POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

A DBscan clustering approach of acoustic emission signals of adhesively bonded joints under mode I fatigue loading

Tesi di Laurea Magistrale in
Mechanical Engineering - Advanced Mechanical Design (ME5)

Author: **Simone Delli Carri**

Student ID: 953070
Advisor: Prof. Michele Ezio Ruggero Maria Carboni
Co-advisors: Ph.D. Rosemere De Araujo Alves Lima
Academic Year: 2022-23

Abstract

The purpose of this master's thesis is to verify the clustering capabilities of the DBSCAN algorithm for data from structural health monitoring by acoustic emission of a adhesive bonded joint. Bonded joints are chosen for light weight, bonding of different materials, and uniform load distribution in structural applications. The data analyzed are from double cantilever beam (DCB) specimen. Acoustic Emission (AE) was used for its ability to detect the signal from a damage source located inside the DCB beam joint. Given the large amount of data acquired through AE, Machine Learning algorithms were chosen for efficient and timely analysis of incoming data. The selected Machine Learning algorithm is the Density-Based Spatial Clustering of Applications with Noise (DBSCAN), a highly promising choice for data clustering, as it has several advantages over better known algorithms: ability to recognize noise, the need to not define the number of clusters a priori as well as the ability to form clusters regardless of shape. However, the methods in literature are based on a limited amount of data. To adjust the two parameters of the algorithm (the minimum number of points, minPts, and the cluster radius, eps), an approach was adopted that reduced the unknowns to be defined (eps as a function of minPts). Next, a method was developed in which, initially, a minimum number of points is assumed, eps is calculated as a function of the minimum number of points chosen, and finally minPts is reevaluated. The resulting clustering reflected expectations and was subsequently verified by waveform representation.

Keywords: adhesive joints, acoustic emission, dbscan

Abstract in lingua italiana

Lo scopo di questa tesi magistrale è verificare le capacità di clusterizzazione dell'algoritmo DBSCAN per dati provenienti da monitoraggio strutturale tramite emissione acustica di un giunto incollato. I giunti incollati sono stati scelti per la leggerezza, l'unione di materiali diversi e la distribuzione uniforme del carico nelle applicazioni strutturali. I dati analizzati provengono da provino di doppia trave a sbalzo (DCB). L'Acoustic Emission (AE) è stata utilizzata per la sua capacità di rilevare il segnale proveniente da difettologia situata all'interno del giunto della trave DCB. A fronte dell'ingente quantità di dati acquisita tramite AE, è stato scelto di adottare algoritmi di Machine Learning per un'analisi efficiente e tempestiva dei dati in ingresso. L'algoritmo di Machine Learning selezionato è il Density-Based Spatial Clustering of Applications with Noise (DBSCAN), una scelta altamente promettente per la clusterizzazione dei dati, poiché presenta diversi vantaggi rispetto agli algoritmi più conosciuti: capacità di riconoscere il rumore, la necessità di non definire a priori il numero di cluster oltre che la capacità di formare cluster indipendentemente dalla forma. Tuttavia, i metodi presenti in letteratura si basano su una quantità limitata di dati. Per adeguare i due parametri dell'algoritmo (il numero minimo di punti, minPts, e il raggio del cluster, eps), è stato adottato un approccio che ha ridotto le incognite da definire (eps in funzione del minPts). Successivamente, è stato sviluppato un metodo in cui, inizialmente, si ipotizza un numero minimo di punti, si calcola eps in funzione del numero minimo di punti scelto, e infine si rivaluta minPts. La clusterizzazione risultante ha rispecchiato le aspettative ed è stata successivamente verificata tramite la rappresentazione di forme d'onda.

Parole chiave: giunti incollati, emissione acustica, dbscan

Contents

Introduction

In recent years, adhesive joints have gained popularity in several industries due to their effectiveness in transferring loads between composites or metals, attributed to a larger contact area than traditional mechanical fastening methods. Adhesive joints offer versatility with even stress distribution and high/low temperature strength-to-weight ratio, design flexibility, damage tolerance, and fatigue resistance. They could have several applications in the automotive, aerospace, petroleum, and construction industries. The strength of these joints is influenced by factors such as joint geometry, mechanical properties of the adhesive and bonding agent, and environmental conditions. Evaluation methods including fracture mechanics and testing with specimens such as double cantilever beam are used to measure adhesive strength, particularly with the DCB test to determine fracture toughness in Mode I.

Non-destructive testing (NDT) methods play a crucial role in assessing the integrity of materials by detecting surface defects, internal problems and examining metallurgical conditions without altering the structure of the material or compromising its fitness for service. NDT enables comprehensive characterization of damage, including surface and internal defects such as cracks, voids, cavities, delaminations and defective welds, effectively preventing premature failure. However, real-time inspection and monitoring using NDT has limitations due to the need to stop the machine during preparation, limiting information acquisition to scheduled inspections. To enable continuous monitoring of facilities in real time, ensuring safety and reliability during service, on-demand inspection techniques become essential. Selecting an appropriate inspection method involves evaluating potential discontinuities in the component under consideration, taking into account factors such as stress history, geometric vulnerabilities, and manufacturing cycle.

Structural health monitoring (SHM) is an emerging area of mechanical engineering that is rapidly becoming popular for its ability to improve operational safety, reliability, and reduce component maintenance costs. SHM achieves these goals by providing real-time diagnostics of a structure's condition during use, maintaining a historical record of its operation, and enabling prognosis to assess deterioration trends and remaining life. This

is critical to avoid irreparable damage and detect signs of development before structural collapse. Among the various SHM methodologies, acoustic emission (AE) emerges as a promising choice for our work. AE involves the evaluation of elastic waves generated within monitored materials during deformation, particularly during the initiation and propagation of deformation or damage. A comprehensive understanding of the physical principles behind AE, including sensors and overall monitoring of the structure using this technique, is essential. When an external force acts on a solid, it undergoes deformation, releasing strain energy during plasticization and cracking. This release of strain energy, expressed as elastic waves, is called "acoustic emission" (AE), which represents the sound produced by a solid material undergoing stress and deformation.

Having obtained the data, it is important to understand how we can obtain useful information to make the most of real-time acquisition through AE. To do this, Machine Learning (ML) algorithms come to our aid. The choice to utilize machine learning is based on the fact that, through structural monitoring methods (as Acoustic Emission or Guided Waves testing), the amount of data one receives are very large because data are captured, not only during a particular phenomenon but throughout the life of the component. To help us manage the large amount of incoming data, the use of artificial intelligence allows us to manage what is being tested or monitored relatively quickly. ML proposes two types of approaches to data: supervised and unsupervised. Both learning methods have dedicated algorithms that have been developed over the years; however, it is important to understand how to use them consciously. After data acquisition through appropriate sensor placement, processing of the data obtained during the test or service is necessary. Therefore, a suitable recognition procedure should be chosen for the case under study. Classification can be done in two ways:

- supervised methods use a dataset who acts as a guide to teach the algorithm how to generate the results. In other words, a known input and output pattern is given to take as an example to classify the data obtained. It is essential to know a priori the number of classes of interest. The algorithm learns how to handle the input data through a set of examples representative of the case in question, known as a "training set."
- unsupervised methods use a more independent approach, in which a computer learns to identify complex processes and patterns without knowing a preset model provided a priori. Objects are classified according to the similarities of their features, and no reference cases are required. This is referred to as clustering, so that "N" objects are grouped into "C" a priori unknown groups. There is then knowledge of the input

but not of the output (or vice versa).

In this thesis, the data utilized were derived from a mode I fatigue crack growth test conducted on metallic adhesively bonded specimens of the "Double Cantilever Beam" (DCB) monitored throught Acoustic Emission. The DCB is a commonly employed experimental technique in engineering for assessing the fracture toughness and material resistance to crack propagation. The data originate from an experimental study, the details of how the DCB specimen and test was performed and data analysis methodology are explained in the article[1].

The purpose of this study was to select the most promising methodology from the multitude of machine learning algorithms available in the literature. This involved a detailed analysis of the operating principles, pros and cons associated with each algorithm, as well as fitting our data to the chosen algorithm to achieve true clustering. The investigation focused primarily on DBSCAN, an increasingly popular algorithm known for its many advantages. The goal was to emphasize that even algorithms under development, originally designed for smaller datasets, can give satisfactory results when applied to "big data". This is especially relevant in scenarios with background noise while maintaining the accuracy of the final result. In conclusion, to validate the clustering performed with the DBSCAN algorithm, waveform comparisons were conducted for each cluster. This validation aimed to ensure that distinct groups really identified different phenomena.

The thesis is organized into five chapters.

- **Chapter 1** State of the art. A study was conducted on adhesive bonded joints and major bonding methods, focusing particularly on the DCB sample. A review of non-destructive testing(NDT), structural health monitoring (SHM), with an emphasis on acoustic emission was also conducted. Finally, explanation was given on machine learning for post-processing and how it works.
- **Chapter 2** Clustering algorithms. In this chapter, the procedure followed to arrive at the selection of the DBSCAN algorithm is delineated. Several algorithms were carefully analyzed and compared with each other in order to identify the most promising one to apply to our data.
- **Chapter 3** Methodology. In this section, the methodology for determining the key parameters needed for DBSCAN to achieve accurate clustering is detailed.
- **Chapter 4** Results. All the results obtained were presented, providing detailed verification that the work led to reliable results.

- **Chapter 5** Conclusions and future work. Final conclusions and possible future work are described in this chapter.

1 | State of the art

1.1. Adhesive bonded joints

In recent years, adhesive joints have become popular in several industries due to their effectiveness in transferring loads between composites or metals. This effectiveness derives from the larger contact area compared with traditional mechanical fastening methods. In addition, adhesive joints demonstrate versatility in various industrial applications due to their homogeneous stress distribution[2], high and low temperature resistance[3] in different field such as automotive[4], aerospace[5], petroleum[6], and even construction[7]. They are chosen for their high strength-to-weight ratio, design flexibility, damage tolerance and fatigue resistance. In essence, they can match or even surpass classic joining methods such as bolts, welding, screws, and riveted joints[8].

1.1.1. Bonding methods

Firstly, we must consider different bonding methods: co-bonding, co-curing, secondary bonding and multi material bonding. The first one is done when only one adherent (which is the part that it will be joint) is cured with the adhesive, which means that only one layer is polymerized, while, when both parts are simultaneously cured, we are talking about co-curing (it can be formed either with or without the use of an adhesive, and the entire laminate undergoes a single curing process[9]). In both cases they can be used to joint composites[10] or metals[11]. Secondary bonding and multi material bonding have the same techniques of joining with the only difference that in the first one the adhesive layer is cured between two pre-cured panels of the same material, in the other we have a combination of different cured substrates of materials[10].

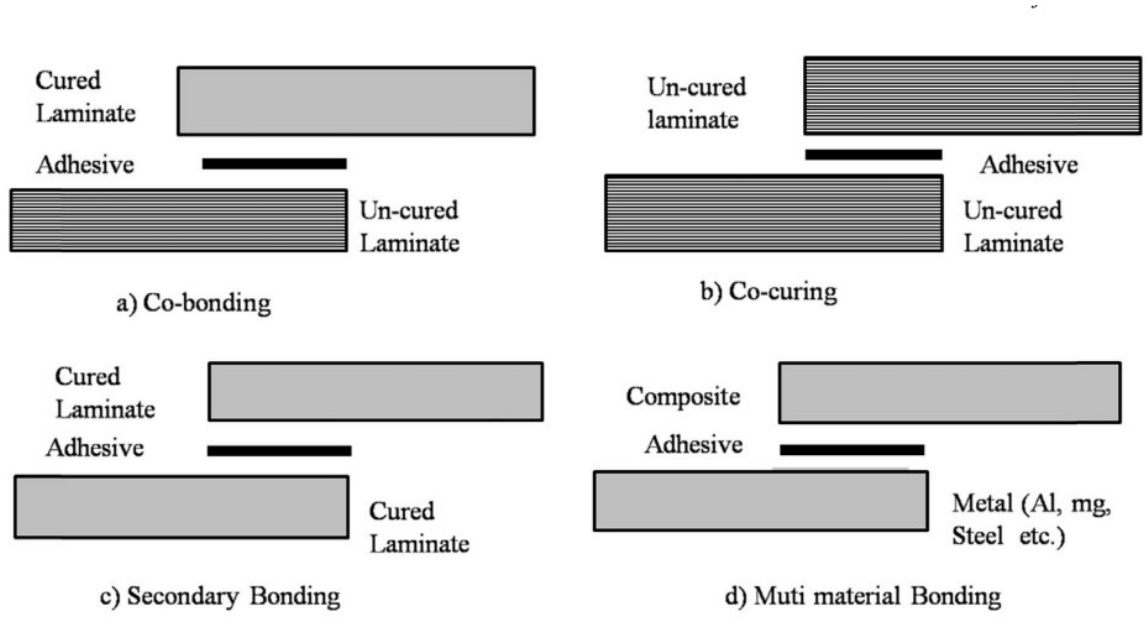


Figure 1.1: Bonding methods[10]

1.1.2. Type of adhesive joint

In industrial applications, different adhesive joints can be used. The strength of a specific joint type, under a particular load, is determined by the stress distribution within the joint. This is influenced by the joint geometry as well as the mechanical properties of both the adhesive and adherend[12]. According to the type of load that is applied to the structure, different ways of bonding can be adopted in other to resist as much as possible avoiding fractures, cracks or collapse of joints, even if the reliability of bonded connections in terms of fatigue and long-term behavior is limited[13]. However, the mechanical resistance of an adhesive joints strongly depend on several factors, that could compromise its structural integrity, including the materials being joined, the working environment, pre-treatment methods[14] but also by surface preparation[15].

Different adhesive joints can be used as for example single lap joint (SLJ), that is one of the most used joint, in which two adherents are joined together with an adhesive overlay. It's easy to fabricate and their results are sensitive to both adhesive quality and adherent surface preparation[16]. In addition, several studies have been conducted to understand its behavior under fatigue[17] or static load[18]. Double Lap Joint (DLJ) compared to the SLJ, has higher efficiency because of duplication of the shear-resistant area. However, we need access to both sides of the structure to obtain this junction [19]. As in the previous case, also for DLJ some studies are done[20]. Adhesive butt joints involve connecting two surfaces along their edges in a butt configuration. In this joint type, the two materials

are positioned end-to-end, and adhesive is applied to bond them together without any overlapping[21]. Adhesive butt strap joints typically involve bonding two surfaces together using an adhesive, and a strap or band is used to add extra reinforcement and support to the connection. These joints can be created in either a single configuration or a double configuration.

In the figure 1.2 are shown various types of structural adhesive bonded joints commonly utilized in industrial applications explained above.

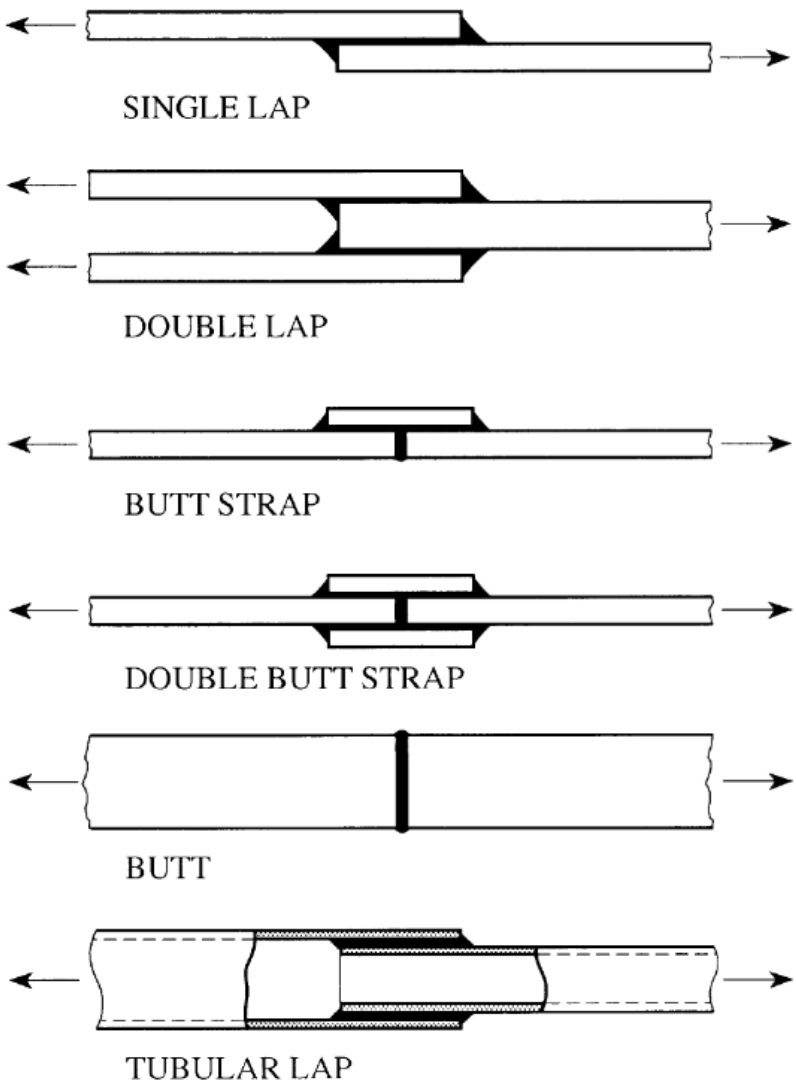


Figure 1.2: Structural adhesive bonded joints[22]

To assess the strength of adhesive joints, a commonly employed approach involves utilizing fracture mechanics. Traditionally, when the crack propagation follows a pure open mode, the use of a double cantilever beam (DCB) has been employed to comprehend the fracture

behavior[23] and so a DCB specimen is created for test method for measuring Mode I fracture toughness[24]. The procedure for preparing the test specimen involves bonding two adherents with adhesive and introducing an initial crack, before you clean up well the surfaces. Subsequently, a cyclic load is applied. A more detailed description of how the sample is prepared, whose data were used in this thesis work, is given in the article[1].

In the following paragraphs, a quick introduction on NDT is done necessary to better introduce the concept of structural health monitoring methods and the specific approach used for acquisition will be detailed.

1.2. Non-Destructive Testing

Non-destructive testing methods (NDT) are used to evaluate the integrity of materials by identifying surface defects, internal or examining metallurgical conditions, all without causing any alteration of the material or compromise its suitability for service. This implies that non-destructive testing allows for the characterization of damage or defects both on the surface and within the materials without employing cutting or any other form of alteration. These methods are designed to identify problems such as cracks, internal voids, surface cavities, delamination, incomplete or defective welds, and any other imperfections that could lead to premature failure [25]. The NDT techniques offer a cost-effective method for testing samples on an individual or can be applied to the entire material for a complete investigation and comprehensive examination.

Inspection and monitoring of materials, components, or structures by NDT, however, cannot take place in real time because, each of the nondestructive testing techniques needs a preparation that must take place while the machine is stopped. Therefore, there is no opportunity to receive information in real time, but only during scheduled inspections. As a result, inspection techniques must be used to monitor the degradation of structures on-demand. Assessing the condition of a facility in real time is essential to ensure its safety and reliability during service [26].

To choose the correct inspection method, and since there are several non-destructive testing methods, it is important to evaluate the types of discontinuities that could be detected in the component under investigation, since there are no NDT methods that can detect all possible imperfections. Knowledge of the previous history of the component is essential, including the stress history, geometric points (notches) susceptible to over-stressing, and the technological cycle used to produce the part itself. A careful preliminary

study in the laboratory and a detailed study of the component's documentation is critical.

