Hence, it is expected that the first IMF will be the one that contains the greatest amount of energy with respect to the original signal and that is exactly what happens. Figure 27 shows that the first IMF (84.4% VCR) has a large amount of energy with respect to the second (13.4% VCR) and third IMF (1.5% VCR), and the 3 IMFs together contain almost all the energy with respect to the original signal (99.3% VCR).

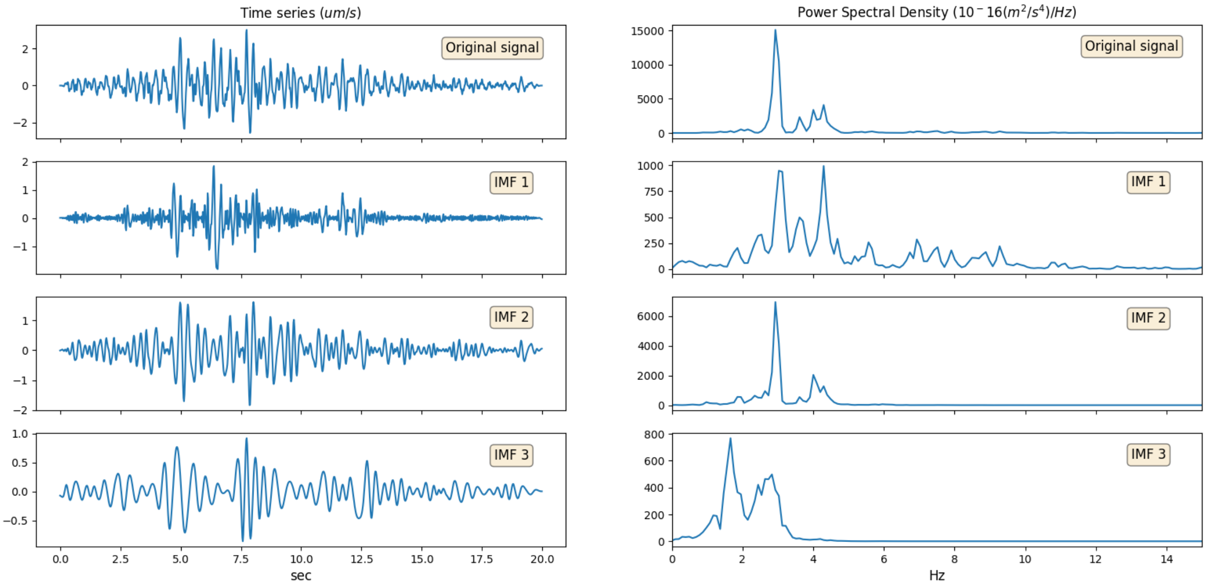
Figure 27 – Example of Variance Contribution Ratio of the three selected IMFs of a signal VT.



Source: Own author.

Other examples of IMFs are illustrated in Figure 28. It shows the three selected IMFs of a LP signal, as well as their respective PSDs. It is known that a signal of the LP type has a spectrum with energy concentrated at low frequencies, generally less than 5 Hz. It can be viewed from this figure that the first IMF has a considerable energy in frequencies between 3 to 5 Hz, and each subsequent IMF has a spectrum concentrated around a lower frequency. Indeed, the second IMF has a high energy concentration around 3Hz, and the third IMF has considerable energy around 1 and 3Hz. From the fourth IMFs onwards, the energies are not considerably high. In comparison to the original signal, there are now 3 signals (IMFs) that reinforce that the signal is LP type. This nature of IMFs helps to differentiate between one class and another. The attributes used by the classifiers will be extracted from the three selected IMFs to take even more advantage of the natural characteristics of these signals, in order to generate a feature vector that represents efficiently each class.

Figure 28 – A sample of an LP signal, its three selected IMFs and their PSD.



Source: Own author.

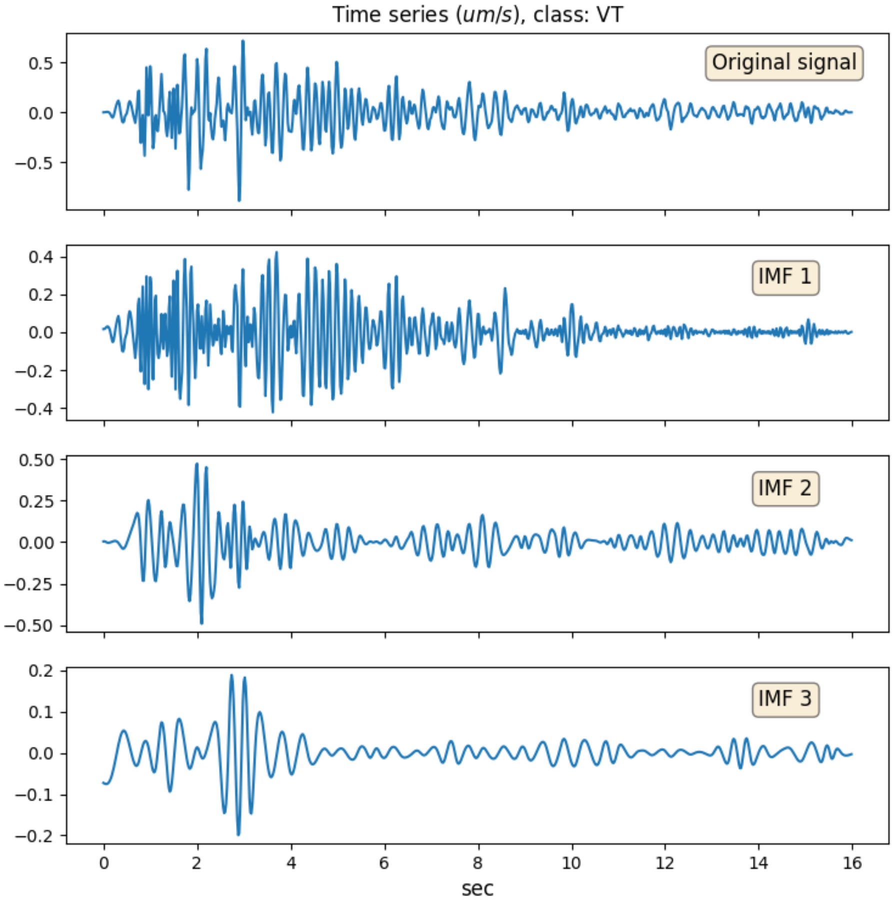
4.2.2 *Calculation of Attributes*

As explained in Section 3.3, our simulations showed that certain types of classes are better distinguished in the time domain, while other signals are better distinguished in spectral or cepstral domains. A good success rate using the 3 aforementioned domains was obtained in (MALFANTE *et al.*, 2018). Due to this reason, the proposed classification system performs the extraction of the attributes in temporal, spectral and cepstral domains, in the following way:

- In the temporal domain, we used attributes obtained directly from the IMFs $I_i[n]$.
- In the spectral domain, we used attributes calculated from the PSDs of the IMFs, using the Welch's method (PARHI; AYINALA, 2014) with a FFT length of N = 512, 75% overlapping and a Hanning window function.
- In the cepstral domain, we used attributes obtained from $F^{-1}\{\log|F\{I_i[n]\}|\}$, where $F\{\cdot\}$ is the Fourier transform, with 13 Mel frequency coefficients (MFCC) being used.

In the following subsections, these three domain in which the attributes are calculated are detailed and illustrated. In these subsections, we use a VT-type example signal to illustrate the physical meaning of the attributes. The VT example signal is shown in Figure 29, with its 3 selected IMFs.

Figure 29 – Timeseries of an example VT signal.



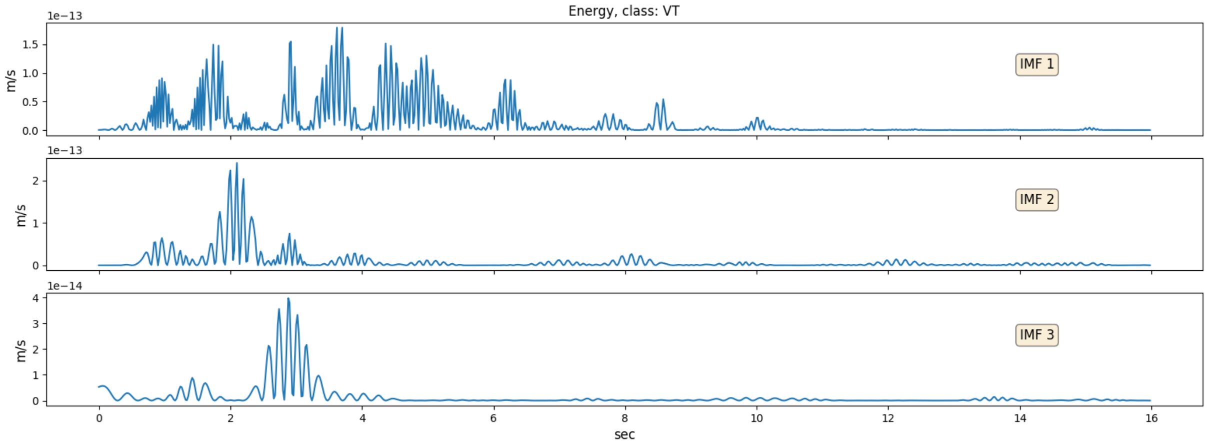
Source: Own author.

