A Machine Learning approach for damage detection and localisation in Wind Turbine Gearbox Bearings

A thesis submitted to the University of Sheffield for the degree of Doctor of Philosophy in the Faculty of Engineering

by

Ian Ericovich Martinez Efremov

Department of Mechanical Engineering

University of Sheffield

June 2020

Acknowledgements

First of all, I want to acknowledge all the people that were involved during my journey as a PhD student and the Panamanian government for supporting my studies through the IFARHU-UTP programme in 2014. This work would not be possible without the guidance of Graeme Manson and Rob Dwyer-Joyce who supervised me after the second year. Thanks to professor Matthew Marshall who guided me during my first year of my PhD and provided me with an initial idea of this project and allowed me to closely collaborate with him — he is a great supervisor and researcher and I am really lucky I had the chance to work with him. I cannot imagine how less exciting this project will be without his guidance and advice even after my first year.

Additionally, I need to mention that the experimental stage of this work would not be possible without the technical skills of Gary Nichols, Benjamin Clarke and David Butcher who supported me for a long time in my PhD. I am very grateful that I also had the chance to collaborate with them as I gained a lot of expertise and friends along this journey. Thanks to the people working in the Leonardo and Lea laboratories their technical support was invaluable.

I have to mention that I am grateful to share the office with some of the most dedicated, talented and bright students in the RD14 office. My journey as a PhD student would not be that enjoyable without them. I am quite frankly convinced that there is no better office in the Mining block. I would like to thank Anthony Reid, Zheng Ai, Jose Daniel Pérez, Luis Urena, Emiliano Ximenes and a list of innumerable people who I have met during my career as a researcher. Finally, I want to acknowledge the support and understanding provided by my family and relatives.

Abstract

Increasing demand for renewable sources requires more cost-effective solutions to mitigate the cost of maintenance and produce more energy. Preventive maintenance is the most normally adopted scheme in industry for maintenance but despite being well accepted has severe limitations. Its inability to intelligently schedule maintenance at the right time and prevent unexpected breakdowns are the main downsides of this approach and consequently leads to several problems such as unnecessary maintenances. This strategy does not justify the additional costs and thereby represents a negative aspect for renewable energy resource companies that try to generate cost-competitive energy. These challenges are progressively leading towards the predictive maintenance approach to overcome these aforementioned issues. Wind Turbine Gearbox Bearings have received a lot of attention due to the high incidence failure rates provoked by the harsh operational and environmental conditions. Current techniques only reach a level one of diagnostics commonly known as the Novelty Detection stage and normally requires the expertise of a skilled operator to interpret data and infer damage from it. A data-driven approach by using Machine Learning methods has been used to tackle the damage detection and location stage in bearing components. The damage location was performed by using non-destructive methods such as the Acoustic Emission technique — these measurements were used as features to locate damage around the bearing component once the damage was detected. The implementation of this stages also led to the exploration of damage generation due to overload defects and proposed a methodology to simulate these defects in bearings — the study of this concept was implemented in a scaled-down experiment where damage detection and localisation was performed. Due to the importance of the implementation of a damage location stage, damage in AE sensors was also explored in this work. Features extracted from impedance curves allowed to train Machine Learning methods to trigger a novelty when a bonding scenario occurred. This ultimately allowed the identification of unhealthy sensors in the network that could potentially generate spurious results in the damage predictions stage.

Abbreviations

Acronyms, Full text

AANN Auto-associative Neural Network.

AE Acoustic Emission.

AIC Akaike Information Criterion.

CFRP Carbon Fibre Reinforced Plastic.

CM Condition Monitoring.

CWT Continuous Wavelet Transform.

EDM Electrical Discharge Machining.

EI Electrical Impedance.

EMI Electromechanical Impedance.

FEM Finite Element Method.

FFT Fast Fourier Transform.

GP Gaussian Process.

HT Hilbert Transform.

LDA Linear Discriminant Analysis.

MAS Maximum Amplitude Spectra.

ML Machine Learning.

MLP Multilayer Perceptron.

NDT Non-Destructive Testing.

NN Neural Network.

PCA Principal Component Analysis.

PLB Pencil Lead Break.

PZT Lead Zirconate Titanate.

RBF Radial Basis Function.

RCF Rolling Contact Fatigue.

SHM Structural Health Monitoring.

SMD Squared Mahalanobis Distance.

SNR Signal-to-Noise ratio.

STFT Short Time Fourier Transform.

SVM Support-Vector Machines.

TOA Time of Arrival.

TOF Time of Flight.

TSA Thermoelastic Stress Analysis.

WEA White Etching Area.

WEAs White Etching Areas.

WECs White Etching Cracks.

WSF White Structure Flaking.

WTGB Wind Turbine Gear Box.

WVD Wigner-Ville Distribution.

List of Figures

List of Tables

Contents

CHAPTER 1

Introduction

1.1 Wind energy: Trend and challenges

The trend of shifting from fossil fuels to renewable energy sources has been progressively growing, this is due to many concerns such as high oil prices and global warming and has led to greater investigations of new alternative sources. Wind power has become the most commercially available alternative to generate electricity. In a local context, the United Kingdom reported an operational capacity of 20.13 GW in 2018 where a contribution of 7.9 GW came mainly from offshore wind turbines. This represents a power supply to around 4.5 million homes annually and expected to generate over 10% of the electricity by 2020 [3].

The increase in wind power capacity has imposed various challenges in terms of design and maintenance; specifically for offshore wind turbines. Placing wind turbines off-shore gives the advantage of high-speed winds and consequently an increasing amount of power per unit installed. This comes as a trade-off, as subjecting them to harsh conditions, such as severe transient loads, high humidity, salt water and higher loads and mixed lubrication regimes, eventually affects the gearbox of the wind turbine. This component is one of the most crucial parts of the wind turbine, and it is responsible for transmitting the low rotational speed of the main shaft to the high-speed shaft coupled to the electric generator. Therefore, a failure in this component has a negative impact on the power generation capabilities of the wind farm and the costs associated with maintenance.

A further understanding of the failure incidence has raised concerns among the governmental institutions such as the U.S. Department of Energy and the National Renewable Energy Laboratory (NREL) where collected information of the most common failure rates showed that 76–80% of turbine gearbox failures were generated by bearings [4, 5].

Additionally, gearboxes have a life expectancy of around 20 years according to the standard DIN ISO 281 [6] but, due to the aforementioned operational conditions, the life expectancy in the planetary bearings is reduced to 2–11 years [5, 7]. This results in a knock-on effect in the maintenance costs as gearbox replacement will normally cost around £300k [8].

In general, it has been reported that the maintenance and operational cost of 750kW wind turbines accounts for 25–30% of the overall energy generation cost and 75%–90% of the investment costs [9, 10], therefore, there is a great necessity to mitigate unexpected failures and improve their efficiency and reliability in order to make wind power a feasible alternative to fossil fuels.

1.2 The wind turbine gearbox

The gearbox is one of the most important components in the wind turbine as its main function is to transmit the load and increase the rotational speed from the input shaft to the output shaft connected to the generator. Figure 1.2.1(a) shows the position of the gearbox in the wind turbine where the rotor is attached to the input shaft of the gearbox. Figure 1.2.1(b) shows the gearbox assembly where the rotational speed from the rotor is transmitted to the input shaft with a speed range between 15–20 RPMs, this rotational speed increases to 1500–1800 RPMs by means of a series of planetary gears and intermediate gears [11, 12].

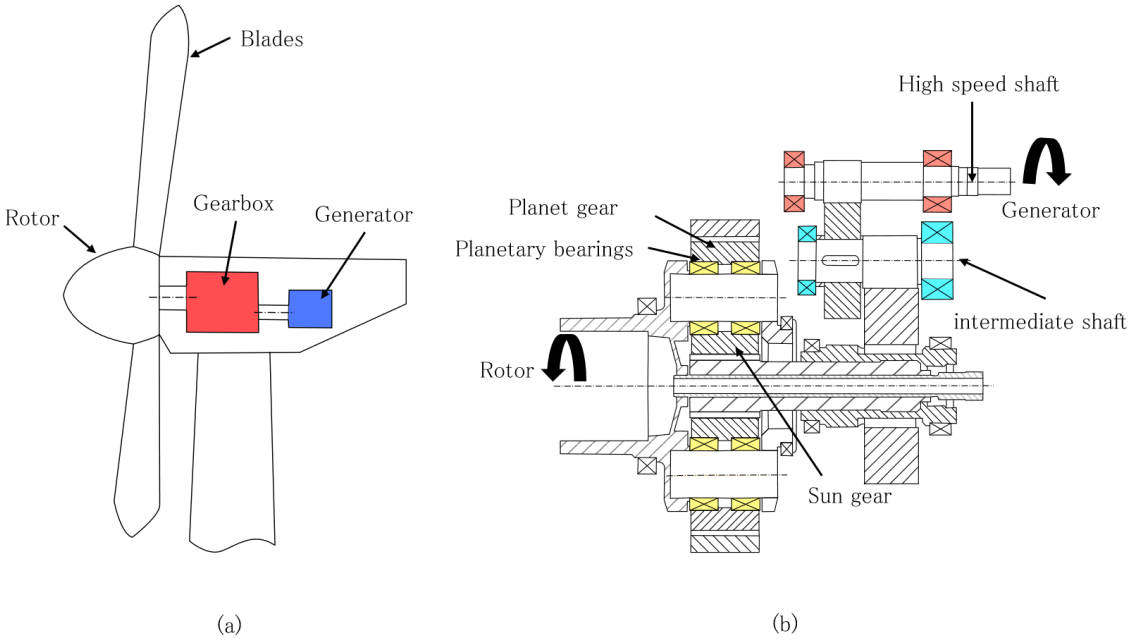


Figure 1.2.1: (a) Wind turbine components and (b) gearbox with a single epicyclic gear stage [1].

1.2.1 Failure modes

Gearboxes are normally subject to harsh conditions. Off-shore wind turbines are a clear example in that they are subjected to severe transient conditions, overloads, corrosion and high speed winds. This ultimately affects the overall life of the gearbox components. Thus, it is not surprising that around 40–50% of wind turbine failures are as a result of general failures in the gearbox [13] due to extreme operational conditions. This leads to failure modes that are predominantly due to fatigue of components, wear, fracture and insufficient lubrication. Moreover, the most common failures are observed on bearings, where an estimated 76%–80% of the gearbox failure correspond to such failures [4, 5]. A taxonomy of the most common bearing failures can be found in the ISO 15243 [14]. This can be summarised as shown in Figure 1.2.2.

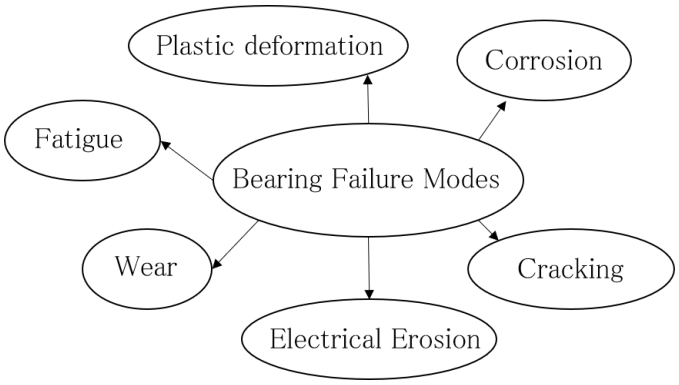


Figure 1.2.2: Bearing failure modes according to ISO 15243.

1.2.2 Life rating and Rolling Contact Fatigue

Rolling Contact Fatigue (RCF) is the main predominant failure mode in bearings where its occurrence only happens under certain circumstances such as correct mounting, lubrication and absence of overload events [15]. The effects of high cyclic shear stresses beneath the contact subsurface initiate the formation and growth of cracks until imminent failure occurs in the component. This led to the development of mathematical models that estimate this life expectancy by obtaining an estimate of the number of revolutions or operating hours under specific operational conditions. One of the most well-known models is a statistical expression based on the Lundberg and Palmgren's theory [16] known as the "L_{10} life" and states the life rating where 90% of the bearings will survive under a particular operational condition. This life rating can be calculated according to ISO 281 as [17]:

$$L_{10} = \left(\frac{C}{P}\right)^{p} \tag{1.2.1}$$

where the term C is the dynamic load rating [kN], P is the equivalent dynamic load [kN] and p is the load-life exponent that depends on the type of bearing ($p = 3$ for ball bearings and $p = 10/3$ for roller bearings).

In the context of Wind Turbine Gear Box (WTGB) bearings, well-controlled conditions are generally difficult to attain, leading to misalignment issues, poor mounting conditions during assembly or even after a scheduled maintenance might happen. Sometimes insufficient lubrication becomes a problem and frequent overload events might occur due to transient events in these harsh operational conditions. It is

enough to say that these factors will eventually lead to a premature failure that is not attributed to RCF, thereby making it more difficult to predict a life expectancy that is generally less than the life calculated under the RCF assumption. Premature failures have often been associated with the formation of White Etching Areas (WEAs) in the subsurface of the bearing rings. Such features have been reported for two decades [5, 18] and have been the motive to study its formation as described in the next section.

1.2.3 White Etching Areas and White Etching Cracks

Unfortunately, life expectancy in WTGB bearings is significantly less than the life predicted by the L_{10} model and some literature suggests a life expectancy of around 2–11 years [19, 20]. Further studies have attributed the premature failure to the formation of WEAs in the subsurface of the bearings [5, 21, 22].

WEAs as its name suggests are known for its characteristic white colour resulting from the localised microstructural change in the material matrix. WEAs also have been described in literature as regions having a locally changed hardness in the material. Evans [5] reported 30% increase in the hardness of the surrounding material, and stated that the White Etching Area (WEA) microstructure has shown to be composed of equiaxed nanoferrite grains of around 10–100nm in diameter being supersaturated with carbon [5]. Surface failure will arise from a subsequent phenomenon as the formation of a WEA is also accompanied with White Etching Cracks (WECs). These microscopic cracks will grow and their complex propagation paths will form crack networks that will lead to White Structure Flaking (WSF).

Currently a few aspects and their implication in bearing failure are partially known but they are still a matter of debate due to their formation mechanisms not being completely understood [5, 23]. The formation WEA has been mostly attributed to a localised plastic deformation beneath the surface. This suggests that, once the yield strength of the material is exceeded, a plastic region will form below the contact zone. This plastic strain will surround some non-metallic inclusions that, under cyclic stress, will lead to a dislocation movement and further plastic accumulation that will consequently lead to a microstructural change [22].

The long-term effects are well-known and damage must be expected at some point in the component's life. Note that the term damage is used vaguely at this point as there is no connotation to the component usability. In a practical situation a suboptimal operational state might be required but to the point that the safety of the gearbox is not fully compromised. This may be achieved by the implementation of Condition Monitoring (CM) strategies.

1.3 The MultiLife concept

The reduction in the WTGB life expectancy led the industrial sector to mitigate the current design challenges in WTGB. A few innovations in this area have been proposed by Ricardo plc by implementing the MultiLife™concept. This concept tries to extend the bearing life expectancy by more than 500% and currently is installed and commissioned on a 600kW turbine at the Barnesmore wind farm owned by ScottishPower [24].

The MultiLife™concept evolved from the idea that bearing life can be extended by rotating a static inner raceway during operation, the rotation will avoid further wear progression within the critical loading zone of the bearing by relocating the defect on a low loading zone. The concept is currently under test in a 600kW turbine in Barnesmore operated by ScottishPower [24]. The bearing is rotated by an indexing mechanism that actuates a series of hydraulic cylinders (3) as shown in Figure 1.3.1. The actuation is achieved by coupling the timing key (1) with a modified inner raceway, a sprag clutch mechanism (2) is used to avoid reverse slip during the indexing operation.

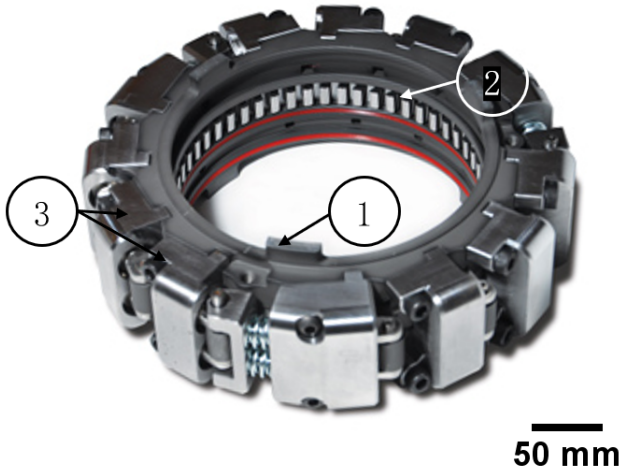


Figure 1.3.1: Multilife Mechanism.

A wide range of indexing strategies have been investigated by Howard [11], whose approach consisted in developing a modified life rating model based on the L_{10} life model. This new model included two parameters defined as indexing period and indexing angle, the indexing period defines the period for each indexing event and the indexing angle defines the radial rotation for each cycle. The varying of these parameters will result in different indexing strategies where the optimum overall life rating can be achieved by optimising these parameters. The results of this analytical model shown that the best optimisation strategy was achieved by applying short indexing periods at short angle intervals leading to a five-fold increase in the bearing life.

1.4 The predictive maintenance approach

Before continuing, it is important to define the term: *damage*. In this work, *damage* is defined as any changes to the material or geometric properties of the structure/machine or the boundary conditions that affect the system's performance [25]. Notice, that this term is not equivalent to the term *failure*, as a damaged structure/machine might operate within suboptimal conditions before reaching a level where its functionality or operation is no longer safe or acceptable. In other words,

an uncorrected damaged state might progress to a more severe state denominated as *failure* that consequently compromises the safety of the structure/machine.

Traditional maintenance approaches such as the run-to-failure and preventive maintenance strategies are generally the most common methods used in industry. Even though they partially mitigate the increasing number of unexpected breakdowns, a more efficient method is required. The run-to-failure has a huge impact on manufacturing efficiency as it leads to long downtimes. This is partially solved by the preventive maintenance approach which performs a more frequent maintenance on a periodic basis based on the bathtub curve behaviour as seen in Figure 1.4.1. Even though this strategy reduces the amount of unexpected break-downs, it is relatively inefficient in terms of unnecessary maintenances and spare part costs [26]. These issues can be addressed by adopting a more sophisticated approach known as predictive maintenance.

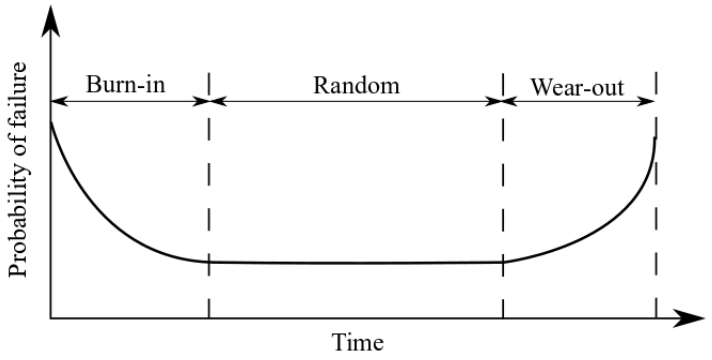


Figure 1.4.1: Bathtub curve showing the probability of failure through time.

The main characteristic is that predictive maintenance is performed in an on-line manner conjointly with Non-Destructive Testing (NDT) in order to determine the machine condition. As sensors can not infer damage by themselves, a diagnosis is based on certain features or parameters drawn from acquired data that give information about the machine conditions. As with any other maintenance strategy, the main goal is to allow the equipment to perform within the required operating condition limits.

These parameters are commonly extracted from different data signals such as vibration, AE, thermal and debris analysis [27]. Thermal analysis has had a great impact to detect/localise damage by monitor the sudden change in temperature. The methods to measure this parameter ranges from simple thermocouples to more sophisticated thermal imaging techniques [28]. A more complex analysis included both thermal and vibrational data in order to improve the discrimination between faulty and healthy states by using Principal Component Analysis (PCA) methods [29]. Debris analysis generally obtains its parameters from the data obtained from the generated debris in machinery and oil conditions. Any change in the debris morphology such as its size, shape and number or particles or change in the oil conditions such as contamination and degradation can indicate wear in bearings, gears and rotor [30, 31]. Several methods such as spectrography have shown capabilities not only on damage detection but also for damage localisation as this technique analyses material composition of the debris and relate it to a specific composition of the component part. The morphological change in the debris has been successfully monitored by ferrographic analysis but its application has been limited due to the high costs of the equipment [32]. Chip detection has also been implemented by giving an alarm when a certain amount of debris generates [33]. Further CM methods have been vastly developed towards the vibration and AE analysis currently becoming as a standard for CM applications.

Predictive maintenance can be performed systematically, and this can be understood by means of a damage identification hierarchy structure which is the framework for Structural Health Monitoring (SHM)/CM strategies. This framework has been described by Rytter with four levels where each level requires information of the previous level [34]. This hierarchy is shown in Figure 1.4.2 where Level One starts with the determination of the presence of any damage then goes systematically through damage location, severity and prediction of the remaining service life.

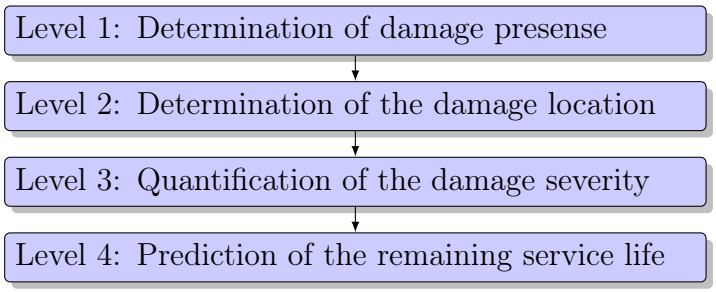


Figure 1.4.2: Rytter's damage identification hierarchy.

As seen in Figure 1.4.2, damage localisation belongs to Level 2 of the hierarchy and requires further information on whether or not damage is present. Level 3 is commonly known as *damage assessment* and it predicts the damage severity after the location stage. Level 4 also known as *damage prognosis* predicts the remaining life of the structure/component. A modified version of the presented damage identification hierarchy includes an intermediate stage between levels 2 and 3 known as *damage classification*. It is important to mention that each level is dependant on and more difficult than the previous stages, also the amount of data and cost increases for higher levels making its implementation impractical, thus for many applications in SHM/CM the implementation of level 1 and 2 can be deemed sufficient.

1.5 Motivation

The MultiLife™concept has been proposed as a solution that might potentially increase the remaining life of WTGB bearings. However, despite its promise for improvement, there is no established method to systematically use this mechanism resulting in the need for further exploration in novelty detection and damage localisation techniques in WTGB bearings.

So far, the life extension mechanism involves extending the life expectancy as much as possible by constantly rotating the inner raceway at shorter intervals regardless of any damage onset or damage position. In practical applications life expectancy is predisposed to have fluctuations due to the variations in the material strength and the random nature of the material inclusions, resulting in certain zones of the bearing being stronger or weaker than others and more susceptible to suffer an early onset of

damage than other locations. The MultiLife™concept can therefore be exploited by optimising its rotating strategy based on the detection and location of the generated damaged by applying CM strategies. In the context of WTGB, the implementation of a Level 2 of diagnostics has not been fully explored, and further investigations might be beneficial in improving the reliability of these components. The present work aims to go towards the second level of detection by applying well-established principles of preventive maintenance, together with Machine Learning (ML) techniques in order to extend the life expectancy of WTGB.

1.6 Aim and objectives

The main aim of this thesis is to implement a damage detection and damage localisation method for wind turbine roller bearings. The techniques will be demonstrated on scaled experiments, in order to understand their performance and reliability. The following objectives can be summarised as:

- Perform a hierarchical damage detection framework to target a novelty detection and damage localisation stage in roller bearings.
- Implement a data-driven approach using machine learning methods to target novelty detection and damage localisation in a scale-down and real-scale bearing.
- Incorporate an integrated approach for sensor self-assessment using ML techniques.
- Understand additional challenges and limitations for the implementation of this technology.

1.7 Scope of the thesis

CM of WTGB has attracted interest due to its advantages over traditional maintenance strategies, even though most of the reported studies have been performed in small-scaled bearings within an outer raceway diameter range of 35 – 140 mm [33].

These studies demonstrated the ability to perform damage detection in very controlled environments and so far have been proven successful for damage detection purposes. However, a damage location stage has not been extensively studied in this field and just a few works have been developed so far but these were mainly targeted at small-scale bearings where only damage detection was deemed necessary [35–37].

This current work tries to implement CM strategies in large scale bearings by implementing a damage detection and location strategies. The implementation of this approach could potentially result in helpful information for performing predictive maintenance and therefore to extend the life of WTGB bearings. Moreover, it should result in a reduction in downtime due to its capabilities for predicting damage location without performing in-situ inspections. This information could also be useful conjointly with the Multilife™concept by implementing a location-based strategy index.

A novelty detection stage will be tackled with machine learning techniques by using an Outlier Analysis and Linear Discriminant Analysis (LDA). The use of these algorithms have been proven useful in the context of SHM and can be potentially used to explore novelty detection in roller bearings. Traditional methods of novelty detection in roller bearings involved more of a subjective analysis in terms of the trend in the vibrational data. This clearly imposes a problem as some damage features cannot be classified by mere human intuition, thus leaving the problem to more sophisticated methods, such as ML techniques.

A damage localisation stage will be implemented using neural networks. The main reason for this choice is that they have been proven powerful for classification and regression problems which makes them suitable for CM applications. Furthermore, this allows the localisation approach to be based on a data-driven approach which simplifies some of the current localisation problems by rather understanding the data characteristics. This allows building predictive models that can learn complex data relationships and therefore mitigate the limitations of localisation on complex geometries, influences of material anisotropy and variability in the propagation speed. From the machine learning perspective, these features will be extracted from the Acoustic Emission (AE) measurements. This NDT method has been widely adopted in industry due to its capabilities to detect, locate and characterise damage.

Additionally, due to the critical role played by sensors in this work, methods of monitoring the sensor integrity have been included as part of an integral approach. This is focused on monitoring the AE sensors in the network. In general the CM framework will be applied as shown in Figure 1.7.1. This provides not only the means of monitoring the bearing integrity but also provides a more reliable monitoring system by including the sensor self-assessment into the CM schema.

