ANNALS OF GEOPHYSICS, 67, 3, SE320, 2024; doi:10.4401/ag-9055
OPEN ACCESS

Use of deep learning to improve seismic data quality analysis

Paolo Casale[1], Alessandro Pignatelli[1]

(1) Istituto Nazionale di Geofisica e Vulcanologia, Roma, Italy

Article history: received December 4, 2023; accepted June 28, 2024

Abstract

Seismic data are influenced by various types of noise, which are typically categorized into two primary classes: anthropogenic and environmental. However, the detection of instrumental noise or malfunctioning stations also plays a crucial role in ensuring the data quality and the efficiency of a seismic network. The visual inspection of seismic spectral diagrams (e.g. power spectral density) enables us to identify issues that could potentially compromise data quality, thereby affecting subsequent calculations such as Magnitude or Peak Ground Acceleration (PGA). However, this process is time-consuming and demands significant human expertise due to the complexity of the diagrams, compounded by the sheer number of stations requiring examination. Therefore, in this paper, we explore the feasibility of transferring human expertise into an artificial intelligence system to create an automated system capable of rapidly performing such detection. More specifically, in the first part of this paper, we use Probability Density Function (PDF) diagrams, enabling an initial assessment of station performance via visual inspection. We describe this plot type and provide examples that reveal whether a station is functioning correctly or if technical issues exist. A table containing the main evaluation criteria is provided. In the second part of this paper, we demonstrate that these plots can serve as input for a neural network, allowing the development of the aforementioned automated system. Through extensive testing under various conditions, we have observed that the trained network consistently achieves an accuracy rate exceeding 85% across all four conducted tests. In the latest and most significant test, the achieved accuracy is approximately 87%.

Keywords: Seismic data quality; Machine Learning; Convolutional Neural Networks; Seismic Noise; Power Spectral Density

1. Introduction

Seismic data serve as an invaluable source of information, enabling investigations into various phenomena such as earthquake behaviour, fault geometries, tomography, and more. Ensuring data quality control is crucial for conducting accurate analyses of signals produced by seismic stations. For examples, precise knowledge and regular checks of instrumentation sensitivity are essential for accurate evaluation of Peak Ground Acceleration (PGA) and magnitude estimation following an earthquake; time marks on seismogram, number of gaps (data transmission efficiency check) are crucial for Early Warnings improvements [Picozzi et al., 2015]; instrumental transfer function

consistency (also known as instrument response [Wielandt, 2012]) and good signal-to-noise ratio plays a fundamental role in analysing low-frequency signals [Morelli et al., 2000; Custódio et al., 2014; Pondrelli et al., 2020].

Typically, the quality of the seismic data retrieved by a data center, especially if it is used for real-time earthquake detections, is assessed by checking various parameters and using available codes (e.g. MUSTANG [Casey et al., 2018] package available thanks to the IRIS-Earthscope consortium). Some of them allows to check the data transmission such as the verification of station connectivity and the real-time latency information; others are measured in the time domain such as time offset, RMS (Root Mean Square of the signal), sum of gaps, number of gaps, number of spikes and others. However, these metrics usually do not check the correctness of the metadata and not always check the signal anomalies (par. 7), which indeed are fundamental to correctly calculate parameter such as PGA, the magnitude of an event and others. Moreover, to be able to declare a signal reliable or not, a more comprehensive criteria should be applied. As identifying large gaps is generally already performed by automatic procedures, the main aim of this work is to develop a more comprehensive approach by utilizing seismic noise power spectra density to automatically capture signal anomalies (even serious and not rare) that often go unnoticed.

Seismic noise analysis, especially when employing various spectral approaches, is a powerful method for evaluating station performance [McNamara and Boaz, 2006]. This approach not only helps in identifying noise sources but also in assessing the quality of station data. It facilitates the detection of operational issues [McNamara and Boaz, 2006], which can be invaluable for managing seismic networks. In this context, various approaches have been applied. For example, [Massa et al., 2022] present a method that enables the real-time validation of data recorded by a strong motion station. This validation involves comparing certain current noise parameters of the station with a noise reference level. This reference level is derived from an analysis of the station's historical noise data, as well as the noise data from stations within the same soil categories. [Wielandt, 2012] suggests several methods for determining the self-noise of seismometers and for distinguishing instrumental noise from seismic noise. [Sleeman et al., 2006] employ three seismometers operating in close proximity and propose a technique for measuring their self-noise through coherence analysis. Additionally, [McNamara and Buland, 2004] demonstrate how, in a noise spectral diagram, visual inspection can reveal certain patterns associated with specific malfunctions, such as an excessive number of calibration impulses, errors in the instrumental response [Wielandt, 2012], data loss (small gaps), frequent re-centering of the seismometer mass, and so on.

Due to the large number of stations and the various types of waveforms and associated spectra, visually identifying different event types (such as landslides, volcanic tremors, small earthquakes, explosions, etc.) or potential precursor signals, and in many cases, diagnosing instrumental malfunctions, demands a significant amount of time and effort. This task is often too demanding for the humans, even if they are expert in the field. So, assistance from neural networks can be highly beneficial, particularly when dealing with noise, given its inherently non-deterministic nature [Scales and Snieder, 1998; Wielandt et al., 2002]. For an overview of neural networks and their current state of the art, refer to Section 4.

In the literature, numerous papers have explored the application of artificial intelligence to the analysis of seismic data [Pignatelli et al., 2021], with some of them focusing specifically on the utilization of neural networks for noise management. In most instances, such studies pertain to related domains, such as denoising [Bekara and Day, 2019] or the identification of particular types of noise or poor data quality in seismic exploration [Mejri and Bekara, 2020; Thorp et al., 2020]. For example, [Thorp et al., 2020] demonstrates how deep learning can classify a seismic stack image in terms of its quality level, considering specific noise types or geological features visibility. While we were in the process of collecting data and fine-tuning an appropriate neural network, [Nugroho et al., 2022] published an article with similar objectives to our current work. They demonstrate how "human knowledge" can be transferred to artificial intelligence, particularly to a convolutional neural network, to automate the verification of seismic station efficiency. Specifically, they explain how seismometer malfunctions (or errors in metadata, as we discuss in sections 2 and 3.2) can be reliably identified by an "expert eye" when examining noise spectral diagrams. This expertise can be utilized to train convolutional neural networks to autonomously perform such detection.

Compared to [Nugroho et al., 2022], our work introduces an enhanced approach aimed at improving the efficiency and accuracy of the automated system. We outline some key upgrades and differences between the two studies, elaborated further in points 1-10 of Section 7: a) during the initial phase, more specific criteria for diagram classification are applied, as described in Section 3.3. b) in the learning phase, a broader range of diagrams is utilized, as detailed in Sections 5 and 6.6; c) this work shows that it is also feasible to prioritize feature optimization using a neural network, which encompasses hyperparameters (as explained in Section 4); d) to validate

the results, various validation methods are employed, as discussed in Section 6 and further elaborated upon in the discussion (Section 7); e) Accuracy improves (Section 6 and 7).

Establishing the groundwork for automating the data quality control system will facilitate the efficient monitoring of seismic station signals, making it a valuable asset for enhancing the data quality of a seismic network.

2. Seismic noise and its spectral representation

Noise analysis is a crucial component of the seismological data quality process. In this paper, we contextualise this analysis within the domain of earthquake seismology. While there has been much discussion over the years regarding the definitions of signal and noise in this research field [Scales and Snieder, 1998; Wielandt et al., 2002], a seismic signal is typically defined as ground motion recordings primarily caused by earthquakes, but also by events like landslides, explosions, and similar sources. In contrast, seismic noise refers to ground vibrations recorded from various non-seismic sources, which may include environmental or anthropogenic factors beyond earthquakes [Holcomb, 1989], such as wind or the passage of a train. Nevertheless, the noise recorded by seismometers, not limited to ground sources, encompasses various disturbances that are frequently induced by instrumentation noise [Wielandt and Steim, 1986; Bormann and Wielandt, 2013], as well as the surrounding microclimate, which can directly affect seismometers, even without generating ground vibrations [Wielandt et al., 2002]. Henceforth, we will use the term "noise" in this broader and more general context, reserving the term "seismic noise" to specifically refer to ground noise.

In the waveforms recorded by a seismic station, the seismological signal of interest spans multiple orders of frequencies, depending on the physical phenomena under investigation or the specific results we aim to extract [Broucke et al., 1972; Wielandt and Steim, 1986; Morelli et al., 2000; Allen, 1982; Vassallo et al., 2012; Ma et al., 2005]. Therefore, understanding the phenomena that generate seismic noise and noise, spanning the entire frequency range, is of paramount importance.

The amplitude of seismic noise varies across different frequencies, as illustrated in the noise spectrum depicted in Fig. 1. The ordinate represents the Power Spectral Density (PSD) [Peterson, 1993; Bendat and Piersol, 2011], measured in absolute decibels (dB) and normalized with respect to the PSD calculated for an acceleration of 1 m/s^2. Specifically, 0 dB corresponds to 1 $(m^2/s^4)/Hz$ [Rastin et al., 2012; Bormann and Wielandt, 2013; Jha et al., 2023]. The figure illustrates the levels of low and high seismic noise (black curves), as per the empirical models for ground acceleration [Peterson, 1993].

The variation in seismic noise sources is not limited to differences in amplitude but also extends to different frequency ranges [Holcomb, 1989]. As depicted in Fig. 1, three primary regions can be discerned. In frequencies higher than 1 Hz, seismic noise primarily arises from anthropogenic activities, such as trains, cars, industrial processes, engines, and turbines [Stutzmann et al., 2000], although contributions from rain and wind are also possible. Within the intermediate frequency range, spanning approximately from 0.04 to 1 Hz, seismic recordings capture sea activity, characterized by two distinct peaks [Darbyshire and Okeke, 1969; Huang et al., 2022]. According to the Low Noise Model (LNM), the higher of these peaks occurs at around 0.2 Hz, termed the secondary frequency peak. For frequencies lower than 0.04 Hz, atmospheric pressure fluctuations become increasingly significant as the frequency decreases [Sorrells, 1971; Alejandro et al., 2020]. Pressure changes, as per Sorrells [Sorrells, 1971], result in ground tilts, exerting a more pronounced impact on the horizontal components of seismometers [Wielandt and Forbriger, 1999; Wielandt et al., 2002] (as illustrated by the green and blue curves in Fig. 1). Furthermore, within this frequency range, temperature fluctuations have a direct impact on the seismometer [Wielandt et al., 2002; Doody et al., 2017]. This serves as an example of "non-seismic" noise, where the effect is more pronounced at lower frequencies.

The INGV receives and stores signals from seismic stations. The seismic waveforms are accessible online on the EIDA database (EIDA Italia, https://eida.ingv.it/it/). Beyond waveforms, the database also encompasses the "metadata" (parameters associated with each station, including transfer function, sampling rate and so on).