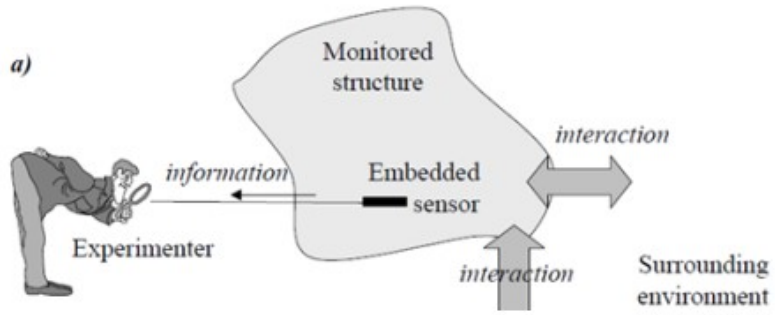
On-demand monitoring of a structure, also known as Structural Health Monitoring (SHM), is not suitable for all forms of non-destructive testing, as it requires that the structure (or specimen) be constantly monitored all along its service or testing and equipped with appropriate sensors. The selection is therefore limited to a few methodologies, among which Acoustic Emission (AE) stands out.

1.3. Structural Health Monitoring

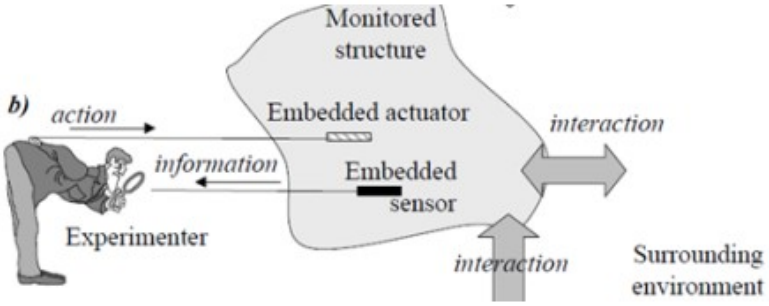
Structural Health Monitoring (SHM) is a very recent area of mechanics that is spreading rapidly because it allows for increased safety in operation, increased reliability, and decreased costs related to component maintenance and repair. These goals are achievable because SHM can provide real-time diagnosis of the state of a structure at any time during its use. The structural monitoring system keeps a record of the past operation of the structure, also allowing prognosis, i.e., assessment of the deterioration trend and remaining life. This capability is critical for preventing irreparable damage and identifying any signs of damage development before it leads to the collapse of the structure. It should be remembered that monitoring using SHM must always remain linked to the structure being monitored because, just as the structure (or component) is subject to wear and aging over time, the sensors associated with structural monitoring can also deteriorate over time. All of this is possible because, unlike other non-destructive testing, SHM allows planned interruptions in service to be replaced by targeted maintenance based on the actual condition of the structure on an instant-by-instant basis [27].

1.3.1. Passive and Active monitoring systems

Monitoring systems can be divided into passive or active, depending on the methodology with which they work. We refer to passive monitoring when monitoring consists of simply observing the behavior of the structure using sensors embedded in the component 1.3a. Instead, we refer to active monitoring when we interact with the structure using actuators to perturb and sensors to detect its response 1.3b.



(a) Passive monitoring



(b) Active monitoring

Figure 1.3: 1.3a "Passive" and 1.3b "Active" SHM methods[28]

Actuators and sensors can belong to the same or different categories, and some types of transducers can perform both functions simultaneously. Sensors can be either placed on the structure or embedded in manufacturing. The advantages of the latter are certainly that the sensor is protected and not exposed to the environmental conditions of operation, there is better sensor-material interaction, and monitoring in incessant areas becomes possible. On the other hand, however, not all materials and technological processes allow the incorporation of elements into structures, in addition to the fact that if the sensor were to be damaged it would prove impossible to replace it, but the entire part would have to be replaced. Regardless of the method chosen, it is important that SHM sensors give a faithful response to the reactions of the embedded structure, faithfully transmit the acquired signals, are unobtrusive to the structure, withstand the operating conditions, and should be easy to handle. The choice of sensor is subjective for a given application; there are no rules that apply to all case studies, and the wrong choice can lead to incorrect data acquisition and post-processing. Infact, the choice of sensors depends on the type of applications, data, and damage we expect from the structure being monitored.

1.3.2. Methothologies for SHM

Every Structural Health Monitoring (SHM) technique varies from others in terms of the measurement component (sensors) and the specific physical phenomenon adopted during the monitoring process by the sensors. The use of each phenomenon has advantages and disadvantages in practical application. Importantly, the choice of one over the others directly and completely defines the monitoring strategy. This decision affects the type of sensors, the setup configuration, and how the data are analyzed. Therefore, once the decision is made, it is difficult to change it and go back. The most popular methodologies are:

1. **Modal-data-based**: relies on the observation that the existence of structural damage leads to a reduction in the stiffness of the structure, alterations in natural frequencies, and shifts in frequency response patterns and structural modes [29]. It is a cheap method, and insensitive to damage, at least until it is of a size that changes the dynamic behavior of the structure.

2. **Electro-mechanical-impedance-based**: built on the premise that the composition of a system adds a specific contribution to its overall electrical-mechanical impedance, and the existence of damage alters the impedance within a high-frequency range, typically exceeding 30 kHz [30]. It is cheap and not very sensitive to damage that needs to be near the sensors.

3. **Static parameter-based**: founded on the observation that the existence of damage induces alterations in displacement and strain distribution as compared to a benchmark [31]. It is easily integrated; however, it detects damage only near the sensor. It is also very expensive.

4. **Acoustic emission**: grounded in the reality that the swift release of strain energy produces transient waves, allowing for the assessment of the presence or progression of damage by detecting acoustic waves emitted due to damage.

5. **Elastic-wave based**:built on the observation that structural damage induces distinctive wave scattering phenomena and mode conversion. The quantitative assessment of damage is attainable by examining the wave signals scattered by the damage [32].

In this thesis, the most suitable technique chosen for structural monitoring of the bonded joint was found to be Acoustic Emission. In particular, the choice of the physical phenomenon to be adopted and the related sensors are strongly influenced by the type of

structure, type of damage we expect on the structure to be monitored, and the boundary conditions.

As a last, it should be mentioned that through SHM it is possible to both monitor the in-service health of the structure, thus having the possibility to have a diagnosis, but it is also possible to anticipate the moment when a structure or system will no longer be able to effectively perform its design function. In order for this to happen it is necessary to have both an initial knowledge about the potential failures of the structure, including their location, mode, cause, and mechanism, to accurately identify the system parameters and to have chosen the correct method of SHM. Such a discipline is known as "prognostics." With prognostics, the future performance of the component is defined relative to nominal performance, as properly schematized in the figure1.4.

An effective prognostic system is based on several fundamental pillars:

- A thorough understanding of the failure modes and mechanisms of the materials involved
- The ability to detect early signs of damage
- The detailed understanding of the conditions leading to system failure

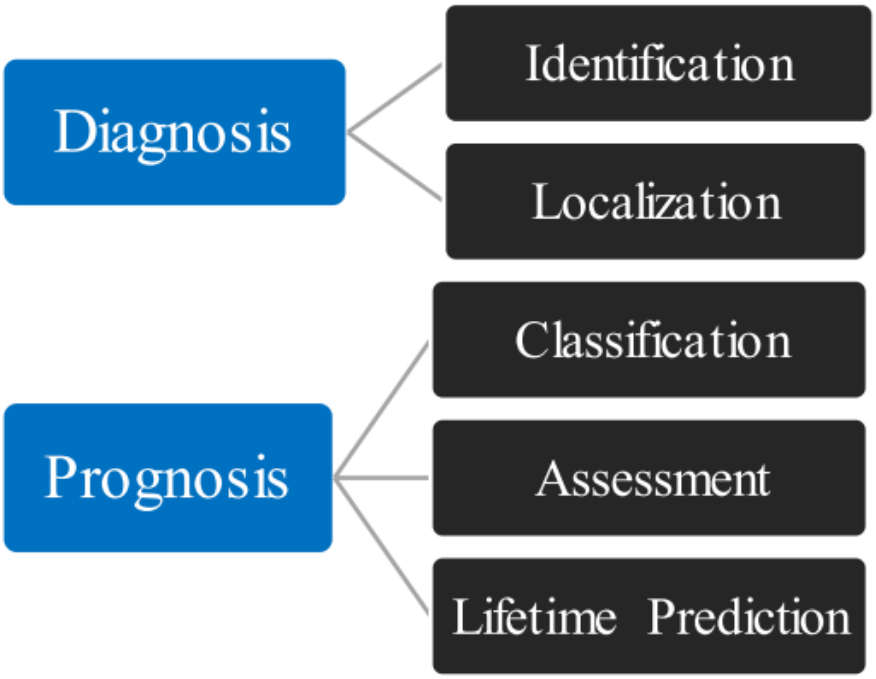


Figure 1.4: Diagnosis and Prognosis scheme[27]

1.4. SHM by Acoustic Emission

Among the various Structural Health Monitoring (SHM) methodologies, Acoustic Emission (AE) emerges as a promising option. This method offers the ability to evaluate elastic waves generated within the monitored material during the reduction of their deformation, that is, when strain energy is released during the initiation and propagation of deformation or damage. It is necessary, however, to explain the physical principle behind acoustic emission, sensors and everything related to structure monitoring by adopting this technique.

1.4.1. Physical principle of AE

When an external force acts on a solid, it undergoes deformation. In the elastic regime, the force generates elastic energy accumulation that is fully recovered at unloading. With the first plasticization, the strain energy is released in the form of plastic deformation, and the recovery at unloading is not complete. By further increasing the force, strain energy is released through nucleation and crack propagation. During plasticization and cracking, some of the strain energy is dissipated in the form of heat and, of considerable interest, sound. This event is called "acoustic emission" (AE) and is the sound expression of a solid material when it is subjected to stress and deformation. Specifically, plastic deformation and cracking release strain energy in the form of elastic waves that propagate both inside and outside the material. In summary, AE represents a phenomenon in which strain energy stored in a solid is released in the form of deformation and cracking, generating elastic waves.

The main categories of waves in acoustic emission (AE) are generally divided into two types:

- **Transient AE Waves**: these waves consist of a single wave packet that fades over time. They are commonly referred to as "events" or "hits" (figure 1.5).

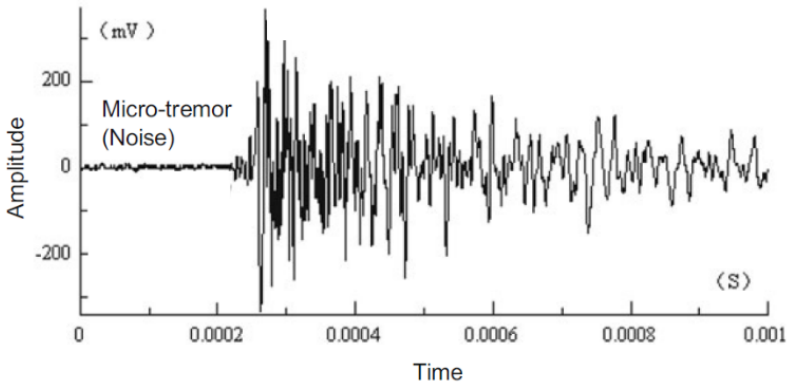


Figure 1.5: Transient wave[28]

- **Continuous AE Waves**: these waves are the result of continuous mechanical phenomena over time, such as friction, or from the superposition of numerous transient events close together in time (figure 1.6).

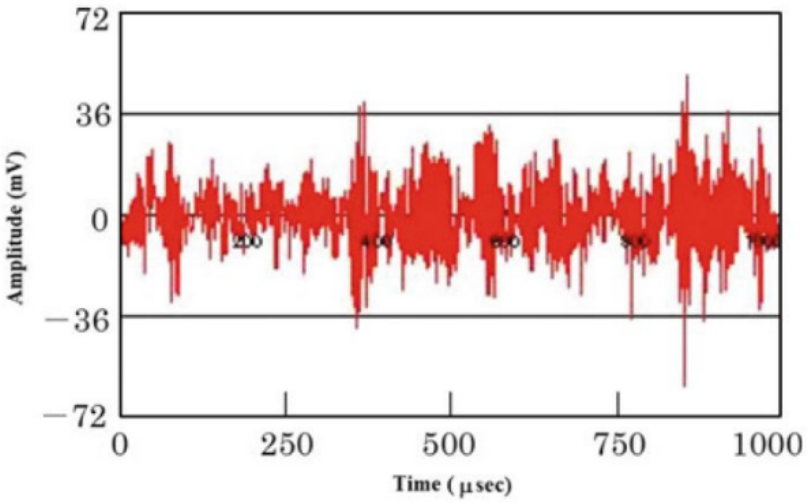


Figure 1.6: Continuous wave[28]

Monitoring by AE involves listening for and capturing these elastic (ultrasonic) waves by means of special sensors, typically piezoelectric, placed on the surface of the structure. Some of these waves are also released into the air in the form of sound waves, which are often distinctly audible. The manifestation of collapse does not occur instantaneously, even though its occurrence can be rapid. It begins at a microscopic scale and progresses gradually, accumulating damage to the point of failure. Structural Health Monitoring (SHM) by Acoustic Emission (AE) allows monitoring of this entire process, following it from beginning to end or until a decision is made to remove the structure from service.

1.4.2. "Kaiser" effect and sources of AE

As mention in the section 1.3.1 there are two types of SHM methods. Acoustic Emission (AE) is a passive method; therefore, the structure under monitoring must be subjected to loads to generate sound waves. Specifically, AE represents an irreversible phenomenon; this implies that after the application of a maximum load, a further higher load must be applied to observe further acoustic emission, a concept known as the "Kaiser effect." This phenomenon is found in certain metals and composites, bringing with it a significant practical consequence: each emission occurs only once, and there is no second opportunity to repeat the test, inspection, or monitoring. Sources of acoustic emission (figure 1.7) include various deformation and cracking mechanisms, known as "primary AE." In the case of metals, this may involve crack propagation, dislocation movement, creep, grain edge sliding, crystal sprouting, fracture and de-cohesion of inclusions, and so on.

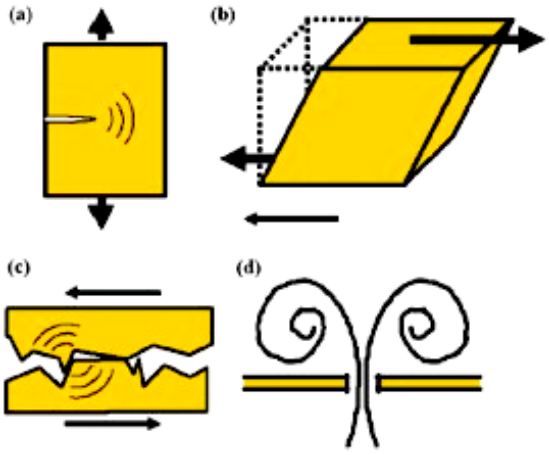


Figure 1.7: Examples of acoustic emission sources: (a) Cracking, (b) Deformation, (c) Sliding or slip, (d) Leakage[33]

In composites, the sources include matrix fracture, fiber fracture, and fiber de-cohesion. In ceramic materials, such as cement, the sources can result from phenomena such as fracture and creep. In addition to what happens internally in the material, there are other clearly audible mechanical phenomena known as "secondary AE." These can result from leakage and cavitation, friction, realignment of magnetic domains, liquefaction and solidification, phase transformations, corrosion, and so on. Depending on the specific context, such phenomena may be desired or represent annoying background noise.

1.4.3. Sensors for AE

Signals from the monitored structure must be equipped with sensors. Unlike many other methods, where the collation of the latter is of fundamental importance for acquisition, in acoustic emission you do not have to be particularly rigid about their placement. This implies that the placement of sensors in the proximity of the damaged area is not as critical as in other methods. Important, however, is the choice of sensors, which allow a mechanical strain to be transformed into a voltage after the sensor has detected and received an AE wave. For this purpose, piezoelectric transducers are used. There are two categories of AE sensors:

- **Resonant sensors**: when the wave is received, the piezoelectric transducer (PZT), which is not damped, begins to vibrate, emphasizing the harmonic components of its resonances, while the others are rapidly damped. The primary resonant frequency of the PZT is determined by the formula (1.1):

$$fr = \frac{V_{piezo}}{2s} \tag{1.1}$$

- **Broadband sensors**: this type is more similar a traditional ultrasonic probe (UT) because the PZT element is damped by a backing material. As a result, the frequency spectrum is flat and broad, but, at the same time, the sensor is less sensitive due to the increased damping

It should be noted that, through the sensors, it is not only possible to capture information about the monitored structure, but it is also possible to proceed to the localization of the sources by triangulating the flight times of the signals to a sensor network. The sensors are part of a more complete set of equipment called a measuring system that typically includes several independent channels, each with its own elements, including sensors and an associated preamplifier located directly on the structure. This positioning is essential as the detected signals are low intensity and require immediate pre-amplification to avoid an unacceptable signal-to-noise ratio before being transmitted.

1.4.4. Type of waves and attenuation

As for AE wave types, they can generally be classified into two main categories. Transient AE waves consist of a single wave packet that fades over time and are often referred to as "events" or "hits". On the other hand, continuous AE waves derive from mechanical phenomena that continue over time, such as friction, or from the overlapping of numerous

transient events that occur close together over time. The AE waves, similar to all elastic waves, undergo a process of attenuation during their propagation through a medium. The phenomena involved include the geometry of the front, in which the energy of the spherical front remains constant but its surface increases with propagation.

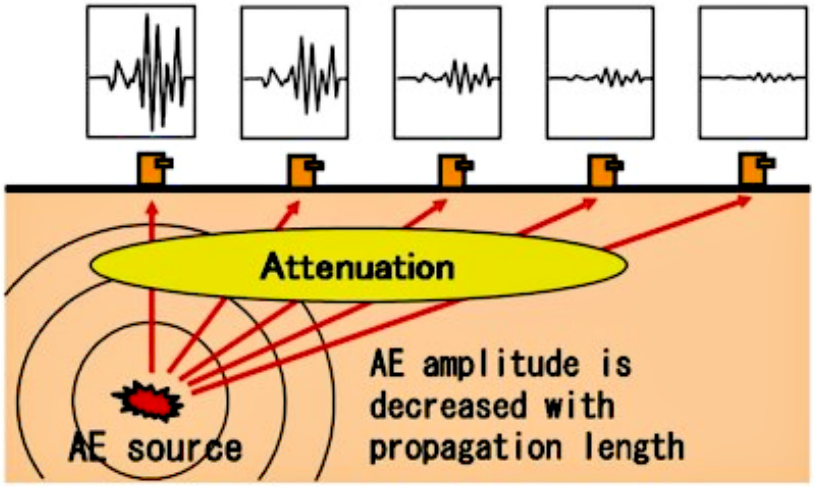


Figure 1.8: Attenuation mechanisms[34]

Diffusion and diffraction occur when the front interacts with the microstructure of the material, fragmenting and spreading without any decrease in energy. In addition, there is absorption, where the energy of the waves is converted into heat through interaction with the material.

1.4.5. Advantages and disadvantages of AE

Acoustic emission (AE) monitoring has several advantages that contribute to its attractiveness as a structural evaluation technique. It is relatively inexpensive and offers a quick and applicable method even for large structures. Its versatility makes it suitable for any type of material, while ensuring a very low invasiveness. One advantage is the ability to distinguish different types of damage, if any, and to pinpoint precisely where the damage occurred by triangulation. However, it is important to also consider some disadvantages associated with this methodology. The structure must be subjected to loads to generate acoustic emission, and the process is irreversible, known as the "Kaiser effect". In addition, monitoring is susceptible to noise interference, both electrical, mechanical, and environmental. The elastic waves used in the AE undergo attenuation, both of a geometric and structural nature, which may require the use of numerous sensors to ensure adequate coverage. Another limitation is the difficulty in sizing the damage detected. In

summary, while offering many advantages, monitoring through AE presents challenges and limitations to be carefully considered in implementation.