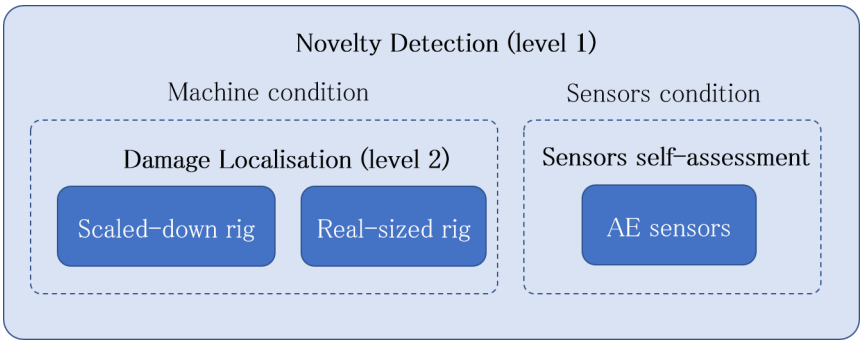


Figure 1.7.1: The CM framework applied in this work for WTGB.

1.8 Thesis outline

Chapter 2 will introduce vibration analysis and AE methods for roller bearings. An overview of signal processing techniques for damage detection using time-domain, frequency-domain, time-frequency and time-scale domain discusses their application for extracting features from these signals. Additionally, some of the novelty detection and damage locations techniques using a ML approach are discussed in the context of SHM/CM. Finally, as a means of incorporating the AE sensors into the CM strategy, an introduction of sensor-self diagnosis will be included.

Chapter 3 introduces ML techniques for novelty detection such as the Outlier detection method and the LDA. A background of neural networks for regression problems is included. This ML method will be used for damage localisation purposes using a dataset of Time of Arrival (TOA) features. Additionally, this chapter describes a different approach for AE source localisation using the beamforming technique.

Chapter 4 introduces the experimental rigs used in this work. A scaled-down rig and the MultiLife rig are described in this chapter together with their instrumentation

and test bearing dimensions. Details about their assembly, loading mechanisms and operational conditions will be explained in this chapter.

Chapter 5 introduces the experimental procedure for the novelty detection and damage localisation stage for the scaled-down and MultiLife rig. A description of the methods used to simulate the bearing damage by generating a geometrical defect and an overload defect is also presented. This chapter introduces a method to simulate an overload defect by using using a Finite Element Method (FEM) approach and its validation using Neutron Bragg imaging. The feature extraction procedure for novelty detection is discussed, together with the signal enhancement techniques used to pre-process the vibrational signal from scaled-down and MultiLife rig.

Subsequently, the procedures to implement a damage localisation stage in roller bearings will be explained. This chapter will detail the feature extraction and signal processing stages to generate a TOA feature from an AE signal. Moreover, the procedures to implement a time-delay beamforming approach are described. Finally, an experimental stage for feature extraction for AE sensor self-assessment is detailed in this chapter.

Chapter 6 will show the results for the novelty detection and damage localisation stage on the scaled-down and MultiLife rig. This chapter also includes the results for the damage detection of AE sensors.

Chapter 7 will include the conclusions, discussion and further work. Implications of the implementations of such techniques will be discussed together with the potential challenges encountered in a real scenario.

CHAPTER 2

Literature review

WTGB bearings have recently received great attention in the CM field [38]. The main reason is that these components are suffering from premature failure and high operational and maintenance costs [10]. Preventive maintenance aims to reduce these unexpected breakdowns by means of exploiting NDT and signal analysis methods. This chapter will summarise the most common CM and signal processing techniques used to monitor roller bearings.

2.1 Vibration analysis methods for roller bearings in condition monitoring

Vibration analysis has become one of the most common and reliable approaches in CM. It consists of measuring the machine vibrational response and monitoring its change over time. As this change in the machine vibrational response is mainly governed by the rotational speed, mass and geometry variations, different types of failures and defects can be related to a specific vibrational response in machines.

The vibrational response can be measured using transducers for displacement, velocity and acceleration. The selection of these sensors relies on the frequency range of interest to analyse, therefore based on the frequency range of operation of the machine and the way that damage affects the signal. Over time different methods have been used to monitor the vibration signal, and many techniques have been explored.

This has led to the broad classification of these approaches as time, frequency, time-frequency/time-scale methods [26].

2.1.1 Time-domain and frequency-domain methods

Time-domain methods are generally considered the most simple analysis method as no further processing is required for analysis. It is based on the fact that any onset of damage or deviation of the machine form its normal behaviour can be observed in a variation of a time-domain signal. The most basic feature used for this purpose is the change in the amplitude levels of the signal. One problem with this approach is that this type of analysis involves capturing a considerable amount of data with high sampling rates, resulting in the storage of large quantities of data. This issue led to the exploration of several different features sensitive to damage. A particular method that has shown great success was based on processing the vibrational response and calculating its statistical parameters over time. This process will be discussed later as its correct definition is feature extraction in the CM field. These parameters are the root mean square, skew and kurtosis, mean, crest factor and peak value methods [30].

Time-domain methods normally rely on a large amount of data for collection and interpretation is not a trivial task and can require a great deal of expertise to produce a reliable diagnostic. An alternative, and often more successful feature domain is based on monitoring changes in the frequency domain. A great example is the use of a Fast Fourier Transform (FFT) algorithm to extract the spectral content of time-domain signals. The advantage of this approach is that it produces a reduced set of often highly sensitive features that allows damage detection to be more straightforward. In the case of bearings, the method has gained great acceptance due to its capabilities to diagnose damage in any of its elements. This can be explained in terms of the periodic nature of the defect interaction with the surrounding components. When the defect interacts with a specific bearing component it releases a short duration pulse that excites the natural frequencies of the bearing, thereby resulting in an increase in the vibrational energy. These vibration levels provide information about the defect presence and location on the bearing.

The damage presence in any of the bearing elements can be related to specific

frequencies denominated *characteristic damage frequencies*. These frequencies can be calculated and are based on the bearing geometry and the rotational speed, these frequencies are defined in Table 2.1.1 [33]:

Table 2.1.1: Roller bearing defect frequencies.

Element	**Symbol**	**Defect Frequency**
Cage	ω_c	$\frac{\omega_s}{2}(1 - \frac{d}{D}\cos\alpha)$
Ball spinning	ω_b	$\frac{D\omega_s}{2d}(1 - \frac{d^2}{D^2}\cos^2\alpha)$
Outer raceway	ω_{or}	$Z\omega_c$
Inner raceway	ω_{ir}	$Z(\omega_s - \omega_c)$
Rolling element	ω_{re}	$2\omega_b$

where ω_s represents the shaft rotation (rad/s), Z the number of rolling elements, D the pitch diameter, d the diameter of the rolling element and α the contact angle as shown on Figure 2.1.1.

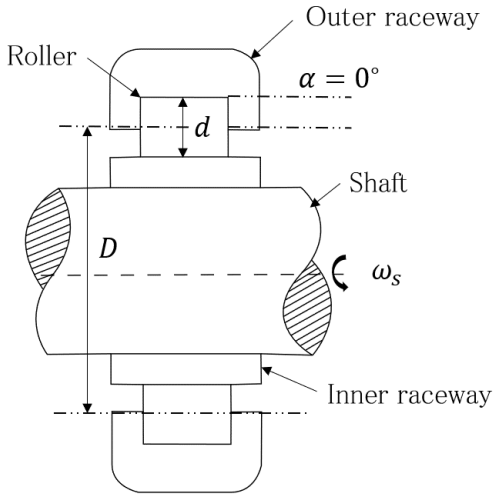


Figure 2.1.1: Roller bearing parts.

Normally these defect frequencies can be detected by applying the Fourier Transform to the vibrational signal but, in certain circumstances such as the existence of low signal to noise ratio in the vibrational signal, it may be difficult to enhance this

defect frequency thereby leading to the use of alternative methods to enhance or obtain this frequency values such as the signal average, bispectrum, bicoherence, power spectrum and enveloping techiniques [39–44].

2.1.2 Time-frequency/time-scale domain methods

Up to this point, the methods described previously are suitable for stationary signals, i.e. signals whose statistical parameters are independent over time. For non-stationary signals, such as continuous and transient signals, these methods are not adequate and more sophisticated analysis tools should be used. This type of analysis involves analysing the signal in a time-frequency domain where more detailed information can be provided in a two-dimensional domain. The most well-known methods are:

2.1.2.1 Short-Time Fourier Transform

The method allows the segmentation of the signal into small time-window slices and the analysis of the frequency content. The Short Time Fourier Transform (STFT) performs a time-localised Fourier transformation on a signal $y(t)$ using a window function $g(t-\tau)$, where τ is the window width which is independent from the time and frequency domain as given by the Equation 2.1.1:

$$STFT(\tau, f) = \int_{-\infty}^{\infty} y(t) g(t-\tau) \exp^{-2i\pi ft}\, dt \tag{2.1.1}$$

A disadvantage of this technique is that high resolution on both frequency and time cannot be achieved together. This means a trade-off between frequency and time resolution as the Heisenberg uncertainty principle states in Equation 2.1.2.

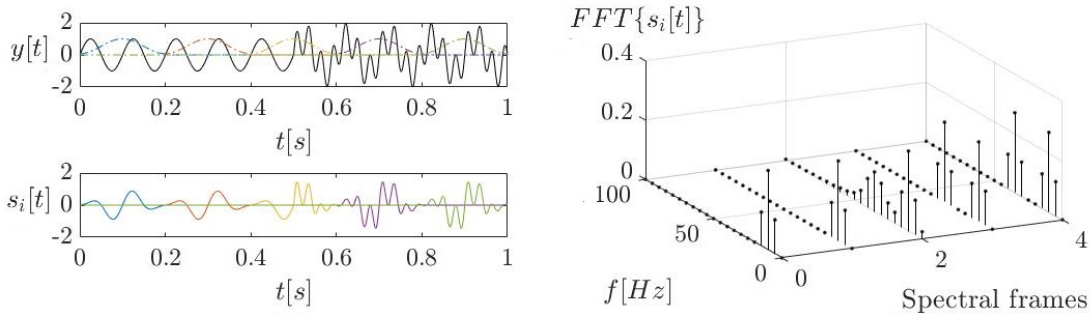
$$\Delta t \Delta f \geq \frac{1}{4\pi} \tag{2.1.2}$$

In practical terms, that means that an event cannot be localised both in time and frequency domain. This situation in terms of feature extraction will constrain the number of features that can be obtained from a signal with a specific frequency resolution Δf.

The STFT can also be understood as the FFT of the individual segments of the original signal multiplied by a window function, as shown in Figure 2.1.2a for a signal defined as:

$$y(t) = sin(2\pi f_1 t) + sin(2\pi f_2(t - t_a))H(t - t_a); \tag{2.1.3}$$

where, in this case, $f_1 = 10Hz$, $f_2 = 35Hz$, $t_a = 0.5s$. $H(t)$ represents the Heaviside function. As shown in Figure 2.1.2a, a second frequency component appears at 0.5 seconds. By using the STFT it is possible to represent this frequency component in a specific time frame. The original signal $y(t)$ was segmented in 5 different parts by multiplying it with a Hanning window without overlapping each other leading to a set of signals defined as $s_i(t)$ for $i = 1, 2, ..., 5$. An FFT was applied to each segment leading to the spectral frames shown in Figure 2.1.2b where it may be seen that a second frequency component f_2 appears in the third spectral frame corresponding to 0.5 seconds of the signal. Another disadvantage is that this method does not have an inverse transformation, but for the context of damage detection, the only interest is to build a set of features that represents unambiguous damage and undamaged states.



(a) Time Domain

(b) Frequency Frames

Figure 2.1.2: STFT applied to a time-domain signal with a Hanning window at 0% overlap.

2.1.2.2 Wavelet Transform

The Wavelet Transform approach is a time-scale method, in contrast to the STFT. It overcomes all the limitations of the STFT due to the addition of the scaling and shifting parameters and the possibility of applying the inverse transform. The

Continuous Wavelet Transform (CWT) can be defined as:

$$CWT(s,\tau) = \frac{1}{|s|^{\frac{1}{2}}} \int_{\infty}^{\infty} x(t)\overline{\Psi}(\frac{t-\tau}{s})dt \tag{2.1.4}$$

The term $\overline{\Psi}$ refers to the complex conjugate of the wavelet function which has two parameters, s and τ, denominated as the scaling and shifting parameter respectively. These parameters dilate and move the wavelet function along the time domain respectively. Additionally, the wavelet function should meet a few conditions such as having a zero mean, a finite power and meet the admissibility criteria to be constructed and used in the transformation. An approach in CM used the CWT to extract features for novelty detection for roller bearings such as [45]. Their work extracted features from the vibrational signature and used the CWT to analyse the signal self-similarity by calculating the variance of the CWT coefficients.

2.1.2.3 Wigner-Ville Distribution

Also based on the Fourier transformation, this technique transforms the auto-correlation function of a signal $x(t)$ and generates contour plots of the energy distribution on the time-frequency domain. The Wigner-Ville Distribution (WVD) can be defined with the following equation [46]:

$$WVD(t,f) = \int_{-\infty}^{\infty} x(t+\frac{\tau}{2})f^*(t-\frac{\tau}{2})\exp^{-2i\pi ft}\,dt \tag{2.1.5}$$

The use of the WVD has been explored as a potential feature for damage detection. Initial exploration in this field was attempted by Forrester *et al.* [47] to analyse the faults generated in a helicopter gearbox. Application in the determination of tooth failure in gearboxes was later studied by Staszewski *et al.* [48]. The use of the WVD in these studies has shown to generate enough visual patterns on the time-frequency domain to perform pattern recognition.

2.1.3 Envelope analysis using the Hilbert Transform

An important signal processing technique involves the envelop analysis of signals. A common reason to use the Hilbert Transform (HT) is to enhance the signal and remove the high frequency oscillations that modulates the signal. This can be either

be approached by using a low-pass filter or a signal enveloping technique. The later technique can be seen as a form of low-pass filter as it leaves the "slowest" component of the signal and removes the "fastest" that modulates the signal. Figure 2.1.3 shows a frequency modulated signal defined as:

$$y_{mod}(t) = \sin(2\pi f_1 t)\sin(2\pi f_2 t) \tag{2.1.6}$$

where f_1 and f_2 represent the "slowest" and the "fastest" frequency component in the modulated signal respectively. In this case, the calculation of the envelope removes the faster oscillations at f_2, that corresponds to the second term in Equation 2.1.6.

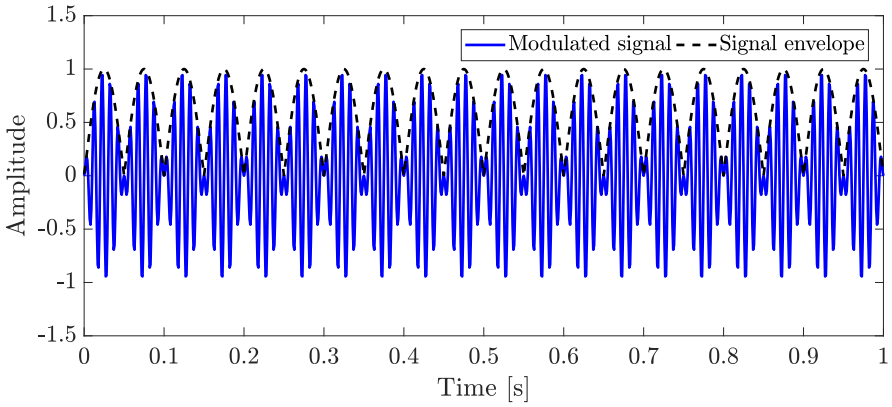


Figure 2.1.3: Signal demodulation using Hilbert Transform.

One of the most standard methods for this purpose involves the calculation of the Hilbert Transform of a signal. It is important to mention that there are other classic methods to perform envelope extraction by using RMS and peak-to-peak values but normally leading to distorted or overly smoothed envelopes [49]. The Hilbert Transform of a time domain signal $u(t)$ can be defined as [44, 50]:

$$\tilde{u}(t) = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{u(\tau)}{t-\tau}d\tau \tag{2.1.7}$$

This integral is a linear operator and leads to a time-domain representation of the transformed function. The transformation becomes useful for envelope analysis as it finds the analytical representation of a signal $u(t)$. An analytical signal can be represented as a complex function with a real and imaginary part defined as:

$$\tilde{u}(t) = u(t) + \hat{u}(t)j \tag{2.1.8}$$

Equation 2.1.8 shows the relationship between the analytical signal $\tilde{u}(t)$ and the Hilbert Transform of the signal $\hat{u}(t)$. It can be seen from Equation 2.1.8, that the

projection of the analytical signal $\hat{u}(t)$ onto the imaginary plane corresponds to the Hilbert Transform of the function $u(t)$. A better understanding can be attained by representing Equation 2.1.8 in polar coordinates as:

$$| A(t) | = \sqrt{u^2(t) + \hat{u}^2(t)} \tag{2.1.9}$$

$$\theta(t) = \arctan\left(\frac{\hat{u}(t)}{u(t)}\right) \tag{2.1.10}$$

where the terms $A(t)$ and $\theta(t)$ represent the envelope and the phase of the analytical signal respectively. So far the concept of envelope analysis has been explained but not how the Hilbert Transform can be calculated in practise. A practical approach uses the Fourier Transform to calculate the Hilbert Transform conjointly with some properties. This approach implies a relationship between the Fourier Transform of the signal $\tilde{u}(t)$ and its original form $u(t)$. This relationship is defined as:

$$\mathcal{F}\{\tilde{u}(t)\} = \mathcal{F}\{u(t)\}\,(-jsgn(\omega)) \tag{2.1.11}$$

Equation 2.1.11 shows that the frequency domain representation of $\tilde{u}(t)$ is equivalent to apply the Fourier Transform to the real function $u(t)$ and shift the phase components by $-\pi/2$ and $+\pi/2$ for the positive and negative frequency components respectively.

2.2 Acoustic Emission and signal processing techniques

The AE phenomenon occurs due to a rapid release of energy by an alteration in the material structure such as dislocations, crack generation and propagation. These high-frequency elastic waves travel as discrete acoustic packets until reaching the surface where they can be detected. This technique is widely used as a NDT to determine damage and also used to localise damage in structures and machinery.

The primary source of AE burst during plastic deformation is attributed to the dislocation motion in the materials and has been proven suitable to monitor these responses [51]. This effect was systematically studied by Josef Kaiser in 1950 were different types of materials were loaded and their response recorded. This experiment led him to discover what is known as the Kaiser effect [52] which describes the

occurrence and absence patterns of AE events when a material undergoes cycling loading conditions.

A basic visualisation of an AE burst can be seen as a particular event inside the material which releases a sudden burst of energy leading to a transient stress response that propagates accordingly to the geometry of the specimen (non-bounded or bounded media) where the pulse is generated, thereby allowing the generation of different wave modes. As the waves propagate through the media, surface displacements can be captured using piezo-electric transducers that convert the displacement response into a voltage response.

Regular AE sensors are mainly composed with a Lead Zirconate Titanate (PZT) patch as shown in Figure 2.2.1. This patch is normally enclosed in a metallic case together with a damper and additional signal condition circuitry. As the sensor is supposed to be reused, a wear plate is attached to the PZT element, thus allowing the safe removal of the sensor. The attachment between the wear plate and the structure is normally performed with an adhesive bonding that also acts as a coupling mechanism.

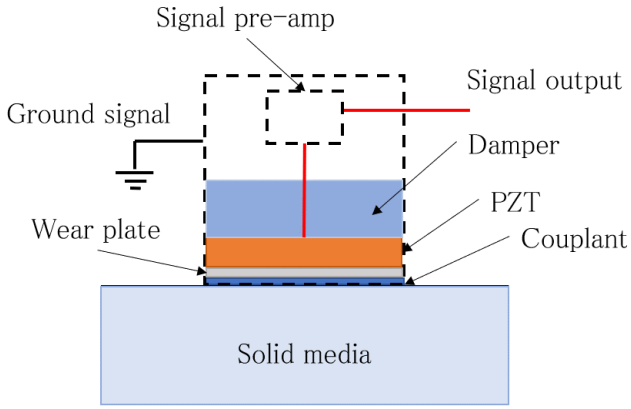


Figure 2.2.1: Diagram of an AE sensor bonded to a component.

Several features in the AE signals can be utilised for SHM/CM purposes. In roller bearings, a localised defect type can be expected due to RCF and traditional vibration methods may be unsuitable due to lower frequency bands tending to become contaminated with noise coming from different parts of the machine [53]. In such cases, AE can be utilised to monitor material degradation and as a tool to monitor faults such as subsurface cracks generated due to RCF. For this reason, the AE method has been successfully utilised to detect and localise defects on roller bearings as considered in previous studies [54–57].

2.2.1 AE signal features

Similarly, features can be extracted for AE signals. These features, as shown in Figure 2.2.2, describe the time-domain signal characteristics by using signal parameters. The most important features used so far in CM are described in Table 2.2.1.

Table 2.2.1: AE time domain features.

Parameters	**Description**	**Units**
Amplitude	Represents the maximum (positive or negative) peak	[V/dB]
RMS	A root mean square from the voltage response of the signal	[V]
Energy	Rectified AE response and integrated over time	[aJ]
Ring Down Counts	The number of times the signal crosses a predefined threshold	–
Events	The moment when the AE is generated	–
ASL	The average signal level	[V/dB]
AE duration	Describes the time length or duration of the AE burst	[s]
Signal Envelope	Curve outlining the signal's extreme values	[V]

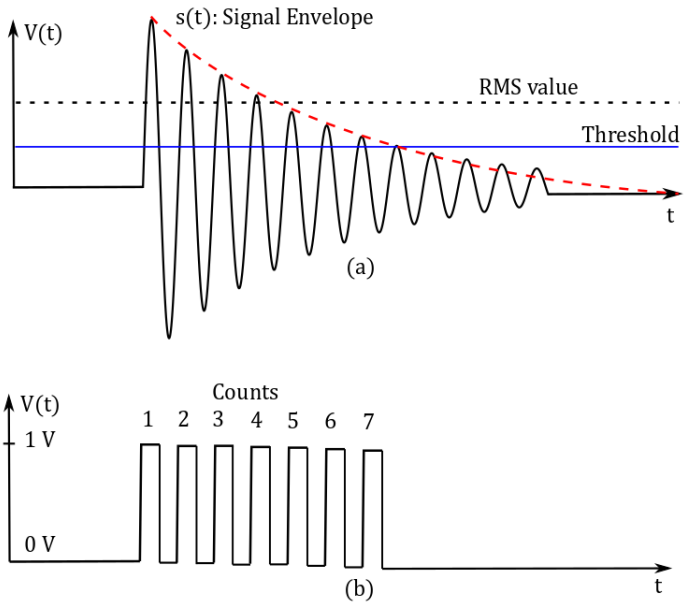


Figure 2.2.2: Time-domain AE features.

2.2.2 The dispersive behaviour of Lamb waves

The AE belongs to the transient signal class type and, in the case of wave propagation on plates or thin structures, will lead to the formation of multiple propagation modes. This will affect two aspects of AE source localisation; first, a threshold approach will find an onset independently of the propagation mode, therefore, making TOA estimation inaccurate and second, the multiple dispersive modes propagate at different speeds as will be shown later. As opposed to bulk waves, such as longitudinal and transverse waves, where a constant propagation speed can be considered, wave propagation on structures commonly occurs in bounded media that leads to variable speed propagation constants. A model of this dispersive behaviour is described by Lamb's equation as [58]:

$$\frac{\tan(qh)}{\tan(ph)} = -\frac{4k^2pq}{(q^2+k^2)^2}^{\pm 1} \tag{2.2.1}$$

where the exponent ± 1 will be $+1$ for a symmetric mode and -1 for an anti-symmetric mode respectively. The term k which is a function of the angular frequency ω is defined as:

$$k = \frac{\omega}{C_p(\omega)} \tag{2.2.2}$$

The term $C_p(\omega)$ represents the phase velocity and is also a function of the angular frequency. The group velocity $C_g(f)$ can be obtained from the phase velocity $C_p(f)$ using the following relationship:

$$C_g(f) = C_p(f)^2 \left(C_p(f) - f \frac{dC_p(f)}{df} \right)^{-1} \tag{2.2.3}$$

Equations 2.2.2 and 2.2.3 show a dependence of the frequency value and show that the wave propagation speed in bounded media is not unique, and it is governed by a complex relationship. Figure 2.2.3 shows the numerical solution of Equation 2.2.1 for the zero-order symmetric and anti-symmetric modes for a 17 mm Aluminium plate, a longitudinal C_L and transverse C_S wave propagation of 6300 and 3130 m/s were used to solve this equation. As seen, the fastest propagation speed occurs at frequencies lower than 300kHz then they abruptly change and stabilise at 2923 m/s for a Poisson's ratio ν of 0.35. This limit is known in literature as the Rayleigh wave speed and can be approximated as [59]:

$$C_R \approx C_S \left(\frac{0.862 + 1.14\nu}{1 + \nu} \right) \tag{2.2.4}$$

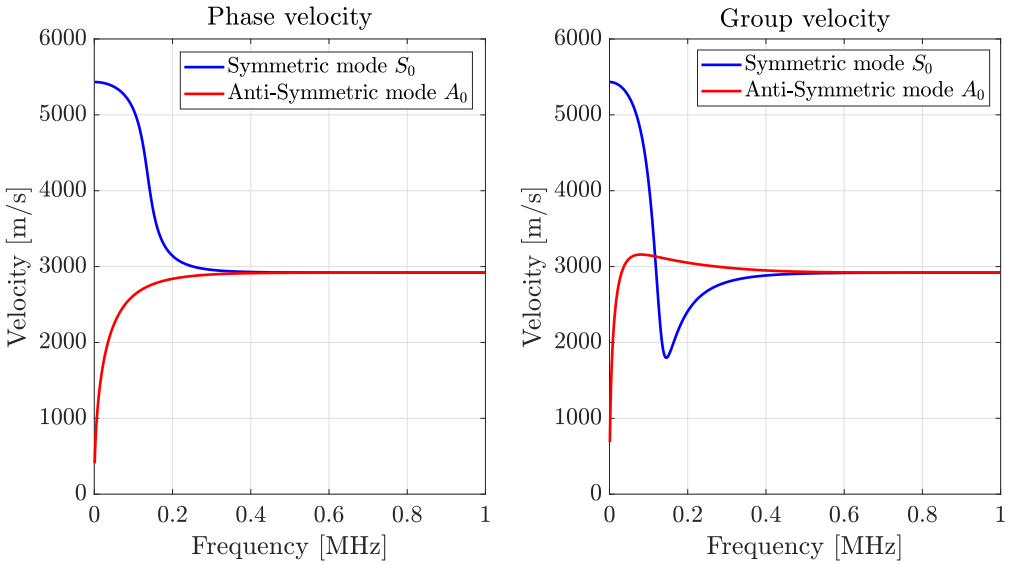


Figure 2.2.3: Zero-order symmetric and anti-symmetric modes for an aluminium plate of 17mm thickness.