The Figure 1 also displays the noise level in terms of Power Spectral Density (PSD) recorded by the AIO seismic station located in Antillo, Sicily. Nonetheless, to gain insight into the statistical distribution of noise over a specific time interval and for a more comprehensive representation of seismic noise, the SQLX (Seismic data Quick Look eXtended, https://sqlx.science/) software package is a valuable tool [Marzorati and Lauciani, 2015; McNamara and Boaz, 2006; McNamara and Buland, 2004]. Essentially, for each station, SQLX takes numerous waveforms (in

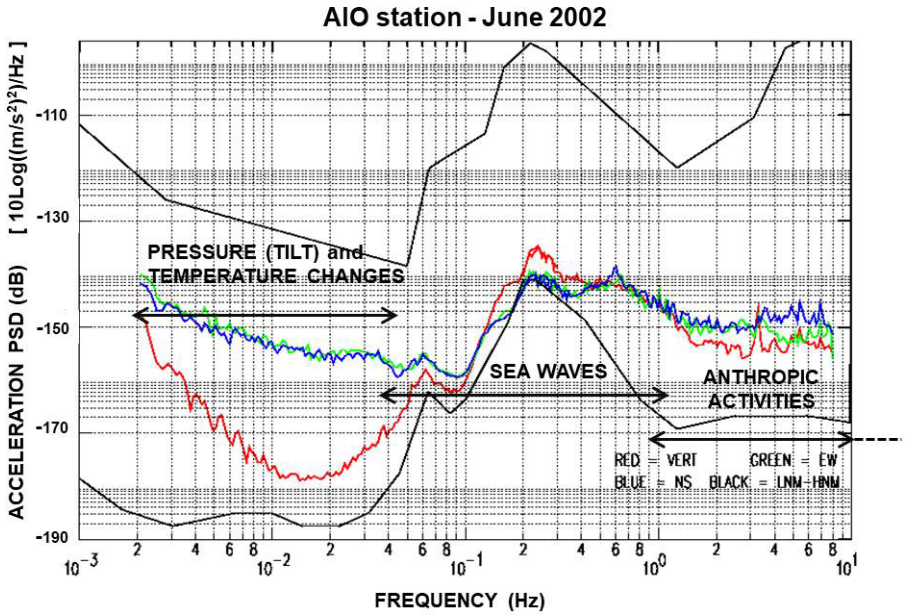

Figure 1. Noise Spectra recorded in Antillo (ME, Italy) station (AIO code). Where thermal and electronic effects are negligible (in this case for f > 0.04 Hz approximately), the coloured curves represent Vertical (Red), N-S (Blue) and E-W (Green) components of ground acceleration level spectrum at various frequencies. In black the empirical noise models of Peterson.

the time domain) as well as the metadata and then generates a set of control diagrams, including spectrograms, Power Spectral Densities (PSDs), noise amplitude behaviour at specific frequencies, trace statistics, and more. In addition to the averaged PSD, similar to the diagrams presented in Fig. 1, SQLX also generates Probability Density Function (PDF) diagrams [McNamara and Buland, 2004]. An example of a PDF is depicted in Fig. 2. While the complete method for calculating the PDF is outlined in [McNamara and Buland, 2004], we will focus on highlighting specific aspects here:

a) The PDF is generated by processing multiple PSDs, and the statistical distribution of noise is visually represented through a colour scale. Each individual PSD corresponds to the spectrum derived from a waveform, such as a 60-minute seismogram.
b) In the PDF calculation process, earthquakes and other transient phenomena, including occasional disturbances and brief anomalies in data flow, are not excluded. However, due to their typically short duration (except for significant and sustained seismic sequences), they manifest in the PDF with colours indicative of a low probability level. Conversely, the presence of quasi-stationary seismic noise or persistent instabilities, such as permanent instrumentation failures, errors in the instrumental transfer function, and similar factors, leads to a high frequency of occurrence [McNamara and Boaz, 2006; Jha et al., 2023]. This heightened occurrence is reflected by a corresponding probability-associated colour within the PDF plot. As a result, as exemplified in Fig. 2, regions displaying colours with higher probabilities of occurrence (red, yellow, green, cyan, etc.) correspond to noise levels that are "closer" to the average. Conversely, segments marked by less probable colours, such as magenta, primarily represent fluctuations.

As an illustration of high-quality data, Fig. 2 showcases the PDF distribution for the East-West (E-W) component at the MedNet TUE station located in Stuetta (Italy), within the central Alps region. The data span from January 1st to February 8th, 2021. To create this diagram, a total of 1822 PSDs were analyzed. The transfer function applied is linked to the specific instrumentation, which, in this instance, comprises a Very Broadband (VBB) station. A seismic station qualifies as VBB when it incorporates a high dynamic digitizer and a seismometer with high and constant sensitivity across a broad frequency range [Wielandt et al., 2002]. A Broadband (BB) station is similar to a VBB station but has a slightly narrower bandwidth [Wielandt et al., 2002]. The instrumentation at the TUE station, which operates at a sampling rate of 100 samples per second (sps), meets the VBB criteria [Kinemetrics Inc., 2005; Mazza et al., 2008; Anon, 2020; Pondrelli et al., 2020]. The Figure 2 also includes the Peterson Low Noise Model (LNM) and High Noise

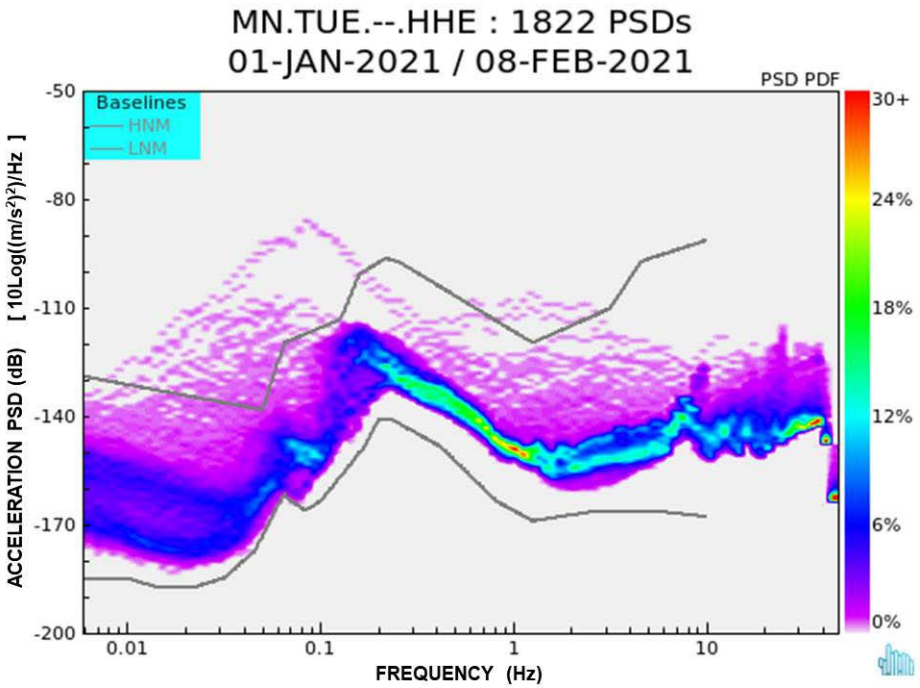

Figure 2. Seismic noise at TUE station (Central Alps) for the horizontal E-W components. At the top right is indicated the number of single trace spectra that formed this distribution plot (1822 in this case). The coloured zone represents the Probability Density Function of the Power Spectral Density. The Peterson curves LNM and HNM are reported in black. The label "30+" refers to a probability greater than 30%. TUE station has a STS2 seismometer [Anon, 2020] and a Q330HR [Kinemetrics Inc., 2005] digitiser/acquisitor. Instrument response is flat in velocity from about 0.01 to about 50 Hz [Pondrelli et al., 2020].

Model (HNM) reference curves in black. Notably, we can observe that the noise behaviour at the TUE site is excellent across the entire frequency range, even during the winter period when environmental seismic noise levels tend to be higher than in the summer [Custódio et al., 2014]. In fact, the PDF diagram demonstrates that the PSD's highest occurrence aligns closely with the LNM curve. No malfunctions in the instrumentation have been observed. However, two minor effects are noticeable, both attributed to the nearby dam for electricity production: A) a slight peak around 8 Hz, corresponding to the frequency of the turbine (8.33 Hz = 500 rpm); B) a low-frequency bimodal trend, observable for frequencies below 0.03 Hz, likely attributed to ground tilting [Wielandt et al., 2002]. In this scenario, the tilt can occasionally be induced by changes in water pressure at the dam's base, particularly when altering water levels are necessary for energy production. In the following Section 3, we will provide examples outlining specific criteria for identifying malfunctions through PDF diagrams and, where possible, suggest potential causes of issues related to the seismic station or metadata.

3. Criteria to detect problems in seismic data using noise spectra

In the preceding paragraph, we presented the noise spectrum of one of the premier Italian stations (Fig. 2). Generally, any station is expected to exhibit a higher noise level than the TUE station. However, as a general criterion, the spectrum shape for a properly functioning station must somewhat follow the trend depicted in Fig. 2. More specifically, the secondary frequency peak at 0.2 Hz, attributed to marine activity, should be evident when analysing a sufficiently extended time interval. Additionally, an excessively high or low noise level can indicate issues, such as problems with instrumental sensitivity.

A noise PDF spectrum can reveal two broad categories of issues:

A) Problems with the station's instrumentation (including gaps in data transmission).
B) Errors in the metadata stored in the database (for example, the sensitivity value; it is important to note that these errors invalidate many subsequent data analyses).

In the following two paragraphs, we present a limited set of examples illustrating how such problems can be identified through the analysis of a noise spectrum, specifically a PDF diagram. In paragraph 3.3, we offer a concise overview of the primary criteria for evaluating good or bad data based on the noise PDF spectrum.

3.1 Problems at the seismic station

An example of a malfunctioning station is illustrated in Fig. 3, depicting the PDF diagram at the RMP station (Monte Porzio, near Rome). The instrumentation is highly similar to the TUE station (see par. 2), employing a 120-second VBB Trillium seismometer (http://support.nanometrics.ca/) and a GAIA digitizer, boasting nearly 24 bits of resolution (https://www.mouser.com/datasheet/2/76/cs5371-72_f3-1160187.pdf). The minor distinctions in instrumentation are not discernible in the spectral representation. In Fig. 3a, the lower boundary of the coloured area consistently traces a gently sloping straight line, with the exception of a slightly pronounced peak around 0.2 Hz and variations in behaviour at the extreme frequencies. Furthermore, from 0.02 Hz to at least 3 Hz, this lower edge essentially represents the mode, i.e., the most statistically populated trend, as indicated by the colours (red, yellow, green, ...) associated with the highest probability of occurrence. Generally, the curve's shape appears significantly different from that of a typical VBB station, such as TUE (Fig. 2). The diagram's configuration suggests issues with the seismometer, likely related to its positioning. The subsequent on-site intervention revealed that the malfunction was attributed to a slight inclination of the seismometer ("out of bubble"). After repositioning the seismometer, the diagrams mirrored those of a typical VBB station (Fig. 3b) with its characteristic shape. Indeed, in Fig. 3b (depicting the PDF of the same station after the intervention), the lower section of the coloured area no longer exhibits a straight edge, and at medium and low frequencies, the trend features both a maximum and a minimum. Especially in the medium-frequency range (0.07-0.4 Hz), it aligns with the LNM curve and effectively resolves the secondary peak. It's worth noting that RMP exhibits a considerably noisy (though still acceptable) pattern at high frequencies, but the diagram's shape and colour shades suggest site noise rather than a significant instrumentation issue.

The Figure 4 (a and b) presents two additional examples of spectra that reveal issues at the seismic station. In Fig. 4a, the PDF of the TERO Broadband station (Teramo, Italy, in Central Apennines) displays a bimodal trend at medium and low-medium frequencies, with the most frequent curve falling significantly below the LNM curve. This indicates that the BB seismometer exhibits non-constant behaviour, suggesting frequent instances of being either not powered or switching from BB to "Short Period" mode [Custódio et al., 2014]. For practical purposes, the two modes are essentially equivalent. In either case, any subsequent analysis utilizing Long Period amplitudes becomes invalidated [Custódio et al., 2014]. The Fig. 4b illustrates the noise at the ARRO station (Arrone TR, Central

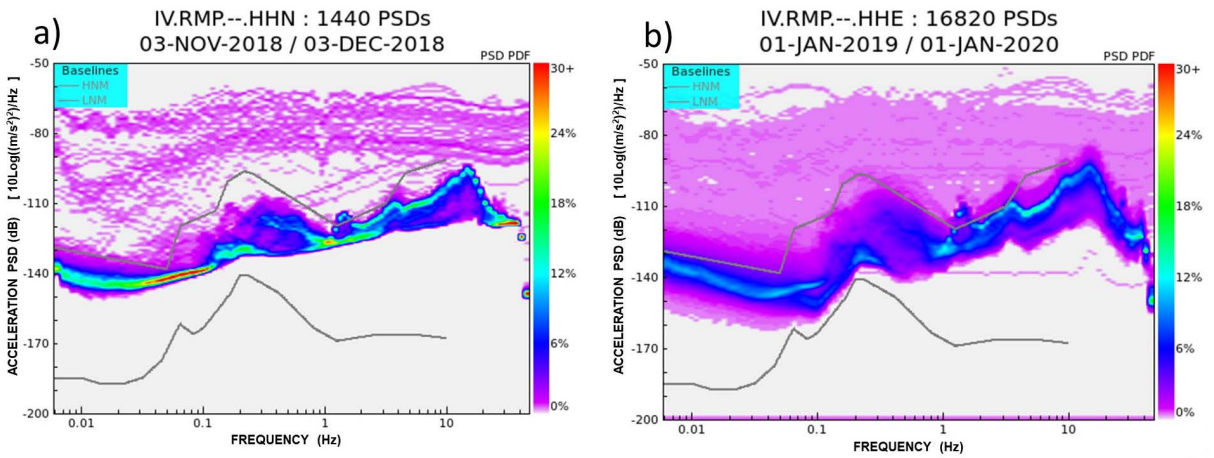

Figure 3. PDF noise level at the RMP seismic station (Monte Porzio, Rome) in the period 3 Nov - 3 Dec 2018 (a, before the intervention) and 1 Jan - 18 Apr 2020 (b, after the intervention). On the left, the lower part of the coloured area (statistically the most frequented) follows a sloped straight line almost everywhere, except for a very slight peak around 0.2 Hz and at extreme frequencies. This abnormal behaviour indicates problems in the seismometer or in its positioning.