4.2.2.1 Time domain

Attributes in the time domain contains important information, such as shape parameters that indicate the time when a signal reaches its maximum peak, the maximum energy or how many times the signal change from positive to negative and vice versa, etc.

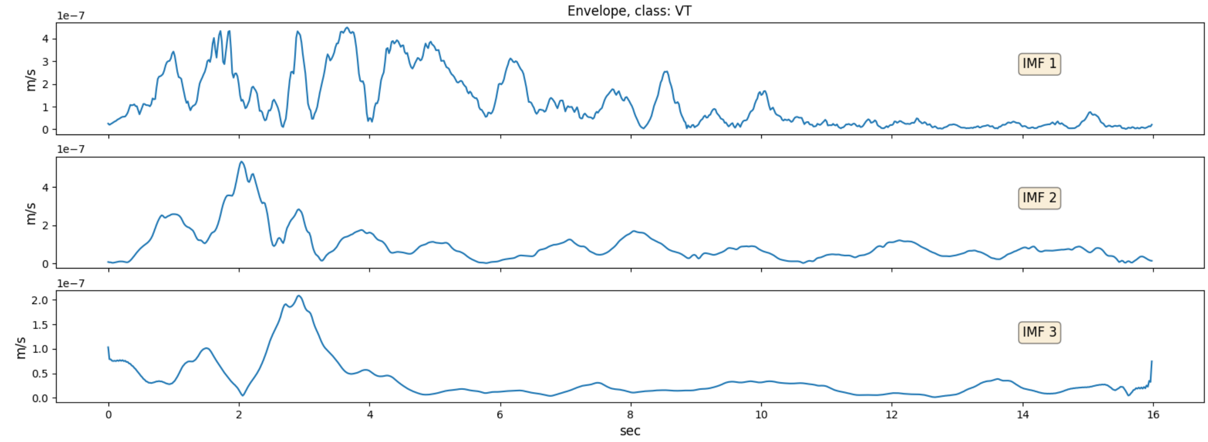
In the time domain, the signals corresponding to the timeseries ($I_i[n]$), the energy ($I_i^2[n]$), and the envelope ($|H\{I_i[k]\}|$), where $H\{\cdot\}$ is the Hilbert Transform and $|\cdot|$ the absolute value), are used to calculate the attributes. The energy of the 3 selected IMFs of the example signal is shown in Figure 30, and in Figure 31 its envelope.

Figure 30 – Energy signal of the 3 selected IMFs of the example signal.



Source: Own author.

Figure 31 – Envelope of the 3 selected IMFs of the example signal.



Source: Own author.

4.2.2.2 *Spectral domain*

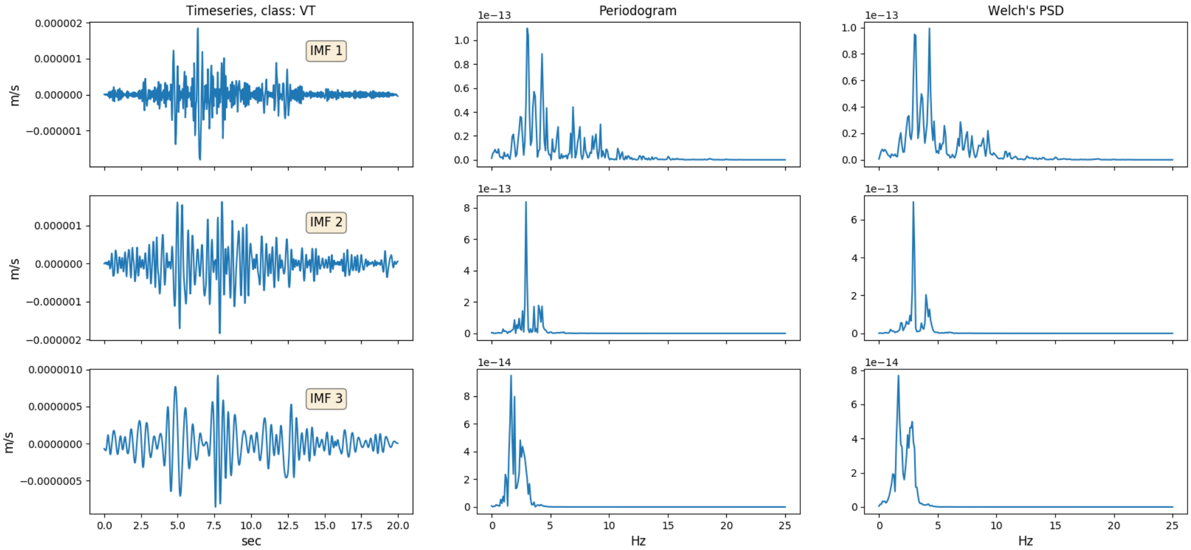
In the spectral domain, attributes of the PSD of $I_i[n]$ are extracted using the Welch's method. This method deepens the frequencies of interest of the volcano-seismic signal, through the use of average periodograms with the use of overlapping signals.

Figure 32 shows the 3 selected IMFs of the example signal with their periodograms and their PSD using the Welch's method. This figure shows that the use of the Welch's method is a smoother curve than the periodogram, emphasizing better the concentration of energies at certain frequencies.

4.2.2.3 *Cepstral domain*

The use of cepstral analysis was originally applied for the detection of echoes in seismic signals in (BOGERT, 1963). Moreover, cepstral analysis had great success in speech

Figure 32 – Timeseries, Periodogram, and Welch's PSD of the 3 selected IMFs of the example signal.



Source: Own author.

signals to determine voice pitch and separating the formants (transfer function of the vocal tract) from voiced and unvoiced sources, as mentioned in (RANDALL, 2016).

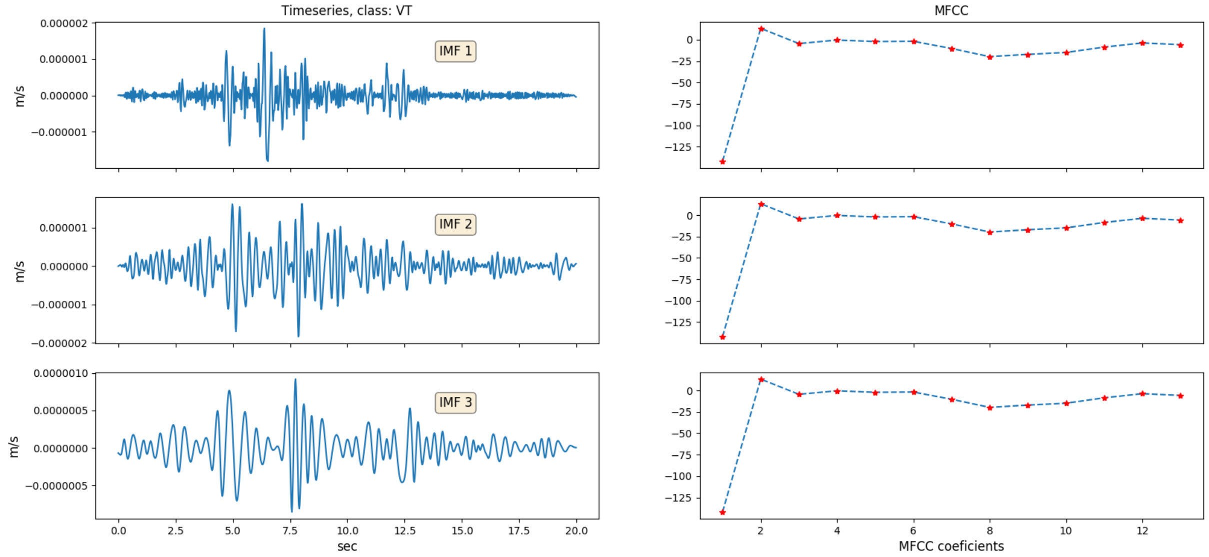
In this sense, we can see the volcano as a great vocal tract, where the voice would be represented by the volcano-seismic events, traveling through the pharynx (propagation paths) and exiting through the mouth (crater). In fact, if we work only with a single domain, between temporal, spectral and cepstral domain, the one that generated the best results, as will be seen in Section 5.4, was the cepstral analysis. In addition, the incorporation of this domain in the use of machine learning applied to seismic signals generated good results, such as in Malfante's work (MALFANTE *et al.*, 2018).