It is important to mention that an infinite number of propagation modes exist in a bounded media, but contrary to the zero-order modes, these modes do not appear at all the frequencies. Additionally, zero-order modes contain a great portion of the signal energy, this characteristic makes them suitable for analysis and to estimate TOA features.

2.2.3 A time-frequency approach for AE signal processing

Frequency domain provides information about the characteristic defect frequencies. This can also be extended to a time-frequency domain where a particular defect frequency can be localised in the time domain by analysing a set of AE pulses. AE bursts have been subjected to studies involving time-frequency decomposition using the CWT approach. The application of this method has shown great value in SHM when it was used to identify propagation modes in structures via the use of the Gabor wavelet [60]. This method states that the CWT of a function of two harmonic waves using the Gabor wavelet allows to identify the frequency component ω_c by finding the maximum values at the parameters a and b of:

$$\mid CWT(x,a,b) \mid = \sqrt{2a} \mid \hat{\psi}(a\omega_c) \mid (1+\cos(2\Delta k x - 2\Delta\omega b))^{0.5} \tag{2.2.5}$$

where $\hat{\psi}$, $\Delta\omega$ and Δk represent the complex conjugate of the Gabor wavelet ψ, the frequency and the wavenumber difference of two harmonic waves respectively. This approach has also shown capabilities for localising damage in simple structures such as plates by identifying the dispersive modes contained in the AE burst. The technique is also known as Single Sensor Modal Analysis (SSMA) and allows the localisation of an AE source using only one single sensor by identifying the time difference between the zero-order symmetric and antisymmetric modes [61]. This is shown in Figure 2.2.4 where the zero-order symmetric S_0 and antisymmetric mode A_0 are visible using a time-frequency representation on a simulated burst using a Pencil Lead Break (PLB) technique known as Hsu-Nielsen source [62].

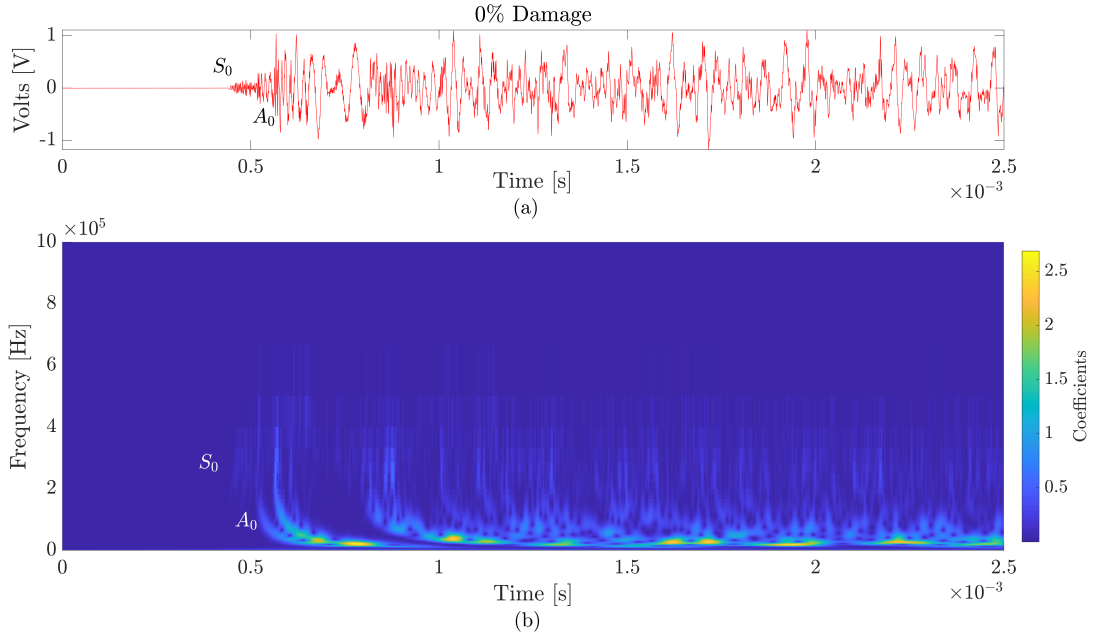


Figure 2.2.4: A Wavelet Transform on AE signal obtained from a PLB excitation on a 5 mm Alumimun plate. A zero-order symmetric S_0 and antisymmetric A_0 mode can be seen on time and time-frequency domain.

2.2.4 AE analysis for roller bearings

The use of AE to monitor roller bearings has been extensively studied and this has demonstrated its capability for detecting damage with results comparable to vibration analysis [63, 64]. This method has also demonstrated great capabilities to detect damage in bearings from sensors installed at a larger distance. Li *et al.* [65] showed that AE signals had a higher Signal-to-Noise ratio (SNR) than vibration signals. Defect frequencies were detected at distances greater than 0.3 meters away from the bearing rig using AE sensors.

In theory, AE can be used in a similar way to a vibrational signal as the defect frequency associated with a specific bearing component is modulated in the signal content, allowing the detection of the defect characteristic frequency. These frequencies are determined with the equations described in Table 2.1.1. For the data obtained from a roller bearing shown in Figure 2.2.5, each burst is spaced every $66\mu s$ which is equivalent to 15Hz. This AE data was obtained from a bearing rotating at 100 RPM and a seeded defect on the inner raceway. The inner raceway defect

frequency was calculated using the equations in Table 2.1.1 giving a value of 14.65 Hz very close to the burst occurrence frequency. In real AE data some variations are expected to occur due to mechanical clearances and roller slipping. Furthermore, secondary AE bursts may also appear due to loose components or friction between surfaces.

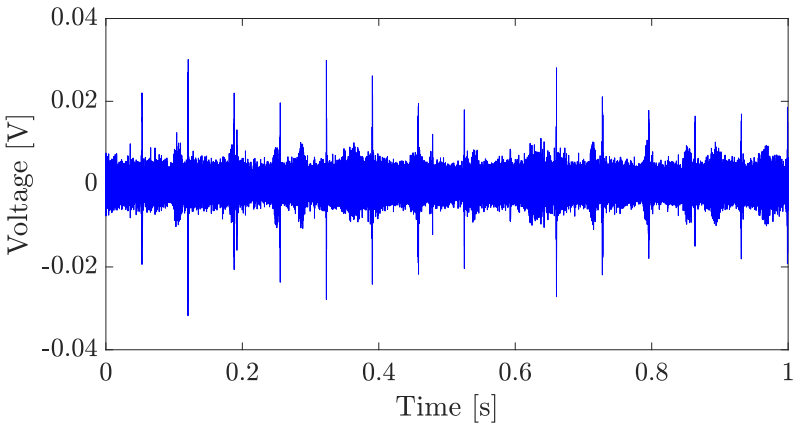


Figure 2.2.5: AE bursts from a bearing in a 1 second interval, the data shows 15 bursts that correspond to an inner raceway defect frequency of 15 Hz at 100RPM.

Even though the defect frequency can be extracted from a raw AE signal, in some cases an excessive amount of noise can difficult the identification of the defect components in the frequency domain. In this case, the characteristic defect frequency in Figure 2.2.5 was estimated by low-pass filtering the signal before applying an FFT algorithm. Another approach is to use the Hilbert Transform to generate an envelope of the signal and post process it using the Fast Fourier Transform to determine the periodicity of the AE events [44, 50, 66]. Even though it is possible to obtain the characteristic defect from the AE signal, one disadvantage arises from the relatively high sampling frequency required to record the signal therefore making this method unsuitable for long term monitoring purposes due to the large amount of data to be stored and processed.

2.3 A machine learning approach for damage detection for roller bearings

It is necessary to address an important aspect of this work, namely, damage detection. A damage localisation stage cannot be performed systematically without first performing a damage detection stage. This first stage in CM involves answering the question of whether or not a machine has potential damage or not. Sometimes damage detection can be achieved by mere intuition such as observables cracks on a structure/component but, in practical applications, this will often fail and more robust methods must be used by analysing features obtained from signals that are sensitive to a particular failure mode such as the vibration and AE signals. Novelty detection is based on collecting these features and comparing them with baseline features in order to detect any statistically significant deviation from normal behaviour. Even though this task may appear trivial, in real conditions there is no single normal behaviour. Features used to build a baseline may contain very different patterns that all still represent normal conditions.

CM in WTGB has been widely studied in terms of trends of statistical parameters and signal features of vibration and AE signals as described in early sections. Even though the variation of these parameters has proven useful for industrial applications, the condition is mostly based on the interpretation of these features using engineering judgement. These parameters, in order to be useful for a pattern recognition algorithm, are combined together as a feature vector and may be represented in a feature space. A representation of features in a two-dimensional space is shown in Figure 2.3.1 for visualisation purposes. A normal condition denoted as NC_1 is graphically represented as blue dots in a two-dimensional feature whereas damage conditions DC_1, DC_2 and DC_3 are represented as red dots. The main aim of any fault detection procedure is to correctly recognise the machine condition based on these features. Clearly, this simple representation can be classified by mere observation/intuition by a human operator but in reality the feature space can be defined with an arbitrary dimension thus limiting intuitive discrimination between different conditions. This is why a machine learning approach is useful to perform a reliable fault detection in CM.

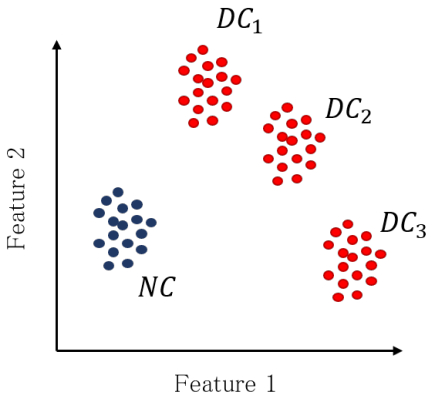


Figure 2.3.1: A two-dimensional feature space where a normal condition NC and damage features $DC1$, $DC2$, $DC3$ are graphically represented.

Earlier works in this field of damage detection in roller bearings have been conducted by Li [67] where the fault detection scheme was performed using short-time signal processing techniques such as the short-time energy function, the short-time average zero-crossing rate and median smoothing. The classification algorithm was based on calculating a discriminant function that separates the data between a faulted and unfaulted condition. Rojas *et al.* [68] applied Support-Vector Machines (SVM) to detect damage on bearings where spectral lines and statistical data were used as features for the classification algorithm. PCA based techniques were also reported being used for rotative machinery, He *et al.* [69] proposed using PCA as a feature reduction technique in order to allow to build a reduced dimension damage sensitive feature. Pirra *et al.* [70] used PCA in order to remove the environmental effects on the bearing damage prediction. The onset of damage was determined by defining a Novelty Index in terms of the residual error of the compressed signal. Neural networks were also successfully implemented, Kateris *et al.* [71] used statistic parameters such as (skewness, kurtosis, etc) and additionally included a line integral feature altogether with an Multilayer Perceptron (MLP). Their results showed that a good level of classification accuracy was achieved around 99%. Ziaja *et al.* [72] used a time-frequency approach to extract the damage features by using the Wavelet Transform approach conjointly with fractal signal processing techniques. These features were used to train a novelty detector based on an Auto-associative Neural Network (AANN) together with a novelty index. A more detailed review of machine learning techniques has been discussed in terms of applicability for SHM/CM [73].

In general all these methods mentioned above share the same schema, that is the damage detection stage normally follows the same workflow as described in Figure 2.3.2. Commonly the two main approaches adopted for damage detection are either analysing the vibrational or AE signals and use the statistical signal parameters as described in early section. These signals can be filtered or enhanced before extracting the features.

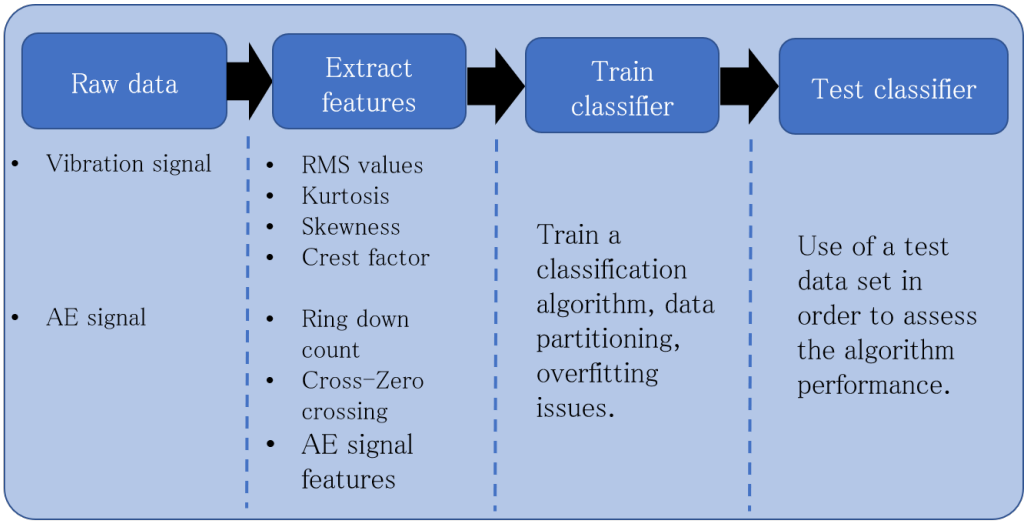


Figure 2.3.2: General damage detection schema.

The extracted features are ultimately used as a training/test set in the classification algorithm where the last step commonly involves stages such as overfitting and performance assessment.

Despite detecting damage, most of these techniques were used under relatively controlled conditions such as in the case of small scale bearings. Additionally, larger defects compared with the bearing size do not reflect the real conditions in a full scale bearing such as the case of a WTGB. Time and frequency domain features have seen difficult to analyse when an early onset of damage occurs, thus only being noticed when the damage progresses to a considerable extent [72, 74]. Additionally, mechanical noise expected on WTGB plays a crucial role in damage detection as it makes the patterns difficult to classify. This leads to an exploration of different methods that allow classifying the data in more realistic conditions.

2.4 Damage localisation using AE

In this section an introduction to the AE localisation techniques is presented. This section aims to provide the different state of art approaches and efforts to circumvent most of the challenges involving AE source localisation. It is important to notice that several methods have recently been successfully implemented in the field of SHM and can potentially be used in CM applications. Moreover, a review of the methods used for localising defects on bearings is presented in this section.

2.4.1 Damage localisation in structures

The next issue to be addressed is the location of the damage. An interesting aspect of this problem is that it has received a lot of attention in the field of SHM whilst current information about localisation techniques applied to roller bearings is relatively scarce. The problem of localisation in SHM has been extensively developed and it should be possible to use the same concepts to monitor specific components in rotative machinery by treating these components as structures subjected to quasi-static conditions.

Starting with the most simple and well-known localisation technique is the linear localisation method [75]. It consists of an arrangement of two sensors separated by a distance D as shown in Figure 2.4.1. The AE source position x is located along a single dimension by measuring the TOA difference Δt_{12} between sensors 1 and 2 and multiplying it with the propagation speed C defined as:

$$x = C\Delta t_{12} \tag{2.4.1}$$

This will probably be the most simple method found in literature [75] which has a limited set of applications due to one-dimensional source localisation capabilities.

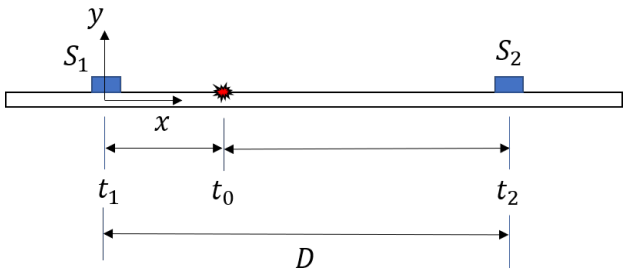


Figure 2.4.1: Linear AE localisation using two sensors.

Geometrical methods consist in localising the AE source by using trigonometrical relationships between a hypothetical source and the sensors surrounding it by measuring the time difference between sensors, the first scenario described using Equation 2.4.1 can be extended to AE source localisation on a two-dimensional plane using polar coordinates as [76]:

$$R = \frac{1}{2}\frac{D^2 - \Delta t_{12}^2 C^2}{\Delta t_{12} C + D\cos\theta} \tag{2.4.2}$$

where θ represents the angle and R represents the radial distance from the source to the reference sensor 1 as shown in Figure 2.4.2.

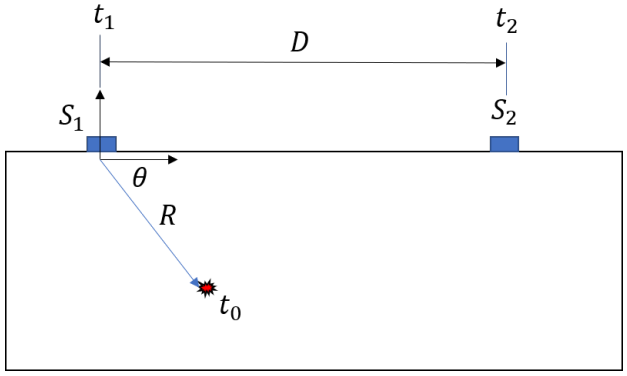


Figure 2.4.2: Linear AE localisation using polar coordinates.

The only problem with this approach is that localising a source in a 2-dimensional plane will require at least a third sensor in order to eliminate the angle variable, as Equation 2.4.2 requires the angle information to give the radial distance to the source. This will lead to the implementation of triangulation techniques where at least three sensors are used to attempt AE source localisation.

The first attempt to localise AE source with triangulation techniques can be found in the analytical approach developed by Tobias [77] whose work proposed an analytical

treatment for the 2-dimensional localisation problem by calculating the AE source using the TOA difference between each sensor. The source position was given based upon the solution of a set of circle equations shown as:

$$x^2 + y^2 = r^2 \tag{2.4.3}$$

$$(x - x_1)^2 + (y - y_1)^2 = (r + \delta_1)^2 \tag{2.4.4}$$

$$(x - x_2)^2 + (y - y_2)^2 = (r + \delta_2)^2 \tag{2.4.5}$$

where the source position is defined in a plane by the coordinates x, y, the sensor position x_i, y_i and the term $r + \delta_i$ can be understood as the radial distance from the ith sensor to the source as shown in Figure 2.4.3.

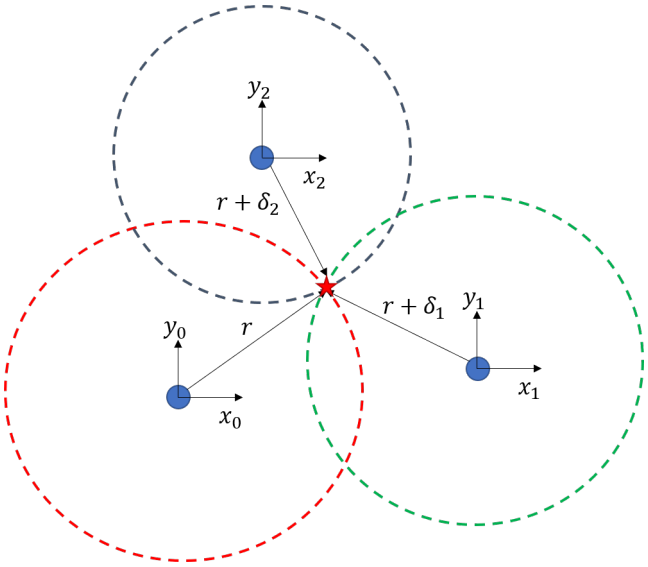


Figure 2.4.3: AE source triangulation using Tobias approach.

The advantage of this analytical method is that it yields an exact solution of the AE source by directly evaluating the derived expressions but has been shown to be susceptible to give multiple solutions and therefore a fourth sensor should be included to improve the ambiguity of the results. The method was also shown to be effective by localising sources within the area covered by the sensors whereas regions close to the sensors shown areas of ambiguity. This work was extended later by Asty [78] to include AE source localisation in spherical surfaces highlighting the same issues with ambiguity in the source location. A similar approach was developed by Barat *et al.* [79] where an analytical approach was developed for cylinders.

Axinte *et al.* [80] implemented an analytical triangulation technique to localise sources of uneven events such as discontinuities, plucking and smearing in machining. Their

localisation approach consisted of a set of spherical equations that were numerically solved by obtaining the TOA differences at each sensor. The source was obtained by the intersection of these spheres similar to the geometrical method proposed by Tobias [77] using circle equations. This technique showed potential to localise uneven defects on workpiece defects and its concepts allow the extension of AE source localisation into a 3-dimensional scenario.

Jingpin *et al.* [61] performed an AE source localisation approach by exploiting the dispersive characteristic of guided waves using AE modal analysis. Their localisation approach involved decomposing the AE signal using the Wavelet Transform and obtaining the TOA of the arrival of the dominant modes as t^l, t^k and their corresponding group velocities C_g^l, C_g^k at specific frequencies f_1, f_2 as:

$$d = \frac{t_1^l(f_1) - t_2^k(f_2)}{{C_g^l}^{-1}(f_1) - {C_g^k}^{-1}(f_2)} \tag{2.4.6}$$

This method is able to calculate the distance d from the AE source to the sensor but is restrained to a narrow set of practical applications as it does not give the orientation information of the AE source. Despite the simplicity of the analytical methods and ability to give relatively accurate results, the geometrical models mentioned above involve assumptions such as a constant propagation speed and isotropic properties. Kundu *et al.* [81] tried to overcome this issue by stating its localisation approach as an optimisation problem of the form:

$$\min_{x,y} E(x,y) = \min_{x,y} f(x, y, x_1, y_1, x_2, y_2, ..., ..., x_N, y_N, t_1, t_2, ..., t_N) \tag{2.4.7}$$

where the objective function E was defined in terms of the AE source positions x, y, the sensor positions x_i,y_i and the term t_i defines the TOA at each sensor. In order to account for anisotropy, the propagation speed C was defined in terms of the wave propagation direction as:

$$C(\theta) = C_0(1 + 1 \times 10^{-3} \mid \theta \mid) \tag{2.4.8}$$

where θ is the direction of wave propagation and C_0 is the propagation in the horizontal direction. The anisotropy was modelled by assuming a 1% increase on the wave propagation speed for every 10° increment in cold-rolled steel. This technique showed superior performance compared with a standard triangulation technique under of the isotropic assumption and also shown agreement with the anisotropic

assumption. This objective function was later modified to overcome some singularities during the optimisation and tested with a graphite-epoxy composite plate [82]. The same technique was implemented by Kundu *et al.* [83] to localise AE sources on a stiffened anisotropic plate. Kundu *et al.* [84] developed a localisation technique based on an array of six sensors arranged in two clusters of three sensors, as the aim was to implement this technique to anisotropic plates the wave propagation speed was expressed in terms of the TOA terms as:

$$C(\Delta t_{ij}, \Delta t_{ji}) = \frac{d}{\sqrt{\Delta t_{ij}^2 + \Delta t_{ji}^2}} \tag{2.4.9}$$

where d represents the distance between the sensors in the cluster. The position of the sensor was obtained by solving this set of equations defined as:

$$\tan\theta_a = \frac{y_a - y}{x_a - x} = \frac{\Delta t_{ai}}{\Delta t_{aj}} \quad ; i \neq j \tag{2.4.10}$$

$$\tan\theta_b = \frac{y_b - y}{x_b - x} = \frac{\Delta t_{bk}}{\Delta t_{bl}} \quad ; k \neq l \tag{2.4.11}$$

where the position of the source was calculated in terms of the angles formed between the origins of the two clusters a and b. These equations were calculated using the TOA between the sensor in the origin of the cluster (a and b) and the surrounding sensors (i, j and k, l). The results showed good agreement with the impact locations on the plate on both isotropic and anisotropic conditions. Aljets *et al.* [85] used an array of three sensors on a Carbon Fibre Reinforced Plastic (CFRP) panel to localise the AE source orientation respect to the array with their approach consisting of an iterative algorithm that exploited the dispersive characteristics of Lamb waves.

McLaskey *et al.* [86] proposed the beamforming technique to monitor large concrete structures by using an array of 4–8 AE sensors on a simulated bridge ramp of $0.3 \times 4 \times 5$ meters. McLaskey *et al.* [86] shown that the method localised the AE events within a tolerance of 14° degrees showing its applicability to localise sources even at larger distances up to 4 metres where severe signal attenuation is expected due to the large size of the structure.

2.4.2 Damage localisation in complex geometries

Until now, the localisation techniques mentioned above were implemented in relatively simple geometries where the waves can be assumed to propagate in a straight line and where there are no discontinuities in the propagation media such as holes, voids, etc. The problem of complex geometries arises as the analytical formulation becomes difficult due to the complex wave propagation paths. Moreover, the uncertainty of wave propagation speed values makes this type of localisation task even more difficult, without mentioning possible anisotropic characteristics of the propagation media. Baxter *et al.* [87] proposed a method where a TOA difference map was generated for each pair of sensors. The maps worked as lookup tables that associated each measurement to a specific position. The Delta T method can be summarised as:

1. *Determine the area of interest*: The area where the hypothetical AE burst can potentially be generated is identified.

2. *Construct grid*: The area of interest is then discretised with a predefined spatial resolution. The smaller the grid the better the resolution that can be obtained.

3. *Conduct Hsu-Nielsen excitations*: Each node on the grid is then excited using a pencil lead break excitation on a specific location of the structure.

4. *Calculate Delta T maps*: The time difference between each pair of sensors is calculated for each node in the grid.

5. *Compare actual data*: The data obtained from the AE sensors can be compared with the Delta T maps recorded from the previous step.

The advantage of this method is that it can locate AE sources by just comparing the data obtained from the maps obtained from the excited nodes. The accuracy can also be improved by performing an interpolation procedure. The method was validated on an aircraft component with several holes and thickness variations and also against a triangulation method [87]. The method showed an accuracy improvement of 1.77% against the traditional geometrical method. The main advantage of this method is that it is completely independent of the wave speed propagation and only depends on the initial sensor configuration, therefore the prediction will only be valid for a

specific sensor position. The Delta T method was also applied for localising AE sources on a composite sample subjected to a fatigue test by Eaton *et al.* [88]. The method was validated using the Thermoelastic Stress Analysis (TSA) that located the damage in an agreement with the Delta T predicted sources.

2.4.3 Damage localisation in bearings

In the context of roller bearings, damage localisation is a relatively difficult task as it deals with many of the problems and limitations encountered in AE source location techniques. The main reason that such tasks are challenging is due to these mechanisms generally being geometrically more complex than structures (e.g. holes, chamfers, etc.). Sometimes they are composed of elements with different materials and shapes. Moreover, clearances can also disrupt the wave propagation path and multiple components are constantly interacting with their surrounding boundaries.