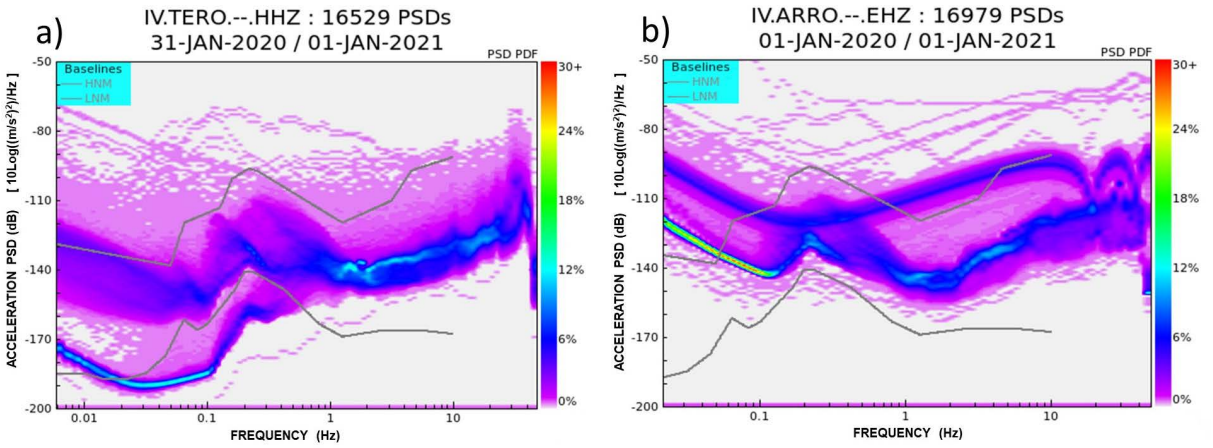

Figure 4. a) Noise spectrum at TERO station (Teramo, Central Italy). The station consists of a BB seismometer, Trillium 40 s (http://support.nanometrics.ca/) and a GAIA digitizer/acquisition system [Michelini et al., 2016; https://www.mouser.com/datasheet/2/76/cs5371-72_f3-1160187.pdf]. At medium and low-medium frequency the noise is much lower (about 20 dB) than the LNM curve. b) Noise at the ARRO station (Arrone, Terni, Central Italy). The station consists of a Short Period seismometer (https://www.lennartz-electronic.de/wp-content/uploads/2021/04/Lennartz-SeismometerManual.pdf) and a GAIA. The bimodal trend on the vertical component denotes a malfunction (sometimes) of the seismometer.

Apennines), where INGV has deployed a short-period seismometer of the Lennartz type, specifically the Le5s model, with a corner frequency [Wielandt et al., 2002] set at 0.2 Hz (5 seconds period, https://www.lennartz-electronic.de/wp-content/uploads/2021/04/Lennartz-SeismometerManual.pdf).

The figure depicts the noise on the vertical component, identified by the code EHZ in the diagram title (codes for seismic data channels are detailed in [Scott Halbert, 1993]). A noticeable split in the blue curve into a bimodal trend is observed. This pattern indicates an intermittent issue in the vertical component of the seismometer, potentially attributed to a railed mass. Indeed, the correct trend would resemble the one depicted in the lower blue curve, representing the typical pattern observed in "short period" stations. This type of anomaly on the vertical component was also observed in other stations equipped with the same seismometer (Le5s), including ATCC (Casa Castalda, [PG], Central Apennines), FOSV (Fossato di Vico [PG]), and so on.

An example of a highly scattered spectrum is apparent in the left side of Fig. 5, illustrating the PDF of station Sersale (South of Italy). Notably, the absence of blue or light blue colours and the dominance of the pink zone across the entire area indicate numerous gaps, likely attributable to transmission or datalogger problems.

3.2 Errors in the metadata (Instrumental Transfer Function)

A deconvolution problem is shown in Fig. 5 (right, station Leskovik, Albania). We note an anomalous relative maximum in the low frequency band, just in correspondence with the corner frequency of the seismometer. Probably the Instrumental Response in the meta data is not correct and so data are not well deconvolved.

A sensitivity error is depicted in Fig. 6a, illustrating the noise trend at GIGS (a VBB station in Gran Sasso, Central Italy) over a time interval spanning more than 5 months. The spectrum's shape is accurate since, for frequencies higher than 0.07 Hz, it adheres to the LNM pattern. However, its value is below the LNM, implying that the correct spectrum would be identical but shifted to higher PSD values. An error in the storage of instrumental sensitivity in the Metadata archive is evident, which was subsequently corrected.

The final example pertains to a short period station: in Fig. 6b, the noise at the BARO station (Barbarano, VT, Central Italy) is displayed. In this case, a 5-second seismometer (https://lunitek.it/seismic/seismic-sensors/tellus-5s/), akin to ARRO's Le5s, is installed. The spectrum's shape significantly deviates from the typical pattern for this 5 s short period seismometer, as typical behaviour is shown in the lower blue curve of Fig. 4b. The spectrum of BARO resembles that of a seismometer with a corner frequency of 1 Hz (1 second). Indeed, an error was identified in

the transcription of the polynomials describing the instrumental transfer function (or Instrumental Response). Following the correction, the diagram exhibited the accurate trend.

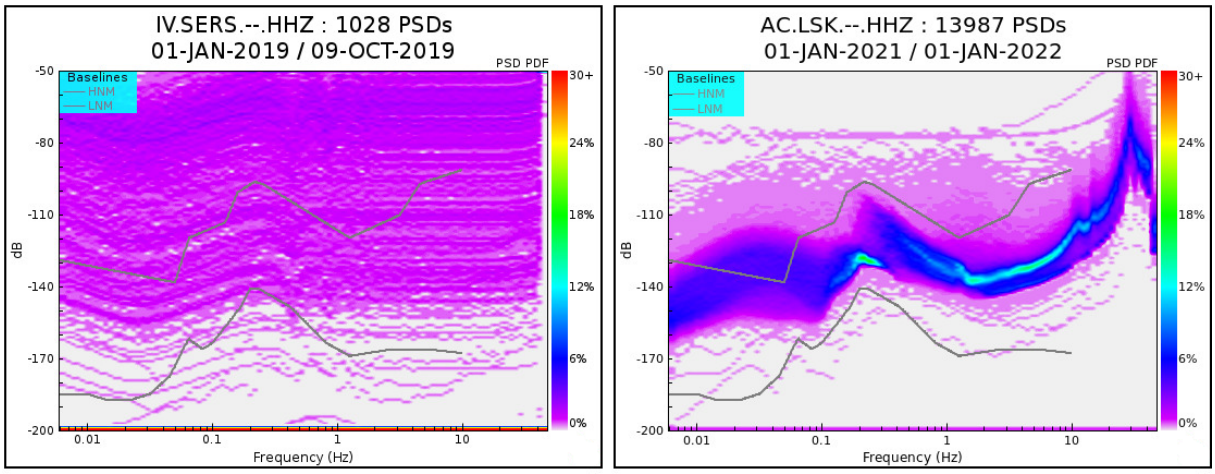

Figure 5. Left: highly scattered spectrum at Sersale (South of Italy): the blue or light blue colours are not present. Right: relative maximum at low frequency at Leskovik (Albania), in correspondence with the corner frequency of the seismometer (0.05 Hz; probably not correct Instrumental Response).

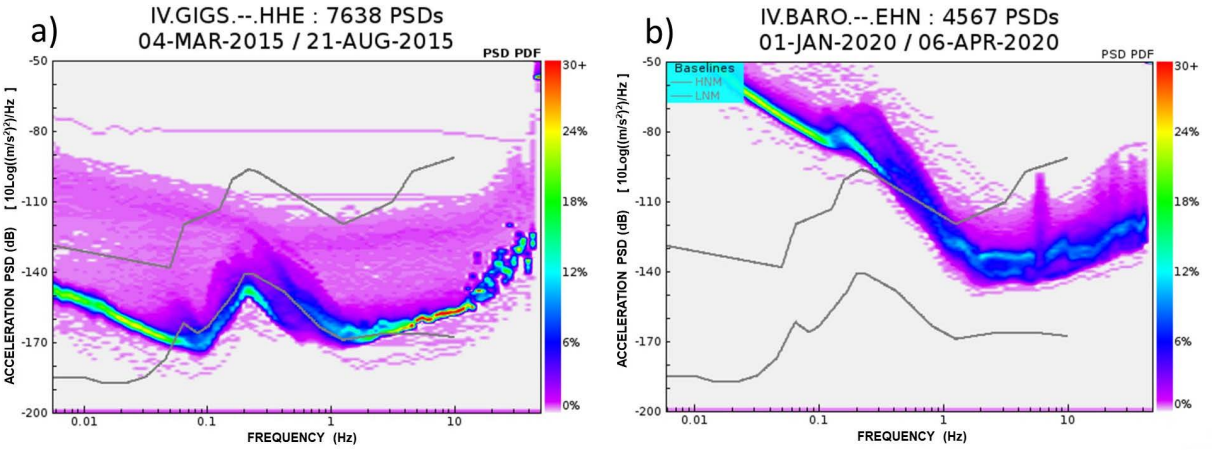

Figure 6. a) Noise at the GIGS station (Gran Sasso, central Italy) for more than 5 months' time interval. The VBB station consists of a Trillium 240 s seismometer (http://support.nanometrics.ca/) and a GAIA digitizer/acquisition system [Michelini et al., 2016; https://www.mouser.com/datasheet/2/76/cs5371-72_f3-1160187.pdf]. The PDF shows a noise lower than LNM on about half of the analysed band. b) Noise at the BARO station (Barbarano, VT, Central Italy) with a Short Period *Tellus 5s* seismometer (https://lunitek.it/seismic/seismic-sensors/tellus-5s/) and a GAIA acquisition system. The shape of the spectrum is very different from the typical trend of a 5 s seismometer which instead follows the lower blue curve of Fig. 4b.

3.3 Main criteria for evaluating noise spectra: consequent considerations

Referring to the preceding paragraphs, we present in Table 1 the main criteria for assessing the PSD-PDF diagrams. In numerous instances, employing these primary criteria allows a station to be categorized as either "OK" (functioning correctly with accurate metadata) or "BAD." Nevertheless, there are also more complex cases where the criteria are challenging to encapsulate succinctly within a table cell. Some issues and problems are elaborated upon both immediately below and in the latter part of Section 6.5 (Mid-way discussion) and Section 7. In the

preceding paragraphs, we demonstrated that visual analysis of PDFs has the potential to identify malfunctions in seismic stations and errors in metadata. However, the shape of the noise spectra is influenced by various variables [Bormann and Wielandt, 2013], encompassing environmental factors [Custódio et al., 2014], anthropogenic sources [Stutzmann et al., 2000], instrumental conditions [Bormann and Wielandt, 2013], and the quality of station installation [Anglade et al., 2015]. Moreover, in certain instances, it becomes important to consider the site characteristics and the specific seismic instrumentation employed, including details about the instrument's manufacturer. Consequently, the spectrum of PDF diagrams is considerably broader than the instances illustrated here. Other examples exist, enabling an expert user to identify malfunctions and, at times, comprehend their underlying causes as well [McNamara and Buland, 2004; McNamara and Boaz, 2006].

Criterion	Assessment
The diagram follows the shape of LNM (in particular the peak at 0.2 Hz is clear) and is included between LNM and HNM (e.g. Fig. 2, Fig. 3b)	OK
The diagram exceeds HNM only at high frequencies and is between LNM and HNM in the rest of the band. Almost surely high noise depends on the site, not on the instrumentation or metadata	OK
The diagram of the most probable colors is found under LNM (even only in a narrow band of frequencies, [Fig. 6a] and even only for a period of the year [Fig. 4a])	BAD/BROKEN
The lower/main part of the diagram is too straight in which the marine secondary peak (0.2 Hz) can hardly be distinguished (see Fig. 3a)	BAD/BROKEN
Clearly a bimodal trend in which one of the two modes has an anomalous trend (see Fig. 4b)	BAD/BROKEN
Shape of the spectrum is very different from the typical one for the in-site seismometer. In some cases, it looks more like that of a seismometer with different characteristics (e.g. higher corner frequency, Fig. 6b)	BAD/BROKEN
The marine primary frequency peak (0.07 Hz) is clear but throughout the year its values are always 15 dB or more above LNM: the Response Function or the sensitivity is wrong (storms cannot last a year)	BAD/BROKEN
Plot is above HNM all over the frequency band (probably incorrect sensitivity in the seismometer or in the metadata)	BAD/BROKEN
Highly scattered data (Fig. 5, on left)	BAD/BROKEN
Relative maximum in low freq, in correspondence with the corner frequency of the seismometer (wrong metadata) see Fig. 5 on right	BAD/BROKEN

Table 1. The table contains the main (major) evaluation criteria of PSD-PDF diagrams. Based on these (ad other) a station is judged "OK" (correctly functioning and with correct metadata) or not.