In the cepstral domain, the computation of the Mel Frequency Cepstral Coeffcients (MFCC) will be carried out using 26 filterbanks and the first 13 coefficients will be chosen. The procedure to calculate the MFCCs is detailed in (PENG *et al.*, 2019). The computation of the first 13 MFCC of the example signal is shown in Figure 33.

4.2.2.4 List of Attributes

With the 3 domains to use well defined, a total of 54 attributes are extracted per IMF. Table 2 shows some of the used attributes, where $s[n]$ is the signal from which the attributes are extracted. In short, for time domain attributes, $s[k] = I_i[k]$, $s[k] = I_i^2[k]$ or $s[k] = |H\{I_i[k]\}|$, where $H\{\cdot\}$ is the Hilbert Transform and $|\cdot|$ the absolute value. For frequency domain attributes, $s[k] = PSD_k(I_i[n])$, where $PSD_k(\cdot)$ is the k^{th} component of the PSD. For attributes in the cepstral

Figure 33 – 13 first MFCC of the 3 selected IMFs of the example signal.



Source: Own author.

domain, $s[k] = F_k^{-1}\{\log|F\{I_i[n]\}|\}$, where $F_k^{-1}\{\cdot\}$ is the k^{th} component of the inverse Fourier transform. Moreover, the function *count(a < b)* returns the number of components for which $a < b$.

As 12 IMFs are used per event, each seismic signal has a feature vector with $12 \times 54 = 648$ attributes. Moreover, one additional attribute was calculated directly from the raw time series: the duration of the observation. That leads to a total of 649 attributes per event. Most of these attributes were used in (MALFANTE *et al.*, 2018) and (ASTAPOV, 2011), and they have proved to be useful in distinguishing signals in classification problems. Appendix A shows the total attributes in detail.

4.3 Principal Component Analisys (PCA)

The very high dimensionality of the above described feature vector (649 dimensions) may cause a dispersion of the data and the well-known problem of "curse of dimensionality" (WANG; SLOAN, 2007). To avoid this issue, the PCA is used for dimensionality reduction.

It was found that the first 200 components obtained with PCA account for 99.7% of the feature vector variance, as shown in Figure 34, which shows the percentage of cumulative variance of the 649 components generated by the PCA. The number of PCA components was then set to 200, leading to a feature vector with 200 dimensions at the input of the classifier. Besides avoiding these problems, the PCA decreases considerably the processing time, which is very important because, in this case, the proposed classification system should be able to be

Table 2 – Attributes extracted from the IMFs

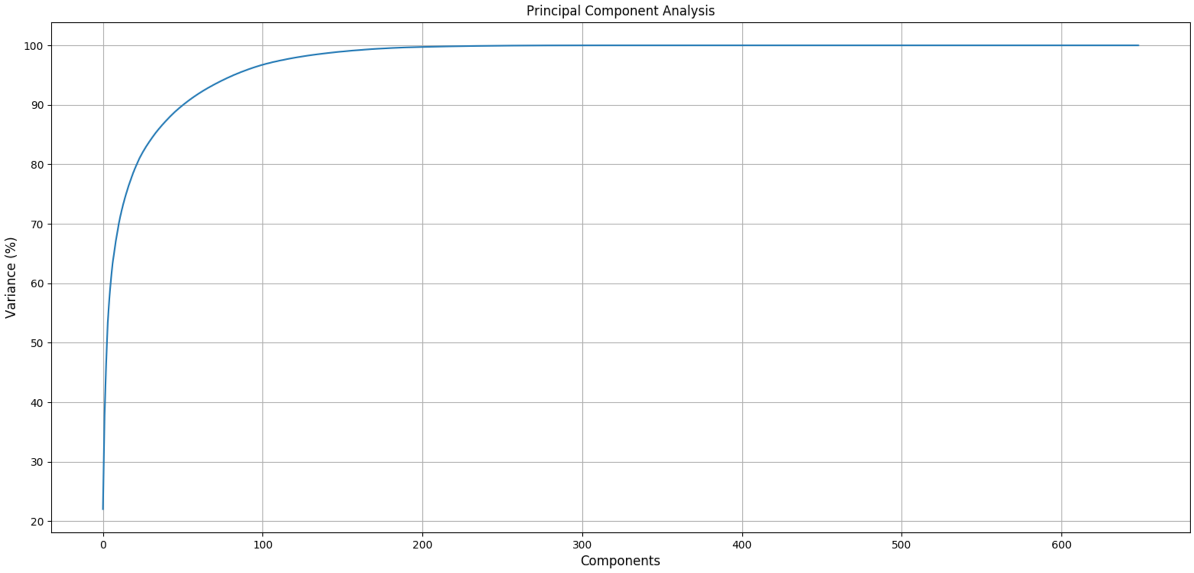
Name	**Formula**	s[k]
Duration	$length(s[k])/Fs$	$I_i[k]$
Zero-crossing rate	$\frac{count(s[k]s[k-1]<0)}{length(s[k])}$	$I_i[k]$
Maximum Energy	$max(s[k])$	$I_i^2[k]$ $PSD_k(I_i[n])$
Maximum index	$argmax(s[k])$	$I_i^2[k]$ $PSD_k(I_i[n])$
Centroid	$\frac{\sum_k k\,s[k]}{\sum_k s[k]}$	$I_i^2[k]$ $PSD_k(I_i[n])$
Skewness	$\frac{1}{length(s)} \cdot \sum_k \left(\frac{s[k]-mean(s)}{std(s)}\right)^3$	$\lvert H\{I_i[k]\}\rvert$ $PSD_k(I_i[n])$
Kurtosis	$\frac{1}{length(s)} \cdot \sum_k \left(\frac{s[k]-mean(s)}{std(s)}\right)^4$	$\lvert H\{I_i[k]\}\rvert$ $PSD_k(I_i[n])$
Increase vs decrease duration	$\frac{t_M - t_{init}}{t_{final} - t_M}$ where $t_M = argmax(s[k])$	$\lvert H\{I_i[k]\}\rvert$
Maximum increment and decrement	$max(s[k]-s[k-1])$ $min(s[k]-s[k-1])$ with $s[k]s[k-1]<0$	$I_i^2[k]$ $PSD_k(I_i[n])$
MFCC	$s[k]$	$F_k^{-1}\{\log\lvert F\{I_i[n]\}\rvert\}$
Others	mean, standard deviation, Shannon and Renyi entropy, etc.	-

implemented in seismic monitoring centers in real time.

4.4 Classification

The last step of the proposed classification system is design a classifying algorithm, for performing the separation of the 5 classes in a space of 200 dimensions given by the PCA components. Four classification techniques were initially tested: as Multilayer Perceptron (MLP), Linear Discriminant Analysis (LDA), Random Forest (RF) and SVM. These methods have already been tested in the context of seismic events. For instance, MLP was used to classify three classes of the Stromboli volcano, in Italy (GIACCO *et al.*, 2009). The LDA was tested for classifying seismic signals with the goal of differentiating earthquakes from man-made

Figure 34 – Principal component analysis of the feature vector.



Source: Own author.

explosions (LINDENBAUM *et al.*, 2018). The RF was used in the classification of earthquake and non-earthquake signals (LI *et al.*, 2018) and the SVM was used to perform classification of volcanic events (MALFANTE *et al.*, 2018).

The four supervised machine learning models above mentioned (MLP, LDA, RF and SVM) will be used to measure the performance of the proposed methodology, and, as it will be viewed in Chapter 5, the SVM technique provided the best results. For this reason, the most part of the simulation results were generated using this method. Several simulations were carried out in order to compare different SVM kernels and penalty parameters. The best results were obtained with a RBF kernel, for a Gaussian parameter of $\gamma = 0.002$, and a penalty parameter equal to $C = 10$. These parameters were used as the default configuration of the SVM.

5 SIMULATION RESULTS

This chapter presents simulation results that evaluate the performance of the proposed method. The database used in the experiments is described in Section 3.4. In order to maintain a balance among the number of samples in each class, the final database has 800 observations from each class, excepting the EX class that has only 592 samples, which leads to a total of 3792 samples. Hold out cross validation is used for the machine learning techniques, with 70% of data used for training and 30% for testing, that is, 2654 samples were used as training data and 1138 as test data. The experiments in this chapter are performed with Scikit-learn (PEDREGOSA *et al.*, 2011).