The first localisation in roller bearings using AE signals was performed by Rogers [89]. That investigation attempted to localise defects in anti-friction slew bearings mounted on slow-speed cranes. Two AE sensors were installed on the inner side of the bearing separated by 180 degrees from each other. The AE source was localised using a linear localisation approach by counting the number of AE bursts captured around the bearing. In the case of a localised defect, its position was determined by an increase of AE bursts at a particular position of the bearing. Unfortunately, in this research, the burst distribution along the bearing resulted in a uniform distribution and was attributed mostly to secondary AE sources generated by the interaction of different elements around the bearing. Yoshioka and Fujiwara [35] developed a localisation technique for a thrust bearing by using a magnetic detector and a single AE sensor. The position of the AE source was calculated by counting the number of gears on a retainer with the magnetic sensor. Their result showed an agreement with the defect position after further inspection. An increase in the AE count at the predicted position corresponded with the actual defect position on the bearing raceway. A few aspects such as the method resolution seem to be limited by the number of teeth on the gear retainer. It is important to mention that this method, despite being relatively simple, was not dependent on physical properties in the material, such as wave propagation speed.

Elforjani and Mba [36] implemented a linear localisation approach on a slow speed thrust bearing. The method consisted of four AE sensors mounted on the bearing ring, and the value was analytically calculated using the time differences between the sensors and using the wave propagation speed of the zero-order symmetric mode obtained from the Lamb's characteristic wave equations. Further work developed by Eftekharnejad *et al.* [37] extended this work by demonstrating the technique on high-speed thrust bearings.

2.4.4 A machine learning approach for AE source localisation

The ML approach has been recently implemented as a solution to overcome the problems with geometrical models, complex geometries, the dispersive nature of waves and anisotropy in materials. Recalling from the previous geometrical model approaches developed so far, specifically the ones that stated the localisation as an optimisation problem in the form of Equation 2.4.7. This may be restated in another form in which is desired that the values of the positions be found that minimise the following expression:

$$\min_{x,y} E(x,y) = \sum_{ij} \left(\Delta t_{ij} - \frac{\mid \langle x,y \rangle - \langle x_i,y_i \rangle \mid - \mid \langle x,y \rangle - \langle x_j,y_j \rangle \mid}{C} \right) \tag{2.4.12}$$

where the term $\mid \langle x,y \rangle - \langle x_i,y_i \rangle \mid$ represents the Euclidean distance between the AE source and the sensor positions [90]. A drawback of this method is that it is not always possible to find a global minimum thus leading to ambiguous results. Furthermore the term C is a defined constant whereas in the case of plates the propagation speed is dependent on the mode of propagation and its frequency. To make the situation worse, all these methods have been proposed for simple geometries and arriving at an analytical expression is not a trivial task. A closed-form solution is not possible, thereby leading to iterative algorithms such as the simplex [91] or genetic algorithm [92]. Geometrical methods are still severely constrained by model assumptions and geometrical complexity that makes them difficult to implement in practical applications.

An extension of Baxter's work [87] was performed by Hensman *et al.* [90]. Their approach was based on the relationship between the TOA difference between each sensor and the position on the structure. The main goal was to learn the inverse relationship between the excitation points to their corresponding maps. The Delta T approach can be seen as a mapping defined as an equation whereas assuming an inverse relationship will lead to the mapping relation 2.4.13:

$$\langle \Delta t_{ij} \rangle \rightleftarrows \langle x, y \rangle \quad i \neq j \tag{2.4.13}$$

Learning the relationship between these variables will result in a regression problem in the context of machine learning. In the inverse mapping the TOA difference will be the input vector and the position values will be the target vector. Assuming an independent Gaussian distribution noise model for each measurement, the regression problem can be stated as:

$$t_n = y_n + \varepsilon_n \tag{2.4.14}$$

where the target vector can be expressed as $\mathbf{t} = (t_1, t_2, ..., t_N)^T$, the model output term $\mathbf{y} = (y(\mathbf{x}_1), y(\mathbf{x}_2), ..., y(\mathbf{x}_N))^T = (y_1, y_2, ..., y_N)^T$ and input vector are defined as $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)^T$. Their approach was to use the map generated in Equation 2.4.13 and learn this relationship by treating it as a regression problem by using a Gaussian Process (GP) approach. For a regression problem such as described in Equation 2.4.14, new target values $\mathbf{t}_\star$ are inferred from points $\mathbf{x}_\star$. Using the Gaussian Process, the joint Gaussian distribution can be represented as:

$$p\left(\begin{bmatrix}\mathbf{t}\\ \mathbf{t}_\star\end{bmatrix}\right) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{C}(\mathbf{x},\mathbf{x}) & \mathbf{k}(\mathbf{x},\mathbf{x}_\star)\\ \mathbf{k}^T(\mathbf{x},\mathbf{x}_\star) & c(\mathbf{x}_\star,\mathbf{x}_\star)\end{bmatrix}\right) \tag{2.4.15}$$

where the matrix $\mathbf{C}$ is composed of elements $C_{nm} = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm}$. The hyperparameter β defines the precision, δ_{nm} is the Kronecker delta function, the elements of the vector $\mathbf{k}$ are defined as $k(\mathbf{x}_n, \mathbf{x}_\star)$ for $n = 1, 2, ..., N$ and the function $k(\mathbf{x}_n, \mathbf{x}_m)$ represents the Kernel function that gives a measurement on how the input vectors $\mathbf{x}_n, \mathbf{x}_m$ are similar. In their study, a Radial Basis Function (RBF) kernel was used, where its parameters were optimised by a cross-validation procedure.

The distribution of the predicted values $\mathbf{t}_\star$ can be defined as a conditional distribution of $\mathbf{t}$ as:

$$p\left(\mathbf{t}_\star \mid \mathbf{x}_\star, \mathbf{x}, \mathbf{y}\right) = \mathcal{N}\left(\mu_{\mathbf{t}_\star|\mathbf{t}}, \mathbf{\Sigma}_{\mathbf{t}_\star|\mathbf{t}}\right) \tag{2.4.16}$$

$$\mu_{\mathbf{t}_\star|\mathbf{t}} = \mathbf{k}^T\mathbf{C}^{-1}\mathbf{t} \tag{2.4.17}$$

$$\mathbf{\Sigma}_{\mathbf{t}_\star|\mathbf{t}} = c - \mathbf{k}^T\mathbf{C}^{-1}\mathbf{k} \tag{2.4.18}$$

Their localisation approach was validated on a perforated plate where clearly the waves will propagate in complex directions. In comparison with conventional node excitation using the H-N sources, each node was automatically excited using a high-intensity laser. Once trained, the model from their experimental mapping resulted in a localisation error of 8mm showing a relatively high localisation accuracy.

Al-Jumaili *et al.* [93], in an attempt to implement the Delta T method, automatically used a classification approach to build the training data by applying an unsupervised clustering algorithm. Hierarchical clustering was used to group events based on a correlation coefficient. Their algorithm was summarised as follows:

1. Assign an AE event to its own cluster.
2. Compute the distance between clusters.
3. Reduce the number of clusters by merging similar ones.
4. Repeat steps 2 and 3 until all items are in one single cluster.

The localisation stage was performed by minimising an error function similar to the one described in Equation 2.4.12 by replacing the analytical time difference with the map generated by the Delta T method as:

$$\min_{x,y} E(x,y) = \sum_{ij} \mid \Delta t_{ij} - \Delta t_{ij}^{map}(x,y) \mid \tag{2.4.19}$$

where Δt_{ij} represents the time difference calculated from each pair of sensors i and j and Δt_{ij}^{map} is the generated Delta T map. Even though this method is not fully a machine learning approach, it reduces the source localisation task from 8 hours to 18 seconds. This was mainly due to an improvement in selecting the best data for the training map with the data clustering approach.

2.5 AE sensor self-assessment using a machine learning approach

SHM/CM relies on sensor measurements to determine the condition of a structure. These sensors are normally attached to the structure/component thereby becoming part of it and therefore require monitoring themselves. It is well known that any sensor malfunctioning will lead to unreliable measurements and consequently might lead to spurious diagnosis results regardless of the novelty detection/localisation algorithm. This issue ultimately leads to the assessment of not only the structure/machine integrity but also the integration of the sensors' health into the diagnostic scheme.

In the AE signals, progressive damage of the sensor might lead to changes in the AE signal features. Sensor debonding is a relatively common issue that affects the acquired response from the structure/machine. These changes are not only a concern for low-level detection stages but for subsequent levels such as the case of the localisation stage.

2.5.1 Sensor self-diagnosis

In SHM/CM, the processing chain for a single sensor may be summarised as shown in Figure 2.5.1 [73]. The processing chain shows a number of steps between the sensor and the decision stage. Evidently, any disruption between the sensing stage and the decision stage can lead to a decision fault and thus to undesirable outcomes such as false alarms or damage in the structure/machine going undetected. This, in the worst-case scenario, can potentially lead to a catastrophic event if no further action is implemented.

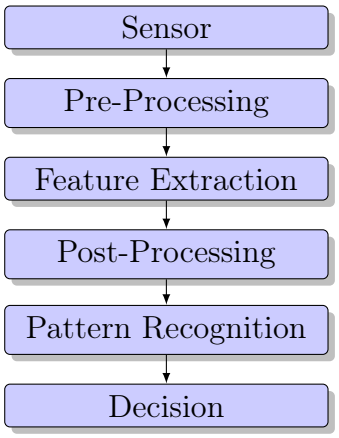


Figure 2.5.1: Single sensor processing chain.

In a practical situation, the sensors might also be considered as part of the structure or mechanism that they are attached and therefore requiring the sensor to be monitored together with the structure/machine. This can be achieved by including a robust health monitoring system that provides sensor validation capabilities to detect damage on the sensors and isolates them from the network [94]. The scheme, despite promoting an integral monitoring system, has not received enough attention in SHM applications, resulting in a wide research gap in this topic [95].

Sensor validation was initially considered in the field of control and chemical engineering by using models and sensor redundancy techniques. Friswell *et al.* [95] proposed two different approaches to the sensor validation problem by exploiting data redundancy on a cantilever beam. Their data redundancy approach was based on the assumption that the number of sensors used on a simulated cantilever beam was higher than the number of excited modes. They also proposed the use of modal participation factors and the subspace response of the system in order to discriminate between working and faulty sensors. Kerschen [96], on the other hand, proposed using PCA to identify, isolate and reconstruct a faulty sensor on a clamped beam.

Recent interest in Electromechanical Impedance (EMI) techniques used in SHM has resulted in sensor validation techniques in PZT elements [97]. The interest in this topic was principally due to the influences of the brittle behaviour of PZT elements and the effects of bonding degradation. Saint-Pierre [98] used Mason's [99] one-dimensional electromechanical model to understand bonding conditions in PZT elements. Giurgiutiu [100] showed that the imaginary part of the impedance can give

a good indication of debonding on PZT patches attached to structures. Their work showed that free vibration resonance during free conditions and its reduction during completely bonded conditions are unambiguous features for debonding detection. Overly [101] proposed the tracking of the imaginary impedance part and was able to iteratively detect faulty sensors in a PZT network. Their work also investigated PZT breakage and temperature variation. An algorithm for sensor damage detection was based on monitoring the imaginary part of the admittance signal and measure their respective standard deviations. An unhealthy sensor was recognised by identifying the sensor that contributed the most to the reduction of the standard deviation in the sensor network. The method identified the faulty sensors but it was constrained by the assumption that the network contained only half unhealthy patches.

2.5.2 Bonding influences on AE signals

Emerging NDT techniques have mainly used Lamb waves to perform structural assessment due to its damage sensitivity characteristics. The reliability of these measurements led to further investigations on the sensor integrity and the effect of bonding degradation [100, 101]. This certainly has an effect on signals propagating in a bounded media such as the case of AE signals, thus affecting their signal features and consequently the diagnostics obtained from these features. Recent studies have shown that bonding and sensor breakage have an effect on the AE signal. Giulia *et al.* [102] experimentally found a change in the amplitude, phase and energy loss of the piezoelectric voltage response. Park *et al.* [103] studied the effect of bonding on PZT sensors and its effects on Lamb waves at different propagation stages on a plate. Their results showed a clear change in the signal magnitudes and TOA between different bonding scenarios. Sathyanarayana *et al.* [104] proposed a methodology to detect different type of sensor debonding based on Time of Flight (TOF) and Maximum Amplitude Spectra (MAS) features. This has a direct implication for the methods used for damage localisation. Altered AE parameters such as changes in the amplitude and signal distortion can lead to erroneous signal features and TOA measurements and ultimately affecting the reliability of the localisation approach. Figure 2.5.2 shows the bonding effects on the AE signal. Influences in the time-domain features seem to decrease the signal amplitude and distort the signal. The

same result also seems to occur in the frequency domain. Three identical sensors equidistantly separated by 565mm were excited with a PLB source in between them on a 5 mm Aluminium plate. The result show a significant change in the measured signal at different debonding levels. Figure 2.5.2 shows an increase in the magnitude at a peak close to the resonance frequency of the PZT when the amount of debonding increases. Additionally, at different bonding levels, the signal's features, such as amplitude and resonance frequency change. A reduction in the bonding surface seems to detrimentally affect the AE signal by reducing its amplitude, shifting and increasing the resonance peaks. On the other hand, perfectly bonded sensors have a similar response both in time and frequency domain where no noticeable peak occurs close to the PZT patch resonance frequency.

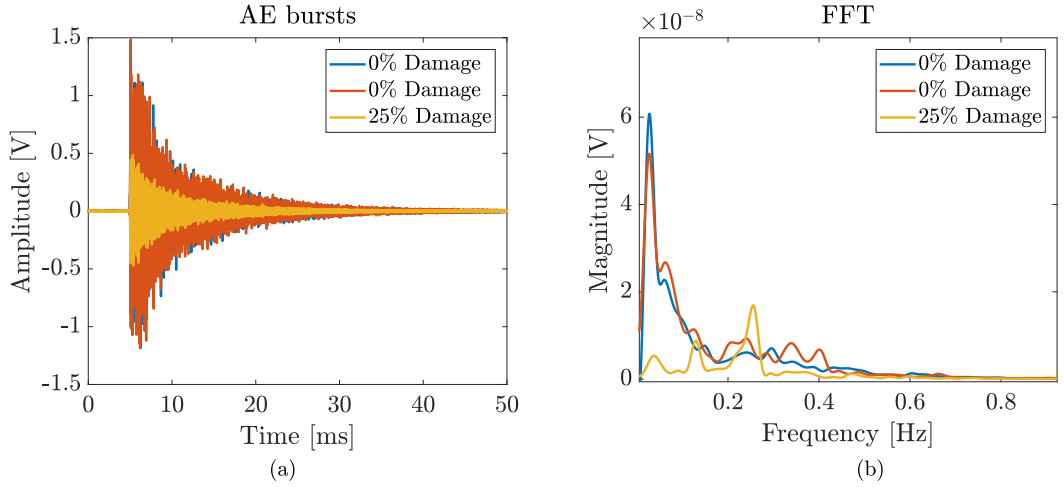


Figure 2.5.2: (a) AE signals on time domain and (b) frequency domain.

A time-domain signal together with its Continuous Wavelet Transform CWT is shown in Figure 2.5.3, 2.5.4 and 2.5.5. These figures show that, in the case of completely bonded scenarios a clear distinction of the arriving modes can be seen in the time-frequency plot. Moreover, the time-frequency signature show a similarity in terms of the coefficients obtained in the CWT, whereas the partially bonded PZT patch shows an unclear and dissimilar signal both in time and time-frequency domain. Additionally, a severe distortion on the time domain signal seems to occur compared with the completely bonded cases. One should also mention the low output signal levels obtained on the partially bonded sensor.

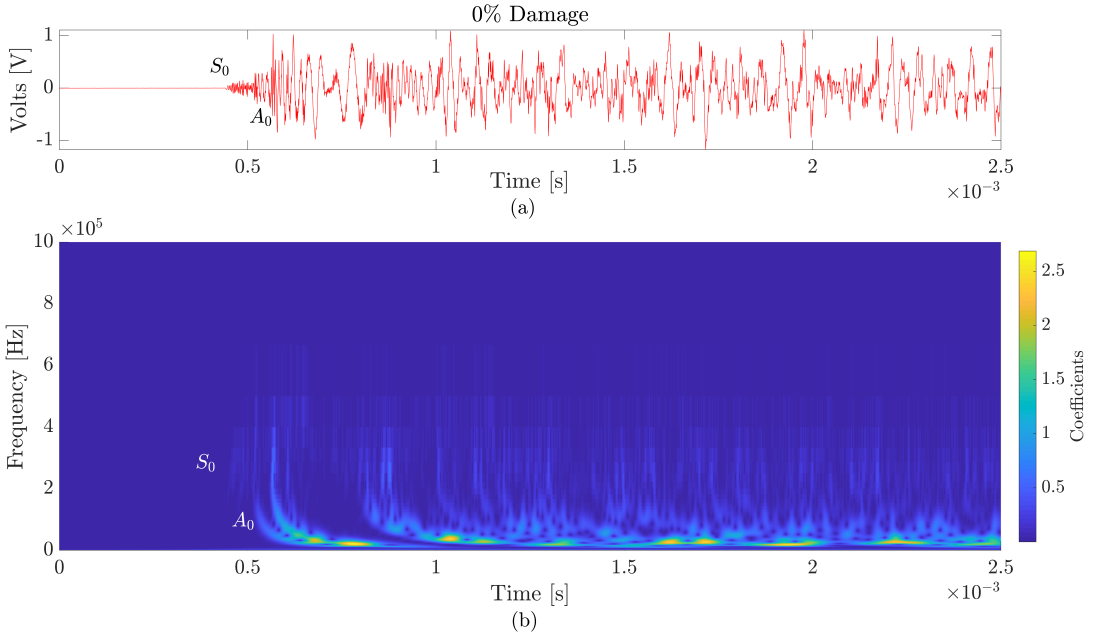


Figure 2.5.3: (a) Completely bonded sensor 1 time and (b) time-frequency plot shows a distinguishable zero-order symmetric and antisymmetric mode.

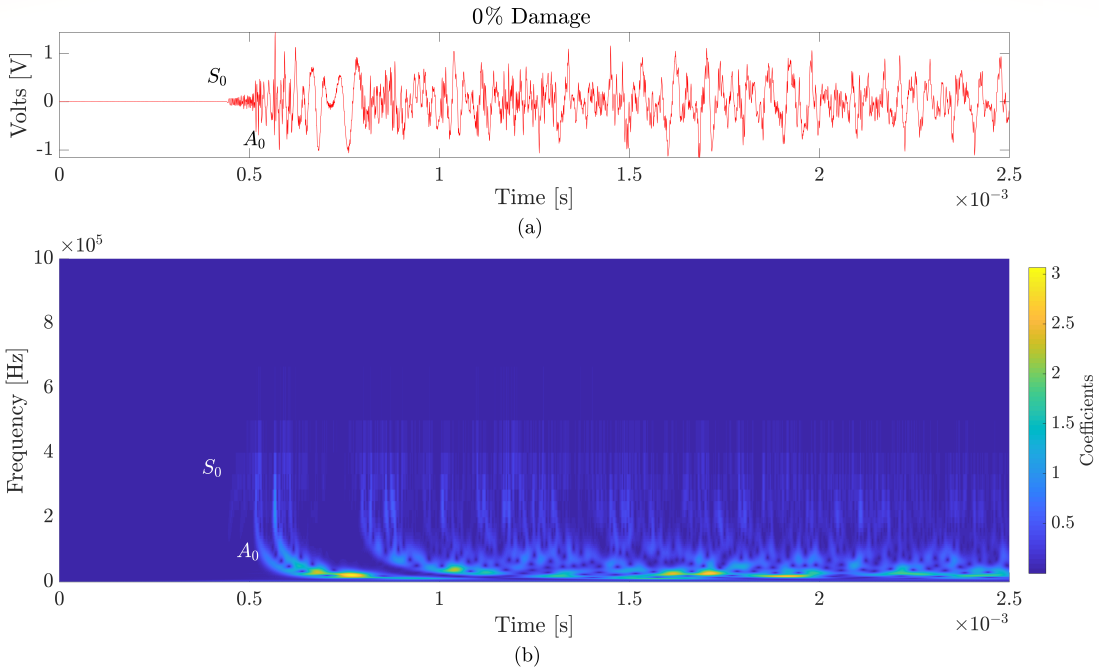


Figure 2.5.4: (a) Completely bonded sensor 2 time and (b) time-frequency plot.

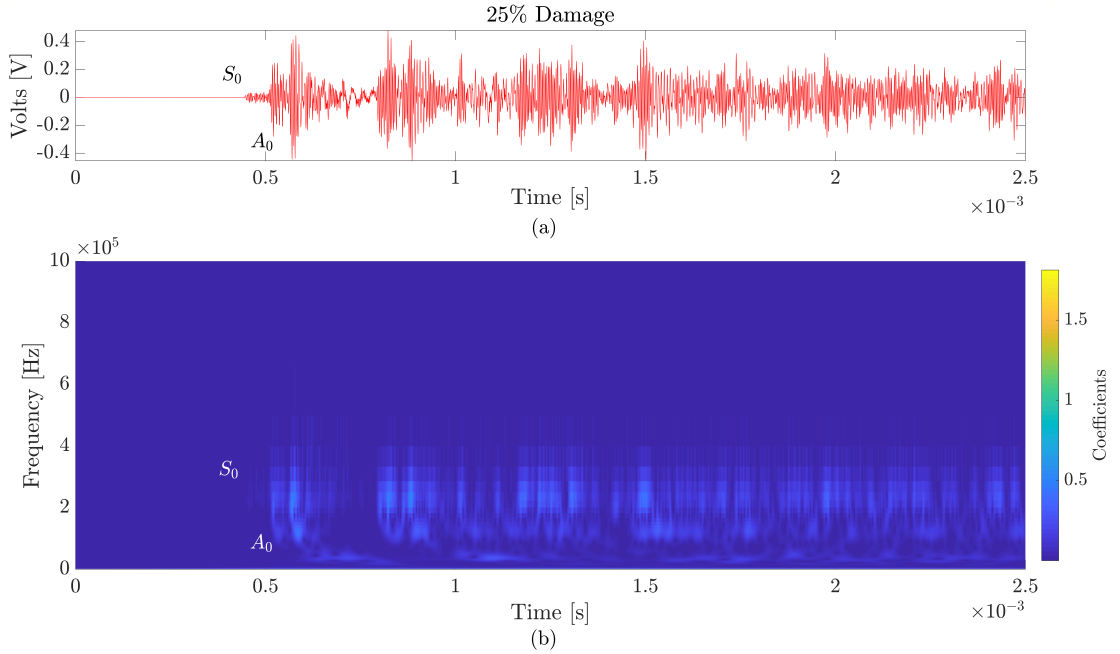


Figure 2.5.5: (a) 25% debonded sensor time and (b) time-frequency plot. Low level amplitudes in the signal obtained on the time and time-frequency representation.

2.5.3 Admittance features for damage detection

Impedance measurements have been exploited to obtain features that are sensitive to both structural and sensor damage. A one-dimensional PZT admittance model initially developed by Lian *et al.* [135] shows the relationship between the measured electrical admittance $Y(j\omega)$ with the patch impedance Z_p and structural impedance Z_s. That means that the measured electrical admittance from the raw signal $Y(j\omega)$ will contain two parts: a part that is sensitive to structural damage and patch damage. The derived relationship can be described in terms of the PZT properties (electrical and geometrical) as:

$$Y(j\omega) = j\omega\frac{wl}{t_c}\left(\varepsilon_{33}^{T}(1-j\delta) - d_{31}^{2}Y_p^E + \frac{Z_p(\omega)}{Z_p(\omega)+Z_s(\omega)}d_{31}^{2}Y^E\left(\frac{\tan(kl)}{kl}\right)\right) \quad (2.5.1)$$

where the parameters ω, I, V, w, l in Equation 2.5.1 represent the angular frequency of the current in [rad/s], the electrical current [A], voltage [V], width [m] and length [m] of the PZT patch respectively. The term k and $\hat{Y}_p^E$ represent the wave number and the dynamic Young's modulus are defined as:

$$k = \omega\sqrt{\frac{\rho}{\hat{Y}_p^E}} \qquad \hat{Y}^E = Y^E(1+\eta j) \qquad \eta = \frac{1}{Q_m} \quad (2.5.2)$$

The terms ρ, η and Q_m represents the density, the mechanical loss factor and the mechanical quality factor of the PZT patch respectively. Equation 2.5.1 can be separated into real (conductance) and imaginary parts (susceptance) as:

$$Y(j\omega) = G(\omega) + B(\omega)j \tag{2.5.3}$$

From Equation 2.5.3, two PZT bonding conditions can be derived; a completely detached and perfectly attached scenarios. The first scenario assumes that there is no structure attached to the PZT patch; meaning $Z_s = 0$. So, the impedance curve for a completely detached PZT can be defined as:

$$Y_{free}(j\omega) = \frac{I(j\omega)}{V(j\omega)} = j\omega\frac{wl}{t_c}\Big(\varepsilon_{33}^{T}(1 - j\delta)\Big) \tag{2.5.4}$$

Redefining Equation 2.5.4 in terms of the real and imaginary parts leads to the following equation:

$$Y_{free}(j\omega) = G_{free}(\omega) + B_{free}(\omega)j \tag{2.5.5}$$

Replacing the term ω by $2\pi f$ in Equation 2.5.4 gives:

$$G_{free}(f) = -\frac{2\pi wl}{t_c}\varepsilon_{33}^{T} f \qquad B_{free}(f) = \frac{2\pi wl}{t_c}\delta f \tag{2.5.6}$$

The completely attached scenario will assume that the PZT path is perfectly bonded to the structure represented as an infinite mechanical impedance. If $Z_s(\omega) = \infty$, the right term of Equation 2.5.1 will be zero giving Equation 2.5.7:

$$Y_{bonded}(j\omega) = \frac{I(j\omega)}{V(j\omega)} = j\omega\frac{wl}{t_c}\Big(\varepsilon_{33}^{T}(1 - j\delta) - d_{31}^{2}Y_{p}^{E}\Big) \tag{2.5.7}$$

where $Y_{bonded}(j\omega)$ can be written in the form:

$$Y_{bonded}(j\omega) = G_{bonded}(\omega) + B_{bonded}(\omega)j \tag{2.5.8}$$

and where the real and imaginary terms are defined as:

$$G_{bonded}(f) = -\frac{2\pi wl\delta}{t_c}f - d_{31}^{2}\eta Y^{E} \qquad B_{bonded}(f) = \frac{2\pi wl\varepsilon_{33}^{T}}{t_c}f - d_{31}^{2}Y^{E} \tag{2.5.9}$$

The mechanical representation of the one-dimensional model can be represented as a PZT patch coupled with a structure with mechanical impedance Z_s with an equivalent mass-spring-damper system represented on Figure 2.5.6. The system parameters m, K_s and C_s represent the mass, stiffness and damping parameters of the structure respectively.