The number of stations and the volume of diagrams generated by automatic spectrum generators are too extensive for routine manual inspection by operators. Conversely, translating the criteria of intricate visual recognition into an effective deterministic algorithm is challenging due to the complex nature of noise spectra, as demonstrated. The simplified criterion of "good/bad," relying on the noise average (single PSD trace) staying within a designated minimum and maximum threshold, fails to capture the intricacies inherent in human visual evaluation. For example, the expert evaluates not just the individual average trace (Fig. 1) but rather distribution diagrams encompassing the entire spectrum of noise (not only single frequency bands) over annual or monthly intervals, such as PDFs (Fig. 2).

Given the complexities outlined earlier, the pursuit of a neural network approach is certainly justified. In the subsequent paragraphs, we detail how we employed a pre-trained convolutional neural network to facilitate the "transfer" of human expertise to an automated system.

4. Machine learning

4.1 Introduction and convolutional neural networks

Machine learning techniques encompass supervised, unsupervised, and reinforcement learning methods that enable algorithms to learn from data, predict outcomes, identify patterns, and make decisions without being explicitly programmed. [Mahesh, 2020; Pandey et al., 2019; Lantz, 2013; Taner et al., 2021; Likas et al., 2003].

To implement a supervised technique, the initial step involves gathering data that includes both the inputs and the ground truth, often referred to as *labelled data*. This dataset is then used to train the algorithm, allowing it to autonomously identify patterns (referred to as features) within the inputs that correspond to variations in the ground truth. This phase of machine learning is known as 'training.' It is often considered the most challenging aspect, as it necessitates the collection of labelled data. However, upon successful training, the algorithm can automatically predict the ground truth for any other dataset that includes the provided inputs.

To assess the model's ability to generalise and accurately predict labels for data not part of the training set, a common practice, as described by [Lantz, 2013; Pandey et al., 2019; Mahesh, 2020], involves initially randomly selecting a portion of the available dataset and excluding it from the training process. This excluded portion, referred to as the "*test data*", contains the labels we are interested in predicting. It serves as a robust means to evaluate the model's accuracy, as it includes both the true labels and, when utilised as input for the model, the labels it predicts. The accuracy percentage is calculated as *Accuracy = (number of matched results)/(total test data)*.

In the realm of machine learning, the most effective way to demonstrate method accuracy is by employing a "confusion matrix," as advocated by [Lantz, 2013], [Taner et al., 2021].

Artificial Neural Networks (ANNs), as discussed by Lantz [2013], are a supervised machine learning subset inspired by the primate cerebral cortex, employing multiple layers of neurons to progressively extract abstract features from input data, with each neuron applying specific transfer functions and adjusting weights iteratively to minimize disparities between true and predicted values [O'Shea and Nash, 2015; Rawat and Wang, 2017; Cao and Parry, 2009].Convolutional Neural Networks (CNNs), elaborated by Cao and Parry [2009], excel in extracting local features from matrices through intricate layers of filters, with each neuron focusing on limited inputs, and comprising convolutional, pooling, fully connected, and softmax layers, as detailed by O'Shea and Nash [2015] and Pignatelli et al. [2021]. Assessing generalization capability in machine learning techniques involves understanding the risk of overfitting, where intricate relationships in training data may hinder performance on test data, as discussed by Srivastava et al. [2014]; Ghojogh and Crowley [2019]. Techniques like "dropout" and "k-fold" validation help mitigate overfitting, described by Ghojogh and Crowley [2019], while data augmentation, though beneficial, may not be applicable to PDF images due to potential classification compromise [Shorten and Khoshgoftaar, 2019].

Another challenge to handle when dealing with convolutional neural networks is selecting the hyperparameters.

4.2 Hyerparameters

A detailed description of the difference between the parameters and hyperparameters and a general hyperparameters description can be found in [Zhu and Beroza, 2019] and all the details would be too long to discuss here, so we will just limit to a general description.

The "parameters" in a neural network are meant to be the weights of the neural connections so basically the numbers "learnt" during the training process. The hyperparameters are basically set before the training process and are related to network architecture, optimization strategy, training test splitting and many others.

The most important ones we want to mention here are:

1) Learning rate and momentum: these hyperparameters determine how fast the training process is. More specifically depending on their values, at each step of the process the information "learnt" in the last examined dataset, will be more or less important compared to information learnt before. If the process is too fast, the risk of a not converging solution increases.
2) Training/test data split percentage: before the training process, it has to be decided how much data has to be used for training and how much data has to be used for the test. Moreover, part of the training data has to be used as validation data (as described in the paragraph 4.2) so there is another percentage to be set. More data

is used for tests, more "generalisation capability" the user should expect but, at the same time, the risk of "not learning enough" increases.

3) Batch size: as the training process is very demanding in terms of computation resources, sometimes it's not convenient using all the training data at each iteration, so the neural connection weights are updated after steps using a limited number of training data. In other words, the use of all training data will require multiple steps. The number of training data used at each step is called "batch size".
4) Dropout percentage: the dropout has been described in paragraph 4.1. This parameter determines how many neurons will be randomly turned off at each iteration.
5) Epochs number: depending on the batch size values, multiple iterations are needed to use all data to update the neural network weights. Such an iteration number is called "epoch". So, every time all input data has been used, it is said that a training "epoch" has been concluded. The maximum epochs' number before closing a training process is another hyperparameter that can be set and that can affect the final result. In fact, if more epochs are used, the probability of learning "more" is most likely, however, the probability of overfitting occurring is more likely too.

In summary, hyperparameters play a critical role in shaping the behaviour of a neural network during training, influencing its learning speed, generalization capability, and overall performance. Careful selection and tuning of hyperparameters are essential to achieving the desired results in machine learning tasks.

4.3 Pre-trained convolutional neural networks

Before training a convolutional neural network, determining its architecture, which includes the sequence of layer types, neurons, and their connections, can be a challenging step. However, previous research by [Pignatelli et al., 2021] has demonstrated the advantages of using images, rather than raw data, by employing pre-trained networks.

In the context of our main objective, which is to develop an automated real-time system for monitoring the operational status of seismic stations and potentially alerting users to issues, using pre-trained networks on images instead of raw data is highly beneficial for several reasons:

1) PDF Plots as Data Representation: PDF plots in 2D are a valuable representation of the quality information within seismic data. While seismograms are typically time series data in 1D, experts have traditionally learned to classify noise and station quality by examining such plots. Producing a PSD-PDF diagram (which is a distribution) necessitates substantial effort and computations, as it requires numerous single-trace spectra, and each of these spectra is derived from a seismogram (e.g., a 30-minute seismogram). The quality of a station's operation is often best assessed by examining the ensemble of all spectra within the plot, as some individual spectra might incorrectly indicate issues with the station. Using these single spectra for data labelling would make the process extremely complex and computationally intensive.
2) Complex Data Handling for the operator: Classifying individual seismograms based on a 1D network would require running a spectrum analysis on each seismogram, distinguishing between good and bad data, and then associating seismograms with their respective classifications. This process would need to be repeated for all PSD-PDF distribution diagrams. The complexity and computational demands of such an approach would be impractical for operators, and it would involve handling a significantly larger volume of data.
3) Increased Computational Requirements: Employing a 1D network would necessitate a more complex network architecture with a larger number of input neurons, significantly extending the training time.
4) Real-Time Data Classification: Implementing real-time data classification using a 1D network would be time and resource-intensive.

The INGV network has been designed to generate these images and make them available on their website, allowing the system to efficiently collect images using web scraping techniques.

Using a pre-trained network offers the advantage of having many of the weights initialised with tested values, and the training process primarily fine-tunes these initial parameters. This leads to a substantial reduction in computing time compared to creating a convolutional neural network from scratch.

In essence, while theoretically possible, a 1D classification network would be impractical for real-time automatic warning systems. Pre-trained networks, such as AlexNet [Han et al., 2017; Krizhevsky et al., 2017; Indolia et al., 2018]

and EfficientNet, with their efficient architectures designed for general image classification, offer more practical and time-effective solutions for the task at hand. AlexNet, for instance, has pre-calculated weights for recognizing basic image features, facilitating its application for image-related tasks like station quality assessment. Comparing the results of different pre-trained networks, such as AlexNet and EfficientNet, can help identify the most suitable model for the specific problem at hand. The architecture of Alexnet is shown in Fig. 7. For a detailed understanding of EfficientNet, the reader can refer to [Taner et al., 2021].

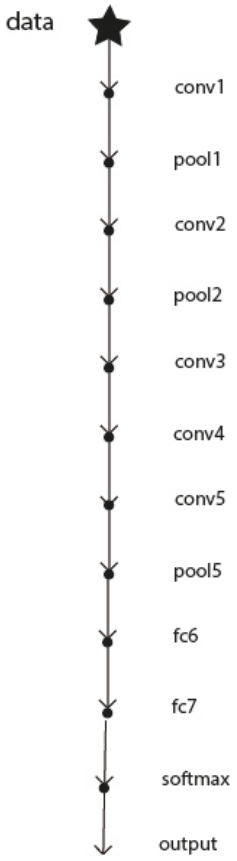

Figure 7. AlexNet classification architecture.

4.4 Evaluation metrics

To assess the performance of the trained network, several evaluation metrics are employed, and these metrics are computed using test data, which were separated from the training data to evaluate the network's ability to generalise its classification. While the accuracy metric has been previously discussed, our experiments also consider the following evaluation metrics:

1) *Precision*: Precision measures the proportion of "positive cases" that the network correctly identifies. In this study, a "positive case" refers to a detected broken station. Therefore, precision is calculated as follows:

$$Precision = \frac{Correctly\ Detected\ Broken\ Stations}{(Correctly\ Detected\ Broken\ Stations + number\ of\ OK\ stations\ classified\ as\ BROKEN)}$$

2) *Recall*: Recall quantifies how many total broken stations are correctly identified by the system. In other words, if the number of false negatives is low, recall increases. It is calculated as follows:

$$Recall = \frac{Correctly\ Detected\ Broken\ Stations}{(Correctly\ Detected\ Broken\ Stations + number\ of\ BROKEN\ stations\ classified\ as\ OK)}$$

3) *F1-Score*: The F1-Score is a metric that combines precision and recall into a single equation. It is particularly useful when both precision and recall are important in a classification problem. The F1-Score is calculated as follows:

$$F1\text{-}Score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)}$$

These metrics provide a comprehensive assessment of the network's performance, considering both the ability to correctly detect broken stations (precision) and the capacity to identify the majority of broken stations (recall). The F1-Score is a balanced measure that takes into account both precision and recall, providing a more complete evaluation of the network's classification capabilities. Further details and explanations of these metrics can be found in [Taner et al., 2021].

4.5 Loss function and K-fold cross validation

When training a neural network, it's crucial to define a loss function, which serves as a mathematical measure of how much the neural network is deviating from correct classifications when processing input data. The choice of the loss function is of paramount importance because during the training process, the neural network's weights are adjusted to minimize this loss. While there are various loss functions available, a commonly used one for neural network classification tasks is "cross-entropy."

In a classification network, the predicted output consists of probabilities assigned to each class. The cross-entropy loss function is expressed by the following equation:

$$Loss = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} -log(pi)qi \tag{1}$$

where N is the number of data, k is the total number of classes for the classification problem, pi is the predicted probability of each class during the training and qi is the real probability of each class (that is one for the real class and zero for the others).

To illustrate this with an example, let's consider a three-class classification network. For a single input, the network produces the following output: [0.1, 0.5, 0.4], where each component represents the probability assigned to each class. If the true class is the third one, then the cross-entropy for this single data point is calculated as follows:

$$Loss = -\log(0.1) * 0 + -\log(0.5) * 0 - \log(0.4) * 1 = 0.92$$

The overall cross-entropy (Loss) is computed by taking the mean of all the individual cross-entropy values across all the data points.

In essence, the cross-entropy loss function quantifies how well the predicted probabilities align with the actual classes, and during training, the neural network aims to minimise this loss to improve its classification performance. For a more comprehensive understanding of various loss functions, you can refer to [Janocha and Czarnecki, 2017].

To address the potential randomness of test data selection and ensure robust evaluation metrics, the K-fold cross-validation technique, outlined mathematically by Ghojogh and Crowley [2019], divides the dataset into K partitions, where each partition serves as the test set in iterations, leading to averaged metrics across K neural networks trained on different datasets and enhancing reliability in performance assessment.

4.6 Class imbalance

During the training of a neural network, a common challenge that may arise is referred to as "class imbalance." This issue occurs when certain classes within the dataset significantly outnumber other classes. As a result, during

training, the neural network may be biased towards favouring the majority classes due to the higher probability of selecting the most numerous classes.