5.1 Multi-channel vs single-channel, and Machine Learning model choice

The first experiment carried out has the objective of comparing the performances of several classification algorithms, and of evaluating the impact of using multiple channels. Table 3 shows the success rates obtained by the proposed methodology with the MLP, LDA, RF and SVM classification techniques, for 1 and 4 channels. Many simulations were carried out to adjust some parameters of these classifiers. For the MLP, the best results were found with 2 layers, 100 neurons in the first layer and 50 in the second, and rectified linear units as activation functions. For the RF method, 750 trees provided the best success rates. For the SVM, as earlier mentioned, the best results were obtained with a RBF kernel, for a Gaussian parameter of $\gamma = 0.002$ and a penalty parameter equal to $C = 10$.

It can be viewed in Table 3 that, for all the tested cases, the multichannel approach provides a higher success rate than the single channel approach. The main difference between these two approaches is observed when the SVM technique is used. In this case, the use of the multiple channels improves the success rates in 7.3%. It can also be viewed in Table 3 that the SVM provided the best results, for both the single-channel and multi-channel cases. The best success achieved by the proposed classification system, obtained with the SVM and multiple channels, is equal to 90.5%.

Figure 35 shows the performance of machine learning models over experience, this is called learning curves. These curves show the convergence of the model in its learning, showing the growth in the success rate as the number of training samples grows, in this case until 2654 samples.

Table 3 – Success rate for several classification techniques - single-channel and multi-channel cases

Classifier	**Success Rate**	
	Single-channel	Multi-channel
LDA	79.3%	83.6%
MLP	80.9%	84.6%
RF	80.8%	86.6%
SVM	83.2%	90.5%

In the case of LDA, Figure 35a shows that the training score is declining significantly, while the cross-validation increases until converging with the training score (separated by a space between 3∼4%). It can be seen that this model converges quickly (around 1600 training samples).

In the case of the MLP, RF and SVM techniques, the training score remains constant, except for the SVM, where it declines a bit, while the cross-validation score tends to converge towards the training score. The learning curve of the SVM (Figure 35d) suggests that, if there were more training data, the performance would be better, but this implies having more data of the Explosion class (class with fewer events). On the other hand, as shown in Figure 35c, the RF model converges quickly (around 1500 training samples) and the MLP learning curve (Figure 35b) shows that a greater amount of training data is needed for optimal performance.

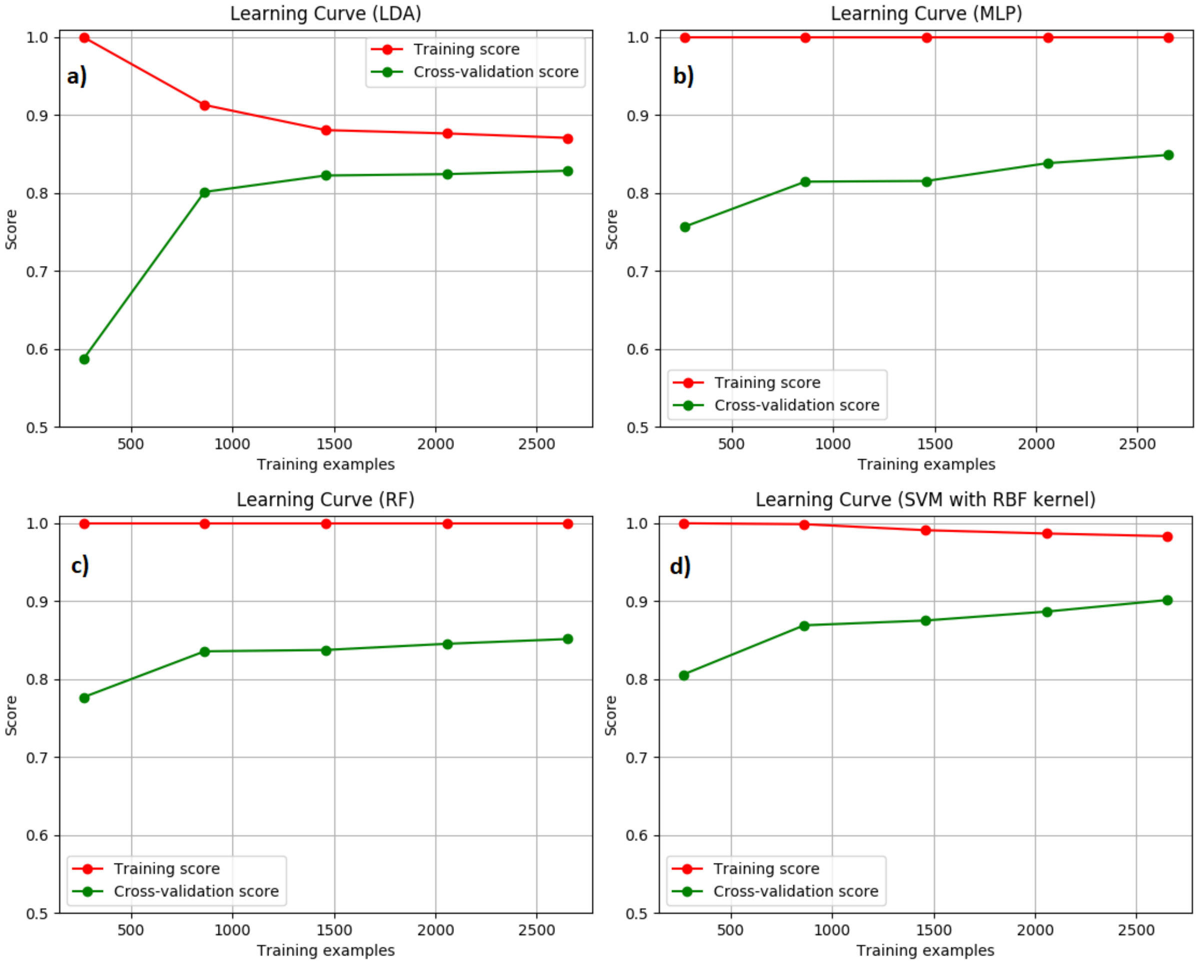
Because of the high success rate presented by the SVM model, this technique will be used as the training model for the next simulation results.

5.2 EMD performance

The next simulation results have the objective of evaluating the impact of the use of the EMD. Table 4 shows the confusion matrix obtained by the proposed classification system using the SVM and multiple channels, using the EMD (in parentheses) and without using the EMD. When the EMD is not used, the attributes are directly calculated from the raw time series (4 signals), obtaining 217 attributes using all the three aforementioned domains. Firstly, it can be concluded from this table that all the classes have balanced success classification rates, the EX and TR classes presenting the best results and the LP providing the worst performance. The best success rate is obtained by the EX class with the use of the EMD (98.8 %). This comes from the fact that the EX class can be easily distinguished from the other classes due to its high energy. On the other hand, the LP class is sometimes mistaken with the VT class.

It can also be viewed from Table 4 that EMD increases the success rate of 4 of the 5

Figure 35 – Learning curves of the machine learning techniques a) LDA, b) MLP, c) RF and d) SVM.



Source: Own author.

classes, the LP class being the most impacted by the use of the EMD. Indeed, the success rate of the LP is improved by 4.2% when the EMD is used. In contrast, the classification rate of the HB class is slightly worse when EMD is used. This is due to the fact that the signals of the HB class share characteristics of the VT and LP classes. As the LP class has generally low-frequency components and the VT signal are characterized by high-frequency components, the HB class contains considerable energies at both high and low frequencies. As a consequence, the EMD of a HB signal has significant energy at first IMFs (high frequencies), as well as the last IMFs (low frequencies). This means that the IMFs of the HB class are often similar to those of the LP and VT classes, which may cause a classification error when a HB event occurs. This behavior is illustrated in Table 4 that shows a significant number of errors from the true class HB to the estimated class VT and, to a lesser extent, to estimated class LP.