1.4.6. Post-processing methodology

Once the data is acquired through AE, a processing is necessary to extrapolate the useful information to the problem being studied. It is therefore important to identify only the AE of interest and derive from them the maximum amount of useful information on the current condition of damage or the behavior of the material and structure.

The first obstacle consists in separating (discriminating) the relevant signals (events) from the background noise, since continuously capturing data without a skimming, is not feasible. The signals of interest are generally distinguished by an amplitude significantly greater than the background noise and a short time (from the order of microseconds to milliseconds).

Consequently, the first step is to define a detection threshold to exclude background noise and allow the acquisition and analysis only of signals with amplitudes greater than this threshold. What falls below the threshold is not considered and will not be available for further processing. Despite this trick, the amount of data to be managed is very high. It is therefore necessary a post-acquisition processing that allows you to quickly and effectively manage the data that is obtained from the sensors.

Regardless of the approach chosen for data management, it is crucial to adopt an effective post-processing method to analyze the signals from the sensors and extract relevant information for monitoring. It is important to note that, in the case of AE, the amount of data to be dealt with is considerable (attributable to the concept of Big Data). Therefore, it is necessary to use a method that can quickly process all incoming data, enabling real-time monitoring of the specimen. A particularly promising choice in this context is Machine Learning (ML), a branch of artificial intelligence that employs algorithms to obtain information.

1.5. Machine Learning

Machine learning (ML) represents a category of artificial intelligence that enables computers to think and learn autonomously. The fundamental concept lies in guiding computers to modify their actions to improve accuracy and produce correct results [35].

1.5.1. Machine learning system

For machine learning methods to work, it is necessary to provide the learning algorithm with experience in the form of "training data," allowing the algorithm to learn from it. There are several strategies for the design of these learning algorithms, but first it is important to understand which is the working principle of each machine learning methods.

Figure 1.9: The pipeline of building a machine learning system, consisting of three major steps of data collection, feature generation, and model training[36]

In the initial phase, it is essential to gather a sufficient volume of training data that accurately represents prior experiences for computer learning. Ideally, this training data should be acquired under conditions mirroring those in which the system will ultimately be deployed.

Moving to the second stage, specific to the domain, procedures are typically applied to extract features from raw data. These features should be concise yet encompass the most crucial information within the raw data.

In the final stage, a learning algorithm is selected to construct mathematical models based on the extracted feature representations from the training data [36].

To gain a clearer understanding of the components of the model stage, let's break down the final step into four distinct stages [37] as shown in the figure :

- **Algorithm Selection**: Not every machine learning algorithm is universally suitable for all problems. Specific algorithms are better suited to types of problems. It is crucial to carefully choose the most appropriate machine learning algorithm for the given problem to achieve optimal results.
- **Model and Parameter Selection**: Most machine learning algorithms require some initial manual intervention to set the most appropriate values for various parameters. In the literature, there are also programs specifically designed to automatically determine the most relevant features, such as Principal Component Analysis

(PCA) and Linear Discriminant Analysis (LDA)[38] which are two commonly used techniques for data classification and dimensionality reduction.

- **Training**:Following the selection of the appropriate algorithm and suitable parameter values, the model undergoes training using a portion of the dataset as training data.
- **Performance Evaluation**: Before implementing the system in real-time, the model needs to be tested against unseen data to assess the extent of learning. This evaluation involves various performance parameters such as accuracy and precision.

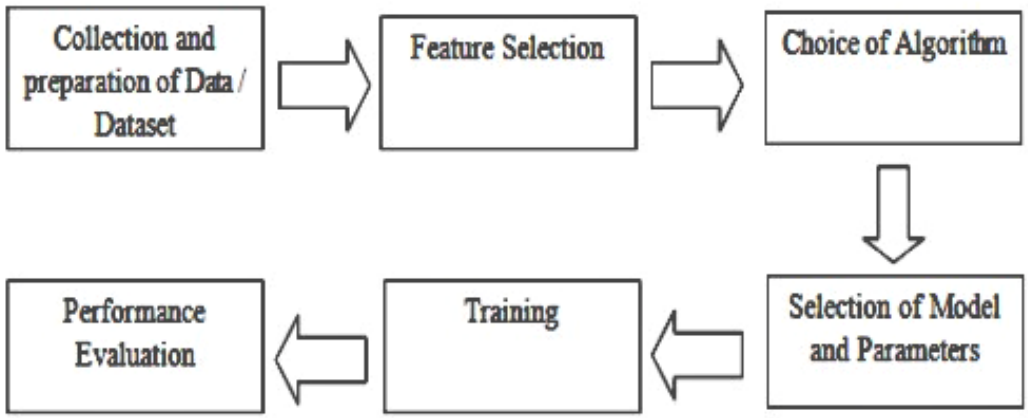


Figure 1.10: Components of a generic ML model[37]

As mentioned above, machine learning algorithms can be of different types. In this study, only supervised and unsupervised machine learning algorithms will be examined, depending on the nature of the input data.

2 | Clustering algorithms

Once we describe the basic concepts needed to understand the work done, we go into the details of the procedure to determine the clustering algorithm and its fundamental parameters.

In dealing with a Structural Health Monitoring (SHM) problem through Acoustic Emission (AE), it is crucial to establish an effective post-processing method. As explained in the previous chapter, Machine Learning algorithms prove to be the best choice to analyze the large amount of data quickly and effectively from the sensors applied on the monitored structure or specimen. However, there is a downside when it comes to Machine Learning algorithms: handling a large amount of data is not suitable for all algorithms.

The first part of the work was to figure out whether the algorithm to be applied to experimental data obtained from previous laboratory experiments was supervised or unsupervised. Having identified the type of algorithm, we proceeded to examine the many algorithms in the literature and figure out, among those present, the most promising one and verify that the chosen algorithm was also applicable to "big data" problems. In this chapter we will deal in detail with the whole procedure followed in choosing the DBSCAN algorithm.

2.0.1. Supervised learning

Observing the nature of the test data, it was decided to focus the work on unsupervised algorithms. In fact, the algorithms supervised, generate a function that maps the inputs to the desired outputs [39]. The working principle of supervised algorithms consist of two distinct phases: training and testing. During the training process, the learning algorithm or learner takes the samples from the training data as input, focusing on acquiring features and constructing the learning model. During the testing phase, the learning model employs the execution engine to generate predictions for the test or production data. Even more significant is the nature of the data derived from the acoustic emission of the specimen under analysis. These signals represent information that requires interpretation to understand what mechanism of fracture, damage, or otherwise generated the struc-

ture's response to stimulation. In this context, one knows the reaction of the system, but not the event that caused it; in other words, one is aware of the output, but the input remains unknown. This peculiarity of the problem excludes a priori all supervised machine learning algorithms, in which it is necessary to provide experimental data in which the nature of the input and its response is known to train the algorithm [36]. Using Supervised learning algorithms, we perform a classification of data.

2.0.2. Unsupervised learning

Unsupervised learning studies how systems can learn to represent input patterns in a way that reflects the statistical structure of the overall collection of input patterns. Unsupervised learning deals with data lacking predefined categories or labels. In our data, necessitating algorithms to independently establish criteria for grouping similar outputs. Determining output similarity when input labels are unknown, it means that we are applying in a correct manner the algorithm. To determine accurate labels (cluster assignments), clustering methods depend exclusively on the intrinsic characteristics of data points. These methods use these intrinsic characteristics to establish an empirical risk associated with a candidate hypothesis. Unsupervised learning is also called Clustering method[40].

There are many clustering algorithms in the literature, which can be divided into two main categories: hard clustering and soft clustering: hard clustering assigns each data point to a single cluster, while soft clustering methods assign each data point to multiple clusters with varying degrees of membership, in contrast to the exclusive assignment of the hard cluster. There are algorithms that can be applied in both contexts, depending on the specific implementations. Some examples include artificial neural network (ANN) algorithms such as the SOM (Self-Organizing Map), as well as hierarchical clustering methods. In addition, there are clustering approaches based on data density, including one of the best known: Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

We now review an overview of the main clustering algorithms in the literature, highlighting how they work and their associated advantages and disadvantages. This analysis was crucial in guiding our choice on which clustering method to focus the work.

2.1. Hard clustering

2.1.1. K-means

The k-means clustering algorithm is distinct as a powerful data mining algorithm widely adopted by the research community. In the algorithm, the process begins by selecting "k" cluster centers to align with "k" randomly chosen patterns or "k" randomly defined points within the volume containing the set of patterns. The strategic location of these centroids is crucial since different positions may produce different results. Therefore, the optimal strategy is to place them as far away from each other as possible. Next, each point in the dataset is associated with the nearest centroid. With the assignment of all the points, the initial phase ends, marking the completion of an initial grouping. The next crucial step is the recalculation of k new centroids. If a convergence criterion is not met, the algorithm will be repeated starting in reassign each pattern to the closest cluster center. Convergence is typically determined by reassignment of patterns to new cluster centers or minimal decrease in quadratic error. This process forms a cycle, and, through successive iterations, it becomes evident that the k centroids progressively change their positions until no more changes occur. Simply put, the centroids reach a steady state, stopping all further movement [41].

The k-means is a simple and easy algorithm to implement, and it is suitable for large data sets, however, despite its popularity, the algorithm has some limitations. These include the problems associated with random initialization of centroids, which often results in unpredictable convergence. In addition, the algorithm requires a priori definition of the number of clusters, contributing to variations in cluster shape and susceptibility to outlier effects. A key disadvantage of the k-means algorithm lies in its inability to effectively handle different types of data [42].

2.1.2. Spectral clustering

The family of spectral clustering algorithms have attracted considerable attention from the academic community in recent years. This growing interest is attributed to its sound theoretical foundation and the admirable clustering performance it demonstrates. Its implementation is simple, solvable efficiently with standard linear algebra software and often demonstrates good performance compared with traditional clustering algorithm [43].

Spectral clustering algorithms exploit the eigenvalues and eigenvectors of Laplacian matrices to divide a graph into clusters. In this partitioning, nodes within the same cluster show stronger connections to each other than nodes in different clusters. The spectral features embedded in Laplacian matrices capture essential information about the

connectivity and structure of the graph, making them valuable in tasks such as clustering and other graph theory-based analyses.

The spectral clustering can be modeled as a process involving three steps. Initially, a similarity graph is constructed for all data points. Next, the data points are placed in a space in which the clusters become more obvious using the eigenvectors of the Laplacian graph. Both steps involve significant computational costs in terms of time. In the case of a dataset involving "n" data points with "m" dimensions, the time complexities for the above steps are $O(n^{2m})$ and $O(n^3)$, respectively. Such computational requirements represent impractical, especially in the context of large-scale applications. Finally, a conventional clustering algorithm is employed to partition the embedding [44].

In cases where data points are observed sequentially, a method was proposed to directly update the clustering without the need to evaluate the entire affinity matrix. In addition, a more versatile algorithm has been developed over the years that allows the similarity between existing data points to be changed by incrementally updating the eigenvectors. This reduces the total computational time, although, compared to many other algorithms, it is very time-consuming [45].

2.2. Soft clustering

In many practical situations, some objects have characteristics that are intermediate between clusters, making clear assignment difficult. In such cases, using the classical (hard) approach to clustering leads to unrealistic assignment, forcing objects to belong exclusively to a single cluster. To overcome this drawback, the soft approach was developed. The fundamental idea is that each data point can belong to more than one cluster. One of the most important soft clustering approaches is known as fuzzy.

Fuzzy logic principle involves allocating data points to clusters based on a specified degree, referred to as a membership degree, ranging from 0 (complete non-membership) to 100 percent (complete membership). This membership degree, also known as a degree of sharing, is computed by evaluating the ratio of the dissimilarity between each object and the closest prototype to the sum of the dissimilarities between each object and all the prototypes [46].

2.2.1. Fuzzy c-means

This algorithm functions by assigning membership to each data point based on its distance from the cluster center. The closer the data point to the cluster center, the higher its membership towards that cluster center. It is evident that the sum of the membership values for each data point should equal one. As an unsupervised clustering algorithm, it enables the construction of a fuzzy partition from the data. The algorithm relies on a parameter, denoted as "m", which represents the degree of fuzziness in the solution. Larger values of "m" result in blurred classes, with elements tending to belong to all clusters. The optimization problem's solutions are influenced by the parameter "m", meaning that different choices of "m" typically lead to different partitions. We can say that by approaching the value of "m" to 0, the Fuzzy C-Means algorithm tends to resemble the K-Means hard clustering algorithm [47].

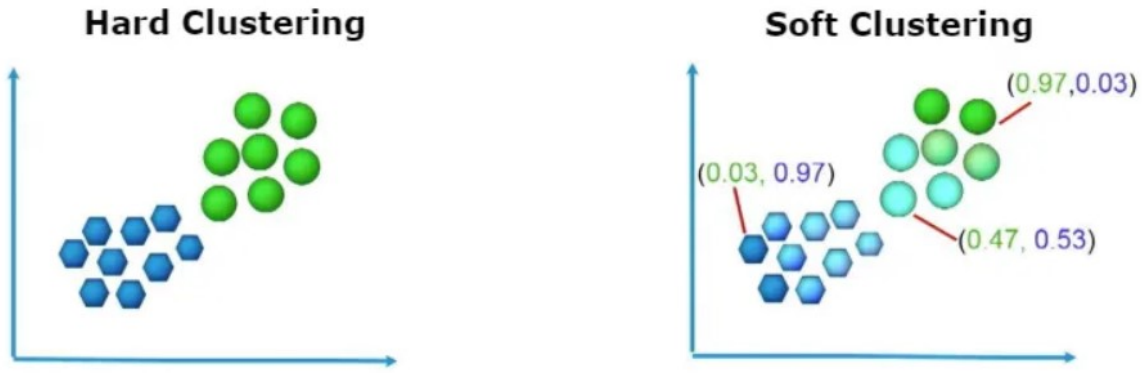


Figure 2.1: Difference betweeen K-means and Fuzzy c-means[48]

The method provides optimal results for data sets with overlaps, with relatively better performance than the K-Means algorithm. Unlike K-Means, in which a data point is forced to belong exclusively to a single cluster center, in this method a data point is granted membership in each cluster center, allowing it to potentially belong to more than one cluster center. However, there are drawbacks, such as the need to specify the number of clusters a priori. A lower value of "m" can improve the results, but at the cost of increased iterations. Euclidean distance measurements may assign unequal importance to underlying factors. In addition, the performance of the FCM algorithm depends on the initial selection of cluster centers and/or initial membership values [47].

2.3. Artifician Neural Network

Artificial intelligence (AI) is a branch of computer science that focuses on creating software capable of performing intelligent computations comparable to those performed by the human brain. It encompasses a range of methods, tools and systems designed to simulate human approaches to acquiring logical and inductive knowledge and solving problems through reasoning. Developments in AI fall into two main categories. The first includes methods and systems that mimic human experience by drawing conclusions from predefined rules, as is the case in expert systems. The second category includes systems that model how the brain works, such as artificial neural networks [49]. Artificial Neural Networks are particularly well suited to address and solve complex challenges involved in real-world scenarios. Inspired by the human brain, these networks use processing techniques to formulate algorithms that can model and understand intricate patterns and prediction problems.

2.3.1. SOM

There are many different types of ANNs, some of which are more popular than others. The self-organizing map (SOM) is an unsupervised ANN's algorithm used for data training to classify and effectively recognize patterns embedded in the input data space [50]. SOM is typically organized in a two-dimensional space; map units or neurons create a mapping from a high-dimensional space onto a plane. For each input vector, the crucial first step is to identify the winning neuron, the one that minimizes the distance between its prototype vector and the input vector. This mapped representation preserves the calculated relative distances between data points. Then, in the second step, a significant improvement is taken: the prototypes of not only the winning neuron, but also its neighbors, are updated. This dynamic process contributes to the self-organization of the map, refining the data representation in a consistent and meaningful way [51].

As a result, Self-Organizing Maps prove to be valuable tools for analyzing clusters within high-dimensional datasets. In addition, self-organizing maps possess the ability of generalization. During this process, the network can identify or characterize inputs that it did not encounter in its training data. The new inputs are assimilated by the map unit, enabling effective mapping and representation.

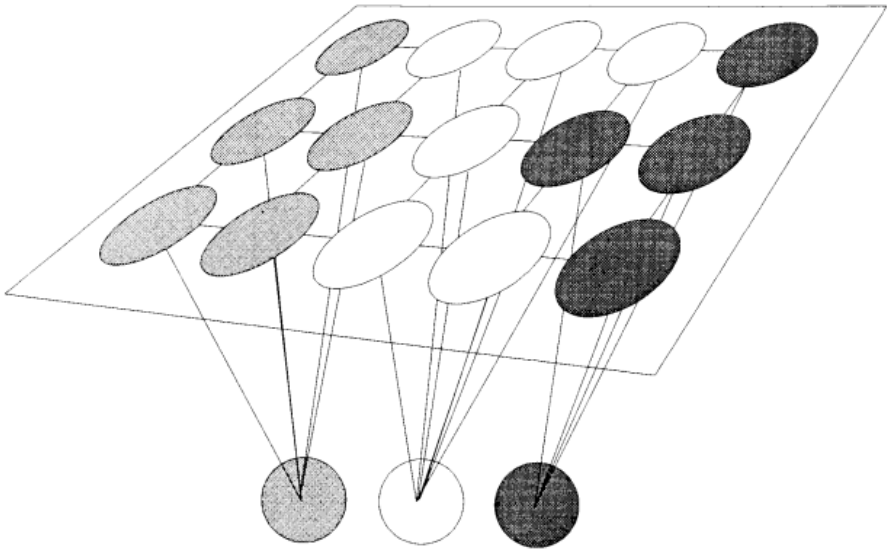


Figure 2.2: SOM graph[49]

However, the model cannot generate a model for the data, which leads to a lack of understanding of the data creation process. In addition, its performance is not optimal when handling categorical data and even more so with mixed data types. Moreover, the model preparation time is remarkably slow, which brings problems when training with gradually evolving data.

2.4. Hierarchical clustering

Hierarchical clustering is an additional approach for conducting exploratory data analysis, falling under the category of unsupervised techniques. Hierarchical clustering means creating a tree of clusters by iteratively grouping or separating data points. There are two types of hierarchical clustering:

- agglomerative clustering
- divisive clustering

Agglomerative clustering

Agglomerative clustering [52] is the process that involves the construction of a binary fusion tree, starting with the individual data elements. The progression involves merging pairs of the "closest" subsets stored at the nodes until the root is reached, which includes

all the elements of set X. The distance between any two subsets of X is denoted as Δ (Xi, Xj) and is called the linkage distance. This method is commonly known as agglomerative hierarchical clustering since it starts with single data elements (the xi) representing leaves of the tree and systematically joins subsets until the root is reached. The dendrogram is a diagram representing tree-based approach and they are used to visualize the relationship among clusters.