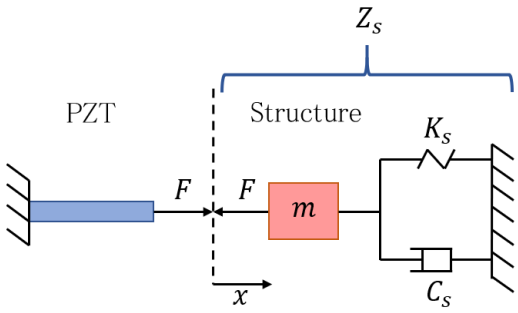


Figure 2.5.6: A one-dimensional impedance model of a PZT patch attached to a structure and exciting the structure with a force F.

Equations 2.5.5 and 2.5.7 represent a characteristic feature in the model shown in Figure 2.5.6. The change in both conditions leads to a change in the susceptance curve as shown in Figure 2.5.7. The curves were obtained using the properties of the PIC181 PZT element as shown in Table 2.5.1 and show a shift on the slope value from the bonded to completely detached scenario.

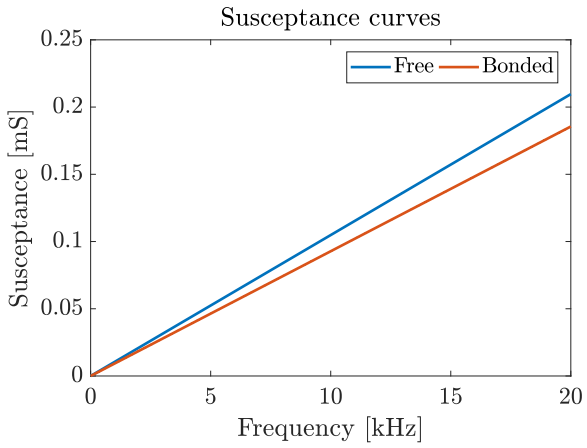


Figure 2.5.7: Free and bonded PZT susceptance features obtained from analytical curves. A change in the slope values indicates two different bonding conditions.

Table 2.5.1: PIC181 electro-mechanical properties.

Properties	**Symbol**	**Values**
Dielectric constant	ε_{33}^{T}	1.062×10^{-8}
Dielectric loss	δ	3×10^{-3}
Piezoelectric constant [C/N]	d_{31}	-120×10^{-12}
Young's Modulus [GPa]	Y_{p}^{E}	84.7
Mechanical loss factor	η	5×10^{-4}
Mechanical quality factor	Q_m	2000
Resonance frequency [kHz]	f_r	227
Radius [m]	r	5×10^{-3}
Thickness [m]	t_c	0.5×10^{-3}

The one-dimensional model discussed in this section describes two different scenarios where free and perfectly bonded conditions occur. This shows two features that permit distinguishing between these two extreme conditions and thus allow them for being processed for novelty detection. This can also be extended to damage progression, as in reality, a wide range of debonding levels can occur due to bond degradation and fatigue. This may lead to an understanding of how the electrical impedance responds to certain structural impedance values produced by progressive debonding.

2.6 Chapter summary

- Damage detection methods based on vibration and AE measurements have been widely implemented in rotary machines. Their damage detection capabilities are well-known in industry and made them suitable for monitoring WTGB. Vibration measurements have been the standard technique for bearing damage diagnosis and have been shown to be robust for incipient damage detection in the bearing components.

- AE signals can provide an additional level of diagnostics, by providing information about the damage location. This can be achieved by extracting the

relevant set of features and exploiting them using networks of AE emission sensors. An AE sensor network can be exploited for WTGB by understanding the response of the sensors to different damage location scenarios.

- Novel damage detection strategies try to replace the human intervention on deciding whether damage had been occurred. This processes involves using sophisticated techniques that automatise this process by analysing features sensitive to the damage stage of the equipment.

- Regardless of the method implied for damage detection, special care on the features type is required for the correct pattern classification. Moreover, this also requires careful consideration of the features and their pre-processing, this ultimately performed for damage detection enhancing purposes.

- Damage localisation in structures is benefitting from sophisticated techniques that overcome the challenges of using AE signals. Analytical methods have initially been proposed to localise AE sources in simple structures and have improved to include anisotropy in materials. Current limitations, such as complex geometries found in practical scenarios and complex wave behaviours, requires analytical models that can truly capture the physics of wave propagation. Such challenges make this approach very difficult, if not impossible to implement and a different approach must be proposed.

- The sensor condition has shown to influence the AE signal characteristics, therefore, affecting the signal extraction features from such signals. Impedance features have shown to be a sensitive feature for bonding degradation and might provide a reliable tool for sensor-self assessment purposes.

- A machine learning approach may solve the challenges mentioned before to some extent but it will require a deep understanding regarding how to train and implement these techniques.

CHAPTER 3

Theory and background

This chapter will provide the background for the techniques used for damage detection and localisation. An introduction to outlier analysis and LDA for novelty detection is detailed. Additionally, a review of Neural Network (NN) for regression is included. This will explain the application of NN for building a regressive model for damage localisation. Aspects will be discussed regarding their implementation, training and model order selection techniques. Finally, an implementation of beamforming as a method for damage localisation in structures is detailed in this chapter.

3.1 Outlier detection

In terms of a statistical definition, an outlier can be defined as a data point that appears inconsistent to the rest of the data and presumably being generated by another mechanism [105]. The outlier can be detected by obtaining a discordancy measure (also known as z-score) where in the case of univariate data is defined as:

$$z_\zeta = \frac{\mid x_\zeta - \bar{x} \mid}{\sigma} \tag{3.1.1}$$

The variable x_ζ represents the data that will be classified or not as an outlier and the $\bar{x}$ and σ are the mean and the standard deviation respectively. These two last parameters can be calculated with the outlier or not, if these parameters are calculated without the outlier then the outlier procedure can be defined as exclusive analysis, conversely it will be defined as inclusive analysis.

For the case of a multivariate scenario the discordancy measure takes the form of the squared Mahalanobis distance defined as:

$$D_\zeta = (X_\zeta - \bar{X})^T \Sigma^{-1} (X_\zeta - \bar{X}) \tag{3.1.2}$$

where X_ζ represents the vector for the potential outlier, $\bar{X}$ the mean vector and Σ the covariance matrix. As in the case of the univariate form, an inclusive and exclusive analysis can be perform in the same manner as described above. Both versions of the method can be understood as ways to describe the distance of each observation from a specific distribution mean value in terms of standard deviations. In order to visualise this, Figure 3.1.1 shows a univariate Gaussian Normal distribution where deviations from the mean $\bar{x}$ are designated in multiples of the standard deviations σ.

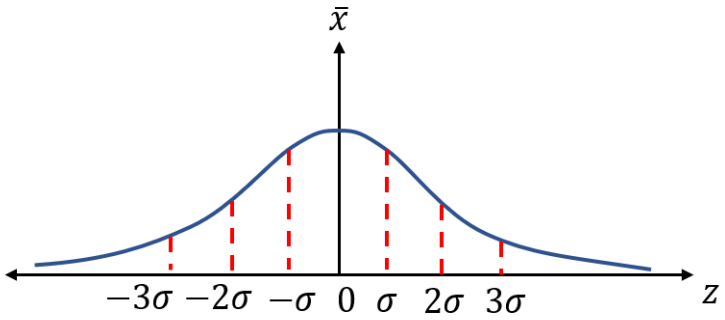


Figure 3.1.1: The Gaussian Normal distribution and the z-score in terms of standard deviations.

An important aspect is that to classify an observation as an outlier or not, a value must be defined. In some cases, a threshold value can be visually estimated and may allow an unambiguous classification for each observation. The following section will discuss a systematic approach to set a threshold level by using a Monte Carlo simulation approach.

3.1.1 Threshold determination

A Monte Carlo simulation based method described in [40, 106] has been used to determine the threshold between the normal and faulty data. The process is defined as follows:

1. A random observation matrix with K rows and D columns in generated where K and D represent the number of observations and the dimensions respectively. Each element of the observation matrix is sampled from a normal Gaussian distribution.

2. The Mahalanobis distance is calculated for each observation and the largest distance is stored in an array.

3. A large number of trials are required, thereby the process is repeated until the the critical values for 5 and 1% tests are found in the array with the largest Mahalanobis distances.

3.2 Linear Discriminant Analysis (LDA)

An introduction to outlier detection has shown a method to detect outlier points from a reference condition by calculating the Mahalanobis distance. Even though a threshold limit is set as in this case with a Monte Carlo approach, the decision of whether or not the measurement is an outlier does not take into account the whole data distribution. In other words, outlier analysis focusses on assigning a single data point as a potential outlier without considering the entire distribution of potential outliers. For instance, it is better to set a boundary between two different conditions and set a decision based on the changes in the data distributions. Another aspect to take into consideration is the amount of overlapping generated between datasets that significantly reduces the ability to distinguish between different conditions. In some cases, the information relating to different conditions might be hidden in certain regions of the multidimensional space. Moreover, this information might be hidden in a subset of this multidimensional space and thereby not evident for certain data representations. This situation leads to the exploration of two problems: the multiple data representations and their separability characteristics.

LDA [107] is a feature reduction/classification method that allows the separation of two or more classes by exploiting feature reduction capabilities. The main concept is that, in order to achieve an optimal separation, the data clusters are projected onto a lower dimensional space that maximises the data separation and displays

the lowest variances for each data class. This has some similarity to Principal Component Analysis (PCA) [108] where the feature dimensions are also reduced but in the directions that the variance is maximised. In general, both techniques seek for a linear combination that better explains the data. Figure 3.2.1 shows two different representations defined as LDA1 and LDA2 of the same classes C_1 and C_2 where the values of d_1 and d_2 represent the distance between the means. The representation at LDA1 shows a higher variance and overlapping between classes than the representation at LDA2 where a minimal variance and larger distance is achieved.

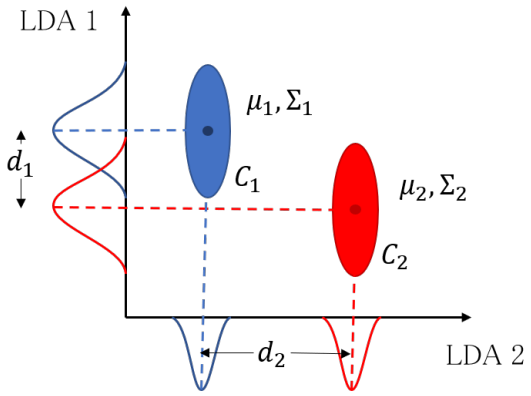


Figure 3.2.1: Linear Discriminant Analysis between two different classes.

In this two dimensional example, a feature reduction is taking place as both distributions are projected onto a one-dimensional space. As mentioned before, the method tries to find a linear combination that best separates two classes. This linear combination can be defined as:

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \omega_0 \tag{3.2.1}$$

where w, w_0, $y(\mathbf{x})$ and $\mathbf{x}$ are a set of weights, the bias, an hyperplane and the feature vector respectively. The separation is based on finding a set of weights $\mathbf{w}$ that maximises the ratio between class variance Σ_b and minimises the within class variance Σ_w as shown in Equation 3.2.2. This form is also known as Canonical Discriminant Analysis (CLA).

$$\mathbf{w}_\star = \arg\max \frac{\mid \mathbf{w}^T\Sigma_b\mathbf{w} \mid}{\mid \mathbf{w}^T\Sigma_w\mathbf{w} \mid} \tag{3.2.2}$$

where Σ_b and Σ_w represent the variance between and within classes and can be calculated using Equations 3.2.3 and 3.2.4 respectively.

$$\Sigma_b = \sum_k (\boldsymbol{\mu}_k - \boldsymbol{\mu}) (\boldsymbol{\mu}_k - \boldsymbol{\mu})^T \tag{3.2.3}$$

$$\Sigma_w = \sum_k \sum_{i \in k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \tag{3.2.4}$$

where the terms $\boldsymbol{\mu}$, $\boldsymbol{\mu}_k$ and $\boldsymbol{x}_i$ represent the global mean of the whole data, the mean for each class k and the observations vectors for each corresponding class respectively. For the purpose of classification, a discriminant function can be derived by using Bayes' theorem. Thus Equation 3.2.1 can be defined as:

$$y(\mathbf{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma_w^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) + \log \mid \Sigma_w \mid \tag{3.2.5}$$

Equations 3.2.2, 3.2.3 and 3.2.5 can lead to view LDA as a feature reduction technique and a classification method. In reality, finding the set of weights to find a projection hyperplane will be equivalent to find the discriminant function that works as the decision boundary between classes as shown in Figure 3.2.2. The hyperplane that separates both classes is represented perpendicular to the weight vector. In the case of a two-dimensional feature space the projection reduces the two-dimensional representation into a one-dimensional space.

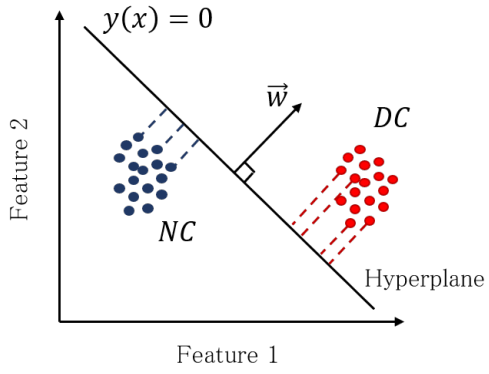


Figure 3.2.2: Linear Discriminant feature reduction where the normal (NC) and damage conditions (DC) are projected on the hyperplane.

3.3 Background to neural networks for regression problems

The use of NN has been widely studied for regression problems and their capabilities have been extensively recognised in the ML field. The use of NN for such problems relies on their capabilities for performing regression and classification problems. They exploit features of real data that regular linear methods fail to do and where also *the curse of dimensionality* becomes a problem. This makes NN a good approach to practical applications [109]. Even though they seem to learn any type of data relationships, special caution should be taken at different stages to provide meaningful features, correct training and more specifically, avoid overtraining. The following sections will treat these points to assure the correct use of NN for this type of problem.

3.3.1 The Multilayer Perceptron (MLP)

A special type of network architecture defined as the MLP is defined as a set of three layers denominated as the input, the hidden and the output layer as shown in Figure 3.3.1. Even though more simple network architectures exist, such as the case of a single-layer perceptron, their applicability to regression problems have been limited. It has been shown that, the single layer perceptron can only represent a limited set of functions, it is necessary to include an additional layer in the network to allow a more generalised representation [109]. The capability of representing any continuous function is normally referenced as the *the universal approximation theorem* [109] and basically states that a feed-forward network with a hidden layer that contains a finite number of units (neurons) can approximate to arbitrary accuracy, any continuous function. Therefore, an MLP network can be taken to be a universal approximator. Even though this theorem constrains the number of layers in the network, it does not mention the number of hidden units required. This will be discussed later on in the model order selection procedure.

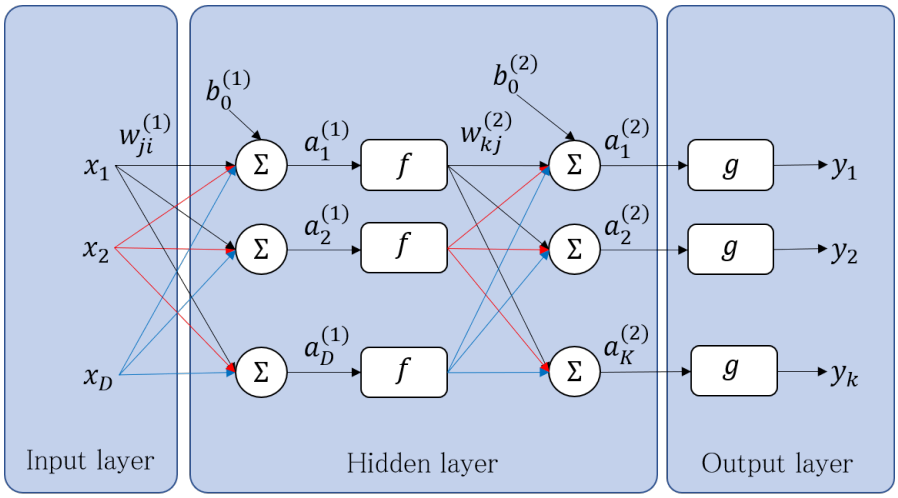


Figure 3.3.1: The multilayer perceptron architecture.

From Figure 3.3.1, each output y_k of the MLP can be represented as a function of the weights of the first and second layer w_{ji} and w_{kj} with each corresponding activation function f and g as:

$$y_k = g\Big(\sum_{j=0}^{K} w_{kj} f\Big(\sum_{i=0}^{D} w_{ji} x_i\Big)\Big) \tag{3.3.1}$$

where x_i are each individual input, w_{ji} and w_{kj} are weights in the first and second layer respectively. Notice that Equation 3.3.1 does not include the bias terms $b_0^{(1)}$ and $b_0^{(2)}$ explicitly in the summation term, it rather assumes the terms $w_{j0}x_0 = b_0^{(1)} \times 1$ and $w_{k0}a_{k0}^{(2)} = b_0^{(2)} \times 1$. The terms $a_j^{(1)}$ and $a_k^{(2)}$ are the output of the activation function in the first and second layer and can be written as a linear combination of the weights of each layer as:

$$a_j^{(1)} = \sum_{i=0}^{D} w_{ji} x_i = \sum_{i=1}^{D} w_{ji} x_i + b_0^{(1)} \tag{3.3.2}$$

$$a_k^{(2)} = \sum_{i=0}^{K} w_{kj} z_j = \sum_{i=1}^{K} w_{kj} z_j + b_0^{(2)} \tag{3.3.3}$$

where the term z_j represents the activation function f evaluated at $a_j^{(1)}$ as:

$$z_j = f\Big(a_j^{(1)}\Big) \tag{3.3.4}$$

The activation functions f and g have the property that they are continuous and differentiable. Normally the activation function f takes the form of a sigmoid function as:

$$f(a_j) = \frac{1}{1 + \exp(-a_j)} \tag{3.3.5}$$

This requirement for such non-linear activation functions is due the need of it being capable of representing any continuous mapping. Generally, for the output layer, g is defined as a linear activation function only for regression problems in order to avoid limiting the output of the sigmoid activation function.

3.3.2 The error function and data modelling

The main task of the NN is to learn from the relationship between the input and output data. This process does not realistically resemble the biological learning mechanism with the ANN, mechanism rather being achieved by the minimisation of an error function derived from the principle of maximum likelihood. For a set of N data samples composed of the input vector $\mathbf{x}$ and its target vector $\mathbf{t}$, the maximum likelihood is defined as:

$$\mathcal{L} = \prod_{n}^{N} = p\left(\mathbf{x}^n, \mathbf{t}^n\right) \tag{3.3.6}$$

The term $p(\mathbf{x}, \mathbf{t})$ gives important information of the data generation and it is modelled with the NN. This input-target joint distribution can be represented as $p(\mathbf{x}, \mathbf{t}) = p(\mathbf{t} \mid \mathbf{x})p(\mathbf{x})$ for the sake of prediction of a new vector $\mathbf{t}$ given an input vector $\mathbf{x}$. Rather than modelling the joint probability the conditional probability $p(\mathbf{t} \mid \mathbf{x})$ is modelled instead. Until this point, no assumption regarding the conditional density $p(\mathbf{t} \mid \mathbf{x})$ have been made. This general concept allows a more generalised way of applying NN for more complex data distributions and in fact specifies a general framework to deal with this particular situation [109, 110]. In this thesis, it is believed that the data does not follow a complex distribution and for instance a Gaussian assumption is made as shown in Equation 3.3.7.

$$p(t_k \mid \mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_k\left(\mathbf{x}, \mathbf{w}\right) - t_k\right)^2}{2\sigma^2}\right) \tag{3.3.7}$$

Equation 3.3.7 describes the likelihood function of the target variable $\mathbf{t}$ where y_k and t_k represents the *kth* network output and target variable respectively. The error function is defined as the minimisation of the negative logarithm of the likelihood function as described on Equation 3.3.6. That finally leads to the sum-of-squares error expression given in Equation 3.3.8. The minimisation of this error function basically seeks for the set of weights $\mathbf{w}$ that minimises this error function.

$$E = \frac{1}{2}\sum_{n}^{N}\sum_{k}^{K} \mid y_k(\mathbf{w}, \mathbf{x}^n) - t_k^n \mid^2 \tag{3.3.8}$$

This error function can be modified in order to improve the networks capability for generalisation and avoid overfitting. An additional aspect that involves the optimisation of Equation 3.3.8 leads to add a regularisation term if the Levenberg-Marquard method is used to optimise this function. The idea behind this approach is to keep relatively small iteration steps during the optimisation process whilst keep the relatively small error values.

So far, the learning process has been defined as an optimisation problem and no mention of any specific algorithm has been made. Even though multiple optimisation algorithms may be used for this problem, it was decided that the Levenberg-Marquardt method can be adopted due to its computational efficiency characteristics.

3.3.3 The Levenberg-Marquardt optimisation algorithm

The learning process in a NN involves minimising an error function through optimisation. Althought this step can be performed with a vast number of general purpose algorithms such as gradient descent and conjugate gradient based methods, their limitations has been extensively discussed in literature [109]. For the case of regression problems, the training involves the minimisation of the sum-of-squares error function and thus requires the calculation of the Hessian matrix. This can be performed efficiently if certain assumptions are made in terms of the linearity of the solution. An initial approach involves the linearisation of Equation 3.3.8 and redefining it in terms of a Taylor series approximation as:

$$E = \frac{1}{2}\sum_{n}^{N}(\epsilon^n)^2 = \frac{1}{2} \mid \epsilon(\mathbf{w}^\tau) + \mathbf{Z}(\mathbf{w}^{\tau+1} - \mathbf{w}^\tau) \mid^2 \tag{3.3.9}$$

where ϵ and τ represent the error of the *nth* pattern and the step increment respectively, the elements of the matrix $\mathbf{Z}$ can be defined as:

$$\mathbf{Z}_{ni} = \frac{\partial \epsilon^n}{\partial w_i} \tag{3.3.10}$$

Equation 3.3.10 represents the error derivative with respect to the network weights, this derivative is calculated using the back-propagation method discussed later. The minimisation of the error function defined in Equation 3.3.9 with respect to the weights $\mathbf{w}^{\tau+1}$ leads to the expression:

$$\mathbf{w}^{\tau+1} = \mathbf{w}^\tau - \mathbf{H}^{-1}\mathbf{Z}^T\epsilon(\mathbf{w}^\tau) \tag{3.3.11}$$

Equation 3.3.11 shows an iterative procedure to calculate the vector of weights $\mathbf{w}$. In practical applications, where a considerable number of network weights are used, the speed at where the Hessian matrix is evaluated becomes an important issue during the training stage. This can be improved by approximating the Hessian matrix $\mathbf{H}$ as:

$$\mathbf{H} \approx \mathbf{Z}^T\mathbf{Z} \tag{3.3.12}$$

The expression in Equation 3.3.12 refers to the Levenberg-Marquardt approximation and it is valid for an infinite dataset. This simplifies the calculation of the Hessian matrix by reducing the number of calculations. Another practical aspect encountered during the implementation of Equations 3.3.11 and 3.3.12 deals with the non-linear behaviour of the error surface. This can be a potential problem if the step size τ becomes sufficiently large during the iterations and thus the linear approximation in Equation 3.3.9 becomes no longer valid. This issue is normally solved by adding a regularisation term to keep the weight value relatively small, thereby Equation 3.3.11 is modified as:

$$\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} + \left(\mathbf{H} + \lambda\mathbf{I}\right)^{-1}\mathbf{Z}^T\epsilon(\mathbf{w}^{\tau}) \tag{3.3.13}$$

where λ and I represents the regularisation parameter and the identity matrix respectively. The importance of this algorithm relies on its efficient way of computing the Hessian matrix and is considered the fastest of similar optimisation algorithms. It is important to mention that, in contrast to other optimisation algorithms used for training, the Levenberg-Marquardt algorithm specifically optimises the sum-of-squares error function defined in Equation 3.3.9.

3.3.4 Error back-propagation

In order to implement this algorithm, some quantities must be calculated, such as of the error derivatives shown in Equation 3.3.10. This is done by using the back-propagation method which is extensively discussed in [109]. The learning process involves finding a set of weights $\mathbf{w}$ that minimise Equation 3.3.8. This is achieved by calculating the error between a target set $\mathbf{t}$ and network output $\mathbf{y}$ and back-propagating it to calculate another set of weights that minimise the error function. This process involves calculating the derivative of the error function with respect to the weights, an efficient method for performing this calculations is known

as the back-propagation algorithm. The derivative of the error function with respect to the weights is defined as:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial a_j}\frac{\partial a_j}{\partial w_{ij}} = \delta_j \frac{\partial a_j}{\partial w_{ij}} = \delta_j z_i \tag{3.3.14}$$

where the difference between the target values and the network output is defined as:

$$\delta_k = y_k - t_k \tag{3.3.15}$$

The derivative of the Error function with respect to the activation functions a_j can be calculated as:

$$\delta_j = \sum_k \frac{\partial E}{\partial a_k}\frac{\partial a_k}{\partial a_j} = h'(a_j)\sum_k w_{kj}\delta_k \tag{3.3.16}$$

Now that all the error derivatives equations are known, the back-propagation method can be iteratively executed as:

1. Apply an input vector $\mathbf{x}$ to the network and obtain all the activation values a_j and z_j of the hidden units using Equations 3.3.3 and 3.3.4.
2. Evaluate the function δ_k for all the output units using Equation 3.3.15.
3. Perform the back-propagation by calculating the values of δ_j using Equation 3.3.16.
4. Calculate the error derivative with respect to the weights using Equation 3.3.14.

3.3.5 The dataset

Even though training a NN with data initially sounds trivial, special care must be given. The main purpose of a dataset is not to provide the means of allowing the NN to exactly learn each point but to learn the underlying relationship that generates these outputs. Initially, these weights need to be adjusted according to the data provided to the NN but further issues such as overfitting and defining the number of hidden units known as hyper-parameters in the network must be addressed. This is generally addressed by the common practise of splitting the dataset into three sets defined as:

- **Training set**: This dataset will be used for training, in other words it will be used to modify the weights of the network. It is important to notice that, for most of the optimisation algorithms, the error of the NN will decrease monotonically as the number of iterations increase, clearly this will eventually will lead to overfitting problems which can be avoided by using a validation set.