To address the problem of class imbalance, one common approach is to use a technique called "class weighting." Class weighting involves assigning higher weights to the errors associated with the minority classes in the loss function. By doing so, the neural network is encouraged to pay more attention to the minority classes and give them greater consideration during the training process. This helps to balance the influence of different classes and mitigates the bias towards the majority classes, resulting in a more fair and effective training of the neural network. So the Eq. (1) can be corrected with the following [Sokolova and Lapalme, 2009]:

$$Loss = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} -w(j)log(pi)qi \tag{2}$$

where:

$$w(j) = \frac{N}{k * \nu i}$$

where N, pi, qi and k have already been described in Section 4.5 and νi is the absolute frequency of the j-class in the data.

For a more in-depth understanding of the class imbalance issue and potential solutions, you can refer to [Abd Elrahman and Abraham, 2013].

5. Method application: data used and network settings

The primary objective of our study was to assess the feasibility of developing a neural network-based system capable of automatically distinguishing between regular signals and those recorded by malfunctioning seismic stations, considering that flawed metadata can also adversely impact spectra. Pursuant to this objective, as detailed in Section 4.3, employing time series would impose significant demands on human and computational resources, as well as time consumption. For these reasons, as outlined in Section 4.3, we aim to investigate whether automatic discrimination can be effectively achieved using images instead of raw data. As the initial step in our study, we curated a dataset comprising two distinct categories, labelled as 'ok' or 'broken.' This categorization was performed based on our expertise, following the criteria outlined in Section 3.3. Specifically, our dataset consisted of PDF spectral diagrams, similar to those presented in Sections 2 and 3 (Figs. 2 to 6).

We recall that these diagrams typically depict the seismic and/or instrumental noise levels of a seismic station (Section 2, point b).

In this study, we utilised diagrams primarily sourced from data within the Italian Seismic Network (ISN, code IV) [Michelini et al., 2016]. Additionally, we incorporated diagrams obtained partially from the Mediterranean Very Broadband Network (MedNet, code MN) [Pondrelli et al., 2020] and, in part, from selected local Italian networks or foreign networks associated with countries neighbouring Italy. The data from these networks either directly contributes to the seismic monitoring facilities of the Istituto Nazionale di Geofisica e Vulcanologia (INGV) or plays a role in the broader INGV surveillance system [Michelini et al., 2016]. These other networks are labelled as: 1J, 3D, AC, CH, CR, FR, GE, GU, HL, IX, NI, OE, OT, OX, RD, RF, SI, SL, ST, TV, VR, XK, Z3, ZH (see https://terremoti.ingv.it/instruments for more detail). As a preliminary step, we selected broadband (BB) stations with a sampling rate of 100 sps and a flat response to ground velocity of at least 40 seconds (i.e., with a frequency band ranging from no more than 0.025 to at least 25 Hz, as described in Section 2), for the three Cartesian components of ground motion named HHZ (Vertical), HHN (North-South), and HHE (East-West) [Scott Halbert, 1993]. The decision to restrict the analysis to BB and VBB stations aimed to simplify the learning process. Diagrams from these stations, when functioning correctly, exhibit similarities, yet they are notably distinct from those of short-period stations (channels EH*, SH*, ...) especially at low frequencies, as evident in the comparison between Fig. 2 and the lower part of Fig. 4b. The BB and VBB stations under consideration constitute a subset of approximately 600 stations out of around 1000 that contribute data to the INGV monitoring centers and to the INGV database. For each station, the SQLX

package (paragraph 2) generates a diverse array of diagrams. From these, in this first study, we have opted for annual diagrams, as seismic noise (and occasionally instrumental noise) can exhibit significant variations not only between day and night but also on a weekly, monthly, or seasonal basis. Utilizing the annual distribution makes it easier to categorize the diagrams into "OK" and "Broken" classes, as the annual diagrams within the "OK" class are more uniform and similar. In summary, the chosen data set for training possesses the following characteristics:

a) Broadband (BB) or Very Broadband (VBB) stations (HH* channels) from the Italian (IV) or Mediterranean (MN) network, as well as local or foreign networks (refer to the list of network codes mentioned above)
b) PDF spectra (similar to Fig. 2-5)
c) Annual diagrams
d) Analysis conducted on all three components of motion (a diagram for each component).

5.1 Experiments description, data used and neural network setting

We conducted four experimental tests to assess various aspects of the training and prediction capabilities of convolutional neural networks. The first experiment aimed to evaluate the method's performance when trained exclusively on data from a specific year (2019). In the second experiment, we investigated the system's ability to generalize when exposed to data from multiple years, necessitating the inclusion of diverse year-specific data in both the training and test sets. The third experiment aimed to assess the extent to which the pre-trained network could generalize its accuracy when classifying images from a year not included in the initial training. Specifically, in the second experiment, we deliberately excluded data from 2018 during training, and in the third experiment we exclusively used data from the year 2018 as a test set. For the fourth test, considering that the network accuracy was high but not flawless (a few questionable diagrams were excluded a priori, as detailed in paragraphs 6.4 and 6.5), we devised the fourth test to enhance the performance of the automatic system by introducing a new class. The primary concern with the automatic system lies in the possibility that a small percentage of faulty data might be categorized as 'good' signals. This implies that there might be undetected broken stations or instances of bad metadata. To minimize this risk, we introduced a third class labelled 'doubt'.

Prior to the description, we provide a summary of the experiments conducted and the data used for training and testing in Table 2. As stated in paragraph 4.3, even if finding the best model was not in the scope of this work, for all the tests, we used two neural networks architectures: AlexNet and EfficientNet. Moreover, we trained multiple networks using different values of the most important hyperparameters. More specifically we used all the possible combinations of the following values for both networks:

Learning rate = [1.0000e-03, 1.0000e-04, 1.0000e-05];
Batch size = [15, 25];
Validation data fraction = [0.1, 0.2];

Experiment Number	Total used data (training + test)	Year of Training data	Number and percentage of training diagrams	Year of Test data	Number and percentage of TEST diagrams
1	280	2019	196 (280 – 30%) or 224 (280 – 20%)	2019	56 (20%) or 84 (30%)
2	450	2019 + 2015 + 2016 + 2020	315 (450 – 30%) or 360 (450 – 20%)	2019 + 2015 + 2016 + 2020	90 (20%) or 135 (30%)
3	1200	2019 + 2015 + 2016 + 2020 (NO 2018)	360 (best result of 2^ experiment)	2018	840
4	1865	2021	1305 (1865 – 30%) or 1492 (1865 – 20%)	2021	373 (20%) or 560 (30%)

Table 2. Summary of experiments and data used for training and testing.

LR	BS	VF	TF	E	NT	accm	precm	recm	f1m	accsd	precsd	recsd	f1sd
10^{-5}	10	0.05	0.1	30	AlexNet	0.92	0.89	0.99	0.94	0.05	0.07	0.04	0.04
10^{-4}	10	0.05	0.1	30	AlexNet	0.95	0.93	0.98	0.95	0.04	0.05	0.04	0.03
0.001	10	0.05	0.1	30	AlexNet	0.75	0.84	0.85	0.82	0.20	0.18	0.19	0.13
10^{-5}	15	0.05	0.1	30	AlexNet	0.92	0.89	0.97	0.93	0.04	0.07	0.04	0.04
10^{-4}	15	0.05	0.1	30	Efficient Net	0.95	0.94	0.97	0.96	0.04	0.04	0.05	0.03

Table 3. Example of results table using all possible hyperparameters combination and the two networks AlexNet and EfficientNet. The results for all the tests have been saved into excel files and can be found in the https://www.kaggle.com/datasets/alessandropignatelli/seismicnoiseexperimentresults?rvi=1. The performance metric values have been expressed ad mean and standard deviation of the K fold procedure. The meanings of the column headers are: LR-Learning rate; BS-Batch Size; VF: Valid Fractio; TS-Test Fractio; E-Epochs; NT-Network Type; accm- accuracy mean; precm- precision mean; recm-recall mean; f1m- F1 Score mean; accsd-accuracy standard deviation; precsd-precision standard deviation; recsd-recall standard deviation; f1sd- F1 Score standard deviation.

Test data fraction = [0.2, 0.3];
Number of epochs = [30, 50].

Additionally, the k-fold technique has been employed to minimize the impact of random test data selection when computing the results. The k-fold number has been set to 5 when test data fraction was 0.2 and to 3 when test data fraction was 0.3. Considering all possible combinations for the two networks, we obtained 96 results for all four tests. Since the results are numerous, we compiled them into Excel files following the same format as Table 3.

For all the tests, we employed the two pre-trained networks without modifying the main network features. So, the existing network includes two dropout layers with a 50% dropout percentage, and the loss function utilized is the cross-entropy, as described in paragraph 4.5. The only modification we introduced is the imbalance weighting, as outlined in paragraph 4.6.

The code to perform the described analysis has been written in matlab. All the training steps, using different data and different parameters has been run on different machines (sometimes using also matlab parallel toolbox) and each training experiment required about ten minutes to run.

6. General results

We have organized the results into the repository https://www.kaggle.com/datasets/alessandropignatelli/seismicnoiseexperimentresults?rvi=1.

A README.docx file is contained in such repository with a detailed description of the material.

6.1 General results summary

There are some preliminary observations to discuss before delving into specific test results. Upon examining the Excel files generated during the analysis, we observed the following points:

1) AlexNet performs better than EfficientNet (at least using the MATLAB implementation of EfficientNet). This indicates that, even though it's an image recognition problem, the success of the system is not inherently obvious. It depends on the information within the data and the network architecture. However, the metrics for AlexNet, including accuracy, precision, recall, and F1 score, appear satisfactory for building the intended automatic system, which is the main objective of this study.
2) The accuracy and all other metrics are consistently high, often exceeding 86%.

3) Once the network architecture has been set, changing the hyperparameters does not significantly impact the results. Calculating the mean and standard deviation of the evaluation metrics across all hyperparameters, we observe that the standard deviations are much smaller than the means.
4) We did not encounter a significant overfitting problem, especially when using AlexNet (slightly more pronounced when using EfficientNet). As mentioned at the end of Section 6.1, we saved the training process figures. In Fig. 8, an example of such figures illustrates the accuracy function of training data in blue and validation data in black. As one can see, there is no evident split between these two lines. This is the most common situation in the training process. Sometimes, especially when using EfficientNet, there is just a slight separation at the end of the process when the accuracy is already satisfactory. Therefore, even in these cases, this problem does not significantly impact the quality of the results.
5) The standard deviation of the evaluation metrics calculated through the k-fold iterations is generally much lower than their mean. This indicates that the k-fold procedure does not significantly impact the training process, and thus the final accuracy is not substantially dependent on the specific random choice of the test data partition.
6) Accuracy and f1 score exhibit significant similarity. Furthermore, all evaluation metrics are consistently high. This indicates that there is no substantial imbalance between classes in the data, and/or the weighting procedure described in Section 4.6 is functioning correctly.

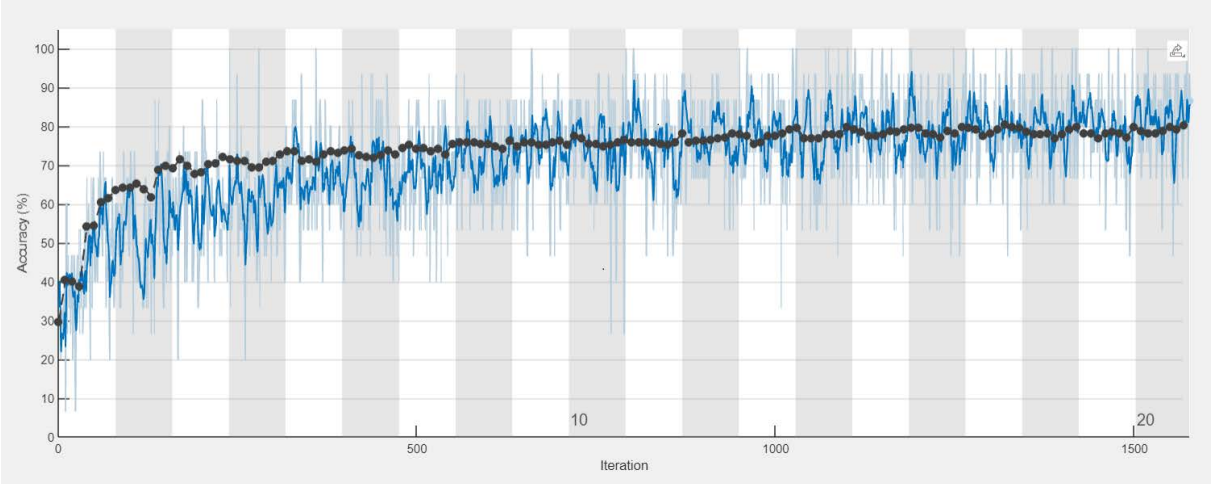

Figure 8. Training process plot produced by matlab. The lines show the accuracy value through the iterations during the process. The blue line represents the value got using training data while the black one represents the one comparing the predicted classification to the validation data. This is the most common behaviour of all the plots and there is not a significant overfitting problem.