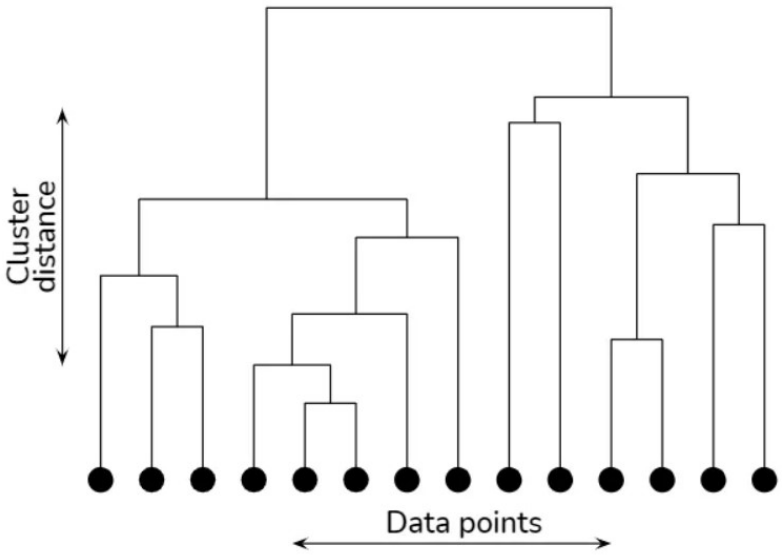


Figure 2.3: Dendrogram[53]

Going up the hierarchy reduces the number of clusters as more samples are agglomerated. At the end of the process, all samples are merged into one large cluster. Advantages of hierarchical clustering include the flexibility of not having to predefine the number of clusters, ease of implementation, and interpretability aided by dendrograms. In addition, it consistently produces the same clusters (unlike the k-means algorithm in which they may be different based on the initialization of centroids). Hierarchical clustering is relatively slower than other clustering methods, especially when dealing with large datasets, resulting in longer processing times [54].

Divisive clustering

Divisive clustering is the other macro category of hierarchical clustering algorithms. Their operation is the same as agglomerative methods, except that they work in reverse.

2.5. Density-based clusters

Density-based clusters can be displayed as sets of data points formed by cutting the probability density function to a specific density threshold. Each cut delineates distinct linked regions in the feature space where the probability density exceeds the threshold. Each of these regions corresponds to a cluster that includes all data points within that region. If the density threshold is set too low, separate clusters can join into a single cluster. On the contrary, selecting too high a density level can cause the loss of clusters with lower density [55].

2.5.1. DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) stands out as an important density-based clustering algorithm. DBSCAN operates without the need to specify the number of clusters as a parameter. Instead, it derives the number of clusters from the data itself, demonstrating the ability to discover clusters of arbitrary shape. To identify clusters, it is essential to configure two key parameters. As we will see later, the correct determination of these parameters is crucial to ensure that the algorithm works properly and achieves effective partitioning into clusters[56].

- ϵ: The radius that determines the neighborhood around a data point "P" and indicates the distance for neighborhood specification. Two points are considered neighbors if the distance between them is less than or equal to ϵ.
- minPts: The minimum required number of data points within a neighborhood to create a cluster [57].

Figure 2.4: Operational dataset[58]

Exploiting the provided parameters, DBSCAN classifies data points into three distinct categories:

- **Center points**: Center points serve as core elements for clusters, relying on the density approximation. A uniform ϵ is used to calculate the neighborhood of each point, which ensures a consistent neighborhood volume. However, the differentiating factor lies in the variable number of other points within each neighborhood. Central points are those data points that meet a minimum density requirement represented by minPts. Cluster construction focuses on these central points, hence the name "core." Adjusting the minPts parameter allows us to precisely adjust the required density for the cores of our clusters.
- **Border Points**: they are elements within our clusters that do not qualify as core points. All points within the neighborhood of point "P" (see figure above) directly reachable from "P" are considered border points.
- **Outliers**:these points do not qualify as core points and are also not near a cluster to be density-reachable from a core point. Outliers remain unassigned to any cluster and, depending on the context, may be deemed anomalous points.

Explained the main actors of algorithm, the fundamental concept of DBSCAN revolves around the notion that, for each object within a cluster, the neighborhood within a specified radius (Eps) must include a minimum number of objects (MinPts). This prerequisite suggests that the number of objects within the neighborhood must exceed a specific threshold (which is represented by the minimum number of points) [59].

Advantages and disadvantages of DBSCAN

Thanks to its operating principle, DBSCAN algorithm has distinctive and advanced capabilities that prove valuable for detecting objects, classes, patterns, or structures of various shapes and sizes. DBSCAN emerges as a strong competitor for discovering clusters and their spatial arrangement within the data space, particularly when these clusters exhibit comparable densities and no prior information about the groups present in a dataset is available [60]. Density-based clustering algorithms prove useful for discovering clusters within datasets characterized by arbitrary shapes and considerable size. These algorithms typically identify dense regions of points in the data space, distinguishing them from low-density regions [59]. However, the time complexity of the DBSCAN algorithm is $O(n^2)$. The running time of the algorithm is greatly affected by the process of identifying neighbors for each data point to determine the data density. As a result, for large data sets,

the DBSCAN algorithm involves substantial calculations, resulting in decreased clustering speed and increased execution time [58]. Several algorithms are proposed to reduce the time needed for the computation: in the article [57], it is explained a procedure to reduce it to O(n).

Despite the difficulties connected with the definition of the main parameters and the considerable execution time, DBSCAN is a promising prospect for unsupervised applications, constituting a prominent alternative to the most well-known clustering algorithms in the literature. Therefore, it was chosen as the subject of study for this thesis work. In the next chapter, we will detail the procedure followed to adapt the algorithm to the available data, using the MATLAB software. The main challenge lay in the efficient determination of eps and minPts parameters, essential to achieve effective clustering and results consistent with expectations.

In the table below, we summarize the principle of operation of every examined method of clustering, and we list the main advantages and disadvantages.

	How it works	Pros	Cons
K-means	A small number of k clusters is established, and subsequentiatly, each data point is allocated to the nearest centroid. After each assignment, the centroid are recalculated.	1) Easy implementation 2) Scalable to handle large datasets 3) Can dynamically re-evaluate centroid position 4) Easily adaptable to other problem domain	1) Manual selection of the number of the cluster (k) 2) Sensitivity to initial values 3)Challenging for clustering data with diverse sizes and densities

Spectral clustering	This approach relies on graph theory. Spectral clustering leverages information extracted from the eigenvalues of specific matrices constructed from either the graph or the dataset. By interpreting these matrices and eigenvalues, we can assign our data to clusters. This method offers a high degree of flexibility compared to other clustering techniques.	1) It doesn't adhere to fixed-shape clusters, eliminating the need to define a radius or a cluster centroid 2) This flexibility implies that the algorithm could be effective for datasets with varying shapes and size 3) It demostrates computational speed, especially for sparse datasets comprising several thousand data points	1) The number of clusters may need to be predermined before initiating the procedure 2) Computing large datasets can be costly due to the calculation of eigenvalues, eigenvectors, and subsequent clusterization
Fuzzy c-means	The operational concept resembles that of the K-means algorithm; however, the key distrinction lies in Fuzzy C-means, where the weighted distance between each cluster is assessed for every point. This evaluation assigns a specific percentage of membership to each cluster. While this method is essentially supevised, it frequently proves to be a viable alternative to the K-means approach.	1) Provides optimal outcomes for datasets with overlapping elements and performs relatively better than the K-means algorithm 2) In contrast to K-means, where a data point must exclusively belong to one cluster center, in this method, each data point is assigned membership to multiple clusters centers. Consequently a data point may belong to more than one cluster center	1) A-priori specification of the number of clusters is needed 2) Euclidea distance measures can unequally weight underlying factors 3) The performance of the FCM argorithm depents on the selection of the initial cluster center and/or the initial membership value

SOM	This technique is employed for dimentionality reduction in extensive datasets. Similar to clustering, it can examine substantial data valumes and generates maps within the input space, where the maps consist of data points that are closely positioned in the input space. Moreover, SOM has the capability to construct maps within the input space, even for input data it has not encountered previously.	1) Understanding and interpreting data is facilitated throught techniques such as dimetionality reduction and grid clustering 2) This approach can effectively address various classification problems, offering a valuable and intelligent summary of the data simultaneously	1) It doesn't generate a model for the data, resulting in a lack of understanding of how the data is generated 2) The model exhibits optimal performance when dealing with categorical data and performs even worse with mixed type of data 3) The preparation time for the model is notably slow, making it challenging to train against slowly evolving data

Hierarchical clustering	This approach involves the iterative cration of a cluster tree, either by progressively grouping smaller clusters into larger ones (agglomerative clustering) or by starting with larger cluster and deviding it into smaller ones (divisive clustering). The clustering process takes into account variuos concepts of similarity. The dendrogram provides a clearer visualization of the clustering	1) There is no need to pre-specify the number of cluster 2) It is straightforward to implement and interpret, aided by dendrogram 3) Consistently produces the same clusters, in contrast to methods like K-means clustering, where different cluster outcomes may arise based on the initialization of centroids	1) Hierarchical clustering exhibits prolonged execution times, particularly when handling large data 2) It may be susceptible to noise in the data

DBSCAN	This algotithm has the capability to determine the number of clusters based on the provided data. Unlike K-means, it identifies clusters of diverse shapes (typically, K-means identifies spherical clusters). In essence, the process involves defining a radius and a minimum point threshold required to establish a cluster. Based on the density distribution of points, three primary point classes can be delineated for each cluster: core points (those within the specified radius), border points (those outside the core but within the defined range), and outliers (points situated far beyond the core or border points)	1) There is no need to specify the numeber of clusters in advance 2) Excels in handling clusters with arbitrary shapes 3) DBSCAN exhibits robusteness to outliers and has the capability to detect them	1) Determining a suitable neighborhood distance (eps) can be challenging and may necessitate domain knowledge 2) DBSCAN is less adept at defining clusters when there are significant variations in-cluster densities. The characteristics of clusters are determined by the combination of eps-minPts parameters. Since a single eps-minPts combination is provided to the algorithm, it struggles to generalize effectively to clusters with markedly differnt densities 3) High computational time of the algorithm

3 | Methodology

3.1. Study case

The DBSCAN algorithm will be used to analyze data acquired by structural monitoring using acoustic emission from a Double Cantilever Beam (DCB) sample from a previous work[1].

Double Cantilever Beam (DCB) test is one of the most common tests used to evaluate the fracture toughness of composite or adhesive materials[61]. Procedures for performing these tests in mode I opening are provided by standard (ASTM D 3433-99[62] and ISO 25217:2009[63]).

A specimen for the Double Cantilever Beam (DCB) test comprises two beams with matching length and thickness. In the standard DCB setup, there exists a segment without adhesive recognized as the pre-crack (a_0). The "adherent thickness" (h) denotes the thickness of the bonding area, while the "adhesive thickness" (t) specifies the thickness of the adhesive applied[64].

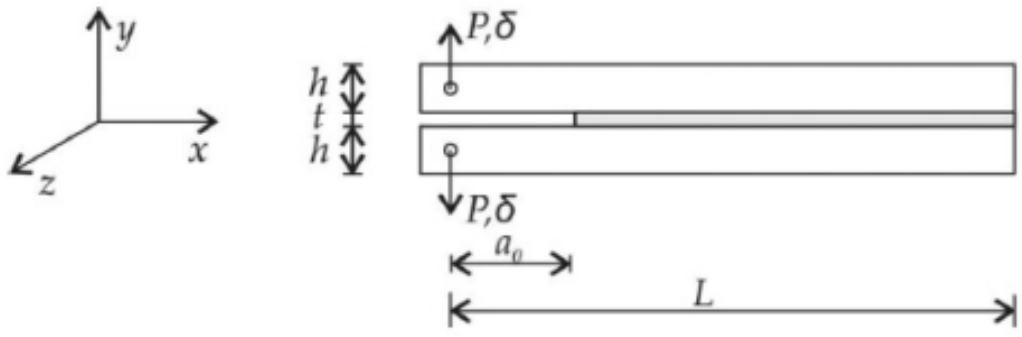


Figure 3.1: DCB specimen[61]

The test was conducted on two specimens both with a_0 pre-crack equal to 65 mm. In figure 3.2 are reported the specification of the specimen. In this thesis, it is used only one specimen in three different moments of the test and it is called "specimen S1".

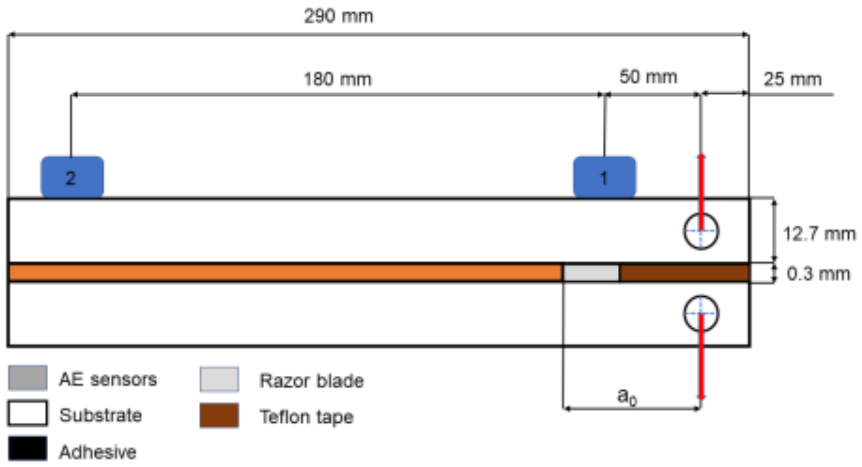


Figure 3.2: DCB dimensions [1]

A fatigue growth test in mode I was conducted using an MTS 810 servo-hydraulic testing machine, capable of handling a maximum load of 15 kN. For this experiment, a constant load test was selected with a fatigue ratio (R, defined as the minimum load divided by the maximum load) set to 0.1. The testing machine operated at a frequency of 5 Hz, applying a maximum load of 850 N. At intervals of 5000 cycles, the tests were temporarily arrested to execute a monotonic load ramp, gradually increasing the load to the maximum level experienced during the fatigue cycles. This ramping was done at a rate of 0.5 mm/min. Next, the peak load was maintained for a duration of 10 seconds to help measure crack length (in this work we are not interested in the length of the fracture). Thereafter, the machine was unloaded to reach the specified minimum load for fatigue testing, and the test cycles were resumed. The experiments continued until the specimen showed a complete crack.[1]. In this thesis, the terminology "cycle n" was adopted to denote a set of 5000 cycles. Therefore, "cycle 1" refers to the first set of 5000 cycles performed on the sample. Similarly, "cycle 9" refers to the ninth iteration in which the cyclic force was applied, i.e., data are analyzed after the sample has been subjected to between 40000 and 45000 cycles, and so on until the sample is completely broken. For detailed information on the sample used to obtain the data that were used in this thesis work and described above, including details on surface preparation, the type of adhesive used and its application, and the parameters of the test performed, see [1].

In order to achieve an understanding of the data from the specimen, data were represented graphically using a cumulative amplitude-time-energy plot. However, the scattering of the data makes it difficult to draw conclusions from this representation. For example considering the graph of specimen S1_cycle 1 (cycle between 0 and 5000 cycles shown in figure 3.3) a concentrated distribution between 25 and 50 dB is present and a more

marked dispersion of points above 50 dB is observed. Consequently, it is difficult to make a clear breakdown of the data.

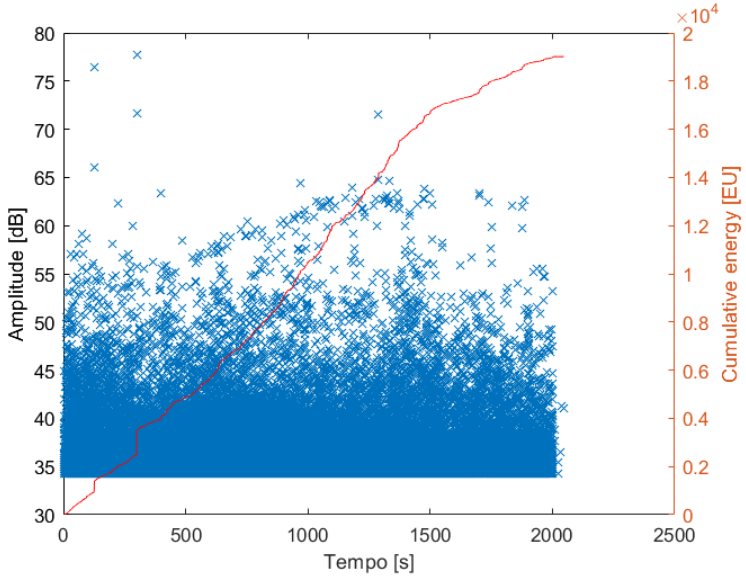


Figure 3.3: Amplitude - Time - Cumulative Energy, specimen S1_cycle 1

Similarly, considering the same specimen but different cycle (between 40000 and 45000 cycles, cycle 9) there is a large concentration of points with an amplitude between 40 and 50 dB and two distinct groups with amplitudes between 65 and 75 dB (figure 3.4), differentiated by the time of initiation of the phenomenon.

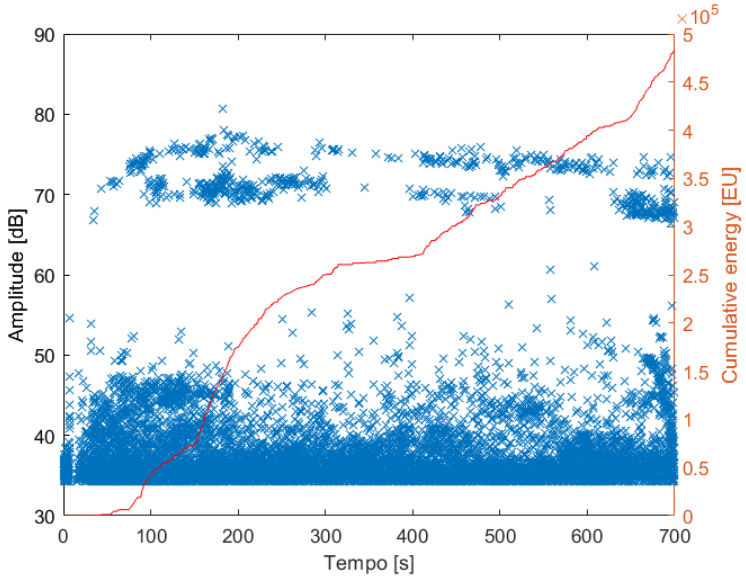


Figure 3.4: Amplitude - Time - Cumulative Energy, specimen S1_cycle 9

An analogous result is obtained by analyzing an ulterior amplitude-time-cumulative energy graph of a cycle (cycle 12) towards the end of the test 3.5.

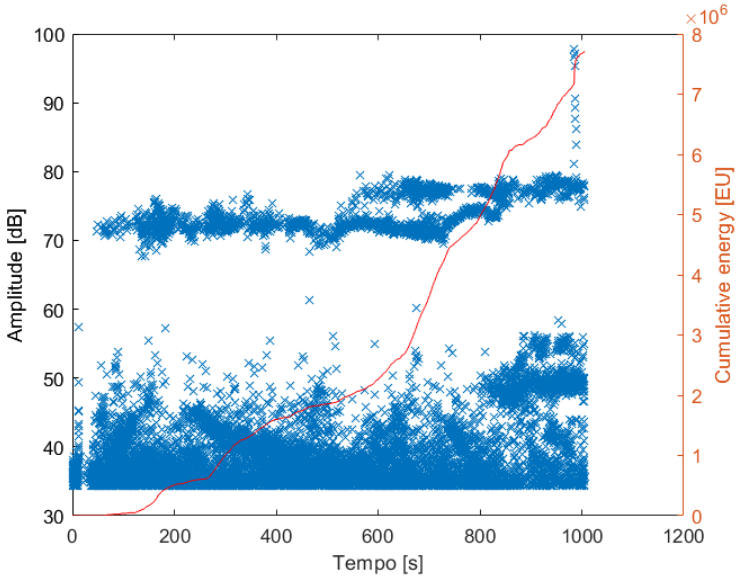


Figure 3.5: Amplitude - Time - Cumulative Energy, specimen S1_cycle 12

From the analysis of these graphs, a proper discrimination of the phenomena and their nature cannot be obtained. Therefore, a more in-depth analysis is needed, using a proper clusterization post-process method.