- **Validation set**: An additional data set is required to determine the network hyperparameters such as the number of hidden units and to prevent the NN from overfitting issues. The error calculated using this dataset is known as the validation error and normally prevents the NN from overfitting by stopping the training process once the validation error starts to increase.

- **Test set**: The test set provides an independent test set that allows to test the network performance.

3.3.6 The size of the dataset

An important aspect to consider during the implementation of a NN deals with the size of the dataset; This ultimately having an effect on the NN generalisation capabilities. The concept of generalisation can be understood as equivalent to fitting a a polynomial curve for NN, a further explanation will be discussed later. This situation leads to the question of how many patterns are required in order to avoid overfitting problems. Fortunately, this topic has been addressed by Baum and Haussler [111], whose work suggested a lower and upper bound defined as:

$$P_l = W \tag{3.3.17}$$

$$P_u = 10W \tag{3.3.18}$$

where P_l, P_u and W represents the lower, the upper bound and the number of weights in the NN respectively, where the number of weights in a MLP can be calculated as:

$$W = (I + 1)J + (J + 1)K \tag{3.3.19}$$

where I, J, and K represents the number of inputs, the number of hidden units and the number of outputs respectively.

3.3.7 Overfitting and model order selection

As the main goal of NN is not to learn the exact representation of the data but rather understand its data generator. In a practical situation, a NN that can correctly predict previously unseen input values is desirable. This generalisation issue addresses the following contrasting scenarios: a NN with little complexity unable to correctly predict due to its inflexibility or a high complexity network that overfits the noise in the data.

As initially said, the number of layers can be fixed by means of the universal approximation theorem but the optimal number of hidden units remains unspecified. More complex models such as NN with a higher amount of hidden units can lead to overfitting whereas models with a low amount of hidden units can lead to models incapable of representing the data relationship.

The problem can be understood as a trade-off between the bias and variance of a model. If one assumes there is a dataset denoted as D, where $E_D[.]$ defines the expected value of the whole dataset, $y(\mathbf{x})$ the network outputs and $\langle t \mid x \rangle$ the optimal network output. The expectation of the ensemble average can be defined as:

$$E_D[((y(\mathbf{x}) - \langle t \mid x \rangle)^2] = \{E_D[y(\mathbf{x})] - \langle t \mid \mathbf{x} \rangle\}^2 + E_D[\{y(\mathbf{x}) - E_D[y(\mathbf{x})]\}^2] \quad (3.3.20)$$

Equation 3.3.20 can be defined in terms of the bias and the variance as:

$$(bias)^2 = \frac{1}{2} \int \{E_D[y(\mathbf{x})] - \langle t \mid \mathbf{x} \rangle\}^2 p(\mathbf{x}) d\mathbf{x} \quad (3.3.21)$$

$$(variance) = \frac{1}{2} \int E_D[\{y(\mathbf{x}) - E_D[y(\mathbf{x})]\}^2] p(\mathbf{x}) d\mathbf{x} \quad (3.3.22)$$

In order to provide an example of two extreme situations where the bias and variance play a role in model order selection, the two output scenarios shown in Figure 3.3.2 will be considered. Assume a smooth function $h(\mathbf{x})$, where the optimal network output can be defined as $\langle t \mid x \rangle = h(\mathbf{x})$. If a fixed output $y(\mathbf{x})$ is generated, the term $E_D[y(\mathbf{x})] = y(\mathbf{x})$ and therefore the variance term defined in Equation 3.3.22 will become zero. In the opposite case, if the network interpolates perfectly between each point, the bias term in Equation 3.3.21 will be zero as $E_D[y(\mathbf{x})] = h(\mathbf{x})$ but it will give a large of error as it includes the noise term.

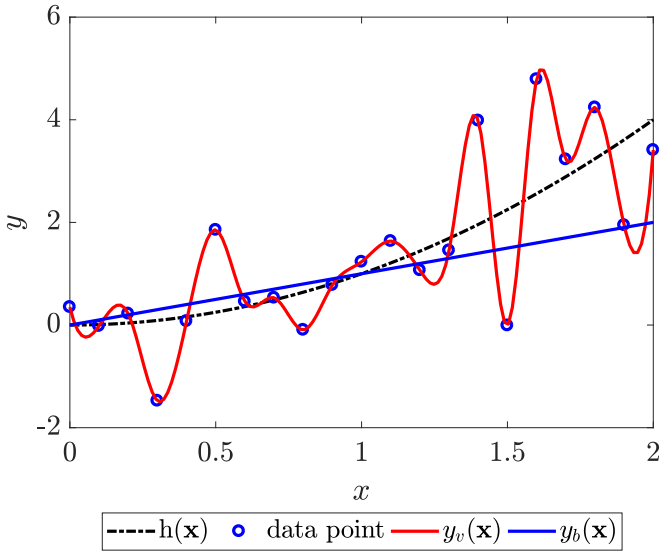


Figure 3.3.2: A fixed function $y_b(\mathbf{x})$ (blue) and a function that perfectly interpolates $y_v(\mathbf{x})$ (red) each data point (blue circles) obtained from the underlying function $h(\mathbf{x})$ (dashed curve).

The optimisation of the complexity of a NN is commonly performed by a cross-validation method or a regularisation technique. A cross-validation technique normally selects the model that has the best performance on a new data set whilst a regularisation approach involves adding an additional multiplicative term in the error function that penalises large weight values. The most common regularisation methods are the L_1 and L_2 (also known as weight decay) regularisation terms defined by Equation 3.3.23 and 3.3.24 as:

$$\Omega_{L1} = \frac{1}{2}\sum_i w_i^2 \tag{3.3.23}$$

$$\Omega_{L1} = \frac{1}{2}\sum_i w_i \tag{3.3.24}$$

In real applications some issues such as the amount of data available become a problem as it may not allow the inclusion of an independent validation set in order to evaluate the performance of different models. A K-Fold cross-validation technique deals with this issue by systematically portioning the data.

3.3.8 K-Fold cross-validation method

The main advantage of using the K-Fold cross-validation method is that a high proportion of the data can be used for training in situations where the data are scarce or limited. The K-Fold method for cross-validation consists of splitting the data into a training and a test set with the latter set being used to select the appropriate model complexity. More specifically, the entire data will be partitioned into S groups allowing the remaining $(S-1)/S$ proportion of the data to be used for training purposes. Each model will be run S times and a model can be selected based on the predictive performance obtained from the test set. As illustrated in Figure 3.3.3, the data are partitioned into $S = 5$ groups. 80% can then be used for training and validation whereas the remaining 20% for the neural network performance testing.

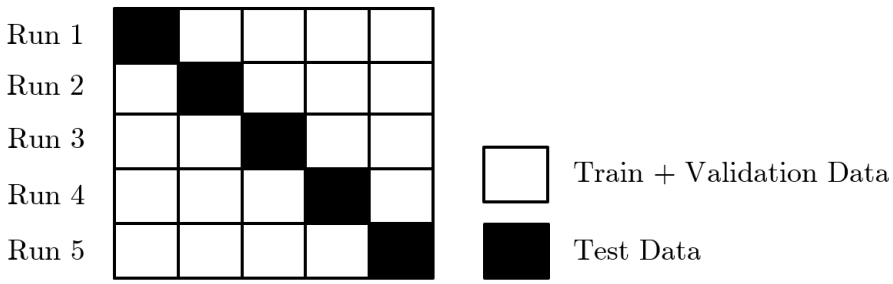


Figure 3.3.3: Data partition for K-Fold cross validation.

In order to summarise the method, the following pseudo-code showed in Figure 3.3.4 was implemented in order to cross-validate the models. The procedure described in Figure 3.3.4 specifies a range of NN models by defining a lower and an upper bound of hidden units defined as J_l and J_u respectively. A main loop initialises each NN with random weights during each iteration. The initialised NN is trained using the remaining $S-1$ training sets whilst retaining the corresponding test set k. The procedure continues the training process until the validation error e_{minval} reaches its minima. The lowest validation error e_{minval} is stored in an array defined as $e_{opt}[k]$. This process continues until all models are trained in the first loop. At the end, a set of candidates networks are selected and the test error evaluated using the test set.

Algorithm 1 K-Fold Cross-validation

```
1: procedure 10-FOLD
2:     J_l = set lower bound for hidden units
3:     J_u = set the upper bound for hidden units
4:     for j = J_l; i ≤ J_u; j = j + 1 do
5:         initialise random weights
6:         for k = 1; k ≤ 10; k = k + 1 do
7:             Train MLP
8:             Stop training when validation error e_minval reaches its minima
9:             if e_minval < e_opt[k] then
10:                e_opt[k] = e_minval
11:    for k = 1; k ≤ 10; k = k + 1 do
12:        Evaluate test error for each optimal network
```

Figure 3.3.4: Algorithm implemented for K-Fold cross validation.

3.4 Time-Delay and Sum Beamforming

So far, analytical and ML methods explored the use of TOA features for damage localisation. Although the extraction of this feature is relatively straightforward when a clear AE onset is visible, this can become difficult when there is a relatively low signal-to-noise level and high attenuation in the signal. Meaning that generating a TOA map becomes impractical in this scenario and might result in ambiguous TOA features. Time-Delay beamforming uses the delay generated between the signals as the main feature to localise sources, this technique is based on the summation of the signals obtained in an array of N sensors, assuming that the signals can be represented as delayed copies of the same event. The beamformer output is defined as [112]:

$$g(t, \vec{r}) = \sum_{i=1}^{N} x_i(t - \tau_i(\vec{r})) \tag{3.4.1}$$

where x_i represents the signal of each sensor delayed by an amount τ_i, this delay value can be calculated analytically and will depend on the position of the steering vector $\vec{r}$. The source position is obtained from the beamformer power defined as:

$$E\{g(t,\vec{r})\} = \int_{T_1}^{T_2} \mid g(t,\vec{r}) \mid^2 dt \tag{3.4.2}$$

where the beamformer output is integrated within a time window defined between bounds T_1 and T_2. This expression is integrated over each point of the spatial domain defined by the steering vector and will result in a set of interference patterns where regions of high intensities can be interpreted as possible source positions. The method can be summarised in this following steps:

1. Identify the region of interest.
2. Discretise the geometry and calculate a time delay
3. The original signals are delayed with the values obtained from the previous step.
4. The signals are summed up in order to generate constructive or destructive interference patterns.

Even though, this method does not rely on the TOA feature, it does rely on the shape of the waves and the generated interference patterns. The main concept of Time-Delay and Sum beamforming involves the analysis of the interference patterns produced by the summation of the amplitude responses of the delayed signals. This can be seen as shown in Figure 3.4.1 where the each signal is delayed and then summed up in order to produce the beamformer output $E_p(t)$ and then processed using Equation 3.4.2 to produce the interference patterns.

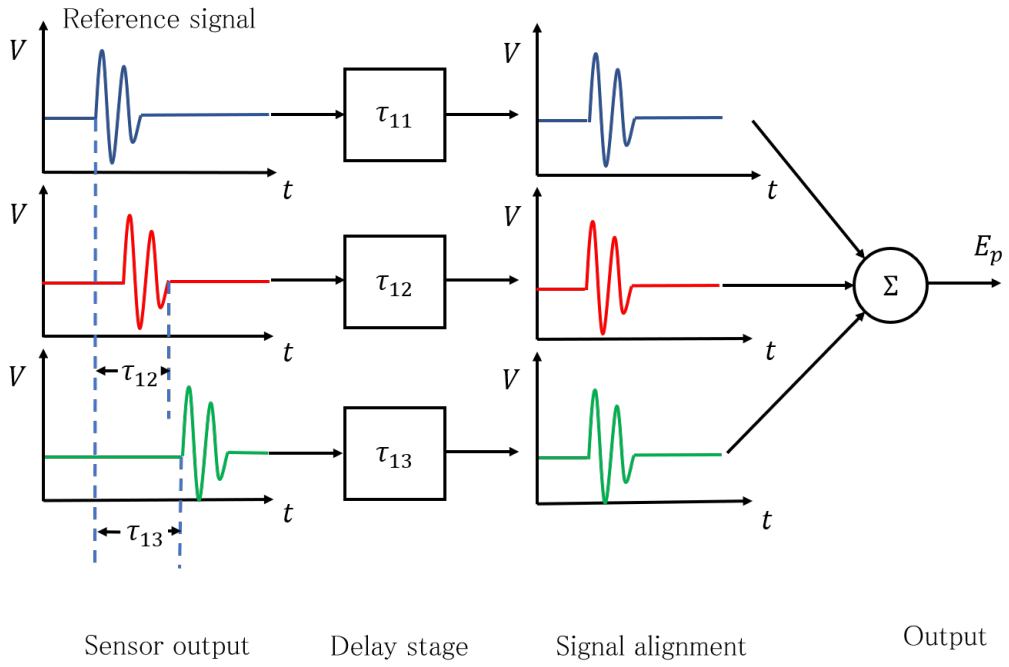


Figure 3.4.1: Beamforming technique using a set of three sensors showing a constructive interference pattern.

3.4.1 Near-Field and Far-Field and Time-Delay calculation

In order to accurately calculate the time delays, an assumption regarding the nature of the front wave during the propagation must be considered. This can be intuitively seen to be relative to the size of the application. Planar wave propagation can be viewed as a spherical wave with an infinite radius and might be suitable for describing AE sources at relatively large distances. Figure 3.4.2 shows the two assumptions where τ_p and τ_s represent the delay for the Far-Field and Near-Field assumptions respectively.

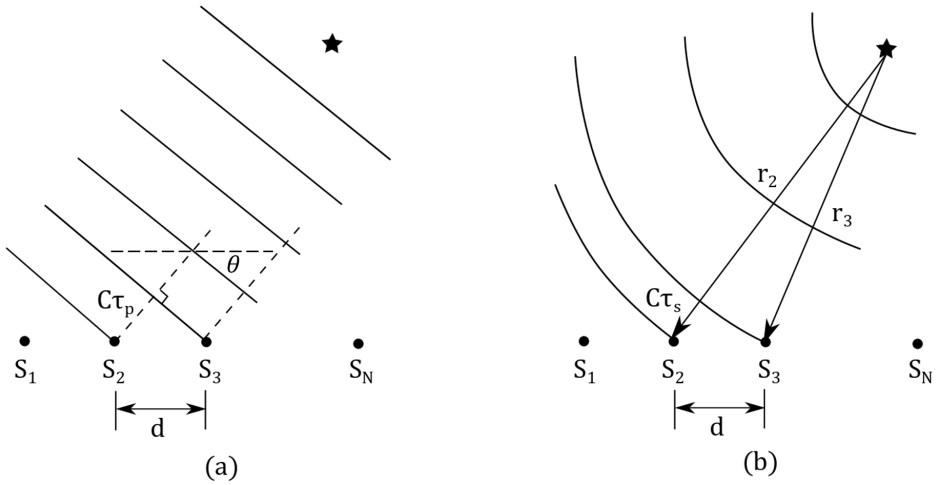


Figure 3.4.2: Far-Field and Near-Field assumptions.

Although this assumption is not trivial, there are a few empirical relationships that define a criteria for Near-Field such as [113]:

$$r \leq \frac{2fL^2}{C} \tag{3.4.3}$$

where L, r, f and C represent the array length, the radius of the wave front, the wave frequency and the wave propagation speed respectively. Finally, the time delays τ_p and τ_s can be calculated as:

$$\tau_p = \frac{d \sin\theta}{C} \tag{3.4.4}$$

$$\tau_s = \frac{|\,\vec{r}\,| - |\,\vec{r_0}\,|}{C} \tag{3.4.5}$$

where d, θ, r and r_0 represent the distance between the sensor array, the angle, the vector of each wave front and the vector of the wave front at the reference sensor.

3.5 Chapter summary

- An introduction to outlier analysis and LDA has been included in this chapter, this consequently will allow the use of these algorithms to perform novelty detection. Although, both methods might be used for the same purpose of damage detection, there is a great difference between the way both methods

deal with outliers. Outlier analysis does not consider the entire distribution of outliers whereas LDA considers the entire data distribution. Additionally, outlier analysis requires determining a threshold as the method by itself does not discriminate between a normal and abnormal condition. An introduction to a Monte Carlo simulation approach has been included as a method to calculate a threshold value based on the dimensions and size of the feature vectors.

- LDA has several advantages compared with outlier analysis, as it tries to optimise an hyperplane that best separates different conditions using a discriminant function. Therefore, no threshold assumption is required to be done to classify a set of measurements as potential anomalies or not.

- An introduction to NN explained the training process and the steps required to produce models that avoid overfitting. The importance of a model order selection using the K-Fold cross-validation method has been explained and its implementation for NN.

- Moreover, a method using the beamforming method has been included in this chapter. Although this method is not a ML approach, it exploits a different type of features that might allow to localise AE sources. Additionally, it is important to mention that this method works under certain localisation conditions, as so far, complex wave propagation paths were not considered.

CHAPTER 4

Test rigs

In this chapter, an introduction to the test rigs used for novelty detection and damage localisation is presented. Description of the scaled-down rig and the MultiLife rig are also introduced at each section to provide details about their working mechanism and instrumentation.

4.1 Scaled-down rig

This rig was originally designed to study the life span expectation on roller bearings by measuring the dynamic strain using X-ray and Neutron Diffraction techniques. This required that the main experiment was performed in the ISIS Neutron and Muon facilities where limited access to the rig and special precautions during the sensor installation were imposed during the experimental stage. The advantage of this rig is its relatively few and simple mechanisms that result in less contamination by secondary noise sources. Therefore, this characteristic made it suitable for an initial investigation into damage localisation. An image of the testing rig is shown in Figure 4.1.1 where the main elements of the loading frame are shown. Figure 4.1.1 also shows the location of the test bearing installed on the main shaft.

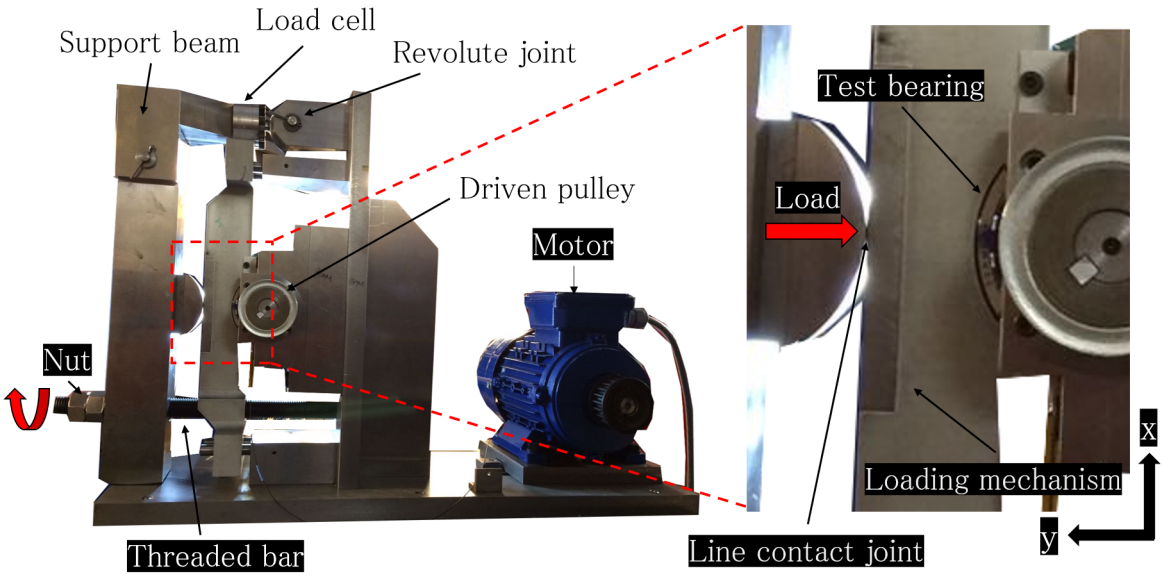


Figure 4.1.1: Scaled-down rig with a zoomed area of the loading mechanism.

This rig was developed to run at a single angular velocity via an electrical motor attached by a belt-pulley transmission. The load applied on the rig was controlled by a threaded bar and a nut, this mechanism allows a constant radial load with fluctuations limited by creep in the material, which in this case, can be considered negligible at room temperature. The load can be increased and decreased by rotating a nut, this consequently rotates the support beam around a revolute joint and pushes the loading mechanism. The load was monitored by using a load cell attached to the revolute joint of the loading frame. In order to apply a concentrated load, a curved support mechanism was situated in the middle of the support beam. This ensures that the load applied to the bearing remains aligned with the bearing centre axis.

Figure 4.1.2 shows the test bearing position in the scaled-down rig. The test bearing is mounted on the main shaft that is supported by two support bearings mounted at both sides of the shaft. This configuration makes the bearing replacing relatively simple as it requires to lift the loading mechanism and remove the housing of the support bearings.

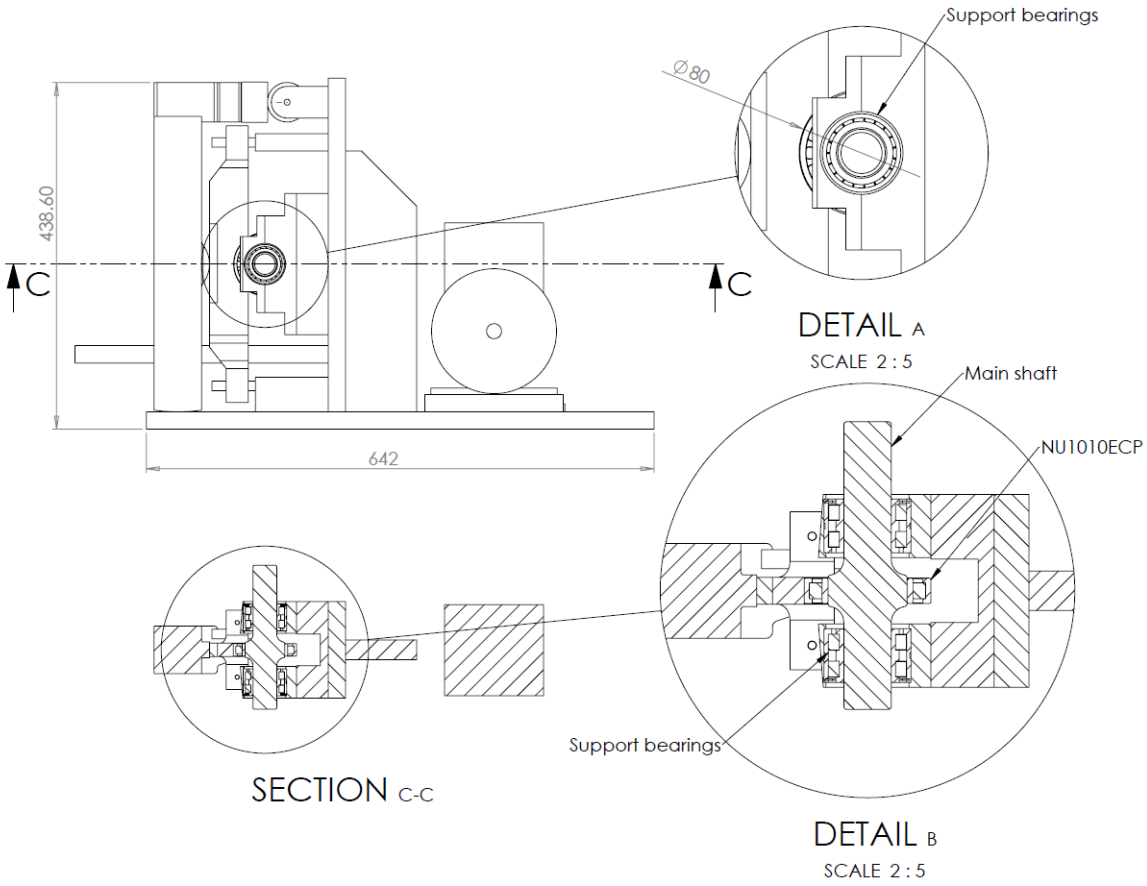


Figure 4.1.2: Scaled-down rig schematic, detail A and B show the NU1010ECP bearing position in the main shaft.

As mentioned before, this rig represents a scaled-down representation of the real size bearing used in WTGB. The bearing used in this test was the NU1010ECP manufactured by SKF and was selected as it geometrically resembles large scale roller bearing used for wind turbine applications. This also allowed the simulation of similar contact conditions to those observed in WTGBs. The dimensions of the NU1010ECP are shown in Table 4.1.1.

Table 4.1.1: Bearing dimensions for the scaled-down rig (NU1010 ECP) and the Multilife rig (NU 2244 ECP).

Parameters	**Symbol**	**NU 1010 ECP**	**NU 2244 ECP**
Outer raceway diameter:	D	80mm	400mm
Inner raceway diameter:	d	57.5mm	259mm
Roller number:	N_b	20	15
Roller diameter:	D_r	8mm	54mm

4.2 MultiLife rig

The MultiLife™Rig was designed by Ricardo UK Ltd to replicate the real conditions on a WTGB bearing and also to prove and validate the indexing mechanism concept. As described before, the main reason for this concept is to increase the remaining life in roller bearings by allowing the periodical rotation of the inner raceway away from the maximum contact load thereby allowing different contact points along the inner raceway during the bearing operation.

The rig as shown in Figure 4.2.1 is built similarly to the scaled-down rig discussed previously. It is composed of two arms (7) actuated by a hydraulic cylinder (8) that allows the application of a radially concentrated load on the test bearing (3) up to 1500kN. The rotation of the outer raceway is achieved by a pair of pulleys (2) with a reduction stage of 2.9:1 and a 7.5kW electric motor. The load is measured by a pair of load cells (6) installed as shear pins in the loading arms (7) at both ends of the stationary shaft (5). This allows controlling the radial load via a closed-loop control system. An inner-sleeve mechanism (4) is installed in between the test bearing (3) and the main shaft (5). This configuration allows monitoring the defects generated on the inner raceway of the test bearing using AE sensors. As this rig recreates the bearings located in the WTGB, the inner raceway is kept fixed. Therefore, a secondary roller bearing (1) is required to allow the rotation of the outer raceway of the test bearing (3). Additionally, this rig provides a temperature regulation system that comprises a radiator and a pump. This allows the oil temperature to remain

stable at 65°C and shut down at temperatures over 90°C. This rig uses a Castrol Hyspin VG 32 oil with a kinematic viscosity of 10cSt and it is recirculated around the test bearing during operation.