Moreover, the overall success rate is improved by 1.5% when the EMD is used, compared with the case where it is not used. This means that the use of the EMD decreases the

Table 4 – Confusion matrix with the true and predicted classes, using multiple channels and SVM - without EMD and with EMD (in parentheses)

Overall (%):		**True class**				
89.0 (90.5)		LP	TR	EX	VT	HB
Predicted class	LP	**198 (208)**	3 (3)	0 (0)	5 (5)	4 (3)
	TR	4 (1)	**231 (232)**	0 (0)	19 (17)	2 (3)
	EX	0 (0)	0 (0)	**171 (176)**	2 (2)	3 (4)
	VT	29 (21)	0 (0)	3 (0)	**203 (205)**	21 (21)
	HB	9 (10)	6 (5)	4 (2)	11 (11)	**210 (209)**
Accuracy (%)		82.5 (86.7)	96.2 (96.7)	96.1 (98.8)	84.6 (88.0)	87.5 (87.1)

error rate from 11% to 9.5%. Although the gain provided by the EMD in the overall success rate is not very high, it should be highlighted that the EMD yielded more significant gains for LP and EX classes (2.7% and 4.2%, respectively). Indeed, the detection of these classes can be considered more relevant for the classification system, as the most violent volcanic events generally fall into the EX class, and the LP is very relevant for the forecast of eruptions (MCNUTT, 2005), (CHOUET, 1996), (AKI; FERRAZZINI, 2000).

5.3 Performance of instrumental correction

The next experiment evaluates the impact of the instrumental correction described in Subsection 4.1. Table 5 shows the confusion matrix obtained without the instrumental correction. Comparing the results of Tables 4 and 5, it can be viewed that the instrument correction has a great impact on the success rate. Indeed, without the preprocessing, all the classes showed a reduction in the success rate in relation to the classification using the instrumental correction. The overall success rate falls to 87.1% without the instrument correction. This is due to the fact that, as earlier mentioned, without the signal conditioning, valuable information of the physical energy of the signals is lost. On the other hand, the instrumental correction gives the energy of the signals a physical sense, providing valuable information.

5.4 Performance of the temporal, spectral and cepstral domains

This experiment evaluates the success rate when the attributes are extracted from different domains, using the SVM with the EMD and multiple channels. Table 6 shows the

Table 5 – Confusion matrix with true classes and predicted classes, using multiple channels and SVM - without instrument correction.

		True class					
		LP	TR	EX	VT	HB	Overall
Predicted class	LP	**193**	1	1	4	0	
	TR	4	**231**	1	18	12	
	EX	1	0	**164**	0	9	
	VT	29	0	4	**199**	14	
	HB	13	7	8	19	**205**	
Accuracy(%)		80.4	96.2	92.1	82.9	85.4	87.1

Table 6 – Success rate using attributes from different domains.

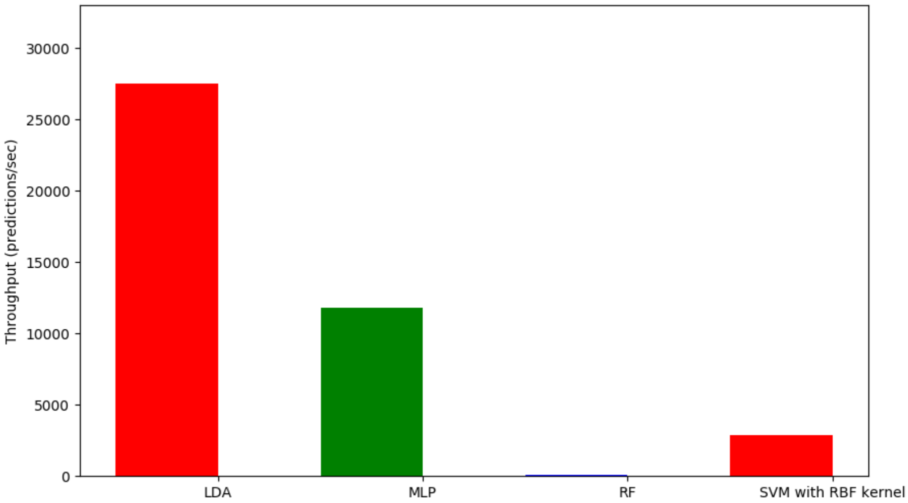
Domains of the attributes	**Success Rate**
Temporal	81.6
Spectral	78.2
Cepstral	85.0
Temporal-Spectral	84.2
Temporal-Cepstral	87.5
Spectral-Cepstral	85.5
Temporal-Spectral-Cepstral	90.5

success rates obtained using seven different combinations of domains. When only one domain is used, the cepstral domain provided the best result, reaching 85.0% of success rate, and the worst performance is obtained by the spectral domain, with a success rate of 78.2%. As expected, using only one domain of attributes leads to worse success rates than using more than one domain. It can also be observed that the use of the three domains together generates the best success rate (90.5%), with a 3% gain over the second best case (temporal-cepstral), which corroborates with the use of the three domains in the proposed classification system.

5.5 Latency in predicting new data

As mentioned earlier, one of the objectives of the proposed classification system is that it could be implemented in monitoring centers in real time. It is then vitally important to know how long it takes to complete the classification for new data. To do this, a simulation is performed with the 4 classification algorithm above mentioned (LDA, MLP, RF and SVM). This simulation consists in making predictions for new data (test data) in a fixed time interval, in order to make a statistic of the number of predictions per second of each model. Figure 36 shows the number of predictions per second of the machine learning models, with LDA being the model that makes the most predictions per second, and RF the model that has the least predictions per second. For the SVM, Figure 36 shows that this model has approximately 2800 predictions per second.

Figure 36 – Number of predictions per second of the machine learning models.



Source: Own author.

While it is true that the LDA makes a huge amount of predictions per second, SVM has a better performance in the success rate of correct predictions, with a significant number of predictions per second of 2800. This number is more than enough for the number of seismic events that occur in a volcano.

6 CONCLUSIONS AND PERSPECTIVES

An automatic classification system for identifying the five most important types of events of a volcano was presented in this dissertation, using the EMD in the feature extraction block. This decomposition, in conjunction with machine learning techniques, has shown to be a promising tool for classification of volcanic-seismic signals. Although the gain provided by the EMD in the overall success rate is not very high, it yielded more significant gains for the LP and EX classes, whose detection can be considered more relevant than the other classes.

Another contribution of this work is the use of multiple seismic channels to perform the classification, contrary to previous works that use only a single channel. The multi-channel approach has provided much smaller error rates when compared to the single-channel case, due to the valuable information added to the classifier. The presented system also performs an instrument correction that helps significantly in the recognition of the classes. This preprocessing standardizes the signals of the seismic sensors to their real values in m/s, making the proposed system independent of the types of sensors used and giving a physical sense to the data. Concerning the classification algorithm, four classification techniques were tested in conjunction with PCA, the SVM providing the best results.

The learning curves of the SVM model have a convergence that can even be improved if a larger database is obtained, which would imply even a larger number of explosion events. In addition, the latency provided by the SVM when predicting new data demonstrates that it would have no problem when implemented in real-time monitoring centers.

The present investigation used a large database from the Ubinas volcano located in Arequipa, Peru. This database is particularly rich in explosion events, when compared with other volcano databases. The simulation showed a good performance of the proposed classifier, with a success rate of 90.5%.

In future works, a complexity analysis of the proposed method will be carried out. Moreover, the presented classification system will be implemented in real time in the volcano monitoring center of the IGP.

REFERENCES

AKI, K. State of the art in volcanic seismology. In: GASPARINI, P.; SCARPA, R.; AKI, K. (Ed.). **Volcanic Seismology**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1992. p. 3–10. ISBN 978-3-642-77008-1.

AKI, K.; FERRAZZINI, V. Seismic monitoring and modeling of an active volcano for prediction. **Journal of Geophysical Research: Solid Earth**, v. 105, p. 16617–16640, 2000.

ALLEN, R. V. Automatic earthquake recognition and timing from single traces. **Bulletin of the Seismological Society of America**, v. 68, n. 5, p. 1521–1532, 10 1978.

ASTAPOV, S. **Feature extraction from band-limited signals and classification of features**. Dissertação (Master's) — Faculty of Information Technology, Tallinn University of Technology, Estonia, 2011.

BACKES, A.; SÁ JUNIOR, J. J. de M. **Introdução à Visão Computacional Usando MATLAB**. [S.l.]: Alta Books, 2016. ISBN 978-85-508-0023-3.

BAILLARD, C.; CRAWFORD, W.; BALLU, V.; HIBERT, C.; MANGENEY, A. An automatic kurtosis-based p - and s -phase picker designed for local seismic networks. **Bulletin of the Seismological Society of America**, v. 104, 02 2014.

BEN-HUR, A.; WESTON, J. A user's guide to support vector machines. **Methods in molecular biology (Clifton, N.J.)**, v. 609, p. 223–39, 01 2010.

BENBRAHIM, M.; BENJELLOUN, K.; IBENBRAHIM, A.; KASMI, M.; ARDIL, E. A new approaches for seismic signals discrimination. **International Journal of Geological and Environmental Engineering**, v. 1, p. 64–67, 2007.