3.2. DBSCAN choice parameters

Because of the large amount of data acquired throught acoustic emission from the tests performed on the sample, the need emerges for a fast and efficient approach to analyze it. The information acquired in a single run (cycle n, consisting of 5,000 cycles) is considerable. DBSCAN-based approaches in the literature usually deal with limited amounts of data. Therefore, it was essential to develop a rigorous method that would allow the algorithm to be adapted to our data specifications, focusing in particular on determining the eps and minPts parameters. Infact, in the DBSCAN algorithm, the "eps" (radius of the cluster) and "minPts" (minimum number of points required to form the cluster) are established based on experiential knowledge of the operator and are subsequently refined in response to clustering results until satisfactory results are obtained[65]. Through this method, however, we proceed by "trial and error," without following a rigorous approach, but rather focusing on the operator's experience. Therefore, a more methodical approach, based on solid scientific foundations, was researched to determine the two basic parame-

ters of DBSCAN. The goal is to obtain a standardized clustering that can fit any type of data.

As explained in the chapter 2, to guarantee the proper functioning of DBSCAN, it is essential to establish two key parameters to identify clusters:

- eps
- minPts

Starting from a specific point in the dataset, the surroundings of that point are considered, defining the limit of the cluster radius as the value of eps. Within this "circle" with center at the considered point, for the set of points to be recognized as a cluster, the predefined minimum number of points must be reached. If this is not done, the cluster is not created. Of course, the algorithm does not stop but continues by considering a second point of the dataset and always checking that the two necessary conditions for the cluster to be formed are respected. The algorithm will stop until all the points in the dataset have been checked. In the figure 3.6 is shown the totatality of the DBSCAN algorithm in a schematic way.

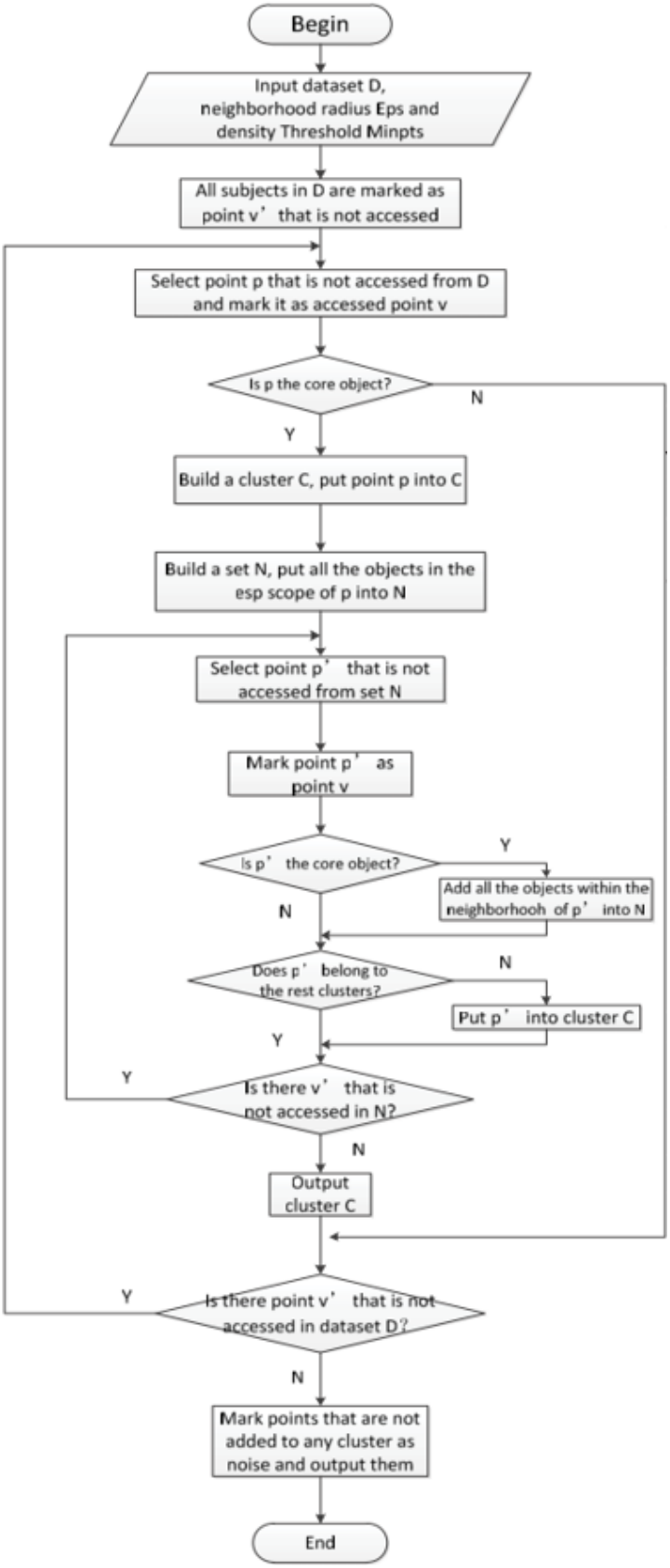


Figure 3.6: Functional blocks of DBSCAN algorithm[65]

The choice of both eps and minpts is critical to prevent distortions in clustering the dataset. Even small variations in the values of eps and minpts can generate very different clustering results. Determining these parameters in advance is one of the main challenge of DBSCAN.

In this chapter, we will review methods proposed in literature for determining these two crucial parameters. It will be elaborate on how, based on already established approaches, a specific method was developed to determine eps and minPts for the data used in this work.

3.2.1. Tournament Selection for detection Eps and minPts

Tournament selection (TS) is a widely used strategy in evolutionary algorithms, used to choose the eps value from a population of stored eps values based on the fitness function (purity)[66]. The value that emerges as the winner is the one with the highest percentage of the purity function. The winner of each tournament is then chosen to run the DBSCAN algorithm on the generated population (minPts). Based on its specification, it generates a more diverse set of eps values until an optimal combination of minPts and eps values is identified[66].

Since the eps parameter can significantly affect the efficiency of the DBSCAN algorithm[67], a combination of an analytical approach[68] for estimating eps and the Tournament Selection (TS) method is employed[69]. During each iteration, the eps parameter is calculated and compared with the stored eps values from previous iterations. Initially, the initial eps value is assigned a high probability by default to increase the probability of being selected in the tournament compared to other values. A scheme of Tournament Selection (TS) in shown in figure 3.7.

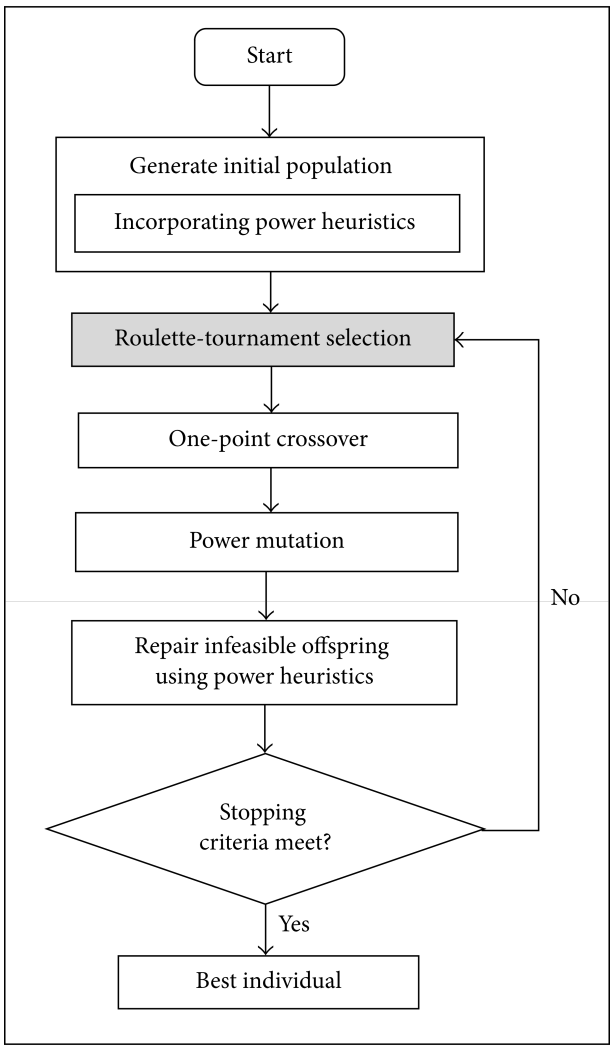


Figure 3.7: Tournament selection graph[70]

The purity metric[66], which serves as both a measure of cluster quality and a fitness function, measures the excellence of the clusters formed. Purity reflects the frequency of the most common category or class within each cluster, with higher purity values near 100 percent indicating more desirable clustering results.

Another important parameter to consider is time complexity[66], that is the processing time of the tournament selection (TS) algorithm for the parameters. The time required to run the algorithm can be evaluated as O(Iterations * Population). As the number of iterations increases, the accuracy of the results also increases. The number of iterations is constrained by the assigned level of purity; until that level is reached, the tournaments continue. The size of the population chosen for the tournament affects the overall duration, as a larger population takes longer to run all the tournaments. This suggests

that the additional time complexity to the DBSCAN algorithm is justifiable in order to simultaneously identify the most suitable minPts and eps values.

However, although promising, it has a computational time directly proportional to the size of the dataset, which in the case study presented in the section 3.1 (but in general for data from AE), is very high because acquisitions were taken until the specimen was broken. Therefore, although accurate for the definition of minPts and eps as shown by the experiment conducted in the article[66], it results in a significant computational slowdown. Consequently, an alternative approach, equally valid but characterized by a significant improvement in reduction of computational time, was adopted.

3.2.2. Procedure for determining the eps and minPts

A three-step process was followed to establish the basic DBSCAN parameters essential for the algorithm to function properly. Each of these steps produced results that helped to enhance the decision-making process in determining the parameters, generating a gradual improvement in the clustering of the data. The validity of these improvements was confirmed through a verification involving the waveforms associated with each cluster.

Choice of features with PCA

The first essential step involved defining the features to apply to the DBSCAN. These features represent the most relevant elements in the acquired signal. In our case, they encapsulate the vital information extractable from the acquired waveform shapes:

- Time-domain: Amplitude, Duration, Energy, Rise Time, Counts
- Frequency-domain: Peak frequency and Centroid frequency

However, applying a clustering algorithm to all features would not only be extremely complex but also time-consuming. For this reason, it is essential to reduce the information set in order to optimize the operation of the algorithm. Feature selection proved crucial not only to conduct a detailed analysis of the algorithm itself, but also to ensure the consistency of the features used in the clustering of the data, which was carried out by other clustering algorithms. This allowed for a meaningful comparison of the results obtained from different methodologies. To accomplish this, we relied on the PCA from previous work[1]. In this work, energy and duration were identified as relevant features in the time domain, and peak frequency in the frequency domain. Therefore, the focus was on using only two of the three features considered relevant: duration and energy.

As previously mentioned, Principal Component Analysis (PCA) was employed for the selection of these features, which is useful in revealing relationships between variables and between samples, such as in the case of cluster formation.[71]. PCA stands as a widely recognized method for reducing dimensionality, with applications spanning various domains such as data compression, image processing, visualization, exploratory data analysis, pattern recognition, and time series prediction[72].

Method 1

Having defined the features for applying DBSCAN, the next step was to choose the two fundamental parameters, eps and minPts, to ensure the proper functioning of the algorithm. A decision was made to express eps as a function of the assigned number of points and determine its value accordingly. This approach reduced the number of parameters to define from two to one. The method employed is known as the "line method"[73].

To apply the "line method," it is necessary to express the two related parameters by a function. The used method of calculating this parameter is based on a function that calculates the distance between each element of a data set and its k-th neighbor. This function is often denoted k-dist, and its parameter k, in our case study, is equal to minPts. The k-nn function evaluates the class or value of a new observation by comparing it with the k closest observations in the training dataset. In the absence of prior knowledge k-nn can be quantified using various metrics, such as Euclidean distance. The class or value of the new observation is then determined by the most prevalent class or value among the k neighbors. The choice of the value of k is a crucial parameter: higher values of k lead to more robust decisions but may be less responsive to local variations, while lower values make the model more responsive to local variations but potentially more susceptible to noise in the data. In summary, the k-th neighbor function is a machine learning approach that exploits proximity in feature space to make predictions on new data[74].

Using the k-th neighbor function, we can express eps as a function of the minimum number of points as shown in figure 3.8. In addition, there is an interval of points known as "knee" which is characterized by significant variation in distances. The challenge is to accurately determine the starting point of the knee, which serves as a crucial factor in identifying abrupt changes in distances (at the practical level abrupt changes in the slope of the curve)[75] and then define the eps parameter for the DBSCAN algorithm. Typically, the first sharp increase in distances occurs towards the beginning of the knee as shown in detail in the figure 3.9.

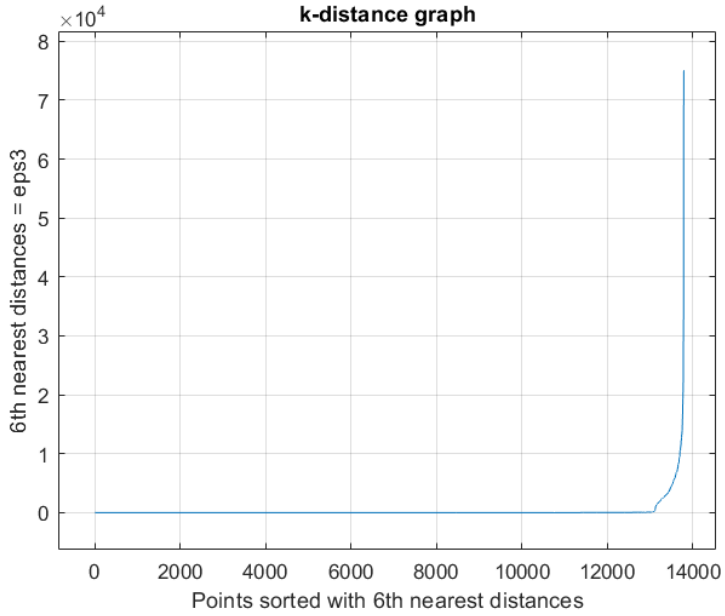


Figure 3.8: k-distance graph of specimen S1_cycle 9, with minPts = 6

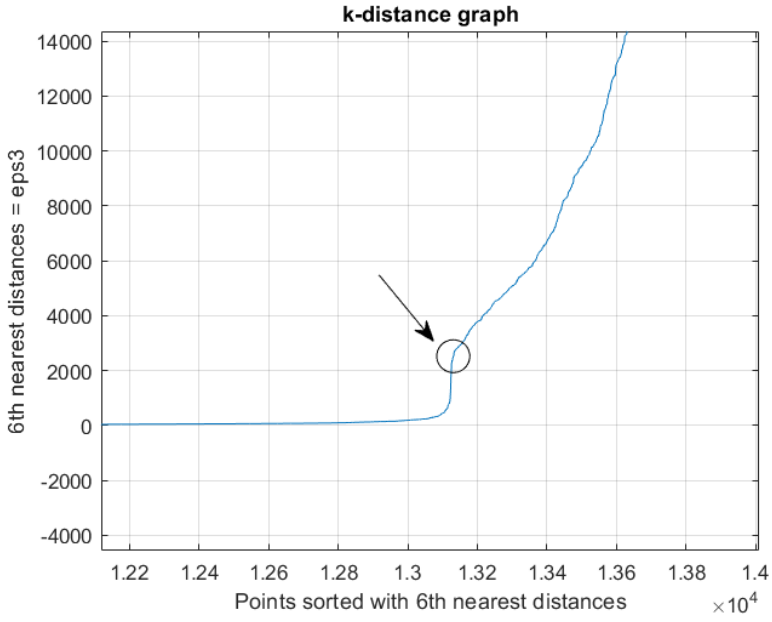


Figure 3.9: k-distance graph of specimen S1_cycle 9 with minPts = 6: highlight of the abrupt change in slope of the curve

The knee size is influenced by the density of clusters and the type of k chosen. Determining the knee point correctly is particularly complex, and variations in the width and slope of the knee depend strongly on the choice of "k" and thus affect its shape.

To handle this issue, a script was developed in MATLAB in order to identify the points at which a change in slope occurs (figure 3.10).

Once a point had been identified as acceptable, it was defined as the first point to determine the coordinates of point one (P_1). Starting from the knee initial position (P_1), identified by considering the intersection of the red line and the graph of K-dist as shown in figure 3.10.

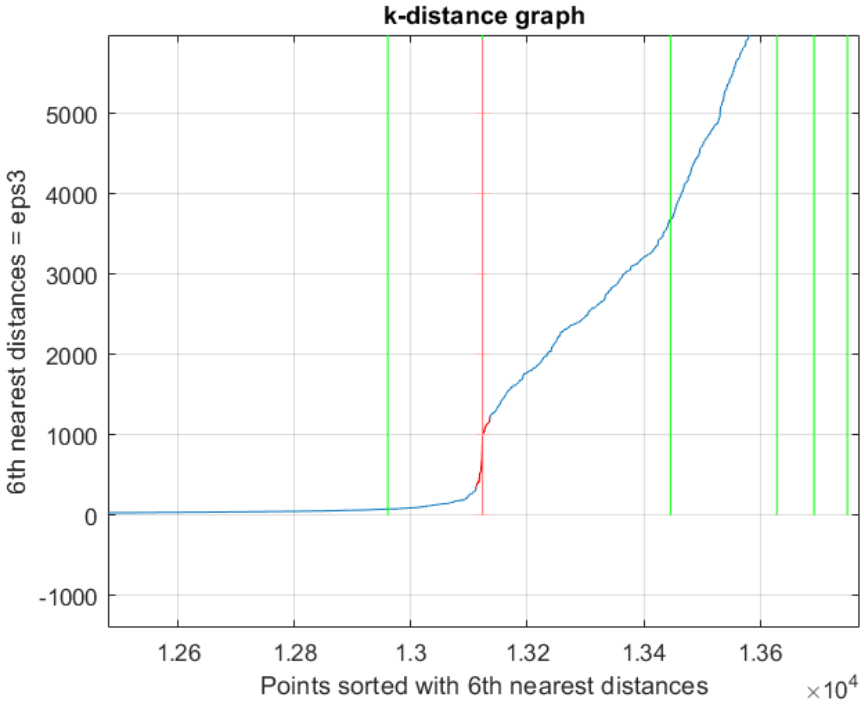


Figure 3.10: K-distance graph of specimen S1_cycle 9, with minPts = 6, with detection of abrupt change of slope

Then, it was essential to identify a second point, designated as point P_2 as shown in the figure 3.11, represented by the maximum limit of the k-dist curve.