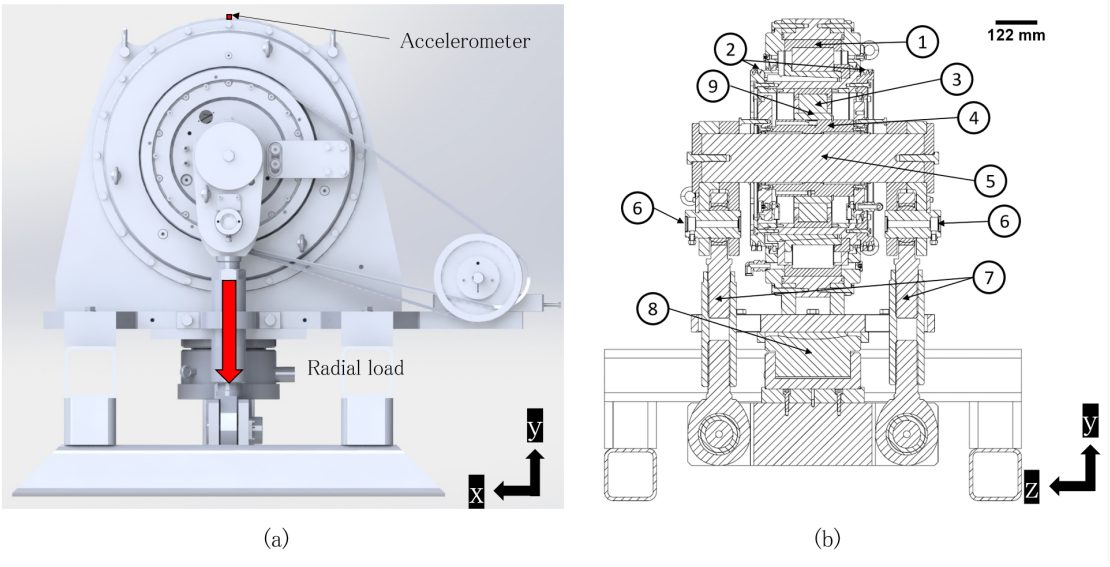
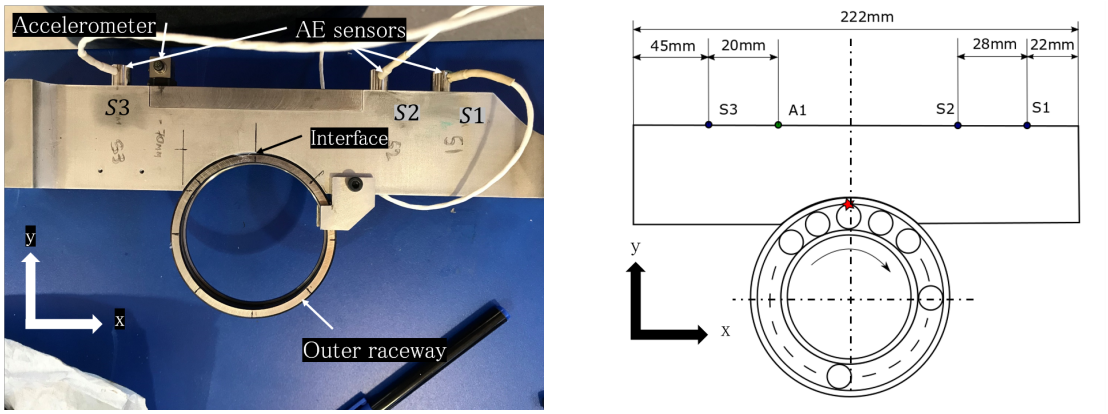


Figure 4.2.1: (a) MultiLife rig and (b) schematic of the internal components.

4.3 Instrumentation

4.3.1 Scaled-down rig

The rig was instrumented with AE and tri-axial acceleration sensors as shown in Figure 4.3.1a. A set of three AE sensors were attached to the loading frame of the rig as it was the closest structure near the bearing and allowed direct measurement of the AE signal. The acceleration sensor was installed on the same mechanism in order to measure the acceleration in all three axes. The acceleration sensor used for the experiment was the PCB 356B21 that allows a frequency range in the y and z-axis up to 10kHz and up to 7kHz for the x-axis.



(a) Fatigue rig instrumentation

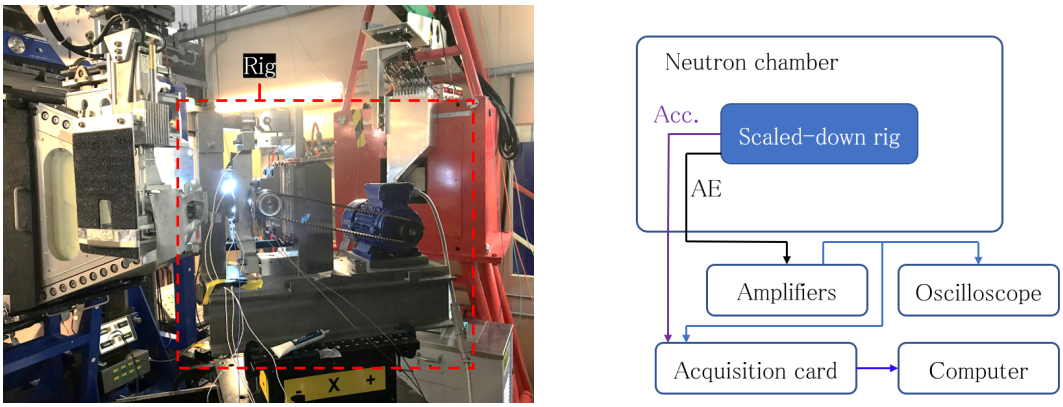
(b) Sensor position schematic

Figure 4.3.1: Accelerometer and AE sensor position on scaled-down rig.

The AE sensors used on this rig were the Mistras NANO30D which are suitable for mid-range applications at around 125–175kHz. These are similar to the NANO30 used in the MultiLife rig, with the main difference being the differential configuration as a method for the noise cancelling in the NANO30D model. The sensors were attached using a thin layer of cyanoacrylate on the top surface of the loading structure as shown in Figure 4.3.1a.

The AE signals were amplified to 40dB using Mistras 2/4/6 amplifiers. This amplification level was found to be the lowest value to amplify the signal during the experimental stage. This amplification value was selected based on the AE signal amplitude, a 20dB value generated signals with low signal-to-noise ratio whereas 60dB generated high clipping in the signals. The position of the sensors was chosen based on accessibility and proximity to the component of interest. This was to provide a clearer signal for further analysis and to allow fewer interruptions of the wave path propagation due to voids and interfaces on the mechanism. A fluoride-based grease was used as a coupling media in order to reduce the reflection coefficient between the outer raceway–loading mechanism interface due to air gaps and avoid inducing noise during the neutron diffraction experiment (neutron scattering, due to the presence of hydrogen in regular grease).

The acquisition system as shown in Figure 4.3.2b consisted of a computer, an oscilloscope, a NI cDAQ 9184 acquisition card with the NI cDAQ 9234 and cDAQ9223 modules for the AE and acceleration sensors respectively and a 28V power supply for the AE amplifiers. The AE signal was captured using a sampling rate of 1MHz which can be deemed acceptable, as a high portion of the energy content of AE signals occurs within a frequency range of 50kHz to 2MHz [114]. The vibrational signal was recorded using a sampling frequency of 51.2kHz, this allows a broad frequency range that contains all the defect frequencies of interest. The AE sensor amplification and accelerometer connections are detailed in Figure 4.3.3. The data was captured using a LabVIEW interface that allowed the data to be saved and monitored online. Due to the restrictions and radiation security requirements, the rig was contained in a security chamber as shown in Figure 4.3.2a. This required the installation of 4 metres of BNC cables at each sensor output to allow the acquisition system to stay outside the neutron chamber.



(a) Fatigue rig inside neutron chamber

(b) Acquisition system

Figure 4.3.2: (a) Rig inside measuring chamber, (b) Acquisition set-up.

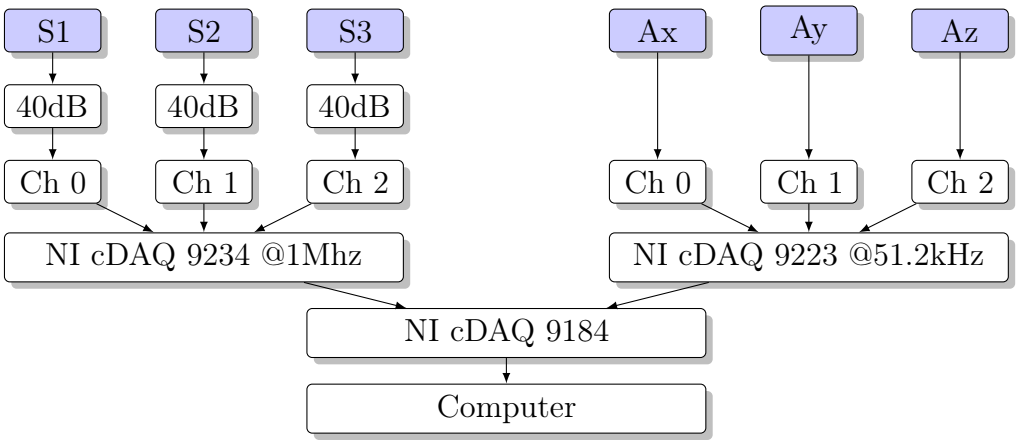
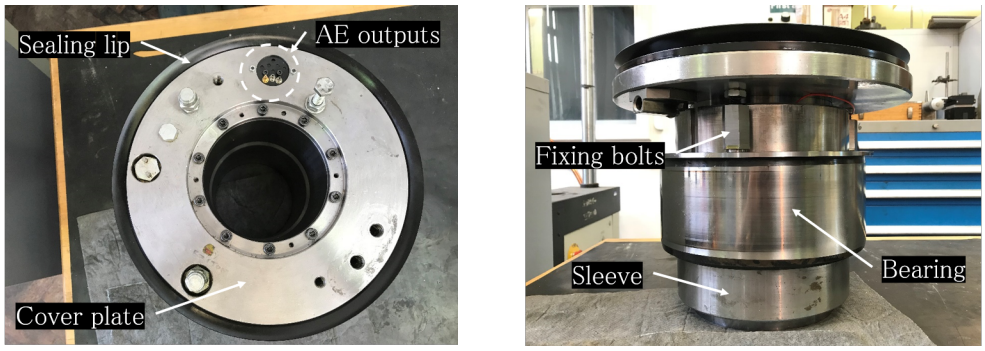


Figure 4.3.3: Amplification values and acquisition set-up.

4.3.2 MultiLife rig

The MultiLife rig had the PCB 356B21 accelerometer installed on the top of the rig at 12 o'clock with the y-axis collinear to the vertical axis of the rig. The data acquisition system is the same one described earlier for the scaled-down rig. The AE sensors were installed on a static sleeve mechanism as shown in Figure 4.3.4a. This component is fixed and is located between the inner shaft and the inner raceway of the bearing. A set of three AE sensors were attached using a spring mechanism that keeps a metallic coupling surface attached to the inner surface of the inner raceway and is separated at 120 degrees from each other as shown in Figure 4.3.5. The sleeve mechanism uses three NANO30 FO82 sensors with similar characteristics as the NANO30D. The signals were amplified by a set of three Mistras 2/4/6 standard AE amplifiers. The sensor amplification set-up was the same as the scaled-down rig set-up and the data were captured using the same equipment as specified previously.



(a) AE sensor output (b) Assebled sleeve and inner raceway

Figure 4.3.4: Sleeve mechanism for sensors attachment.

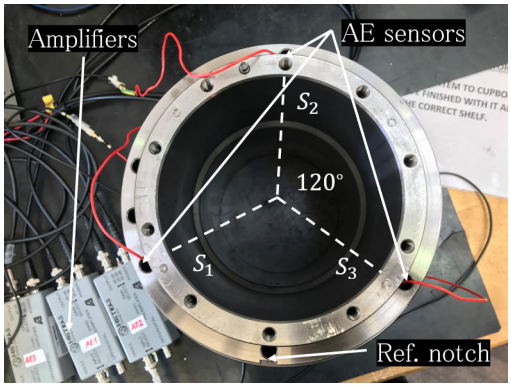


Figure 4.3.5: Installed AE sensors in the sleeve mechanism.

4.4 Chapter summary

- An introduction to the rigs used in this work was presented in this chapter. Although the MultiLife rig is more complex than the scaled-down rig in terms of size and surrounding elements, the working principle is the same as both rigs permit the application of a radial load on the test bearing and actuate it via an electric motor.

- Due to the simplicity of the scaled-down rig, the number of components involved is low, therefore, it allows the test bearing to be changed in less than 20 min, whereas the MultiLife rig requires at least 6 hours.

- A description of the sensors installed on the rigs was also presented. These configurations were used to perform novelty detection and damage localisation during operation. Moreover, the position of these sensors remained fixed once installed as any change in the current position might potentially induce different features.

CHAPTER 5

Experimental procedures

In this chapter the experimental procedures for novelty detection and damage localisation is presented. This chapter will explain the features sensitive to the damage position and their processing from the raw AE signals. Details about the signal preprocessing for both levels of detection are described together with the features used to define a machine/sensor baseline data. Additionally, some technical aspects are discussed such as sensor placement, TOA extraction techniques and most importantly a method to encode the features to allow the implementation of these features in a NN algorithm.

5.1 Seeding of a late onset of damage

As a means to provide a clear and realistic damage signature, a late onset of damage was simulated by machining a geometrically induced defect on the bearing surface. This defect simulates the surface damage caused by the progression of RCF, which ultimately leads to a progressive material loss in the bearing. Ideally, RCF is generated at highly localised stressed volumes beneath the contact area and with a width of up to $1000\mu m$ [115]. Certainly, once surface damage occurs, it might further extend, as depending on the loading conditions it might lead to higher damage sizes and consequently higher amounts of material loss. Additionally, further damage progression might lead to a distributed defect type around the contact area. In this work, a localised damaged type with a size of less than $1000\mu m$ has been used

to provide a clear vibrational and acoustic emission signature by assuring that the excitation is generated by a single source. Certainly, a more advanced damage extent might generate a more complex signature due to the interaction of multiple excitation points [116]. A range of 100–300 μm has been selected for the NU 1010 ECP and the NU 2244 ECP bearing respectively, as such values are within the expected ranges for RCF damage initiation and are representative for an early surface damage initiation. Additionally, it has been shown that a defect size with a similar range has generated a clear vibrational and AE response for a NU 2244 ECP using defects within that range [11, 117]. For the NU 1010 ECP, a defect of 100 μm has been selected, as in contrast with the NU 2244 ECP, the former is 5 times smaller and given that it will be subjected to a high rotational speed of 927 RPM, a defect around that size might produce an unclear AE signal.

The main reason is to detect, and later localise the defect. A localised damage was simulated on the surface of the bearing by inducing a small geometrical variation on the bearing surface. This was performed by using an Electrical Discharge Machining (EDM) procedure as this method allows the introduction of small defects of which normal machining tools are not capable. For the scaled-down experiment, an EDM notch of $100\mu m \times 1.75mm$ depth was machined on the surface of the bearing. The EDM notch was machined using a circular profile wire of $\varnothing 100\mu m$. This would allow a controlled defect to interact with the rollers and generate repetitive AE bursts during each interaction with the rollers by simulating a late onset of damage on the bearing surface. Figure 5.1.1 shows the defect schematic on the outer raceway of the NU1010 ECP, the machined bearing is shown in Figure 5.1.2.

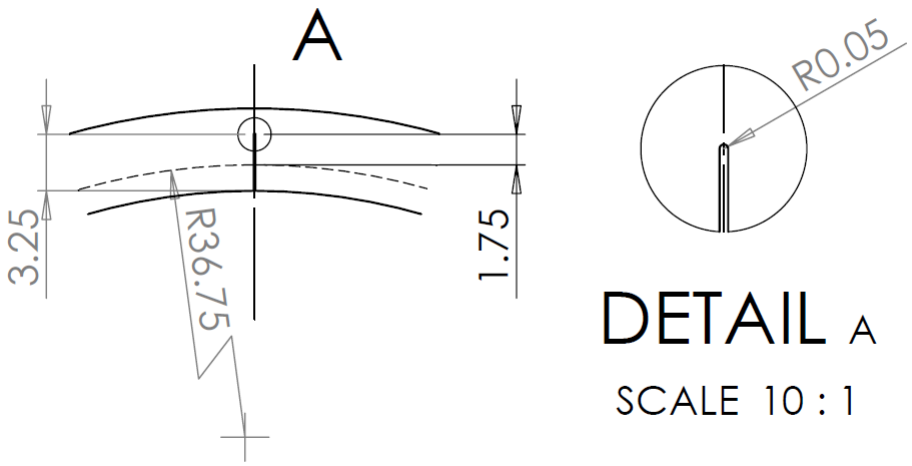


Figure 5.1.1: EDM notch schematic for the NU1010ECP used in the scaled-down rig, detail A shows the notch dimensions.

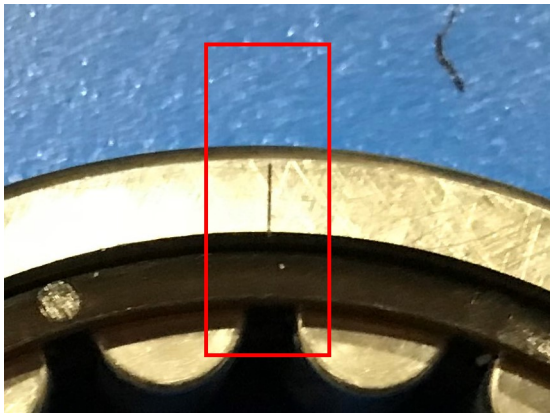


Figure 5.1.2: EDM notch on NU1010ECP.

A $300\mu m \times 600\mu m$ EDM notch was also introduced on the surface of the NU 2244 roller bearing for the MultiLife rig using a circular profile wire of $\varnothing 300\mu m$. A schematic of the bearing defect is shown in Figure 5.1.3. A measurement of the defect using a profilometer across the mid-section of the bearing is shown in Figure 5.1.4.

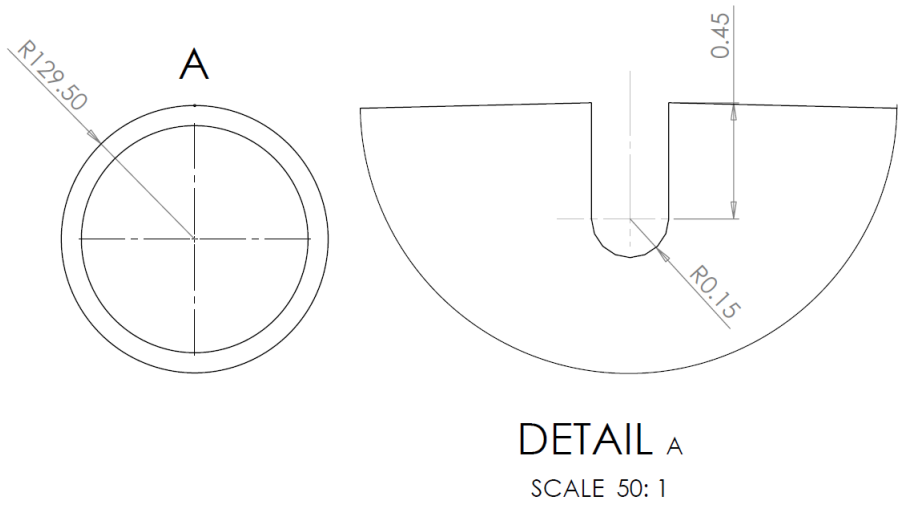


Figure 5.1.3: EDM notch schematic for the NU2424ECP used in the Multilife rig, detail A shows the notch dimensions.

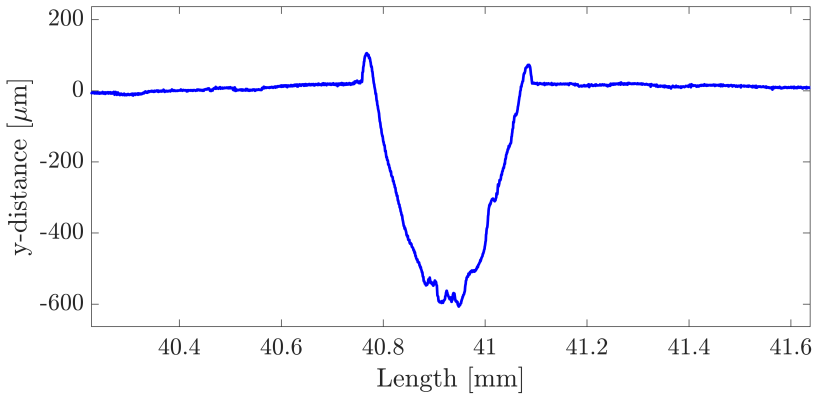


Figure 5.1.4: EDM notch on NU2244 measured on the mid-section.

Additionally, this work also allowed the testing of a new bearing defect concept by seeding a subsurface plastic deformation beneath the surface of the bearing whereas an EDM notch only simulates a very late damage state. This concept was implemented to simulated damage due to overload events mostly generated by transients and sudden load peaks, this being the main reasons for bearing lifetime reduction. As this defect intends to accelerate damage in roller bearings, it led to the design of a defect based on the severity of the induced damage. This quantification led to the definition of the defect size based on the amount of plastic deformation under the bearing surface. The design involved developing a FEM model where the amount of plastic deformation was validated using Neutron transmission imaging.

As a result, the implementation of this methodology showed good agreement with the level of damaged induced in the subsurface of the bearing [2]. The methodology for this concept is described in the following section.

5.2 Seeding of an overload defect

So far, most of the damage detection and localisation methods on roller bearings have been tested with a late damage onset without considering a method that allows to represent more realistic damage. This section is a study of bearing damage due to overload events and provides a means to further apply this technique and test the robustness of damage detection and localisation stages by reducing the life expectancy of the bearing. The study highlights the development of a model that predicts the plastic deformation beneath the surface; thus allowing the control of the damage extension by systematically increasing the plastic deformation. It is important to mention that this work was the result of a collaborative study that focused on measuring the plastic deformation generated during overload events on roller bearings, these deformations were measured by using non-destructive techniques and used to understand the effects of overload events in the remaining life in bearings [2]. The results obtained from the model were validated with his experimental data using Neutron Bragg Imaging techniques in the ISIS Neutron and Muon Source facilities.

Overload events have been associated with a reduction in the life expectancy of bearings due to sudden transient events. It is well-known that accumulation of plastic strain beneath the surface will lead to a premature failure in bearings. Even though current models such as the L_{10} model take into account RCF as the main failure mode in bearings, the effects of overloading leads this model to overestimate the life expectancy. Currently, the methods for determining the evolution of plastic deformation are relatively limited thereby restricting the development of reliable models that predict plastic deformation. The current study develops a FEM of a scaled-down bearing where progression of the plastic zone is validated using Neutron Bragg Imaging. This method allows the generation of a 2-D map of the broadening parameter that has been shown to be an indicative of plastic deformation in materials [118–120].

5.2.1 Finite Element Model

A FEM of the NU1010ECP bearing was developed using the FEM package Abaqus 6.14-1. The model tried to simulate the elastic-plastic behaviour governed by the distortion energy hypothesis. The model simulates the bearing overload using the Fixed Bearing Loading Frame (FBLF) shown in Figure 5.2.1 by predicting the contact stresses in the contact region between the roller and the outer raceway. In order to reduce the computational time, the model was simplified by assuming symmetry on both x and θ axis as showed in Figure 5.2.2. This allowed the modelling of only one-quarter of the bearing and consequently, the reduction of the number of elements in the model.

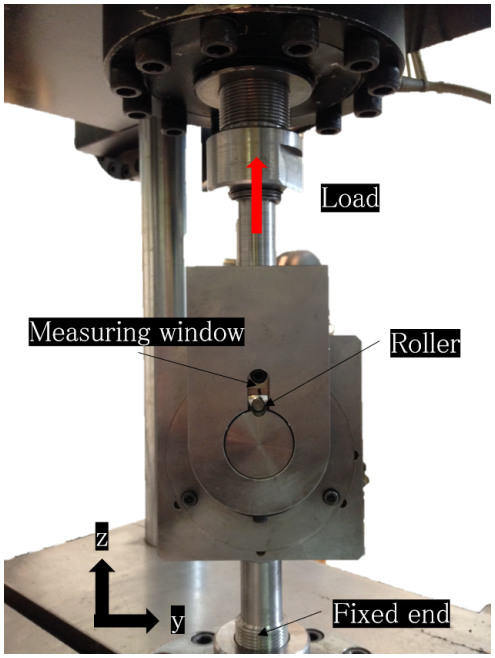


Figure 5.2.1: Loading frame with NU1010 ECP installed in a 40kN Instron tensile machine.

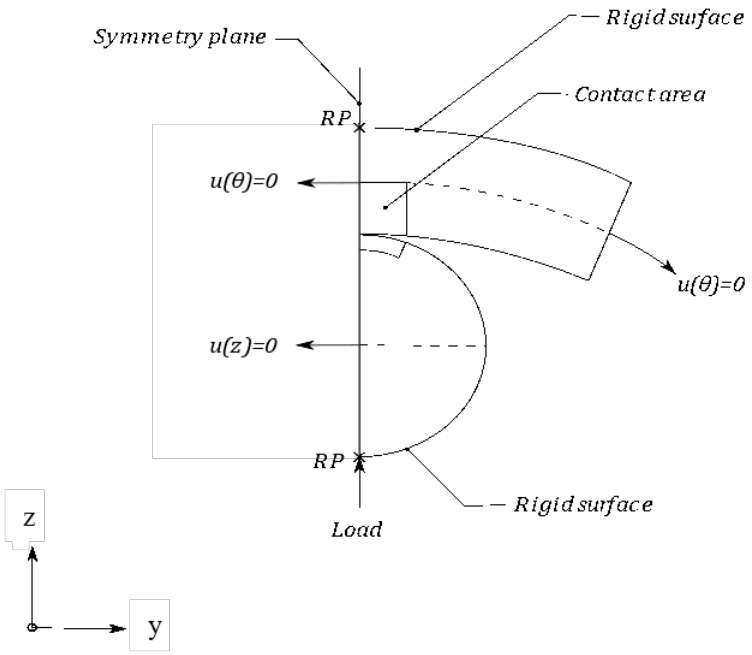


Figure 5.2.2: Boundary conditions in FE model.