6.2 Training and first expeditious experiment: 2019 data

This was a preliminary test conducted to assess whether a neural network could be trained using data from a single year to automatically classify PDFs from the same year. We collected 280 PDF spectral diagrams from the year 2019, comprising 140 labelled as 'OK' and 140 as 'Broken.' The 'Broken' category includes trends that, according to visual inspection and the criteria outlined in Section 3 and Table 1, clearly indicate defective stations. After collecting this data, we utilized it to train the network based on the types, hyperparameters, and strategies detailed in Section 5.1. The complete results of the test are contained in the excel file Experiment1ResultsTable.xlsx in the repository at https://www.kaggle.com/datasets/alessandropignatelli/seismicnoiseexperimentresults?rvi=1 and the results are shown according to the format shown in Table 3.

The mean and standard deviation of the evaluation metrics across all hyperparameters are presented in Table 4. These results affirm the general observations.

Metric	Mean AlexNet	Mean EfficientNet	Max AlexNet	Max EfficientNet	Min AlexNet	Min EfficientNet
AccuracyMean	0,90	0,75	0,95	0,90	0,78	0,54
PrecisionMean	0,88	0,75	0,93	0,89	0,84	0,59
recallMean	0,95	0,89	0,99	0,94	0,79	0,82
f1scoreMean	0,91	0,81	0,95	0,91	0,83	0,69
accuracyStd	0,04	0,04	0,14	0,09	0,02	0,02
PrecisionStd	0,04	0,03	0,07	0,06	0,00	0,00
recallStd	0,05	0,05	0,14	0,08	0,02	0,02
f1scoreStd	0,03	0,03	0,09	0,06	0,01	0,01

Table 4. First experiment results averaged trough hyperparameters values.

BestNetType	Best LearningRate	Best miniBatchSize	Best ValidFractio	Best TestFractio	Best MaxEpochs
AlexNet	0,0001	15	0,2	0,2	30

Table 5. First experiment. Hyperparameters corresponding to the best F1 score.

Additionally, Table 5 displays the hyperparameters associated with the maximum F1 score. We chose the hyperparameters that yielded the highest F1 score as the optimal outcome. Subsequently, we generated the confusion matrix, averaged as a percentage across the k-fold iterations, and presented it in Fig. 9 (Accuracy = 95%).

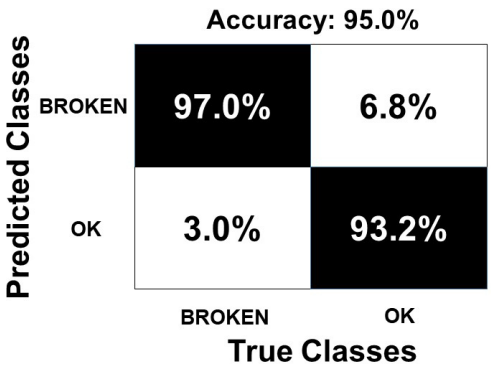

Figure 9. First experiment confusion matrix and total accuracy of trained AlexNet neural network applied to the first set of test data. The results are expressed in percentage average through the k-fold iterations for the case of hyperparameters set showing the best f1 score.

6.3 Second experiment: generalisation to different years by increasing the data

This second experiment was designed to check if, during different years, there are noise trends adding too much variety to build up the automatic system. Therefore, additional labelled data have been provided to improve the learning phase and to test if the method also works for data from different years. More specifically, 61 additional "OK" and 106 additional "Broken" diagrams belonging to years other than 2019 were provided (see Table 2), including their labels. So, the total number of diagrams used for the learning phase was 447 (201 "OK" and 246 "Broken"). The reason why we preferred to increase mostly the number of "Broken" diagrams is that they present greater variety as there are many possible causes of malfunction in a seismic station, while broadband "OK" diagrams are more homogeneous. Additionally, in this experiment, we employed the "k-fold" method. The results corresponding to the best f1 score are summarized in the confusion matrix in Fig. 10. As observed, the augmentation in the variety and quantity of data (introducing different years) resulted in a marginal decrease in test accuracy (93.7%), but the accuracy remains very high. The general results for this experiment align with our previous findings.

The comprehensive results of the test can be found in the Excel file Experiment2ResultsTable.xlsx in the repository at [https://www.kaggle.com/datasets/alessandropignatelli/seismicnoiseexperimentresults?rvi=1], following the format shown in Table 3. The mean of the evaluation metrics across all hyperparameters is presented in Table 6. As one can see, the general results are confirmed. Furthermore, the hyperparameters corresponding to the maximum f1 score are displayed in Table 7. The corresponding confusion matrix, averaged through the k-fold iterations, is presented in Fig. 10.

Metric	Mean AlexNet	Mean EfficientNet	Max AlexNet	Max EfficientNet	Min AlexNet	Min EfficientNet
accuracyMean	0,91	0,79	0,94	0,91	0,85	0,55
precisionMean	0,90	0,80	0,94	0,92	0,85	0,60
recallMean	0,96	0,91	0,98	0,94	0,92	0,86
f1scoreMean	0,93	0,85	0,95	0,93	0,89	0,71
accuracyStd	0,03	0,04	0,11	0,07	0,01	0,01
precisionStd	0,04	0,04	0,12	0,07	0,01	0,02
recallStd	0,02	0,03	0,11	0,05	0,00	0,01
f1scoreStd	0,02	0,03	0,07	0,04	0,01	0,01

Table 6. Second Experiment results averaged trough hyperparameters values.

Best NetType	Best LearningRate	Best miniBatchSize	Best ValidFractio	Best TestFractio	Best MaxEpochs
AlexNet	0,001	25	0,1	0,2	50

Table 7. Experiment 2. Hyperparameters corresponding to the best F1 score.

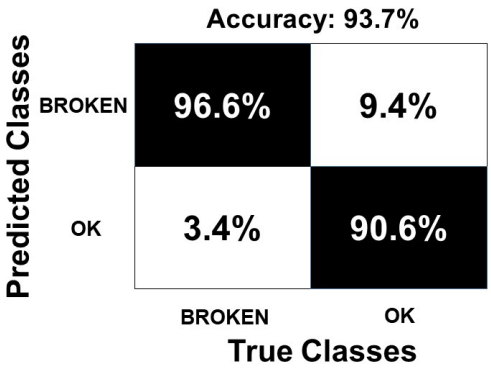

Figure 10. Second experiment confusion matrix and total accuracy of trained AlexNet neural network relative to second learning test for the best f1 score hyperparameters set. The percentages are the mean of the occurrences through the k-fold iterations.

6.4 Third Experiment: evaluation with extensive data from an unanalysed year (2018)

In this test, we did not train any new network but instead conducted a more robust evaluation using the previously trained networks. We selected 840 diagrams from the year 2018, which had not been analysed by the two neural networks before. Initially, the human operator examined 877 diagrams from 2018, discarding 37 of them as they were deemed 'uncertain'. Out of the remaining 840 diagrams, 494 (58.8%) were classified as 'OK,' by the human operator while 346 (41.2%) were classified as 'Broken.' This classification process was time-consuming. Then all these diagrams were evaluated by the previously trained network that exhibited the highest f1 score in the previous experiment. The results are depicted in the confusion matrix in Fig. 11, with an accuracy of 86.7%, and the evaluation metric statistics are presented in Table 8.

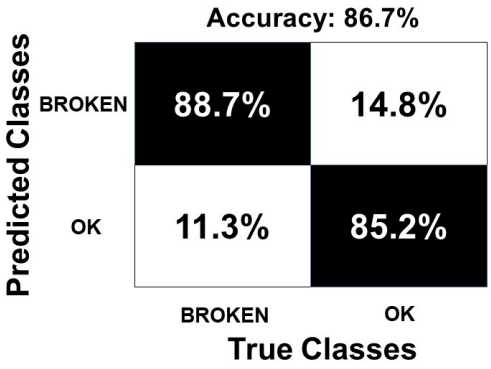

Figure 11. Experiment 3 confusion Matrix for the test on 840 diagrams. The main diagonal indicates the percentage of successes.

accuracy	precision	recall	f1score
0,867	0,84	0,91	0,87

Table 8. Experiment 3. Hyperparameters corresponding to the best F1 score.

6.5 Midway discussion: introducing a third class

In the last test, we observed a high success rate of approximately 87%, with varying percentages of false positives (14.8%) and false negatives (11.3%), indicating a somewhat cautious trend in the neural network. It is important to note that a positive case refers to the detection of a broken station (see par. 4.4). The decision to opt for a 'precautionary' or 'unscrupulous' neural network hinges on whether one prefers a slightly suspicious station to be flagged as Bad/Broken (with the risk of generating a significant number of false alarms = false positive) or if the aim is to capture stations that are unequivocally malfunctioning, with the risk of missing notifications for stations with minor issues (false negative). This decision often depends on external conditions, with various factors influencing the choice of one strategy over another. Considerations include the total number of stations, the seismic network's density, the relative positions of damaged stations, and the ability for prompt intervention on-site (or remote intervention). Opting for a precautionary network necessitates strict classification of training data. The degree of precaution in the neural network is influenced by how we train it, including considerations on how we handle 'uncertain' diagrams. Each choice comes with its advantages and disadvantages. For instance, concerning 'uncertain' diagrams, we can:

a) discard them entirely during the learning phase—however, the neural network may encounter 'uncertain' spectra among the numerous diagrams provided for analysis in the future, and in such cases, how will the network perform?;
b) include them all in the 'OK' spectra—a less precautionary choice; this will probably produce more "false negative" than "false positive";
c) include them all in the 'Broken' spectra-a more precautionary choice; this will probably produce more "false positive" than false negative;
d) decide and classify diagram by diagram-a more demanding choice for the human operator, potentially yielding different results from operator to operator;
e) create a third intermediate class of 'uncertain' diagrams, but in the years up to 2020, they were not numerous, and machine learning would be based on only a few diagrams.

The INGV database also receives data from other local and foreign networks. Recently, the number of stations received has significantly increased, along with the percentage of uncertain diagrams. In many cases, this is due to temporary failures resulting from the non-connection between the station and INGV or datalogger problems, both leading to a loss of data continuity. In these cases, the decision of whether to declare a malfunctioning station or not depends on the quantity and length of the gaps (time intervals without data or telemetry drop-out). However, the evaluation remains subjective, and there will always be a percentage of questionable cases. Furthermore, in local networks, the occurrence of non-standard installations is more frequent than in the ISN, and sometimes inexpensive instruments are used. In these cases, the documentation to control all parameters is not always complete or clear (see also par. 3.3). So, from 2021 onwards, the spectra dataset shows a number of 'uncertain' diagrams adequate to train AlexNet/Efficientnet with three classes: two extremes (OK and BAD) and one intermediate (Dubious).

6.6 Fourth experiment (three classes)

For the three-class learning, we utilized a significantly larger dataset than in the previous steps. We focused on the year 2021 and gathered a total of 1865 PDF spectral diagrams. The classification process, carried out by a single human expert, spanned approximately three weeks, resulting in 554 diagrams labelled as "OK," 476 as "Dubious," and 835 as "BAD." We trained a series of networks using the same hyperparameters employed in the previous tests and assessed the evaluation metrics on the test data using the k-fold procedure. The comprehensive results, detailing the variations in evaluation metrics based on different hyperparameters, can be found in the file Experiment2ResultsTable.xlsx in the repository at https://www.kaggle.com/datasets/alessandropignatelli/seismicnoiseexperimentresults?rvi=1. The mean evaluation metrics across all hyperparameters are presented in Table 9, confirming the general results. Additionally, Table 10 displays the hyperparameters corresponding to the maximum f1 score. The corresponding confusion matrix averaged across the k-fold iterations is depicted in Fig. 12.

Metric	Mean AlexNet	Mean EfficientNet	Max AlexNet	Max EfficientNet	Min AlexNet	Min EfficientNet
accuracyMean	0,82	0,71	0,87	0,84	0,66	0,47
precisionMean	0,89	0,81	0,93	0,92	0,82	0,63
recallMean	0,94	0,90	0,97	0,94	0,76	0,83
f1scoreMean	0,91	0,85	0,94	0,93	0,79	0,71
accuracyStd	0,03	0,03	0,25	0,04	0,00	0,01
precisionStd	0,04	0,01	0,18	0,02	0,00	0,00
recallStd	0,03	0,02	0,28	0,03	0,00	0,00
f1scoreStd	0,02	0,01	0,16	0,02	0,00	0,00

Table 9. Fourth experiment results averaged trough hyperparameters values.