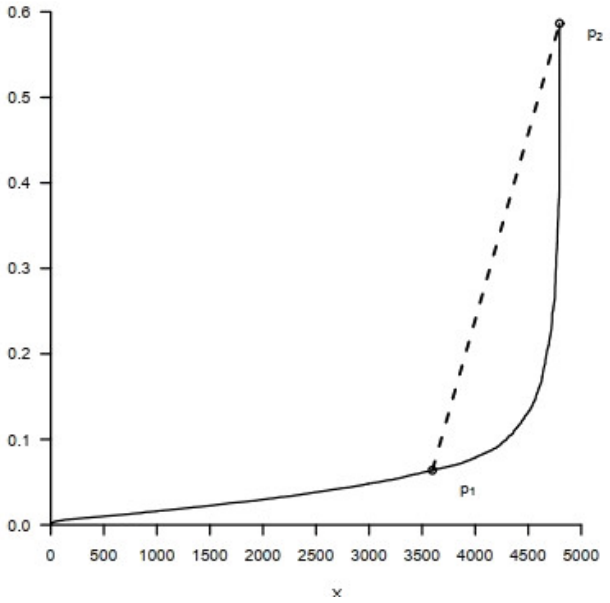


Figure 3.11: Line method[73]

Afterwards, by performing a symmetry passing through the middle of the line and parallel to the y-axis, the corresponding line was generated as shown in figure 3.12. The intersection of this straight line and the k-dist curve made it possible to locate P_3. It can be observed that the straight line determines the point $P_3(x_3, y_3)$, located at the top of the knee. The y-coordinate of point P_3 then defines the value of $y_3 = \text{eps}$ [73].

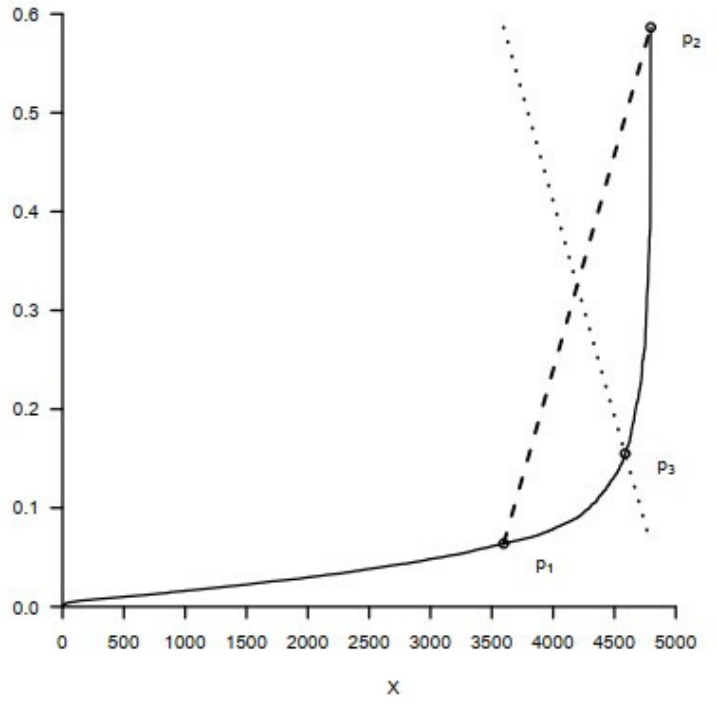


Figure 3.12: Line interpolation for detection of eps[73]

However, the choice of minPts parameter remains crucial. Indeed, analyzing articles in the literature[73, 75, 76], it was chosen to select the value of k (i.e., minPts) as 4, 5 or 6, depending on the amount of data to be processed (usually dataset considered are about 1000 data), or using formulas based on feature size to conduct clustering (remember that for each value of "k", the K-dist graph will change)[77]. Since the data used for this analysis present a considerable number of points for the type of structural monitoring employed, the decision was to opt for the k=6 value (in our work the mean dataset is about 10000-15000 data).

In developing the methodology, the focus was on cycle 9 of specimen S1. This choice was motivated by the fact that the volume of data of specimen S1 in each run was of the order of magnitude of 10^4, lower than, for example, specimen S2 (of the order of magnitude of 10^5). This selection made it possible to reduce the computational time of the algorithm in the Matlab script. It should be noted that this methodology could have been applied to any sample, however, this selection was motivated also by the fact that, visually, the division of specimen S1_cycle 9 into clusters was quite evident. In fact, three broad aggregations of clearly distinguishable points can be observed in figure 3.13, which are likely to correspond to distinct phenomena with differentiated physical characteristics,

thus likely to be subdivided into separate clusters. Of course, the validity of such claims will have to be supported by thorough scientific analysis.

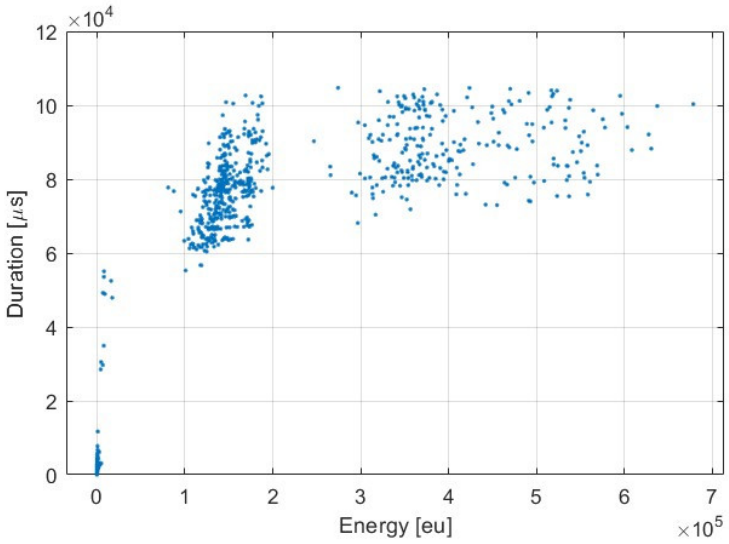


Figure 3.13: Duration - Energy, distribution of dataset without clustering, specimen S1_cycle 9

Consequently, it was possible, upon initial analysis, to assess the correctness of the methodology for the determination of eps and minPts (figure 3.14).

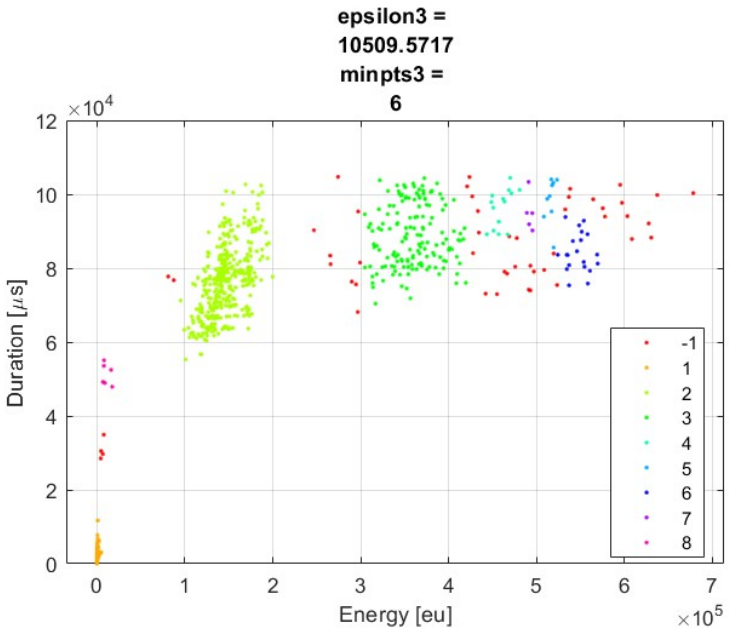


Figure 3.14: DBSCAN: Duration - Energy, specimen S1_cycle 9, method 1

Initially, the subdivision seems to be effective. Indeed, looking at the above figure 3.14 from left to right, two well-defined clusters clearly emerge. The first cluster, represented by the orange dots, is distinguished by significantly lower duration and energy than the other two clusters of data. On the other hand, the second cluster, which is clearly defined and colored light green, shows remarkable compactness and density; as we will see in the following sections, this cluster maintains a uniform coherence in all the tests that will be performed. However, in the aggregation of dots on the right, six distinct clusters are identified, despite the fact that the dots appear similar in energy and duration. Another group is designated with the number -1. According to the theory of the algorithm, the points identified in this group should be considered outliers, that are points that do not belong to any cluster and they are interpreted as "noise." However, it is essential to conduct a more in-depth evaluation to confirm the validity of this assumption. Therefore, the need was felt to further investigate the methodology for assigning the minimum number of points, as the one just described did not seem consistent with the nature of the data under consideration.

Method 2

Since the cluster diameter was determined based on minPts, it was essential to develop a rigorous approach to assign only the minimum number of points. The reasons for this are clear; grouping partitioning did not meet expectations and, in addition, the examples cited in the literature were based on datasets with limited size (not referable to big data)[68],for datasets containing a limited number of objects, it is suggested the following heuristic formula 3.1 (m = object of dataset):

$$minPts = interger(m/25) \tag{3.1}$$

As an example, for a dataset with m = 50 objects, set minPts=2, for m = 100, set minpts=4, and so on, according to the formula 3.1. For datasets with a large number of objects, is recommended using minPts set to 20.

Then, as suggested by the article[68], the K-dist diagram was plotted based on a minimum number of points for large dataset (= 20). Next, the beginning of the knee was identified to determine the point P_1 and consequently the point P_2. Finally, through symmetry with respect to the axis passing through the centerline of the line just created, the point P_3 could be determined. The procedure to determine the eps is explained in detail in 3.2.2.

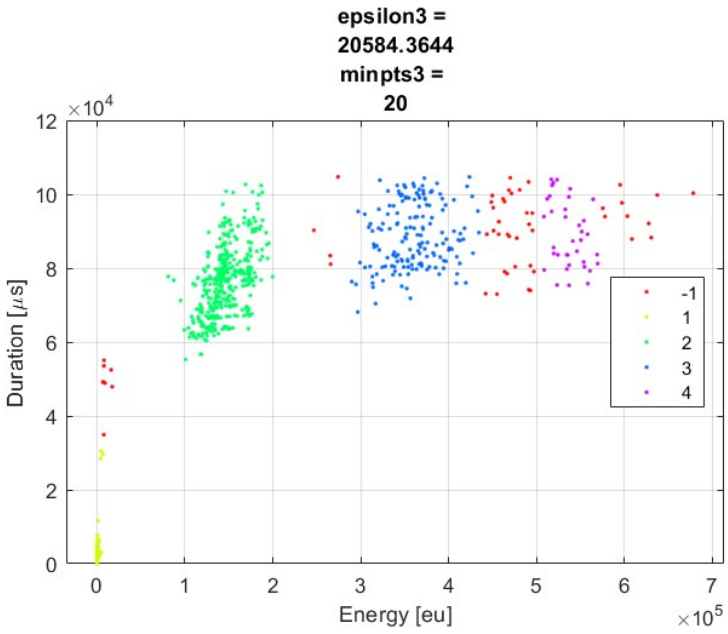


Figure 3.15: DBSCAN: Duration - Energy, specimen S1_cycle 9, method 2

Cluster partitioning occurs as expected for the yellow zone (cluster 1) and the green zone (cluster 2) looking at figure 3.15. However, the scatter of points containing two clusters and points considered as noise raises questions about the correctness of the algorithm. We expected the point cloud in the upper right corner of the graph to represent a single cluster, indicating that these points should have similar feature values. Therefore, it was necessary to develop an improved methodology for choosing minPts to obtain the expected subdivision of clusters.

Method 3

To increase the accuracy of the minPts number value, it was necessary to take a step back.

In the article[73], a formula is provided based on the size of the clustered vector and a ratio of distances to estimate the minimum number of points represented by the formula 3.2:

$$minPts = \begin{cases} round(d_p + 0.5) & for \quad dim(X) == 2 \\ round(d_p - 0.5) & for \quad dim(X) > 2 \end{cases} \tag{3.2}$$

The methodology is based on considering the size of the feature vector used for clustering by evaluating the minimum number of points in two separate ways. If the dimension

is equal to two (the minimum required for the algorithm's applicability), the minimum value of points is defined by adding 0.5 to d_p. In case the dimension exceeds the value of two, the formula calculates its value reduced by 0.5. In the case studied, having applied DBSCAN to two features, the first formula was applied.

Therefore, it was necessary to define the value of dp represented by the formula 3.3:

$$d_p = \frac{d(P_2, P_3)}{d(P_1, P_3)} \tag{3.3}$$

This value represents the ratio between two distances, specifically between the distance between point P_2 and P_3 and the distance between point 1 and 3. Through this approach, it was possible to rigorously estimate the minimum number of points. To determine the points, an initial minpts of 20 is set, applying the concept explained in previous paragraph for large datasets. This provided a baseline for the k-dist graph, allowing P_1 to be defined and then the other two points to be defined by evaluating the value of eps for large datasets and then redefining the minimum value of points for clustering.

The results obtained by applying method 3 are in accordance with the expected cluster partitioning, as highlighted in figure 3.16

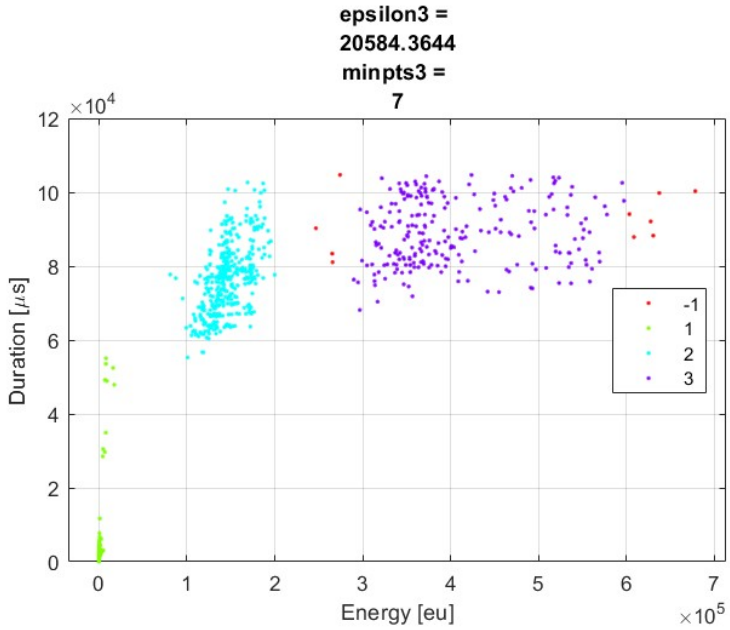


Figure 3.16: DBSCAN: Duration - Energy, specimen S1_cycle 9

There are actually three clusters, each corresponding to one of three clear aggregations of points, which reflect similar characteristics considering the two features used for DBSCAN.

In conclusion, we can say that the three methods differ in a few steps, however, these steps are crucial to the success of the clustering process. Specifically, in Method 1, in which a minimum number of points of 6 is established (as suggested in the literature), using the k-dist function we are able to represent the eps distance as a function of the minimum number of points. At the beginning of the knee of the k-dist graph, indicated by the sharp change in slope of the curve, the points P_1 and then P_2 are defined. Then, by symmetry with respect to an axis parallel to the y-axis and passing through half of the line just created, a symmetrical line is defined that intersects the graph at a point P_3.The y-coordinate of point P_3 delineates the value of eps. The procedure for determining the value of eps by the method of straight lines will remain unchanged for the other methods as well, since we chose to focus on determining a single parameter, namely expressing eps as a function of minPts. However, the results obtained with this approach did not allow satisfactory partitioning into clusters (with very cohesive clusters), and due to the presence of a large number of outliers, it was decided to implement an alternative method.In the article [68], a minimum number of points, set at 20, is provided to handle large datasets. The procedure of method 2 follows the same working logic as method 1, the only difference being that the initial number of points is set at 20. Again, the clustering phase did not produce the visually expected results. For this reason, an additional modification to the method was adopted. Specifically, an additional step was added to method 2: once the eps value is evaluated, the minimum number of points is recalculated using the formula given in the article[73]. The method is based on the ratio of the distances between P_2 and P_3 over P_1 and P_3 points. Based on the results obtained through this methodology to determine the crucial parameters for the proper execution of the clustering step, we can say that method 3 emerges as the most effective and promising approach for clustering large datasets. Infact, adopting method 3, the number of outliers is greatly reduced, as user intervention is minimized and the determination of parameters through the trial-and-error method is avoided.

4 | Results

4.1. Clustering result

Having obtained a result congruent with expectations for cycle 9 (figure 3.16) considering method 3 for determination of eps and minPts parameters explained in the section 3.2.2 of the chapter 3, the same procedure was extended to two other cycles in the same sample to substantiate the hypothesis that the proposed method was adaptable and worked not only for the cycle for which it was originally developed, but also for generic cycles. Accordingly, cycle 1 and cycle 12 of specimen S1 were selected, and application of the proposed procedure to determine eps and minPts produced results that appear consistent with a significant division into clusters.

To ensure the validity of the results derived from the clustering procedure shown in the previous chapter, it is important that clusters should be considered valid if each point in the cluster has similar characteristics.

In our case, the information related to each point is contained in the waveforms associated with the points. To verify this, it was necessary to establish that the waveforms associated with each represent phenomena are intrinsically similar. To test this hypothesis, the waveforms associated with each cluster subdivision were represented graphically. This approach led to interesting results that support the idea that the clusters identified from DBSCAN are not simply the result of random groupings, but rather are indicative of distinct phenomenological modes within the data considered.

Using a script in Matlab, the waveforms related to each cluster were displayed after the clustering operation performed by the DBSCAN algorithm, using the duration and energy features.

Let's start analyzing, at first, cycle 1 of specimen considered for this work, mention as Specimen S1.

4.1.1. Cycle 1

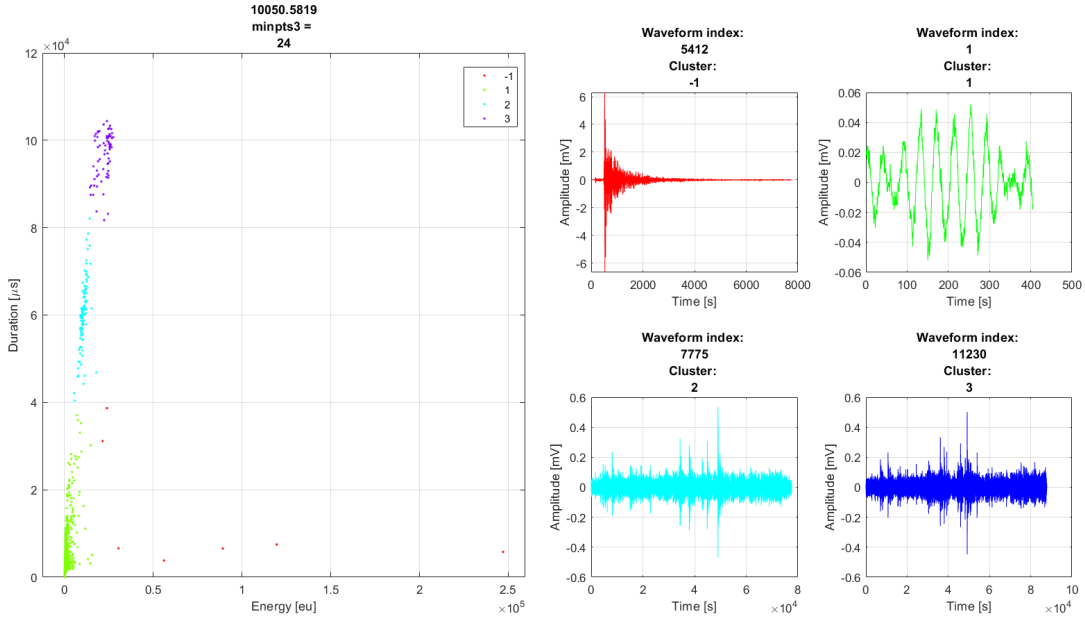


Figure 4.1: DBSCAN: Duration - Energy of specimen S1_cycle 1, waveforms of each cluster

The first cycle examined shows a subdivision into only four groups: the first cluster (green dots) shows phenomena with a duration of less than 3×10^4 μs, while the second cluster (light blue dots) is distinguished by a duration of more than 4×10^4 μs but lower than 8×10^4 and a third group with a duration above 8×10^4(figure 4.1). However, the associated energy is remarkably similar. This could be attributed to the fact that, being at the beginning of the test, no particularly significant phenomena emerge within the bonded joint. As a result, as expected, the waveforms are similar in terms of amplitude but they are different in terms of duration, as shown in the figures 4.1. Due to their shape, can be suppose that they are continuos wave associated with noise or some friction during the tests. However, several amplitude peaks are evident in cluster 2 (light blue waveform) and cluster 3 (blue wavelenght) that could indicate the presence of phenomena unrelated to simple noise or friction during the test. For a more thorough assessment of the nature of this cluster, a more accurate acquisition may need to be performed applying a higher threshold level to eliminate excess noise present in the acquisition. However, if the acquisition has already been completed, it may be of interest to apply a filter to reduce the noise present in the acquired signal, thus improving the ability to detect phenomenology.