BENITEZ, C.; RAMIREZ, J.; SEGURA, J. C.; RUBIO, A.; NEZ, J. M. I.; ALMENDROS, J.; GARCIA-YEGUAS, A. Continuous hmm-based volcano monitoring at deception island, antarctica. **2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings**, v. 5, p. V–749–V752, 5 2006.

BENITEZ, M. C.; RAMIREZ, J.; SEGURA, J. C.; IBANEZ, J. M.; ALMENDROS, J.; GARCIA-YEGUAS, A.; CORTES, G. Continuous hmm-based seismic-event classification at deception island, antarctica. **IEEE Transactions on Geoscience and Remote Sensing**, v. 45, p. 138–146, 1 2007.

BOGERT, B. P. The quefrency analysis of time series for echoes : cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. **Proc. Symposium on Time Series Analysis**, John Wiley & Sons, p. 209–243, 1963.

BORMANN, P.; WIELANDT, E. **IASPEI New manual of seismological observatory practice 2 (NMSOP-2)**. 2. ed. [S.l.]: Potsdam : Deutsches GeoForschungsZentrum GFZ, 2012.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory**. [S.l.]: ACM Press, 1992. p. 144–152.

BREIMAN, L. Random forests. **Mach. Learn.**, Kluwer Academic Publishers, Norwell, MA, USA, v. 45, n. 1, p. 5–32, out. 2001. ISSN 0885-6125.

BUENO, A.; TITOS, M.; GARCÍA, L.; ÁLVAREZ, I.; NEZ, J. I.; BENÍTEZ, C. Classification of volcano-seismic signals with bayesian neural networks. **2018 26th European Signal Processing Conference (EUSIPCO)**, p. 2295–2299, 9 2018.

CAMP, J. B.; CANNIZZO, J. K.; NUMATA, K. Application of the hilbert-huang transform to the search for gravitational waves. **PHYS REV D**, 1 2007.

CHAN, C.-C.; LIN, C.-J. LIBSVM: A library for support vector machines. **ACM Transactions on Intelligent Systems and Technology**, v. 2, p. 27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

CHILO, J.; KINSER, J. M.; LINDBLAD, T. Discrimination of nuclear explosions sites by seismic signals using intrinsic mode functions and multi-modal data space. In: **IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium**. [S.l.: s.n.], 2008. v. 2, p. II–895–II–898.

CHOUET, B. A. Long-period volcano seismicity: its source and use in eruption forecasting. **Nature**, v. 380, p. 309–316, 1996.

CUEVA, R. L.; SALAZAR, V. P.; CHALUISA, M. V. Towards an automatic detection system of signals at cotopaxi volcano. **Dyna**, v. 84, p. 176–184, 2017.

CURILEM, G.; VERGARA, J.; FUENTEALBA, G.; NA, G. A.; CHACÓN, M. Classification of seismic signals at villarrica volcano (chile) using neural networks and genetic algorithms. **Journal of Volcanology and Geothermal Research**, v. 180, p. 1–8, 2009.

DU, W.; ZHANG, Z. Application of wavelet analysis in seismic data processing of volcanic rock region. **2010 International Conference on Intelligent System Design and Engineering Application**, v. 1, p. 371–374, Oct. 2010.

DU, W.; ZHANG, Z. Application of wavelet analysis in seismic data processing of volcanic rock region. **2010 International Conference on Intelligent System Design and Engineering Application**, v. 1, p. 371–374, 10 2010.

GIACCO, F.; ESPOSITO, A. M.; SCARPETTA, S.; GIUDICEPIETRO, F.; MARINARO, M. Support vector machines and mlp for automatic classification of seismic signals at stromboli volcano. **Proceedings of the 2009 Conference on Neural Nets WIRN 2009**, p. 116–123, 2009.

GROSS, J. C.; RITTER, J. R. R. Time domain classification and quantification of seismic noise in an urban environment. **Geophysical Journal International**, v. 179, p. 1213–1231, 11 2009.

GRULIER, V.; DEBERT, S.; MARS, J. I.; PACHEBAT, M. Acoustic and turbulent wavenumbers separation in wall pressure array signals using emd in spatial domain. **2008 IEEE International Conference on Acoustics, Speech and Signal Processing**, p. 333–336, 2008.

GUTIÉRREZ, L.; NEZ, J. I.; CORTÉS, G.; RAMÍREZ, J.; BENÍTEZ, C.; TENORIO, V.; ISAAC, . Volcano-seismic signal detection and classification processing using hidden markov models. application to san cristóbal volcano, nicaragua. **2009 IEEE International Geoscience and Remote Sensing Symposium**, v. 4, p. IV–522–IV–525, 2009.

HAN, J. **Empirical Mode Decomposition for Seismic Aplications**. Tese (Ph.D) — Department of Physics, University of Alberta, Canada, 2014.

HAN, R.; LI, J.; CUI, G.; WANG, X.; WANG, W.; LI, X. Seismic signal detection algorithm based on gs transform filtering and emd denoising. **2018 IEEE 4th International Conference on Computer and Communications (ICCC)**, p. 1213–1217, 12 2018.

HAVSKOV, J.; ALGUACIL, G. **Instrumentation in Earthquake Seismology**. 2. ed. [S.l.]: Springer, 2004.

HOSOKAWA, M.; JEONG, B.; TAKIZAWA, O. Earthquake intensity estimation and damage detection using remote sensing data for global rescue operations. In: **2009 IEEE International Geoscience and Remote Sensing Symposium**. [S.l.: s.n.], 2009. v. 2, p. II–420–II–423.

HUANG, N. E.; SHEN, Z.; LONG, S. R.; WU, M. C.; SHIH, H. H.; ZHENG, Q.; YEN, N. C.; TUNG, C. C.; LIU, H. H. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. **Proceedings of the Royal Society of London**, v. 454, p. 903–995, 3 1998.

INZA, L. A. **Understanding magmatic processes and seismo-volcano source localization with multicomponent seismic arrays**. Tese (Ph.D) — Earth Sciences, Université de Grenoble, France, 2013.

INZA, L. A.; MARS, J. I.; MÉTAXIAN, J. P.; O'BRIEN, G. S.; MACEDO, O. Seismo-volcano source localization with triaxial broad-band seismic array. **Geophys Journal International**, p. 371–384, 2011.

INZA, L. A.; MÉTAXIAN, J. P.; MARS, J.; BEAN, C.; O'BRIEN, G. S.; MACEDO, O.; ZANDOMENEGHI, D. Analysis of dynamics of vulcanian activity of ubinas volcano, using multicomponent seismic antennas. **Journal of Volcanology and Geothermal Research**, v. 270, p. 35–52, 1 2014.

JIA, Y.; MA, J. What can machine learning do for seismic data processing? an interpolation application. **GEOPHYSICS**, v. 82, p. V163–V177, 5 2017.

KHAN, F.; ENZMANN, F.; KERSTEN, M. Multi-phase classification by a least-squares support vector machine approach in tomography images of geological samples. **Solid Earth**, v. 7, p. 481–492, 03 2016.

KISLOV, K. V.; GRAVIROV, V. V. Use of artificial neural networks for classification of noisy seismic signals. **Seismic Instruments**, v. 53, p. 87–101, 1 2017.

KOMATY, A.; BOUDRAA, A.; DARE, D. Emd-based filtering using the hausdorff distance. **2012 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)**, 12 2012.

LI, W.; NAKSHATRA, N.; NARVEKAR, N.; RAUT, N.; SIRKECI, B.; GAO, J. Seismic data classification using machine learning. **2018 IEEE Fourth International Conference on Big Data Computing Service and Applications**, 3 2018.

LINDENBAUM, O.; BREGMAN, Y.; RABIN, N.; AVERBUCH, A. Multiview kernels for low-dimensional modeling of seismic events. **IEEE Transactions on Geoscience and Remote Sensing**, v. 56, p. 3300–3310, 2 2018.

LINS, A. P. S.; LUDERMIR, T. B. Hybrid optimization algorithm for the definition of mlp neural network architectures and weights. In: **Fifth International Conference on Hybrid Intelligent Systems (HIS'05)**. [S.l.: s.n.], 2005. p. 6 pp.–.

LOWRIE, W. **Fundamentals of Geophysics**. 2. ed. [S.l.]: Cambridge University Press, 2007.

MACEDO, O.; MÉTAXIAN, J. P.; TAIPE, E.; RAMOS, D.; INZA, L. A. Seismicity associated with the 2006-2008 eruption, ubinas volcano. **The VOLUME Project**, p. 262–270, 2009.

MALFANTE, M.; DALLA-MURA, M.; MÉTAXIAN, J. P.; MARS, J. I.; MACEDO, O.; INZA, L. A. Machine learning for volcano-seismic signals: Challenges and perspectives. **IEEE Signal Processing Magazine**, v. 35, p. 20–30, 3 2018.