BestLearningRate	Best miniBatchSize	Best ValidFractio	Best TestFractio	Best MaxEpochs
0,001	25	0,2	0,2	50

Table 10. Fourth experiment: Hyperparameters corresponding to the best F1 score.

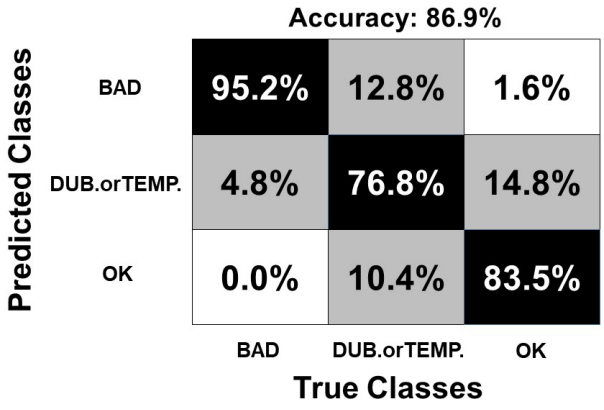

Figure 12. Experiment 4: Confusion matrix and total accuracy of AlexNet trained neural network relative to the fourth data set. The percentages are the mean of the occurrences through the k-fold iterations.

7. Final Discussion and comparison with previous studies

With the introduction of the 3rd class (intermediate), the overall accuracy reached 86.9%. A particularly encouraging result is that no 'BAD' diagrams were misjudged as 'OK,' and only a small percentage (1.6%) of 'OK' diagrams were misclassified as 'BAD' by the neural network. Hence, there were very few 'absolute' false positives and no 'absolute' false negatives in the processing. Most errors occurred when the network predicted a station in

the adjacent class as the correct one, but few or none were placed in the opposite class (OK/BAD). The overall result is deemed useful, as the network demonstrated its capability to narrow down the set of diagrams requiring further analysis and priority intervention (few OK diagrams were incorrectly judged as BAD). Consequently, operators will initially focus on resolving issues with stations classified as BAD. Subsequently, based on various factors, such as the number and density of stations, they will determine when to investigate the doubtful cases. The reasons behind the rise in the number of damaged stations in 2021 mirror those contributing to the increase in uncertain diagrams (par. 6.5): an expansion of connected stations and a rise (both in absolute numbers and percentage) in stations experiencing degraded data transmission or data acquisition issues in the datalogger or wrong metadata. While awaiting improvements in data acquisition and transmission systems, the ability to pinpoint malfunctioning stations remains valuable for planning maintenance interventions.

In the initial phase of classifying the 1860 diagrams from the 4th experiment, the expert observed a certain number of BAD diagrams attributed to probable errors in the metadata. In theory, this occurrence should be avoided, as the station is expected to undergo testing at the outset, ensuring accurate metadata. However, in practice (see the last part of Section 6.5), data from new stations belonging to local or foreign networks, over which INGV has no direct control, are currently being integrated into the INGV monitoring room, and their metadata is not always readily available or reliable. Furthermore, it can happen that the instrumentation is replaced without updating the database, or that a simple (hard) reset of the instrumentation changes sensitivity values (see below). For these reasons, extending control to metadata has become increasingly important over time.

We observe that the Machine Learning model consistently exhibits superior performance on the "BAD" class across all experiments (see par. 6.5 and Fig. 9-12). As stated in par 4.6, this behaviour appears to be independent of the distribution of BAD and OK diagrams during the training phase: specifically, in the 1st experiment, OK and BAD diagrams are equally represented, while in the second and fourth experiments, there are more BAD diagrams, and in the third experiment, there are more OK diagrams. However, in all experiments, the model demonstrates a higher frequency of classifying OK instances as BAD (false positive) compared to misclassifying BAD instances as OK (false negative).

There are fewer errors in the BAD class, which demonstrates a higher success rate. To elucidate this trend, we can refer to the discussions presented in Section 6.5) and consider the nature of precautionary or unscrupulous networks. Our findings suggest that in our experiments, the expert classified the "uncertain" diagrams with a slightly precautionary approach. For instance, in the first two experiments, it's plausible that the majority of "uncertain" plots were classified as BAD. These choices have given the network a slightly precautionary character and resulted in a difference between the accuracies of the BAD ad OK classes. However, all this does not affect the general good accuracy of the method and it is also better in terms of realizing an automatic system as what we mostly want to detect bad signals.

To enrich the discussion, below we present some points that characterize this work and distinguish it from others. In the last part of this paragraph, we will expand the discussion to include a comparison with deterministic methods.

1) In this work, we utilized pre-trained networks that were already tested and optimized for image classification. This offers significant advantages, including:
   a) Significantly less demanding in terms of computation time. As described in Section 4.3, training a pre-trained network involves tuning the already-tested initial values of the network weights.
   b) Relatively few images are needed to achieve high accuracy with test data.
   c) It is unnecessary to crop images and use grayscale. Furthermore, instead of considering image size as hyperparameters that affect efficiency, we utilized the resolution required by the specific pre-trained network input layer. Ultimately, we did not pre-process the diagrams.
2) A lot of training was run and a regular grid "hyperparameters" tuning variation has been performed before any training to compare the results.
3) Validation data are used to explore the overfitting problem.
4) The weighting strategy (describe in Section 4.6) makes the method robust to different populations for input classes.
5) The test reliability/accuracy is evaluated using more metrics than just the accuracy.
6) The K-fold procedure has been implemented to check if the accuracy was dependent on the random choice of test data.
7) Different experiments checking if the neural network accuracy was affected by the temporal distance between plots (years 2015-2021) have been performed.

8) For training the neural network, we exclusively utilized diagrams without the (unstable) mode curve. Both training and testing were conducted using diagrams containing only the PSD distribution (Probability Density Function or PDF). This approach reduces ambiguity in the diagrams, enhancing training efficiency.
9) For the classification of diagrams, over 1800 were utilized, and most of the criteria are detailed in Table 1. This set of pre-classified diagrams constitutes the dataset used for training the neural network in the fourth experiment.
10) A multitude of stations and networks contributed data to form the dataset, with the network code (IV, MN, etc.) indicating the seismic networks that provided their stations.

Some authors [Thorp et al., 2020; Nugroho et al., 2022] adopt a classification using many classes. In our work, we opted for two or three general classes because it is not always straightforward to precisely identify the subclass for categorizing the type of issue; for example, we may not always be able to determine if the error is instrumental or related to metadata. In fact, there are cases where the spectral values are out of scale, and initially, we can only infer that the metadata values differ from the instrumental parameters (e.g., sensitivity). Indeed, at the outset, we do not know if the metadata is incorrect or if, for example after a reset, the instrumental sensitivity values have changed. This 'reset problem' has been encountered in some older digitizers. In these cases, knowledge of the dates of the last station intervention and the last metadata update would certainly be useful. We consider dividing BAD diagrams into subclasses as a subsequent step, when it will be possible to integrate the results of spectral analysis with information regarding quality metrics and recent interventions on the station or metadata. Especially when it becomes feasible to automatically and quickly process and evaluate together the station data set and the 'historical' data on interventions, using a neural network.

Some current routines (for example, https://eida.ingv.it/it/getdata#) favor the assessment of the data quality providing parameters such as RMS, Max, Min. etc. Specifically, if there are significant changes in these values over time or if certain established thresholds are exceeded, the signals can be classified as bad. Unfortunately, these approaches have some drawbacks:

a) The thresholds can be subjective, and the values of parameters can depend on the instrumentation. Indeed, parameters as RMS, Min, max etc. are given in counts, without units of measurement. So, for example, the value of RMS also depends on the sensitivity of the instrumentation: higher sensitivity [count/(m/s)] implies higher counts for the same microtremor (in [m/s]). Instead, PSD or PDF diagrams are expressed in absolute dB that are referred to specific units of measurement (par. 2) as they are calculated after the deconvolution of the data. Furthermore, the spectral LNM diagram is an absolute reference so, if a station PDF diagram is under LNM, we can certainly say that either the station has problems (Fig. 4a), or the metadata are incorrect (Fig. 6a). On the contrary we have no absolute reference value in the case of parameters expressed in counts.
b) Basing the evaluation of a parameter on its changes over time should assume a previous period when the signal was not bad. We do not always can be sure of such assumption.
c) Parameters such as RMS, Min-Max, etc., do not consider the varying 'standard' values of the noise level across different frequencies (see LNM, HNM). To explain better, if we divide the entire Very Broad Band into 6 or 7 bands and filter the seismic signal within these bands, we will see that, for example, the RMS value is very different for each band. So, calculating RMS over the entire band (without filtering) gives us a general value and lacks detail in the information, making it impossible to determine if a threshold breach is due to a specific frequency interval. For instance, high value of RMS at low frequencies can be due to the insulation degradation or drift and in this case an intervention is required. High values of RMS at high frequency could be due to variation of the anthropic noise and in this case no intervention on the instrumentation is required.

In general, for many of the metrics that are calculated analytically, if they are such as to compromise the quality of the seismic signal, they also compromise the spectra, which consequently exhibit poor quality. So, checking and evaluating the spectrum allows us to evaluate most quality issues, or at least noise spectra exhibit bad diagrams for the most frequent problems. In fact, some analytical metrics are somehow linked to the criteria set out in Table 1. However, certain less frequent issues are not identifiable through spectral analysis. For instance, the orientation of horizontal components, the polarity, the accuracy of the time mark. In the future, integrating spectral analysis with intervention information and metrics capable of identifying problems that are not discernible through spectra will enable a comprehensive and reliable assessment of data quality.

8. Conclusions

The collection of high-quality seismic data is crucial for advancing geophysical research. In this context, spectral analysis, in its diverse forms, stands out as a vital investigative step for assessing the quality of seismic data. Unfortunately, the checking and maintenance of numerous stations demand significant human effort to ensure the proper functioning of the system. In this paper, we have demonstrated that visual inspection serves as a powerful tool for achieving a high standard of checks and how artificial intelligence can contribute to such tasks. More specifically, the experience and knowledge of an expert in this field can be 'transferred' into a neural network capable of automatically discriminating signals from malfunctioning stations and those collected from operational ones, achieving an accuracy of 86.7%, as demonstrated in the third test on 840 diagrams. To further narrow down the set of stations requiring intervention, we introduced a third intermediate class containing diagrams indicating questionable operations of the station (the number of such diagrams has increased in the last three years). The concept is to initially focus on stations labelled as BAD, deferring further consideration of stations labelled as 'uncertain' for future investigation. For the three-class learning, we utilized a significantly larger number of diagrams compared to previous steps. The introduction of the 3rd class resulted in a total accuracy of 86.9%. No BAD diagram was mistakenly labelled as OK, and very few OK stations were misclassified as BAD. The majority of errors involve placing an element in a class adjacent to the correct one. This result demonstrates the reliability of AlexNet and EfficientNet, enabling us to focus initial interventions solely on stations classified as BAD and subsequently investigate the Dubious ones. Restricting the number of highly suspicious stations for intervention is a valuable outcome for scheduling maintenance activities.

The applicability of the trained network may extend beyond the Italian network to any international network equipped with BB or VBB seismometers. The overall results are highly promising for future developments. We aim to expand the analysis to include short-period stations and conduct assessments on shorter time intervals, possibly on a monthly basis. Additionally, we anticipate incorporating information on various quality metrics and recent interventions on seismic stations, including metadata, into the neural network analysis in future studies.

Data availability statement. Data used to get the show results have been compressed and put at the following link: https://www.kaggle.com/datasets/alessandropignatelli/seismicnoiseexperimentresults?rvi=1

Acknowledgements. Many thanks to Valentino Lauciani who keeps the SQLX package efficient and functioning and this allows us to extract the spectral diagrams. We also would like to thank Stefano Chiappini, Andrea Morelli, Elisabetta Giampiccolo and Silvia Pondrelli for the suggestions and very useful discussions.

References

Abd Elrahman, S.M. and A. Abraham (2013). A review of class imbalance problem: Journal of Network and Innovative Computing, 1, http://ias04.softcomputing.net/jnic2.pdf.

Alejandro, A.C.B, A.T. Ringler, D.C. Wilson, R.E. Anthony R.E. and S.V. Moore (2020). Towards understanding relationships between atmospheric pressure variations and long-period horizontal seismic data: a case study. Geophys. J. Int., 223, 1, 676-691, https://doi.org/10.1093/gji/ggaa340.

Allen, R. (1982). Automatic phase pickers: Their present use and future prospects, B. Seismol. Soc. Am., 72, https://doi.org/10.1785/BSSA07206B0225.