Another group is identified during clusterization. Infact, some points are considered as "noise". They belong to group named as -1, and their waveforms are represented by color red in the figure 4.1. To understand the nature of the group, it is necessary to consider how the cluster works. When a point does not meet the conditions necessary to be part of a cluster, that is, if it does not fall within the radius of a point or, if it does fall, if it is not part of a group large enough relative to its minimum value, it is excluded from clustering and considered an outlier. However, being classified as an outlier does not necessarily imply that it is noise. As can be seen from the clusteritation graph in figure 4.1 the energy related to this points (red dots) is considerably high compared to clusters 1, 2 and 3. The wave shape looks like a transient wave, generally associated to events. They could result from events generated by internal phenomena to the adhesive bonded joint or simply represent disturbances detected by the sensors. Therefore, a more detailed analysis should be conducted in order to understand their nature by analysing its main characteristic such as frequency, period, phase, crest factor, RMS and so on.

The characteristics of waveforms can be useful in identifying the properties of a given phenomenon, but they are not sufficient to define its type. To associate similar waveforms with similar characteristics, it is essential to acquire controlled signals from the same model used for testing, thus creating a reference for comparison. This approach allows the construction of a defect history in which the causes generating the specific waveform are known. Obtaining a catalog of signals representative of noise, damage or defects allows them to be used as templates to identify defect phenomenology after clustering.

Considering the clustering based on the characteristics of duration and energy, graphs related to other waveform quantities were subsequently generated.

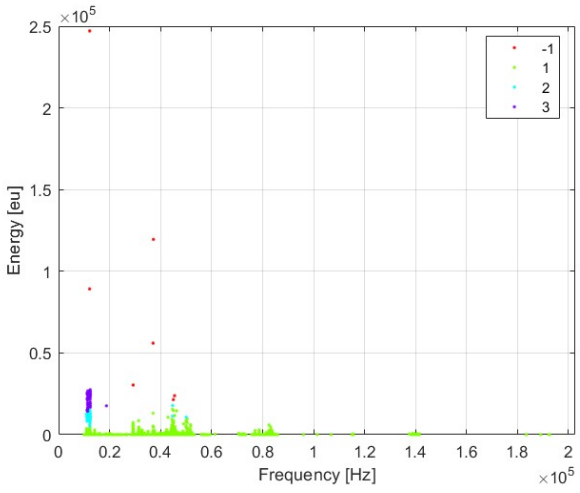


Figure 4.2: DBSCAN: Duration - Energy, specimen S1 cycle 1, energy-frequency

As an example, considering the graph 4.2 , in the light green cluster (cluster 1) the frequency associated with the waveforms is heterogeneous, but a very low energy distribution is observed, suggesting the possibility of noise. In contrast, cluster 2 (light blue) and cluster 3 (blue) have, for all points, a very similar frequency associated with a limited frequency range. Therefore, it can be assumed that this are event that do not just represent noise, but could indicate the presence of similar phenomenon characterize by a frequency of 0.1×10^5 Hz but, according to clusters, with different energy. This assumption is supported by the waveforms previously shown (figure 4.1), where peaks were present in clusters.

Recalling the graph cited in chapter 3 and shown in figure 3.3, a comparison with amplitude-time graph after clusterization is done (figure 4.3).

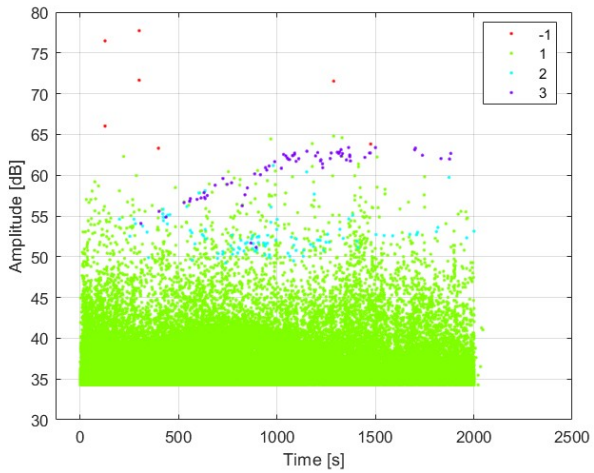


Figure 4.3: Amplitude - time, specimen S1_cycle 1

It is evident that a clearer subdivision of the points is possible due to clustering based on the characteristics of duration and energy, so we can clearly distinguish points of different colors. This is not possibile without clusterization. So, as described earlier, the clustering reflects the waveforms taken as examples. In fact, the average amplitude of the waveforms in cluster one is generally lower than the shapes associated with cluster 2 (light blue) and cluster 3 (blue). This could be attributable to the fact that the waveforms in cluster 2 and 3 have different peaks that contribute to their average value. However, since this is the initial cycle of the test, it is possible that there is a high presence of noise, making the results less reliable. Therefore, it was essential to examine additional cycles to evaluate the effects of method 3 for clustering more completly.

4.1.2. Cycle 9

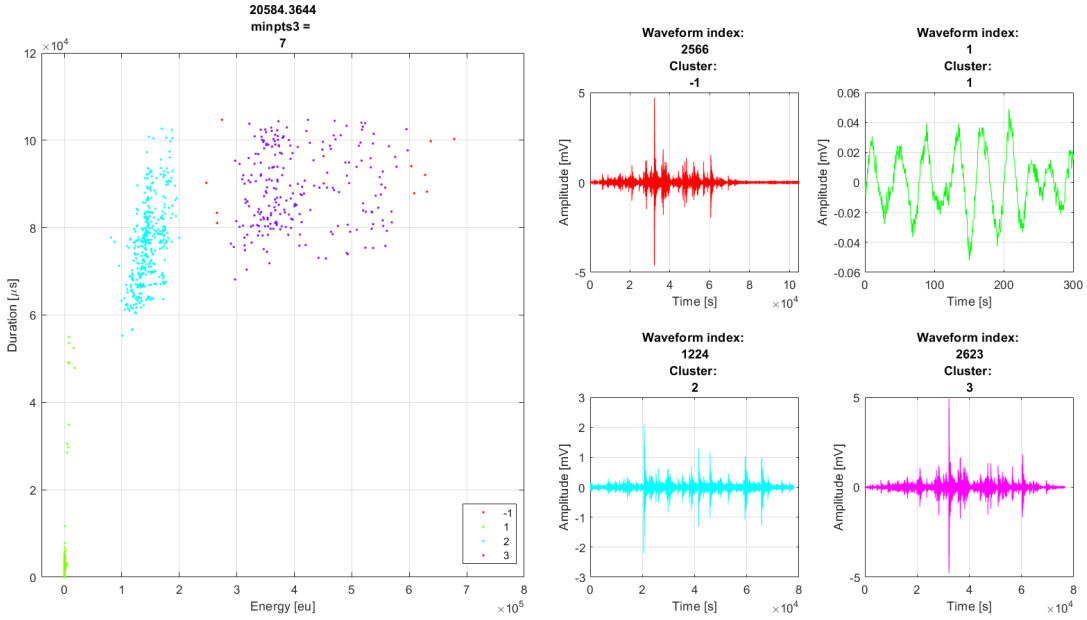


Figure 4.4: DBSCAN: Duration - Energy of specimen S1_cycle 9, waveforms of each cluster

As seen in the figure 4.4 there are three clusters and a group considered as noise:

- -1 (red dots)
- 1 (light green dots)
- 2 (light blue dots)
- 3 (purple dots)

Group -1 includes all points that do not meet the requirements to be included in a cluster, regardless of their nature. However, the fact that they are classified as "noise" by the algorithm does not exclude the possibility that they represent fracture, deformation or damage phenomena within the sample, as can be observed from the red wave(figures 4.4). Indeed, at a time instant of about 3.7×10^4, a peak of amplitude about 5 mV is observed, which could mean that, appling a filter to reduce noise, it can be easily identified a phenomenon in correspondence of the peak.

The remaining three groups differ clearly in duration and energy as a result of the clustering process. This suggests that, when looking at the distribution of clusters, the

phenomena associated with cluster 1 (green waveform: probabily, due to its energy, it is noise) is very different from clusters 2 (light blue wave) and cluster 3 (purple wave), while clusters 2 and 3 are related to different events that may exhibit similar characteristics in terms of duration (about $6-10\times10^4$ μs), while, considering the energy, purple cluster has double energy of light blue cluster (about 4×10^5 eu). In comparison with the previous cycle (4.1.1), a more defined distribution of data is evident in cycle 9. This phenomenon can be attributed to the fact that cycle 9 represents an advanced stage of testing, where most of the possible faults may already be developing, thus generating significantly more energy than the previously analyzed cycle, where most of the events could be associated with only background noise due to low threshold before acquisition.

To conclude the analysis and confirm the effectiveness of clustering by DBSCAN using method 3 for determining the eps and minPts parameters, it was decided to extend the analysis to a cycle near the end of the specimen's life. The cycle selected is 12 (i.e., 55000 to 60000 cycles), which corresponds to two cycles before failure.

4.1.3. Cycle 12

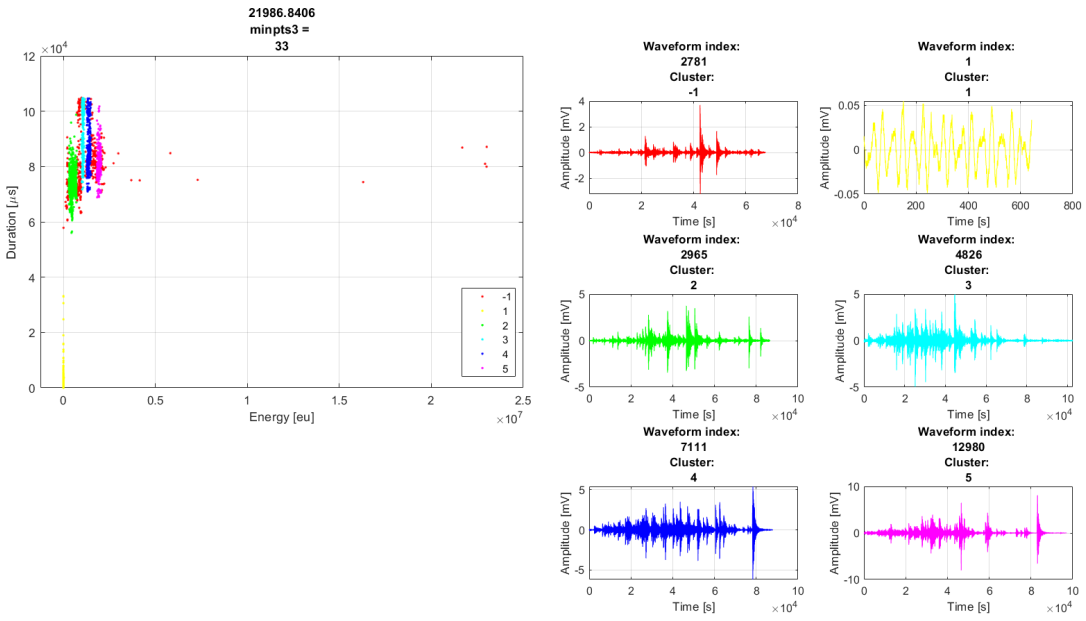


Figure 4.5: DBSCAN: Duration - Energy, specimen S1_cycle 12, waveforms of each cluster

The clustering of cycle 12 consists of six clusters. The yellow cluster has limited energy and duration, as well as a significantly lower amplitude than the other clusters, situating itself between 30 and 45 dB. Analyzing the previous cycles, one constant emerges: the

presence of a cluster whose amplitude never exceeds 50 dB and which, from the energy point of view, is significantly lower than the others. These are clearly noise-related and unremarkable phenomena that require adequate filtering to ensure the input signal is clean. This hypothesis is further confirmed by observing the waveforms associated with these clusters, which exhibit essentially the same characteristics.

Examining the other clusters, it becomes apparent that in this cycle they are all in close proximity to each other, with very similar characteristics in both duration and amplitude and energy. The presence of outliers, i.e., points that do not fit into any cluster and require in-depth investigation to be determined, is also evident

4.1.4. Discussion

Clustering was performed using DBSCAN considering features of duration and energy. Other characteristics of the waveforms were subsequently displayed graphically to support the hypothesis that method 3 used for selection of eps and minPts parameters and subsequent clustering produced promising results.

It should be noted that for cycles 1 there is a distribution of points with similar characteristics. Infact, most of the points could be considered noise because of their limited energy due to the fact that we are at the beginning of the test. By graphically representing all the characteristics of the waveforms (Figure 4.6), it is evident that the clustering, were it not for the colors, would be difficult to interpret since the phenomena are very similar to each other.

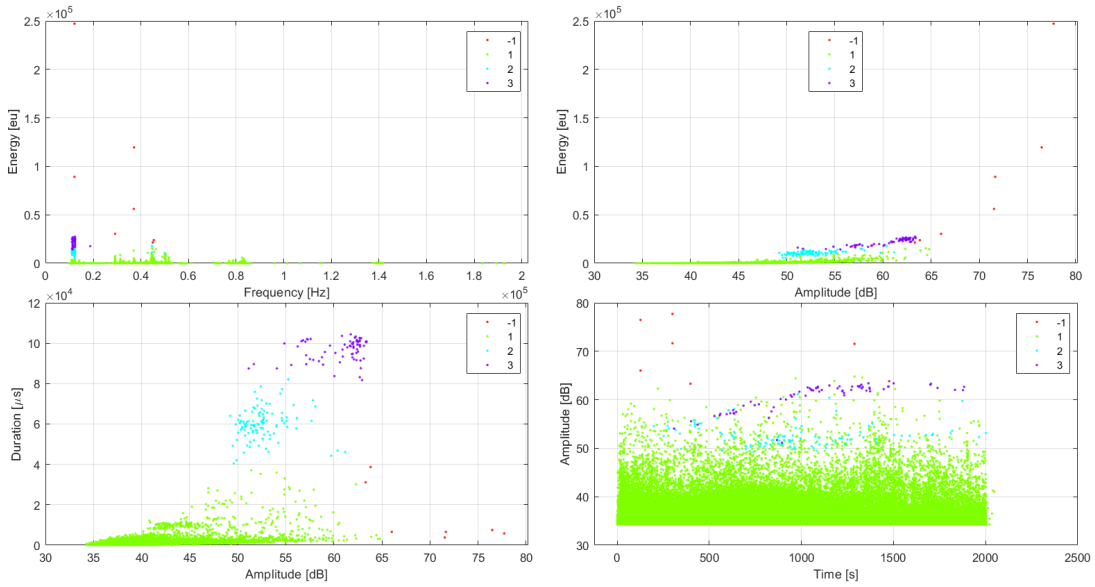


Figure 4.6: cycle 1: Energy - Frequency, Energy - Amplitude, Duration - Amplitude, Amplitude - Time

In contrast, in cycle 9 (Figure 4.7), there is a clear and sharp division of events, the algorithm easily identifies the differences between the points, generating very good quality clustering.

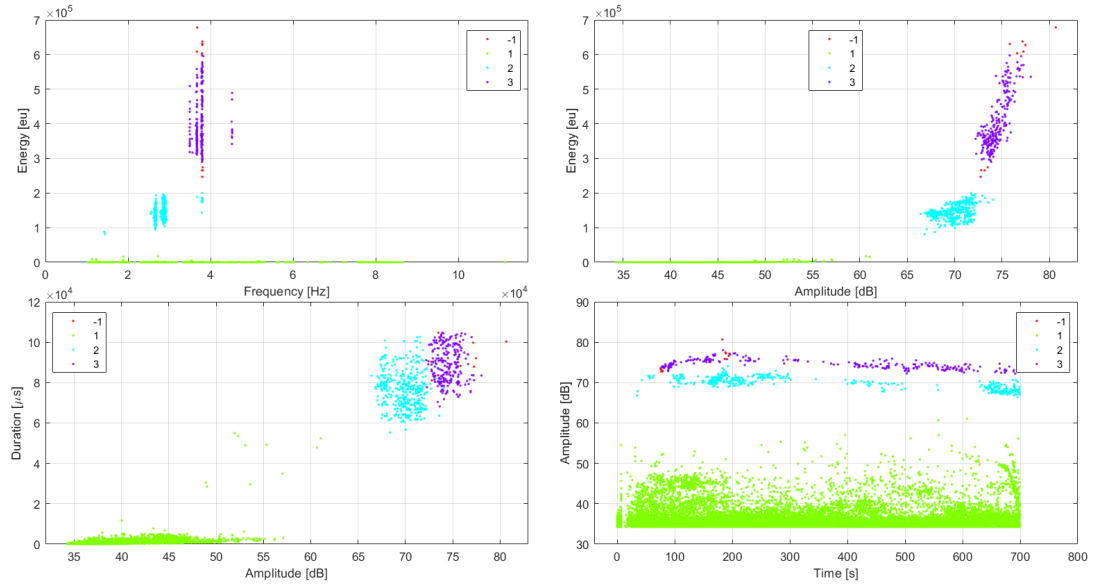


Figure 4.7: cycle 9: Energy - Frequency, Energy - Amplitude, Duration - Amplitude, Amplitude - Time

As can be seen in the graph above (figure 4.7), the three clusters generated are clearly distinguishable from each other. The light green cluster (cluster 1) has very limited energy and a frequency range from 1 to 9×10^4 Hz, but also in duration and amplitude. This cluster could indicate the presence of background noise detected by the sensors. To reduce or eliminate it, it might be appropriate to increase the threshold level so as to start the acquisition with a higher event energy level. The other two point clouds correspond to two clearly distinct clusters, as evidenced by the arrangement of the clouds in well-defined areas of the graphs. This suggests that each cluster could be associated with a specific type of defect or with a family of similar damages, characterized by particular features.

Cycle 12 (figure 4.8), on the other hand, represents the final phase of the test, just before breakup, with few points identified as noise and many characterized by high energy. The latter is due to the nature of the test and acquisition by acoustic emission, in which the kaiser effect plays a key role (a higher force must be applied to make the structure emit relevant sounds). However, in the presence of poorly delineated phenomena, the division obtained with the DBSCAN algorithm is more chaotic than in well-delineated events.