MCNUTT, S. R. Volcanic seismology. **Annual Review of Earth and Planetary Sciences**, v. 33, p. 461–491, 2005.

MILGRAM, J.; CHERIET, M.; SABOURIN, R. "one against one" or "one against all": Which one is better for handwriting recognition with svms? 10 2006.

NUHA, H. H.; SUWASTIKA, N. A. Fractional fourier transform for decreasing seismic data lossy compression distortion. In: **2015 3rd International Conference on Information and Communication Technology (ICoICT)**. [S.l.: s.n.], 2015. p. 590–593.

NUNES, J.; DELÉCHELLE, E. Empirical mode decomposition: Applications on signal and image processing. **Advances in Adaptive Data Analysis**, v. 1, p. 125–175, 2009.

OHRNBERGER, M. **Continuous automatic classification of seismic signals of volcanic origin at Mt. Merapi, Java, Indonesia**. Tese (Ph.D) — Faculty of Mathematics and Natural Sciences, University of Potsdam, Germany, 2001.

OKUBO, P. G.; NAKATA, J. S.; KOYANAGI, R. Y. Chapter 2 the evolution of seismic monitoring systems at the hawaiian volcano observatory. **U.S. Geological Survey**, 2014.

PANAGIOTA, M.; JOCELYN, C.; ERWAN, P.; PHILIPPE, G. A support vector regression approach for building seismic vulnerability assessment and evaluation from remote sensing and in-situ data. In: **2012 IEEE International Geoscience and Remote Sensing Symposium**. [S.l.: s.n.], 2012. p. 7533–7536.

PARHI, K. K.; AYINALA, M. Low-complexity welch power spectral density computation. **IEEE Transactions on Circuits and Systems I: Regular Papers**, v. 61, p. 172–182, 1 2014.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PENG, P.; HE, Z.; WANG, L. Automatic classification of microseismic signals based on mfcc and gmm-hmm in underground mines. **Shock and Vibration**, v. 2019, p. 1–9, 1 2019.

RAHMAN, M. B.; NURHASANAH, I. S.; NUGROHO, S. P. Community resilience: Learning from mt merapi eruption 2010. **Procedia - Social and Behavioral Sciences**, v. 227, p. 387–394, 2016.

RAMIREZ, J.; MEYER, F. G. Machine learning for seismic signal processing: Phase classification on a manifold. **2011 10th International Conference on Machine Learning and Applications and Workshops**, v. 1, p. 382–388, 2011.

RANDALL, R. A history of cepstrum analysis and its application to mechanical problems. **Mechanical Systems and Signal Processing**, v. 97, 12 2016.

ROWELL, D. **Signal Processing: Continuous and Discrete**. [S.l.]: Massachusetts Institute of Technology: MIT OpenCourseWare, 2008.

SAEED, B. S. **De-noising seismic data by Empirical Mode Decomposition**. Tese (Ph.D) — Department of Geosciences, University of Oslo, Norway, 2011.

SCARPETTA, S.; GIUDICEPIETRO, F.; EZIN, E. C.; PETROSINO, S.; PEZZO, E. D.; MARTINI, M.; MARINARO, M. Automatic classification of seismic signals at mt. vesuvius volcano, italy, using neural networks. **Bulletin of the Seismological Society of America**, v. 95, p. 185–196, 2 2005.

SUZUKI, K. **ARTIFICIAL NEURAL NETWORKS – ARCHITECTURES AND APPLICATIONS**. [S.l.: s.n.], 2013. ISBN 978-953-51-0935-8.

VAEZI, Y.; BAAN, M. Van der. Comparison of the STA/LTA and power spectral density methods for microseismic event detection. **Geophysical Journal International**, v. 203, n. 3, p. 1896–1908, 10 2015.

WALD, D.; WORDEN, B.; QUITORIANO, V.; PANKOW, K. Shakemap manual: Technical manual, user's guide, and software guide. **USGS, Techniques and Methods. Advanced national seismic system.**, 2006.

WANG, X.; SLOAN, I. H. Brownian bridge and principal component analysis: towards removing the curse of dimensionality. **IMA Journal of Numerical Analysis**, v. 27, p. 631–654, 10 2007.

WASSERMANN, J. **IASPEI New manual of seismological observatory practice 2 (NMSOP-2)**. 2. ed. [S.l.]: Potsdam : Deutsches GeoForschungsZentrum GFZ, 2012.

WEN, J.; SUN, J. Research on target localization method based on characteristic frequency of empirical mode decomposition. In: **2010 2nd International Conference on Signal Processing Systems**. [S.l.: s.n.], 2010. v. 2, p. V2–118–V2–122.

YILDIRIM, E.; GÜLBAG, A.; GÜNDÜZ, H.; EMRAH, D. Discrimination of quarry blasts and earthquakes in the vicinity of istanbul using soft computing techniques. **Computers & Geosciences**, v. 37, p. 1209–1217, 9 2011.

YU, J.; ZHANG, Z. Research on the seismic signal denoising with the lmd and emd method. In: **2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)**. [S.l.: s.n.], 2017. p. 767–771.

ZHANG, Y.; LI, S.; YIN, Z.; CHEN, B.; CUI, H.-L.; NING, J. Fiber-Bragg-grating-based seismic geophone for oil/gas prospecting. **Optical Engineering**, SPIE, v. 45, n. 8, p. 1 – 4, 2006.

ZHOU, Q.; TONG, G.; XIE, D.; LI, B.; YUAN, X. A seismic-based feature extraction algorithm for robust ground target classification. **IEEE Signal Processing Letters**, v. 19, p. 639–642, 10 2012.

APPENDIX A – ATTRIBUTES

This appendix details all the attributes used in this work, with their corresponding domain and signal. Most of these attributes were used in (MALFANTE *et al.*, 2018) and (ASTAPOV, 2011).

(1) Duration of the event: This attribute is extracted from the original time domain signal $x[n]$ and it is the only attribute that is not extracted from the IMFs $I_i[n]$. Fs being the sampling frequency and N the number of samples, the duration (in seconds) is calculated as follows:

$$Duration = \frac{N}{Fs}. \tag{A.1}$$

(2) Maximum temporal energy: Maximum value of the signal $I_i^2[n]$.

$$E_{max} = max(I_i^2[n]). \tag{A.2}$$

(3) Index of maximum Temporal energy: Time in which the temporal energy $I_i^2[n]$ is maximum.

$$n_{E_{max}} = argmax(I_i^2[n]). \tag{A.3}$$

(4) Total temporal energy.

$$Ene = \sum_{n=1}^{N} I_i^2[n]. \tag{A.4}$$

(5) Temporal energy RMS.

$$Ene_{RMS} = \sqrt{\frac{Ene}{N}}. \tag{A.5}$$

(6) Temporal centroid: Centroid of time weighted by its energy $I_i^2[n]$.

$$Centroid_t = \frac{\sum_{n=1}^{N}(n.I_i^2[n])}{Ene}. \tag{A.6}$$

(7) Bandwidth in time: Variance around the Temporal centroid of $I_i^2[n]$.

$$BW_t = \sqrt{\frac{\sum_{n=1}^{N}((n-Centroid_t)^2.I_i^2[n])}{Ene}}. \tag{A.7}$$

(8) Skewness around bandwidth in time.

$$SK_{pre} = \frac{\sum_{n=1}^{N}((n-Centroid_t)^3.I_i^2[n])}{Ene.BW_t^3}, \tag{A.8}$$

where skewness around bandwidth is:

$$Skewness_{BW_t} = \begin{cases} \sqrt{SK_{pre}}, & \text{if } SK_{pre} \geq 0 \\ -\sqrt{SK_{pre}}, & \text{otherwise.} \end{cases}$$

(9) Kurtosis around bandwidth in time.

$$Kurtosis_{BW_t} = \sqrt{\frac{\sum_{n=1}^{N}((n-Centroid_t)^4.I_i^2[n])}{Ene.BW_t^4}}. \tag{A.9}$$

(10) Maximum increment of temporal energy.

$$TP_{t_{max}} = max(I_i^2[n] - I_i^2[n-1]) \qquad with: I_i^2[n]I_i^2[n-1] < 0. \tag{A.10}$$

(11) Maximum decrement of temporal energy.

$$TP_{t_{min}} = min(I_i^2[n] - I_i^2[n-1]) \qquad with: I_i^2[n]I_i^2[n-1] < 0. \tag{A.11}$$

(12) Threshold count rate of the envelope signal: How many times the signal exceeds 80% of its maximum amplitude per second, signal $s[k] = |H\{I_i[k]\}|$.