Anglade, A., A. Lemarchand, J.M. Saurel, V. Clouard, M.P. Bouin, J.B. De Chabalier, S. Tait, C. Brunet, A. Nercessian and F. Beauducel (2015). Significant technical advances in broadband seismic stations in the Lesser Antilles, Advances in Geosciences, 40, doi:10.5194/adgeo-40-43-2015.

Anon (2020). STS-2 Legacy Product datasheet, https://streckeisen.swiss/en/products/sts-2/.

Bekara, M. and A. Day (2019). Automatic QC of denoise processing using a machine learning classification, First Break, 37, https://doi.org/10.3997/1365-2397.n0055.

Bendat, J.S. and A.G. Piersol (2011). Random Data: Analysis and Measurement Procedures, John Wiley and Sons, http://dx.doi.org/10.1002/9781118032428.

Bormann, P. and E. Wielandt (2013). Chapter 4: Seismic signals and noise, New manual of seismological observatory practice 2 (NMSOP2), Deutsches GeoForschungsZentrum GFZ, 1-62, https://doi.org/10.2312/GFZ.NMSOP-2_ch4.

Broucke, R.A., W.E. Zürn and L.B. Slichter (1972). Lunar tidal acceleration on a rigid Earth: Washington DC American Geophysical Union Geophysical Monograph Series, 16, https://doi.org/10.1029/GM016p0319.

Cao, Q. and M.E. Parry (2009). Neural network earnings per share forecasting models: a comparison of backward propagation and the genetic algorithm, Decision Support Systems, 47, https://doi.org/10.1016/j.dss.2008.12.011.

Casey, R., M.E. Templeton, G. Sharer, L. Keyson, B.R. Weertman and T. Ahern (2018). Assuring the quality of IRIS data with MUSTANG, Seismological Research Letters, 89, 630-639, https://doi.org/10.1785/0220170191.

Darbyshire, J. and E.O. Okeke (1969). A study of primary and secondary microseisms recorded in Anglesey, Geophysical Journal International, 17, https://doi.org/10.1111/j.1365-246X.1969.tb06379.x.

Doody, C.D., A.T. Ringler, R.E. Anthony, D.C. Wilson, A.A. Holland, C.R. Hutt and L.D. Sandoval (2017). Effects of thermal variability on broadband seismometers: Controlled experiments, observations, and implications, Bull. Seism. Soc. Am., 108, 1, 493-502, https://doi.org/10.1785/0120170233.

Ghojogh, B. and M. Crowley (2019). The theory behind overfitting, cross validation, regularization, bagging and boosting, Tutorial, arXiv Preprint arXiv, 1905.12787, https://doi.org/10.48550/arXiv.1905.12787.

Haji, S.H. and A.M. Abdulazeez (2021). Comparison of optimization techniques based on gradient descent algorithm: A review, PalArch's Journal of Archaeology of Egypt/Egyptology, 18, 2715-2743.

Han, X., Y. Zhong, L. Cao and L. Zhang (2017). Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification, Remote Sensing, 9, 848, https://doi.org/10.3390/rs9080848.

Holcomb, L.G. (1989). Seismic noise, The Encyclopedia of Solid Earth Geophysics, 1089-1092.

Huang, B., M. Xue, Z. Guo and W. Song (2022). Exploring the deep ocean single-frequency microseisms southwest of Japan in northern Philippine Sea, Geophys. Res. Lett., 49, 2021097444, https://doi.org/10.1029/2021GL097444.

Indolia, S., A.K. Goswami, S.P. Mishra and P. Asopa (2018). Conceptual understanding of convolutional neural network-a deep learning approach, Procedia computer science, 132, 679-688, https://doi.org/10.1016/j.procs.2018.05.069.

Janocha, K. and W.M. Czarnecki (2017). On loss functions for deep neural networks in classification, arXiv,1702.05659,10.48550/arXiv.1702.05659.

Jha, K., B. Padma Rao, C. Sribin and S. Silpa (2023). Analysis of seismic noise of broadband seismological stations installed along the Western Ghats, J. Seismol., 27, 325-342, 76, 84-85, 10.1007/s10950-023-10138-8.

Krizhevsky, A., I. Sutskever and G.E. Hinton (2017). Imagenet classification with deep convolutional neural networks, Communications of the ACM, 60, 84-90, 10.1145/3065386.

Lantz, B., (2013). Machine Learning with R, Packt publishing ltd.

Likas, A., N. Vlassis and J.J. Verbeek (2003). The global k-means clustering algorithm, Pattern Recognition, 36, 451-461, https://doi.org/10.1016/S0031-3203(02)00060-2.

Liu, Y., Y. Gao and W. Yin (2020). An improved analysis of stochastic gradient descent with momentum, Advances in Neural Information Processing Systems, 33, 18261-18271.

Ma, Y., X. Zhu, T. Guo, T. Rebec and K. Azbel (2005). Reservoir characterization using seismic data after frequency bandwidth enhancement, Journal of Geophysics and Engineering, 2, 213-221.

Mahesh, B. (2020). Machine learning algorithms-a review, International Journal of Science and Research (IJSR), 9, 38-386, https://doi.org/10.21275/ART20203995.

Marzorati, S. and V. Lauciani (2015). SQLX: Test di Installazione e Funzionamento, Rapporti Tecnici INGV, 297, 25.

Massa, M., D. Scafidi, C. Mascandola and A. Lorenzetti (2022). Introducing ISMDq-A Web Portal for Real-Time Quality Monitoring of Italian Strong-Motion Data, Seismological Research Letters, 93, 241-256, 10.1785/0220210178.

Mazza, S., M. Olivieri, A. Mandiello and P. Casale (2008). The Mediterranean broad band seismographic network anno 2005/06, Earthquake monitoring and seismic hazard mitigation in Balkan countries, 133-149.

McNamara, D.E. and R.P. Buland (2004). Ambient noise levels in the continental United States: Bulletin of the Seismological Society of America, 94, 1517-1527, 10.1785/012003001.

McNamara, D.E. and R.I. Boaz (2006). Seismic noise analysis system using power spectral density probability density functions: A stand-alone software package, US Geol. Surv. Open-File Rept, 2006.

Mejri, M. and M. Bekara (2020). Application of Machine Learning for the Automation of the Quality Control of Noise Filtering Processes in Seismic Data Imaging, Geosciences, 10, 475, 43, https://doi.org/10.3390/geosciences10120475.

Michelini, A., L. Margheriti, M. Cattaneo, G. Cecere, G. D'Anna, A. Delladio, M. Moretti, S. Pintore, A. Amato and A. Basili (2016). The Italian National Seismic Network and the earthquake and tsunami monitoring and surveillance systems, Advances in Geosciences, 43, 31-38, https://doi.org/10.5194/adgeo-43-31-2016.

Morelli, A., G. Ekström and M. Olivieri (2000). Source properties of the 1997-98 Central Italy earthquake sequence from inversion of long-period and broad-band seismograms, Journal of Seismology, 4, 365-375, https://doi.org/10.1023/A:1026587817690.

Nugroho, H.A., S. Hasanah and M. Yusuf (2022). Seismic Data Quality Analysis Based on Image Recognition Using Convolutional Neural Network, JUITA, Jurnal Informatika, 10, 67-75, 10.30595/juita.v10i1.11261

O'Shea, K. and R. Nash (2015). An introduction to convolutional neural networks, arXiv Preprint arXiv, 1511.08458, https://doi.org/10.48550/arXiv.1511.08458.

Pandey, D., K. Niwaria and B. Chourasia (2019). Machine Learning Algorithms: A Review, Int. Res. J. Eng. Technol., 6.

Peterson, J. (1993). Observations and modeling of seismic background noise. U. S. Geol. Surv., Open File Rept., 93, https://doi.org/10.3133/ofr93322.

Picozzi, M., L. Elia, D. Pesaresi, A. Zollo, M. Mucciarelli, A. Gosar, W. Lenhardt and M. Živčić (2015). Trans-national earthquake early warning (EEW) in north-eastern Italy, Slovenia and Austria: first experience with PRESTo at the CE 3 RN network, Advances in Geosciences, 40, 51-61, https://doi.org/10.5194/adgeo-40-51-2015.

Pignatelli, A., F. D'Ajello Caracciolo and R. Console (2021). Automatic inspection and analysis of digital waveform images by means of convolutional neural networks, Journal of Seismology, 25, 1347-1359, https://doi.org/10.1007/s10950-021-10055-8.

Pondrelli, S., F. Di Luccio, L. Scognamiglio, I. Molinari, S. Salimbeni, A. D'Alessandro and P. Danecek (2020). The first very broadband Mediterranean network: 30 yr of data and seismological research, Seismol. Res. Lett., 91, 787-802, https://doi.org/10.1785/0220190195.

Rastin, S.J., C.P. Unsworth, K.R. Gledhill and D.E. McNamara (2012). A detailed noise characterization and sensor evaluation of the North Island of New Zealand using the PQLX data quality control system, Bulletin of the Seismological Society of America, 102, 98-113, https://doi.org/10.1785/0120110064.

Rawat, W. and Z. Wang (2017). Deep convolutional neural networks for image classification: A comprehensive review, Neural Computation, 29, 2352-2449, 10.1162/NECO_a_00990.

Scales, J.A. and R. Snieder (1998). What is noise?, Geophysics, 63, 1122-1124, https://doi.org/10.1190/1.1444411

Halbert, S. (1993). Appendix A: Channel Naming, Standard for the Exchange of Earthquake Data, Seed Reference Manual (IRIS Consortium).

Shorten, C. and T.M. Khoshgoftaar (2019). A survey on image data augmentation for deep learning, Journal of Big Data, 6, 1-48, https://doi.org/10.1186/s40537-019-0197-0.

Sleeman, R., A. Van Wettum and J. Trampert (2006). Three-channel correlation analysis: A new technique to measure instrumental noise of digitizers and seismic sensors, Bull. Seismol. Soc. Am., 96, 258-271, 10.1785/0120050032.

Sokolova, M. and G. Lapalme (2009). A systematic analysis of performance measures for classification tasks, Inf. Process. Manag., 45, 427-437, https://doi.org/10.1016/j.ipm.2009.03.002.

Sorrells, G.G. (1971). A preliminary investigation into the relationship between long-period seismic noise and local fluctuations in the atmospheric pressure field, Geophys. J. Int., 26, 71-82, 10.1111/j.1365-246X.1971.tb03383.x.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov (2014). Dropout: a simple way to prevent neural networks from overfitting, The Journal of Machine Learning Research, 15, 1929-1958, 10.5555/2627435.2670313.

Stutzmann, E., G. Roult and L. Astiz (2000). GEOSCOPE station noise levels, Bull. Seismol. Soc. Am., 90, 690-701, 10.1785/0119990025.

Taner, A., Y.B. Öztekin and H. Duran (2021). Performance analysis of deep learning CNN models for variety classification in hazelnut, Sustainability, 13, 6527,10.3390/su13126527.

Thorp, J., K. Davies, J. Bluteau and P. Hoiles (2020). Implementation of seismic data quality characterisation using supervised deep learning, The APPEA Journal, 60, 784-788, 10.1071/AJ19040.

Vassallo, M., C. Satriano and A. Lomax (2012). Automatic picker developments and optimization: A strategy for improving the performances of automatic phase picker, Seismol. Res. Lett., 83, 541,10.1785/gssrl.83.3.541.

Wielandt E. (2012). Seismic sensors and their calibration, New Manual of Seismological Observatory Practice 2 (NMSOP-2), Deutsches GeoForschungsZentrum GFZ, 1-51, https://doi.org/10.2312/GFZ.NMSOP-2_ch5.

Wielandt, E. and J.M. Steim (1986). A digital very-broad-band seismograph, Ann. Geophys. Ser. B., 4, 227-232.

Wielandt, E. and T. Forbriger (1999). Near-field seismic displacement and tilt associated with the explosive activity of Stromboli, Annals of Geophysics, 42, 10.4401/ag-3723.

Wielandt, E., P. Bormann and J. Bribach (2002). New Manual of Seismological Observatory Practice (NMSOP), Chapter 5, Seismic sensors and their calibration, 62, Jahrestagung Der Deutschen Geophysikalischen Gesellschaft, https://moodle2.units.it/pluginfile.php/294221/mod_resource/content/1/manual_seismological_observatory-2002.pdf.

Zhu, W. and G.C. Beroza (2019). PhaseNet: a deep-neural-network-based seismic arrival-time picking method, Geophys. J. Int., 216, 261-273, 10.1093/gji/ggy423.

***CORRESPONDING AUTHOR: Alessandro PIGNATELLI**,
Istituto Nazionale di Geofisica e Vulcanologia (INGV), Via di vigna murata 605, 00142 Roma, Italy
e-mail: alessandro.pignatelli@ingv.it