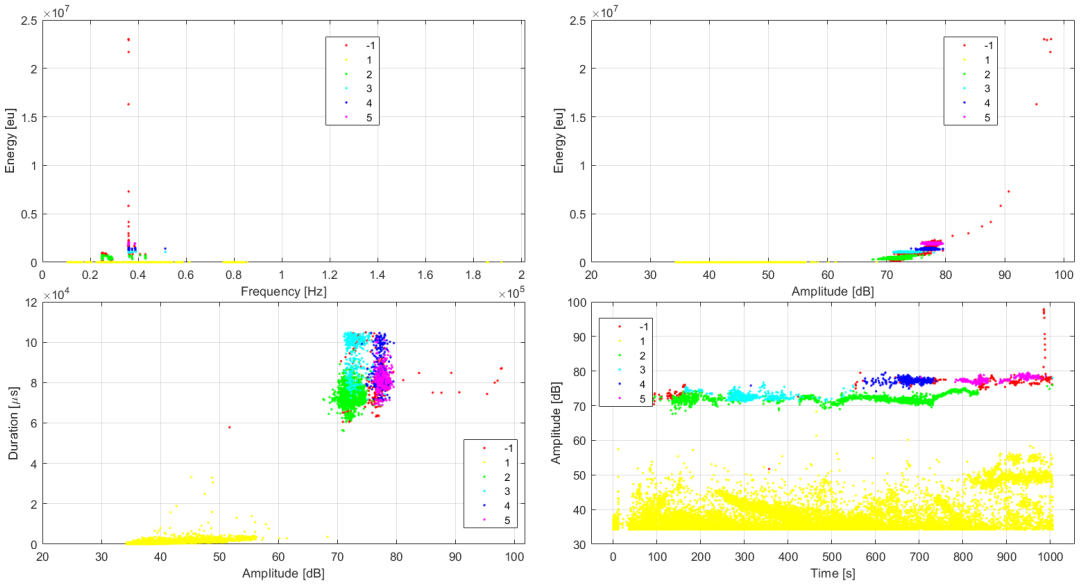


Figure 4.8: cycle 12: Energy - Frequency, Energy - Amplitude, Duration - Amplitude, Amplitude - Time

5 | Conclusions and future work

5.0.1. Conclusions

The introductory study presented in the chapter on the state of the art was essential to gain the knowledge necessary for the development of this work. Among the clustering algorithms described in the literature, we focused in particular on the DBSCAN algorithm, which was chosen for its advantages over its competitors, including the absence of the need to define the number of clusters a priori, its excellence in defining clusters of arbitrary shape, and its ability to identify outliers. To apply the algorithm to the data in this study, a rigorous definition of the fundamental parameters (eps and minPts) was required. The results of this phase led to the definition of a three-step methodology: the minPts = 20 assumption, the definition of eps based on the minimum number of points via the k-dist function, and finally the recalculation of the minimum value of points. This approach, referred to as "method 3" in the thesis, resulted in effective clustering of the data by reducing outliers (i.e., points that do not belong to any cluster) and minimizing user influence in clustering, avoiding parameter determination through trial-and-error methodology, known for its poor performance. In contrast, methods 1 and 2 did not perform up to expectations in terms of the number and distribution of clusters and outliers.

Next, we focused on verifying the consistency of the clustering obtained by "method 3" with the actual distribution of phenomenology in the test. Analysis of three separate cycles of the same specimen revealed that the algorithm is able to cluster regardless of the diversity of phenomena. However, in the presence of similar phenomena, the division is less obvious, suggesting the parallel use of a second clustering algorithm.

In addition, to confirm that the grouping did actually reflect different phenomenologies, the waveforms associated with each grouping were represented graphically. It turned out that the cluster identified as -1 (outliers) did not represent only noise, as some waveforms had energy, amplitude, and frequency typical of transient events. On the other hand, at least one cluster consisted mainly of waveforms with low energy and amplitude (continuous waves), identified as background noise. In addition, many clusters with typical features of phenomenology had a lot of noise but representative peaks of events. The obtained

clustering produced the expected results, but further development is needed to improve the work.

Summarizing:

- DBSCAN algorithm can be used to clusterize even large datasets (big data) as long as the eps and minPts parameters are chosen correctly
- DBSCAN algorithm represents a good clustering method when there are well distinct events, however, its performance is reduced when the events have similar characteristics
- Method 3, implemented to obtain a correct combination of eps and minPts, reduces user intervention. This implies that the assignment of eps and minPts is rigorous and trial and error is not performed. It thus guarantees good clustering of the dataset.
- Outliers, points that do not belong to any cluster, do not just represent noise, but may represent a group of events or single event and need more in-depth study

5.0.2. Future developments

From the data in use it emerges that the noise present in the acquired signals can influence the clustering of the data, compromising the optimal clustering of the events based on their real characteristics. On the basis of data used in this thesis, it is recommended to establish an appropriate threshold value. In particular, it is suggested to increase the threshold value during the pre-acquisition phase to avoid noise acquisition. By doing this, it can also reduce the amount of data, which is particularly significant in AE acquisitions. In the dataset in which the threshold value applied before the acquisition is too low, it is suggested to apply, in post-processing, before clustering, a filter to reduce the noisy signal from the real data.

Regarding the determination of a more precise and detailed definition of the defectology, it would be advisable to analyze in detail the fundamental characteristics of the waveforms associated with each cluster. A recommended approach might involve a controlled study of the materials to obtain a known history of the defects, which could then be compared with the data obtained from clustering.

Summarizing for future work/improvements:

- it could set a higher threshold to filter data from acquisitions via AE. Alternatively, apply a pre-cluster filter to eliminate noise, if the highest threshold is not possible.

- it could generate a history of defectology appropriate for each material, so as to obtain a model to compare for future

Bibliography

[1] Rosemere De Araujo, Alves Lima, Andrea Bernasconi, and Michele Carboni. Acoustic emission applied to mode i fatigue damage monitoring of adhesively bonded joints, 2022.

[2] Pedro Galvez, Alejandro Quesada, Miguel Angel Martinez, Juana Abenojar, Maria Jesus L. Boada, and Vicente Diaz. Study of the behaviour of adhesive joints of steel with cfrp for its application in bus structures. *Composites Part B: Engineering*, 129:41–46, 11 2017.

[3] Lucas F.M. da Silva and R. D. Adams. Adhesive joints at high and low temperatures using similar and dissimilar adherends and dual adhesives. *International Journal of Adhesion and Adhesives*, 27:216–226, 4 2007.

[4] M. D. Banea, M. Rosioara, R. J.C. Carbas, and L. F.M. da Silva. Multi-material adhesive joints for automotive industry. *Composites Part B: Engineering*, 151:71–77, 10 2018.

[5] G. Scarselli, Carola Corcione, F. Nicassio, and A. Maffezzoli. Adhesive joints with improved mechanical properties for aerospace applications. *International Journal of Adhesion and Adhesives*, 75:174–180, 6 2017.

[6] C. E. Moraes, L. F.P. Santos, T. P.F.G. Leal, E. C. Botelho, and M. L. Costa. Influence of surface treatment on the mechanical and viscoelastic properties of adhesive joints applied to the oil and gas industry. *Materials Research*, 26, 2023.

[7] K. V. Machalická, M. Vokáč, P. Pokorný, and M. Pavlíková. Effect of various artificial ageing procedures on adhesive joints for civil engineering applications. *International Journal of Adhesion and Adhesives*, 97, 3 2020.

[8] R. A.A. Lima, R. Perrone, M. Carboni, and A. Bernasconi. Experimental analysis of mode i crack propagation in adhesively bonded joints by optical backscatter reflectometry and comparison with digital image correlation. *Theoretical and Applied Fracture Mechanics*, 116:103117, 2021.

[9] C. B.G. Brito, R. C.M. Sales, and M. V. Donadon. Effects of temperature and moisture on the fracture behaviour of composite adhesive joints. *International Journal of Adhesion and Adhesives*, 100(March):102607, 2020.

[10] S. Budhe, M. D. Banea, S. de Barros, and L. F.M. da Silva. An updated review of adhesively bonded joints in composite materials. *International Journal of Adhesion and Adhesives*, 72:30–42, 2017.

[11] Jozef Kuczmaszewski. *Fundamentals of metal-metal adhesive joint design.*

[12] M. D. Banea and L. F.M. Da Silva. Adhesively bonded joints in composite materials: An overview. *Proceedings of the Institution of Mechanical Engineers, Part L: Journal of Materials: Design and Applications*, 223:1–18, 2009.

[13] A. J. Curley, H. Hadavinia, A. J. Kinloch, and A. C. Taylor. Predicting the service-life of adhesively-bonded joints. *International Journal of Fracture*, 103(1):41–69, 2000.

[14] Antonino Valenza, Vincenzo Fiore, and Livan Fratini. Mechanical behaviour and failure modes of metal to composite adhesive joints for nautical applications. *International Journal of Advanced Manufacturing Technology*, 53(5-8):593–600, 2011.

[15] Adeela Nasreen, Muhammad Kashif Bangash, Khubab Shaker, and Yasir Nawab. Effect of surface treatment on the performance of composite-composite and composite-metal adhesive joints. *Polymer Composites*, 43(9):6320–6331, 2022.

[16] F. Mortensen and O. T. Thomsen. Analysis of adhesive bonded joints: A unified approach. *Composites Science and Technology*, 62:1011–1031, 2002.

[17] S. Kumar and P. C. Pandey. Fatigue life prediction of adhesively bonded single lap joints. *International Journal of Adhesion and Adhesives*, 31(1):43–47, 2011.

[18] Rouhollah H. Goudarzi and Mohammad Reza Khedmati. An experimental investigation of static load capacity of AL-GFRP adhesively bonded single lap and double butt lap joints. *Latin American Journal of Solids and Structures*, 12(8):1583–1594, 2015.

[19] N. G.C. Barbosa, R. D.S.G. Campilho, F. J.G. Silva, and R. D.F. Moreira. Comparison of different adhesively-bonded joint types for mechanical structures. *Applied Adhesion Science*, 6:1–19, 2018.

[20] Jin Hwe Kweon, Jae Woo Jung, Tae Hwan Kim, Jin Ho Choi, and Dong Hyun Kim. Failure of carbon composite-to-aluminum joints with combined mechanical fastening and adhesive bonding. *Composite Structures*, 75(1-4):192–198, 2006.

[21] Do Won Seo and Jae Kyoo Lim. Tensile, bending and shear strength distributions of adhesive-bonded butt joint specimens. *Composites Science and Technology*, 65(9 SPEC. ISS.):1421–1427, 2005.

[22] I. J.J. Van Straalen, J. Wardenier, L. B. Vogelesang, and F. Soetens. Structural adhesive bonded joints in engineering - drafting design rules. *International Journal of Adhesion and Adhesives*, 18:41–49, 1998.

[23] Moon Sik Han, Hae Kyu Choi, Jae Ung Cho, and Chong Du Cho. Experimental study on the fatigue crack propagation behavior of DCB specimen with aluminum foam. *International Journal of Precision Engineering and Manufacturing*, 14(8):1395–1399, 2013.

[24] A. Pirondi and G. Nicoletto. Fatigue crack growth in bonded dcb specimens. *Engineering Fracture Mechanics*, 71:859–871, 2004.

[25] Sandeep Kumar Dwivedi, Manish Vishwakarma, and Prof Akhilesh Soni. Advances and researches on non destructive testing: A review. *Materials Today: Proceedings*, 5:3690–3698, 2018.

[26] Patryk Kot, Magomed Muradov, Michaela Gkantou, George S. Kamaris, Khalid Hashim, and David Yeboah. Recent advancements in non-destructive testing techniques for structural health monitoring. *Applied Sciences (Switzerland)*, 11, 2021.

[27] Vahid Reza Gharehbaghi, Ehsan Noroozinejad Farsangi, Mohammad Noori, T. Y. Yang, Shaofan Li, Andy Nguyen, Christian Málaga-Chuquitaype, Paolo Gardoni, and Seyedali Mirjalili. A critical review on structural health monitoring: Definitions, methods, and perspectives. *Archives of Computational Methods in Engineering*, 29:2209–2235, 2022.

[28] Michele Carboni. Lecture notes, introduzione al monitoraggio strutturale.

[29] Robert James Barthorpe. On model and data based approaches to structural health monitoring. 2010.

[30] S. Park, C. B. Yun, and D. J. Inman. Structural health monitoring using electromechanical impedance sensors. *Fatigue and Fracture of Engineering Materials and Structures*, 31:714–724, 2008.

[31] Nirvan Makoond, Luca Pelà, Climent Molins, Pere Roca, and Daniel Alarcón. Automated data analysis for static structural health monitoring of masonry heritage structures. *Structural Control and Health Monitoring*, 27:1–25, 2020.

[32] Adam Stawiarski, Marek Barski, and Piotr Pająk. Fatigue crack detection and identification by the elastic wave propagation method. *Mechanical Systems and Signal Processing*, 89:119–130, 2017.

[33] Michele Carboni. Lecture notes, introduzione ai controlli non distruttivi.

[34] Michele Carboni. Lecture notes, shm mediante emissione acustica.

[35] Alexander Jung. *Machine Learning The Basics*. Springer, 2022.

[36] Hui Jiang. *Machine Learning Foundamentals*. Cambridge University Press, 2022.

[37] Jafar Alzubi, Anand Nayyar, and Akshi Kumar. Machine learning from theory to algorithms: An overview. *Journal of Physics: Conference Series*, 1142:0–15, 2018.

[38] Selçuk Kaya, Buket Cicioğlu Aridoğan, and Mustafa Demirci. Hepatit B ve C virus enfeksiyonu olan hastalarda hepatit G virus prevalans*ı*. *Mikrobiyoloji Bulteni*, 38(4):421–427, 2004.

[39] Lior Rokach and Oded Maimon. Clustering methods. *Data Mining and Knowledge Discovery Handbook*, pages 321–352, 2006.

[40] Phiroz Bhagat. *Pattern recognition in industry*. Elsevier, 2005.

[41] Dibya Jyoti Bora and Dr. Anil Kumar Gupta. A comparative study between fuzzy clustering algorithm and hard clustering algorithm. *International Journal of Computer Trends and Technology*, 10:108–113, 2014.

[42] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics (Switzerland)*, 9:1–12, 2020.

[43] Hongjie Jia, Shifei Ding, Xinzheng Xu, and Ru Nie. The latest research progress on spectral clustering. *Neural Computing and Applications*, 24:1477–1486, 2014.

[44] Jiawei Han Jialu Liu. *Data clustering*, chapter 8. Chapman and Hall/CRC, 2014.

[45] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

[46] Maria Brigida Ferraro and Paolo Giordani. Soft clustering. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12:1–12, 2020.

[47] L. V. Tilson, P. S. Excell, and R. J. Green. A generalisation of the fuzzy c-means clustering algorithm. *Remote sensing. Proc. IGARSS '88 symposium, Edinburgh, 1988. Vol. 3*, 10:1783–1784, 1988.

[48] Satyam Kumar. C-means clustering explained, 2022.

[49] S. Agatonovic-Kustrin and R. Beresford. Basic concepts of artificial neural network (ann) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, 22:717–727, 2000.

[50] Richard Olawoyin, Antonio Nieto, Robert Larry Grayson, Frank Hardisty, and Samuel Oyewole. Application of artificial neural network (ann)-self-organizing map (som) for the categorization of water, soil and sediment quality in petrochemical regions. *Expert Systems with Applications*, 40:3634–3648, 2013.

[51] F. Rossi, F. Rossi, B. Conan-Guez, B. Conan-Guez, a. El Golli, and a. El Golli. Clustering functional data with the som algorithm. *Proceedings of ESANN*, page 305–312, 2004.

[52] Pedro Contreras and Fionn Murtagh. Hierarchical clustering. *Handbook of Cluster Analysis*, pages 103–124, 2015.

[53] Prasad Pai. Illustration of analysis and procedures used in hierarchical clustering in a simplified manner, 2021.

[54] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2:86–97, 2012.

[55] Hans Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1:231–240, 5 2011.

[56] Dingsheng Deng. Dbscan clustering algorithm based on density. *Proceedings - 2020 7th International Forum on Electrical Engineering and Automation, IFEEA 2020*, pages 949–953, 2020.

[57] P. Viswanath and V. Suresh Babu. Rough-dbscan: A fast hybrid density based clustering method for large data sets. *Pattern Recognition Letters*, 30:1477–1488, 2009.

[58] Nooshin Hanafi and Hamid Saadatfar. A fast dbscan algorithm for big data based on efficient density calculation. *Expert Systems with Applications*, 203:117501, 2022.

[59] Simon Fong, Saif Ur Rehman, Kamran Aziz, and Information Science. Dbscan : Past , present and future. pages 232–238, 2014.

[60] Thanh N. Tran, Klaudia Drab, and Michal Daszykowski. Revised dbscan algorithm

to cluster data with dense adjacent clusters. *Chemometrics and Intelligent Laboratory Systems*, 120:92–96, 2013.

[61] Mykola Drobiazko. An acoustic emission approach to monitor crack propagation in adhesive joints under mode i, 2020/2021.

[62] ASTM ASTM. Standard test method for fracture strength in cleavage of adhesives in bonded metal joints. *Adhesives, American Society for Testing and Materials*, 1999.

[63] R.M. Lopes, R.D.S.G. Campilho, F.J.G. da Silva, and T.M.S. Faneco. Comparative evaluation of the double-cantilever beam and tapered double-cantilever beam tests for estimation of the tensile fracture toughness of adhesive joints. *International Journal of Adhesion and Adhesives*, 67:103–111, 2016. Special Issue on Adhesion, Surface Preparation and Adhesive Properties.

[64] A. Biel and U. Stigh. An analysis of the evaluation of the fracture energy using the DCB-specimen. *Archives of Mechanics*, 59(4-5):311–327, 2007.

[65] Jin Yu Song, Yi Ping Guo, and Bin Wang. The Parameter Configuration Method of DBSCAN Clustering Algorithm. *2018 5th International Conference on Systems and Informatics, ICSAI 2018*, (Icsai):1062–1070, 2019.

[66] Amin Karami and Ronnie Johansson. Choosing DBSCAN Parameters Automatically using Differential Evolution. *International Journal of Computer Applications*, 91(7):1–11, 2014.

[67] Hongfang Zhou, Peng Wang, and Hongyan Li. Research on adaptive parameters determination in dbscan algorithm. *Journal of Information & Computational Science*, 9(7):1967–1973, 2012.

[68] M. Daszykowski, B. Walczak, and D. L. Massart. Looking for natural patterns in data. part 1. density-based approach. *Chemometrics and Intelligent Laboratory Systems*, 56:83–92, 2001.

[69] Mohammed Azmi Al-Betar, Ahamad Tajudin Khader, Zong Woo Geem, Iyad Abu Doush, and Mohammed A. Awadallah. An analysis of selection methods in memory consideration for harmony search. *Applied Mathematics and Computation*, 219(22):10753–10767, 2013.

[70] Rosshairy Abd Rahman, Razamin Ramli, Zainoddin Jamari, and Ku Ruhana Ku-Mahamud. Evolutionary algorithm with roulette-tournament selection for solving aquaculture diet formulation. *Mathematical Problems in Engineering*, 2016, 2016.

[71] Rasmus Bro and Age K. Smilde. Principal component analysis. *Analytical Methods*, 6:2812–2831, 2014.

[72] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 61:611–622, 1999.

[73] Artur Starczewski, Piotr Goetzen, and Meng Joo Er. A new method for automatic determining of the dbscan parameters. *Journal of Artificial Intelligence and Soft Computing Research*, 10:209–221, 2020.

[74] Shiliang Sun and Rongqing Huang. An adaptive k-nearest neighbor algorithm. *Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010*, 1(Fskd):91–94, 2010.

[75] Artur Starczewski and Andrzej Cader. *Determining the eps parameter of the DBSCAN algorithm*, volume 11509 LNAI. Springer International Publishing, 2019.

[76] Wenhao Lai, Mengran Zhou, Feng Hu, Kai Bian, and Qi Song. A New DBSCAN Parameters Determination Method Based on Improved MVO. *IEEE Access*, 7:104085–104095, 2019.

[77] Lu Zhu, Jiang Zhu, Chongming Bao, Lihua Zhou, Chongyun Wang, and Bing Kong. Improvement of DBSCAN algorithm based on adaptive EPS parameter estimation. *ACM International Conference Proceeding Series*, 2018.

List of Figures