$$TCR_t = \frac{count(\frac{s[k]}{max(s[k])} \geq 0.8)}{duration}. \tag{A.12}$$

(13) Ratio of maximum amplitude envelope to its mean, signal $s[k] = |H\{I_i[k]\}|$.

$$RMM_t = \frac{max(s[k])}{mean(s[k])}. \tag{A.13}$$

(14) Mean envelope, signal $s[k] = |H\{I_i[k]\}|$.

$$Mean_{env} = \frac{\sum_{k=1}^{N} s[k]}{N}. \tag{A.14}$$

(15) Standard deviation of the envelope signal, signal $s[k] = |H\{I_i[k]\}|$.

$$STD_{env} = \sqrt{\frac{\sum_{k=1}^{N}(s[k] - Mean_{env})^2}{N}}. \tag{A.15}$$

(16) Skewness envelope, signal $s[k] = |H\{I_i[k]\}|$.

$$Skewness_{env} = \frac{\sum_{k=1}^{N} (\frac{s[k] - Mean_{env}}{STD_{env}})^3}{N}. \tag{A.16}$$

(17) Kurtosis envelope, signal $s[k] = |H\{I_i[k]\}|$.

$$Kurtosis_{env} = \frac{\sum_{k=1}^{N} (\frac{s[k] - Mean_{env}}{STD_{env}})^4}{N}. \tag{A.17}$$

(18) Increase vs decrease duration, signal $s[k] = |H\{I_i[k]\}|$.

$$IncDec_{env} = \frac{t_M - t_{init}}{t_{final} - t_M} \qquad where : t_M = argmax(s[k]). \tag{A.18}$$

(19) Increase vs total duration, signal $s[k] = |H\{I_i[k]\}|$.

$$IncTot_{env} = \frac{t_M - t_{init}}{t_{final} - t_{init}} \qquad where : t_M = argmax(s[k]). \tag{A.19}$$

(20) Number of points ratio that do not exceed a threshold of 80% of its maximum, signal $s[k] = |H\{I_i[k]\}|$.

$$mTCR_{env} = length(s[k]/max(s[k])) \geq 0.8)/N. \tag{A.20}$$

(21) Shannon entropy, signal $s[k] = |H\{I_i[k]\}|$, $bins = 200$.

$$Shannon_{env} = \sum_{i=1}^{bins} -Prob[i].\log_2(Prob[i]), \tag{A.21}$$

where:

$$Prob[i] = Histogram(s[k], bins). \tag{A.22}$$

(22) Renyi entropy, signal $s[k] = |H\{I_i[k]\}|$, $bins = 200$, $\alpha = 2$.

$$Renyi_{env} = \frac{\log_2 \sum_{i=1}^{bins} Prob^{\alpha}[i]}{1 - \alpha}, \tag{A.23}$$

where:

$$Prob[i] = Histogram(s[k], bins). \tag{A.24}$$

(23) Zero crossing rate: How many times per second the signal $I_i[n]$ changes sign.

$$ZCR_t = \frac{count(I_i[n] < 0)}{Duration} \qquad with : I_i[n]I_i[n-1] < 0. \tag{A.25}$$

(24) Maximum spectral energy: maximum value in signal $s[k] = PSD_k(I_i[n])$.

$$PSD_{max} = max(s[k]). \tag{A.26}$$

(25) Index of maximum spectral energy: frequency in which the PSD ($s[k] = PSD_k(I_i[n])$) is maximum.

$$f_{PSD_{max}} = argmax(s[k]). \tag{A.27}$$

(26) Spectral centroid: Centroid of frequency weighted by its PSD, signal $s[k] = PSD_k(I_i[n])$.

$$Centroid_f = \frac{\sum_{k=1}^{N}(k.s[k])}{\sum_{k=1}^{N} s[k]}. \tag{A.28}$$

(27) Bandwidth in frequency: variance around the Spectral centroid of $s[k] = PSD_k(I_i[n])$.

$$BW_f = \sqrt{\frac{\sum_{k=1}^{N}((k - Centroid_f)^2.s[k])}{\sum_{k=1}^{N} s[k]}}. \tag{A.29}$$

(28) Skewness around bandwidth in frequency, signal $s[k] = PSD_k(I_i[n])$.

$$SK_{pre} = \frac{\sum_{k=1}^{N}((k - Centroid_f)^3.s[k])}{BW_f^3.\sum_{k=1}^{N} s[k]}, \tag{A.30}$$

where skewness around bandwidth in frequency is:

$$Skewness_{BW_f} = \begin{cases} \sqrt{SK_{pre}}, & \text{if } SK_{pre} \geq 0 \\ -\sqrt{SK_{pre}}, & \text{otherwise.} \end{cases}$$

(29) Kurtosis around bandwidth in frequency, signal $s[k] = PSD_k(I_i[n])$.

$$Kurtosis_{BW_f} = \sqrt{\frac{\sum_{k=1}^{N}((k - Centroid_f)^4.s[k])}{BW_f^4.\sum_{k=1}^{N} s[k]}}. \tag{A.31}$$

(30) Maximum increment of spectral energy, signal $s[k] = PSD_k(I_i[n])$.

$$TP_{f_{max}} = max(s[k] - s[k-1]) \qquad with: s[k]s[k-1] < 0. \tag{A.32}$$

(31) Maximum decrement of spectral energy, signal $s[k] = PSD_k(I_i[n])$.

$$TP_{f_{min}} = min(s[k]-s[k-1]) \qquad with: s[k]s[k-1] < 0. \tag{A.33}$$

(32) Mean PSD, signal $s[k] = PSD_k(I_i[n])$.

$$Mean_{PSD} = \frac{\sum_{k=1}^{N} s[k]}{N}. \tag{A.34}$$

(33) Standard deviation of the PSD signal, signal $s[k] = PSD_k(I_i[n])$.

$$STD_{PSD} = \sqrt{\frac{\sum_{k=1}^{N} (s[k]-Mean_{PSD})^2}{N}}. \tag{A.35}$$

(34) Skewness PSD, signal $s[k] = PSD_k(I_i[n])$.

$$Skewness_{PSD} = \frac{\sum_{k=1}^{N} (\frac{s[k]-Mean_{PSD}}{STD_{PSD}})^3}{N}. \tag{A.36}$$

(35) Kurtosis PSD, signal $s[k] = PSD_k(I_i[n])$.

$$Kurtosis_{PSD} = \frac{\sum_{k=1}^{N} (\frac{s[k]-Mean_{PSD}}{STD_{PSD}})^4}{N}. \tag{A.37}$$

(36) Shannon entropy, signal $s[k] = PSD_k(I_i[n])$, $bins = 50$.

$$Shannon_{PSD} = \sum_{i=1}^{bins} -Prob[i].\log_2(Prob[i]), \tag{A.38}$$

where:

$$Prob[i] = Histogram(s[k], bins). \tag{A.39}$$

(37) Renyi entropy, signal $s[k] = PSD_k(I_i[n])$, $bins = 50$, $\alpha = 2$.

$$Renyi_{PSD} = \frac{\log_2 \sum_{i=1}^{bins} Prob^{\alpha}[i]}{1-\alpha}, \tag{A.40}$$

where:

$$Prob[i] = Histogram(s[k], bins). \tag{A.41}$$

(38) Ratio of maximum amplitude PSD to its mean, signal $s[k] = PSD_k(I_i[n])$.

$$RMM_f = \frac{max(s[k])}{mean(s[k])}. \tag{A.42}$$

(39) Threshold count rate of the PSD signal: how many times the signal exceeds 40% of its maximum amplitude, signal $s[k] = PSD_k(I_i[n])$.

$$TCR_f = count(\frac{s[k]}{max(s[k])} \geq 0.4). \tag{A.43}$$

(40) Number of points ratio that do not exceed a threshold of 40% of its maximum, signal $s[k] = PSD_k(I_i[n])$.

$$mTCR_{PSD} = length(s[k]/max(s[k])) \geq 0.4)/length(s[k]). \tag{A.44}$$

(41) Total Spectral energy, signal $s[k] = PSD_k(I_i[n])$.

$$PSD_{total} = \sum_{k=1}^{N} s[k]. \tag{A.45}$$

(42) Spectral energy RMS, signal $s[k] = PSD_k(I_i[n])$.

$$PSD_{RMS} = \sqrt{\frac{PSD_{total}}{length(s[k])}}. \tag{A.46}$$

(43) 13 first MFCC, signal $s[k] = F_k^{-1}\{\log|F\{I_i[n]\}|\}$.

$$MFCC = MFCC_{function}(s[k])[13\ coefficients]. \tag{A.47}$$