Contact stresses are generated in relatively small and localised volumes and normally exhibit high gradients in the subsurface of the component in contact. The analysis of such problems requires a large number of elements that result in a high computational effort. This situation can be mitigated by using a dense mesh of the area of interest while keeping a coarse mesh on the rest of the component. In this case, a small volume in the bearing of $1.5 \times 0.8 \times 5$ mm was refined in the contact zone and isolated in order to represent the gauge volume measurement. This allowed the analysis of the yield zone and the stress profiles and comparison with experimental data. The model was simulated by neglecting the elastic deformation on the surfaces of the loading frame, allowing them to be modelled as rigid bodies. The material stress-strain curves and properties of the AISI 52100 are shown in Figure 5.2.3 and Table 5.2.1 respectively [121].

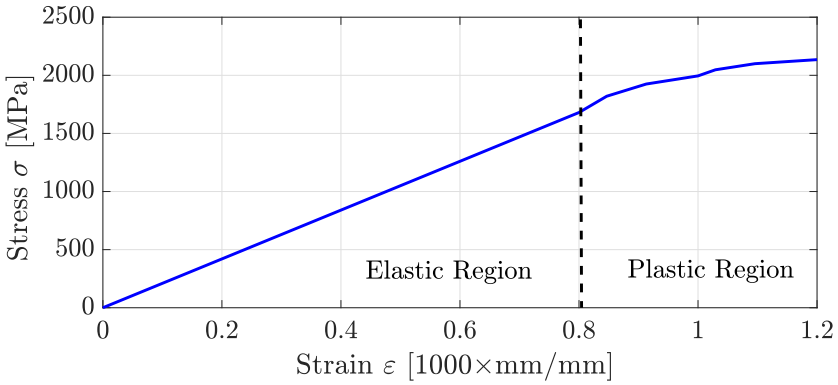


Figure 5.2.3: AISI 52100 Engineering stress-strain curve.

Table 5.2.1: NU1010ECP material properties.

Property	**Symbol**	**Value**
Density [kg/m^3]:	ρ	7827
Young's Modulus [GPa]:	E	201.33
Poisson's ratio:	ν	0.3
Yield Stress [MPa]:	σ_Y	1410.17

The solid components were modelled using quadratic elements of second order (C3D20R) whereas the rigid bodies were modelled with bilinear rigid quadrilateral elements (R3D4). The use of quadratic hexahedral elements have been used in this problem in order to accurately capture the high gradient stresses generated in the contact zone. The contact zone was refined using $20\mu m \times 20\mu m$ elements where the elements of the roller where twice that size. This refinement allowed a high resolution on the stresses generated beneath the contact and allowed extracting the values from the elements using with the same resolution used in the experiments. The surface of the roller was considered as a master surface in order to avoid penetrations. This stems from the assumption that the roller surface is relatively hard compared with the outer raceway. The contact constraint enforcement used a linear penalty method with a default stiffness factor of 10, at this value contact penetrations will not have significant influences in the results [122]. The overload simulation assumed quasi-static conditions and was executed in the ABAQUS/Standard solver. This required including parameters such as the density of the material even though the

inertial effects are going to be neglected and solving convergence issues during contact initialisation problems in load-controlled models.

5.2.2 Elastic and Plastic Analysis

During the overload simulation, two different regions are expected to occur: an elastic and a plastic behaviour. Theoretically, the minimum load that generates plastic deformation and its generated contact pressure can be calculated by assuming Hertzian contact by using Equations 5.2.1 and 5.2.2 [123]. It is important to mention that, beyond these limits, the Hertzian assumption is no longer valid as elastic behaviour is no longer expected.

$$F^Y = \frac{L\pi R^\star}{E^\star}(p^Y)^2 \tag{5.2.1}$$

$$p^Y = 1.67\sigma_Y \tag{5.2.2}$$

where the terms L, $E^\star$,$R^\star$ are the contact length, the equivalent Young's modulus and the equivalent radius respectively. As both roller and outer raceway are assumed to be manufactured using the same material, the equivalent Young's Modulus can be defined as:

$$E^\star = \frac{E}{2(1-\nu^2)} \tag{5.2.3}$$

$$R^\star = \left(\frac{1}{R_1} + \frac{1}{R_2}\right)^{-1} \tag{5.2.4}$$

These limits were assessed using both analytical forms and the FEM resulted in the predictions shown in Table 5.2.2. These results show good agreement in terms of the elastic limit. Beyond that, the plastic evolution, as shown in Figure 5.2.4, cannot be described analytically and therefore needs to be validated experimentally.

Table 5.2.2: Elastic limits calculation using Hertzian contact theory and Finite Element.

	FEM	**Analytical solution**	**Absolute error** [%]
p^Y[MPa]	2894.92	2805.6	3.18
F^Y[N]	9894.92	9285	3.90

In order to produce a quantitative measure of the induced damage, the elements that reached the plastic region described by the Von Mises criteria were flagged in the model. Figure 5.2.4 shows the evolution of the plastic zone beneath a contact point. As the results show, the plastic zone begins just below the surface at 10.3 kN and progressively expands until it reaches the surface of the contact at 34kN. Notice that only the right half of the contours are shown in 5.2.4 as these results are expected to be symmetric due to geometry and load symmetry conditions.

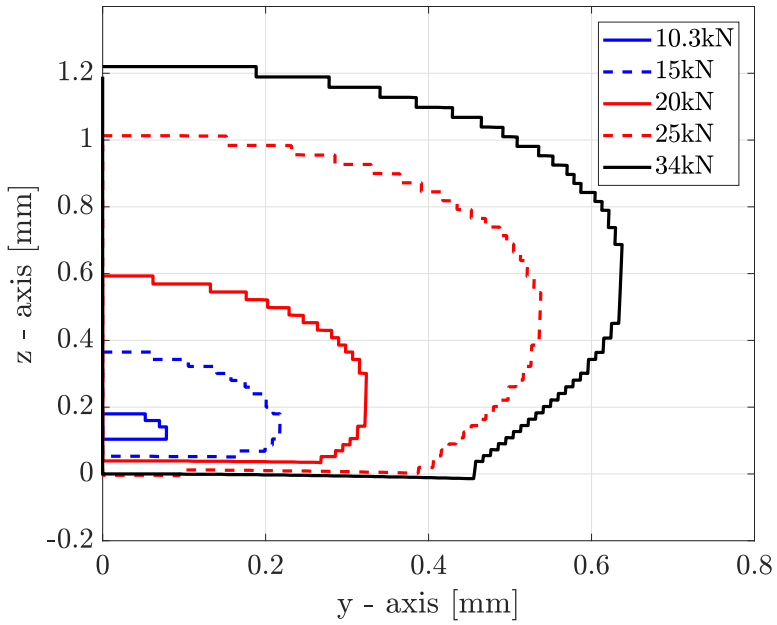


Figure 5.2.4: Overload zone in the outer raceway using FE Model, the constant contours shows the area of the plastic formation at different loads.

5.2.3 Validation using Neutron Bragg Imaging

Even though the method described in this section is beyond the scope of this thesis, as this was mainly the work developed by Anthony Reid, the details can be found in [2]. The main goal was to monitor different plastic deformation amounts using an indicative parameter for plastic deformation defined as broadening parameter σ [124]. Figure 5.2.5 shows the Bragg edge where the shift in edge position λ and broadening of the parameter σ are associated with the elastic and plastic deformation respectively. The measurements were performed on the ENGIN-X instrument that features an array of detectors for each pixel, therefore generating a map of σ values in the area of interest.

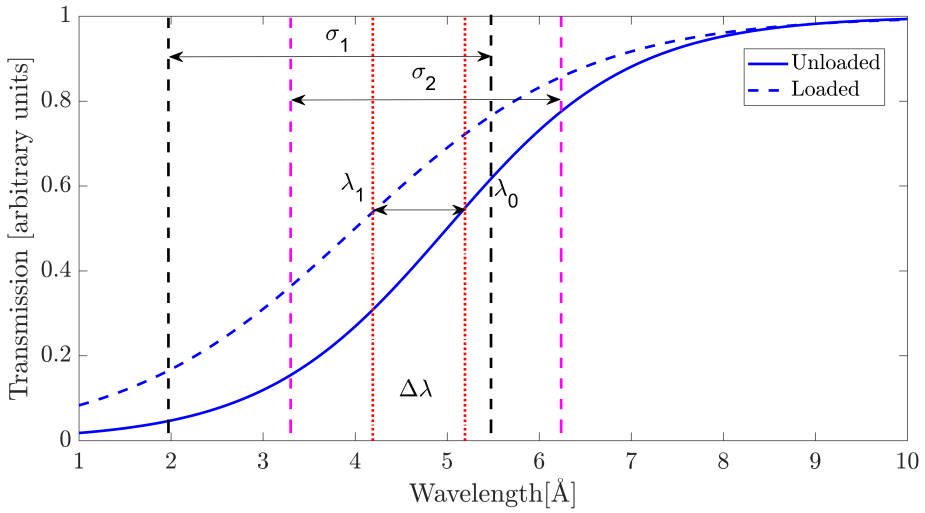


Figure 5.2.5: Transmission spectra where the σ_1 and σ_2 parameters correspond to an unloaded and loaded scenario respectively.

A 50kN Instron tensile machine was placed inside the instrumentation chamber to allow measuring a map for an unloaded and loaded state. In order to allow a concentrated load to be transferred on to the contact zone of the bearing, a fixed raceway loading frame (FRLF), as shown in Figure 5.2.1, was conjointly designed. This allowed the installation of a unique roller on the shaft that allowed it to press against the outer raceway installed inside the main frame.

In order to perform a comparison between the experimental data and the FEM, three different loads beyond the elastic limit were selected. In this case, maps at loads of

15kN, 25kN and 34kN were measured and compared against the model as shown in Figure 5.2.6. Even though there is considerable agreement between the experimental data and the model, there are still a few discrepancies, notably at the volumes outside the contact region. The main reasons are mostly speculative as there were not enough samples to provide repeatable data, thereby attributed to experimental error or changes in microstructure due to overload or during manufacturing.

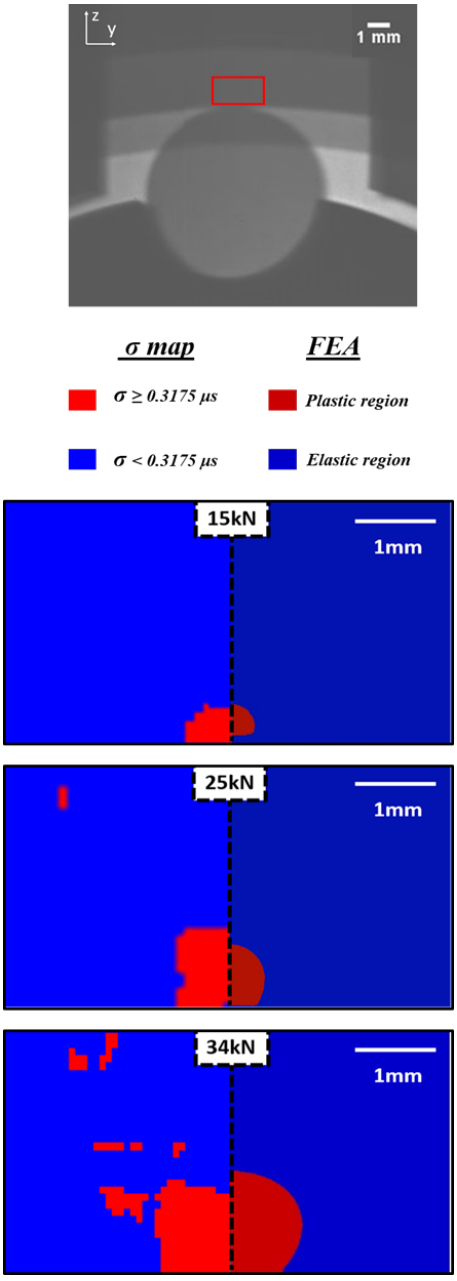


Figure 5.2.6: Finite Element vs Sigma parameter. The Finite Element solution was compared with the experimental data where the blue and red zone represent the elastic and the plastic region [2].

5.2.4 Observations on fatigue life

A pristine bearing was overloaded with 34kN which was chosen to be the most severe loading condition in both experimental and model predictions. This bearing was later on tested in a fatigue machine with a cyclic load at 15Hz with a load of $F_{max} = 10$ kN. This load was based on the elastic limit of the bearing as calculated using Equation 5.2.1. The lower limit F_{min} was selected to have a relatively low value giving a load ratio F_{max}/F_{min} of 1×10^{-3}. After 5 million cycles which is equivalent to 2.78×10^5 revolutions the bearing showed an evident crack that propagated from the contact area and reached the outer surface of the outer raceway. Considering a dynamic load rating C of 46.8kN and a radial load P of 36kN obtained from Stribeck's relation defined as [125]:

$$F_r = \frac{5P}{Z} = F_{max} \tag{5.2.5}$$

where F_r and Z represent the maximum load at the bearing contact zone and the number of roller in the bearing respectively. For a pristine bearing the L_{10} prediction model gives a life estimation of 2.39×10^6 revolutions, therefore this resulted in a life reduction of 88.36% due to the induced overload. Figure 5.2.7 shows a cross-sectional view of the bearing with a shear stress profile being plotted along the contact depth. A band of white etching area that coincides with the higher range of shear stress in the subsurface (-1600–1700 Mpa) at 300μm may be observed. Further investigation using the SEM microscope suggested a crack initiation from the subsurface to the contact area.

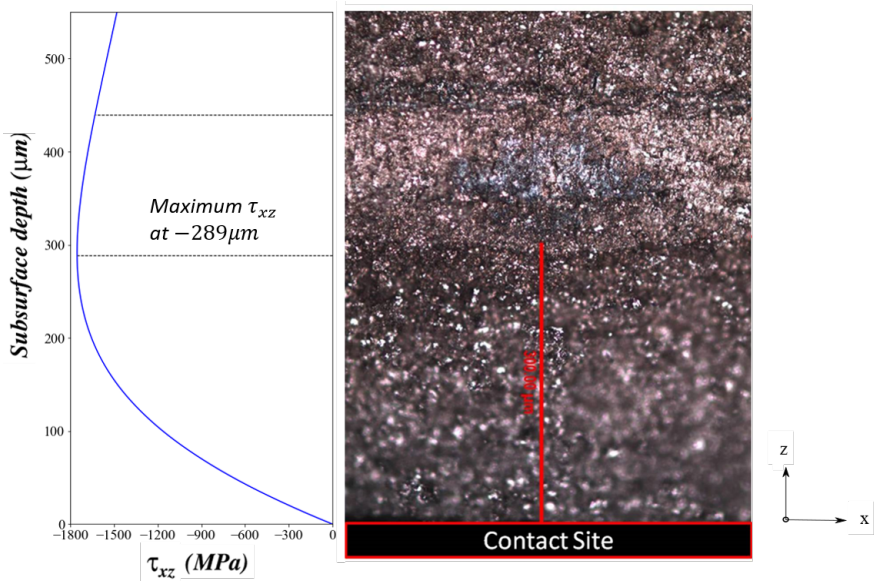


Figure 5.2.7: Shear stress profile (left) and cross-sectional area of the bearing with a 20X magnification (right), courtesy of Anthony Reid.

5.3 Data processing and feature extraction for novelty detection

Sensors by themselves cannot measure damage but rather damage may be inferred from the data they acquire. In this section, the data processing and feature selection process for the scaled-down and MultiLife rig will be explained. This is one of the most important aspects of novelty detection as the features selected should be sensitive to damage conditions. In this thesis, vibrational data was used as it has been shown to be the most robust method of determining damage in bearings through the analysis of defect frequency components in the frequency domain.

The processing stage can be represented as a series of levels as described in Figure 5.3.1 where the raw signal is refined into features that can be used for damage detection purposes. This refinement, might involve a domain transformation, as the domain used to extract this features can bring a significant impact on the novelty detector performance. A good example can be seen with time domain features and frequency domain features, where normally frequency domain features give more information than time domain features for less data. The data processing stages

described in Figure 5.3.1 requires measuring a raw signal from a sensor. Even though the raw signal might show apparent features, in practise, further signal processing might be required due to potential noise and data redundancy encountered in the raw signal. The preprocessing stage comprises two tasks: data cleansing and dimensional reduction. The data cleansing stage removes the noise, spikes, and outliers whereas the dimensional reduction removes the data redundancy by transforming the data into a set of sensitive features. Consequently, these features are assembled into patterns representing normal and damage conditions and ultimately used in the novelty detector.

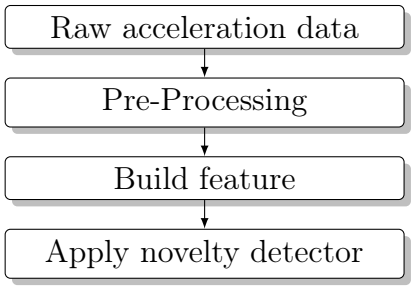


Figure 5.3.1: Data processing stages.

5.3.1 Data pre-processing

In this case, the time domain acceleration signal obtained in the direction of the applied radial load was pre-processed using a low-pass filter and a Hilbert transform approach in order to enhance the signal. The HT was used in the Multilife rig to accentuate the periodicity of the peaks at the characteristic defect frequency due to the low amplitude response at the frequency range of interest. Previous use of low-pass filtering methods resulted in relatively ambiguous features whereas the use of the HT showed to improve them. This has been highlighted in many studies in CM applications for WTGB and have suggested the use of envelope analysis for demodulation purposes [44, 126].

Figure 5.3.2 shows the signal preprocessing steps for feature extraction. The features were extracted by using a STFT method, this method allows the analysis of the signal in the time-frequency domain by performing a homogeneous segmentation of the signal in a finite time interval. Although this sounds practical, additional precautions

have to be taken into account before processing the signal. In a practical context, the use of a raw vibrational signal might have a detrimental effect on the features due to noise, additionally, the effect of sampling a signal within a finite time interval in order to produce features might induce frequency leakage issues in the spectral signatures. This ultimately has a "smearing" effect on the spectral signature, therefore, having a negative impact on the novelty detection algorithm. Even though multiple options for windowing are available, a Hanning window was selected based on its capabilities to reduce the frequency leakage and accurate frequency resolution [127]. The condition is therefore obtained from the spectral content of the signal and any defect on the bearing should generate a spectral signature that shows the corresponding peak at the characteristic defect frequency.

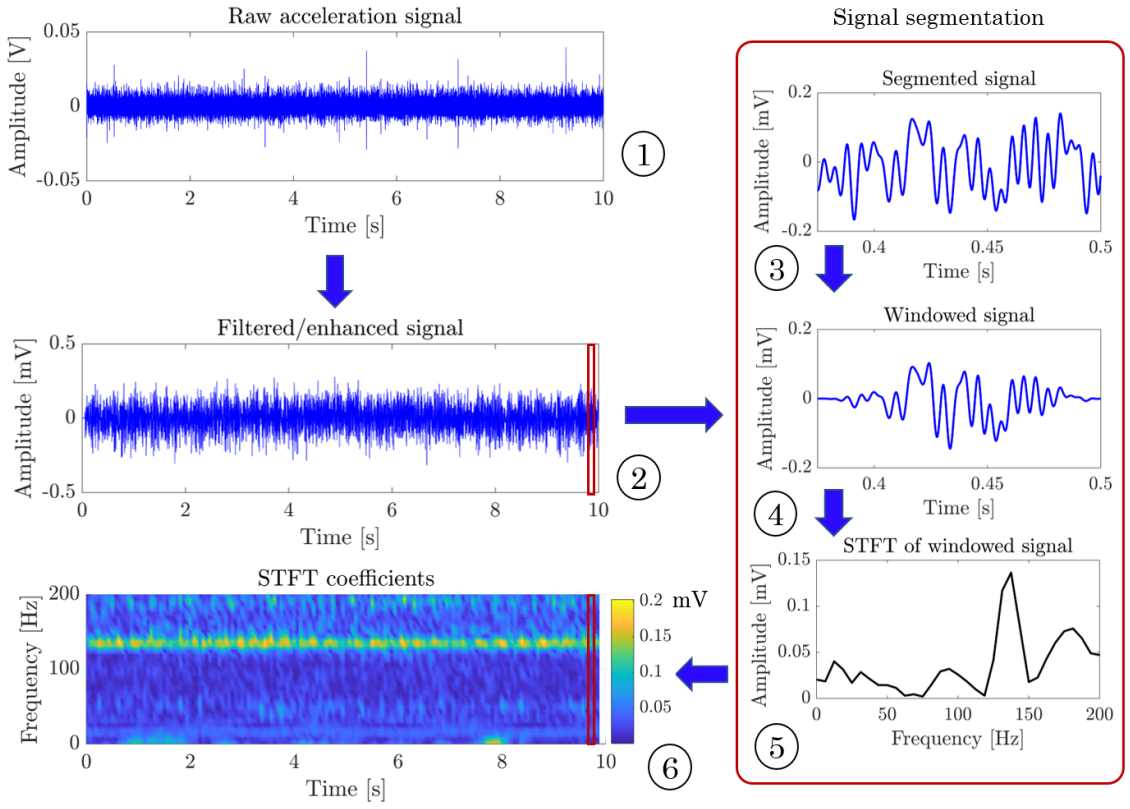


Figure 5.3.2: Step by step approach for feature extraction: (1) Raw signal sampling, (2) Signal enhancement, (3) Signal segmentation, (4) Windowing, (5) STFT of signal segment and (6) Feature assembly.

5.3.2 Feature selection

Defects in bearings have been widely studied in terms of their vibrational signature and this approach provides information of the presence of defects in their components. This is the traditional method to perform CM based on changes in the vibrational data. As in the case for bearings, any presence of damage in the inner raceway will be represented as the spectra data shown in Figure 5.3.3 where a component associated with an inner raceway damage onset appears in the vibration signature.

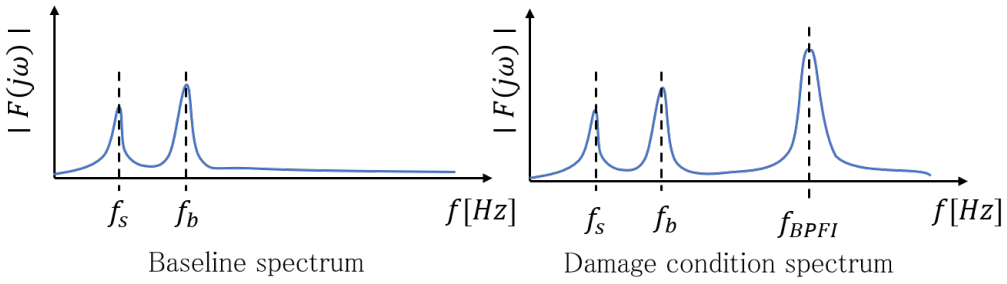


Figure 5.3.3: Two different spectral signatures from a baseline and damaged condition.

A feature vector can be obtained by selecting the frequency components that spans the characteristic defect frequency components. This will finally lead to feature vectors that represents the undamaged and damaged states as shown in Figure 5.3.4. The appearance of this peak at that characteristic frequency is related to the periodic interaction between the defect and the rollers. The use of these set of features allows to monitor changes at the frequency components that are related to a specific characteristic frequencies that can be calculated using the equations described on Table 2.1.1.

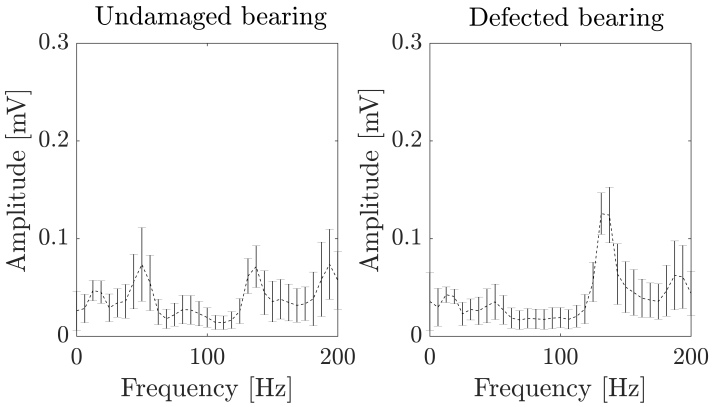


Figure 5.3.4: Feature vector for fatigue rig where a characteristic outer raceway defect frequency appears at 135 Hz.

The novelty detection stage was performed by monitoring the amplitude features obtained from the STFT where a baseline and damage state were obtained by using a bearing under pristine and damaged conditions. This set of data was later encoded as a feature vector with a frequency resolution specified on Table 5.3.1. These frequency resolutions are allowed to have a minimum time resolution of $79\mu s$ as calculated from Equation 2.1.2. In order to provide a meaningful feature vector the frequency range of the feature was defined as a range that contains the characteristic defect frequency. The feature vector used in this work was defined as:

$$\mathbf{x} = (x_0, x_1, ..., x_D)^T \tag{5.3.1}$$

where the elements of this vector will correspond to the equally spaced frequency components and the dimension of the vector will be determined by the number of frequency components D.

Table 5.3.1: Summary pre-processing stage.

Parameters	**Scaled-down rig**	**Multilife rig**
Bearing model:	NU1010 ECP	NU2244
Signal processing:	Low-pass filtering	Hilbert-Transform
Frequency range:	25–150Hz	1–20Hz
Frequency resolution:	6 Hz	1 Hz
Feature dimension:	20	20
Number of samples per condition:	954	1590

5.3.3 Scaled-down rig features

This rig was only designed to run at a specific rotational speed of 927RPM. This speed was defined according to the specifications of the experiment performed in parallel as it was the speed required to accelerate the bearing failure. This allowed the calculation of the outer raceway defect frequency component as 135.7Hz. These features were obtained after filtering the vibrational signal with a low-pass filter which removed the noise at higher frequencies. The filter parameters are shown in Table 5.3.2

Table 5.3.2: Low-Pass filter parameters.

$f_s[Hz]$	$f_{pass}[Hz]$	$f_{stop}[Hz]$	$A_1[dB]$	$A_2[dB]$
51200	200	210	1	40

where f_s, f_{pass}, f_{stop}, A_1, A_2 represent the sampling frequency, the band-pass frequency, the band-stop frequency, the ripple attenuation level and the attenuation level respectively.

As may be seen in Figure 5.3.5, there is a clear distinction between the damaged and undamaged state in the spectrogram. A high narrow peak component appears around the expected frequency of 135Hz showing a periodic defect interaction between the rollers and the outer raceway defect. Several other frequency components are also visible such as the shaft rotation frequency f_s shown as a dashed white line in Figure 5.3.5. Both set of damage features have shown to be consistent over time and remained easily distinguishable. Additionally, a peak at 50Hz appears in the features, this can potentially suggest a variation induced by the support bearings installed on both ends of the shaft.

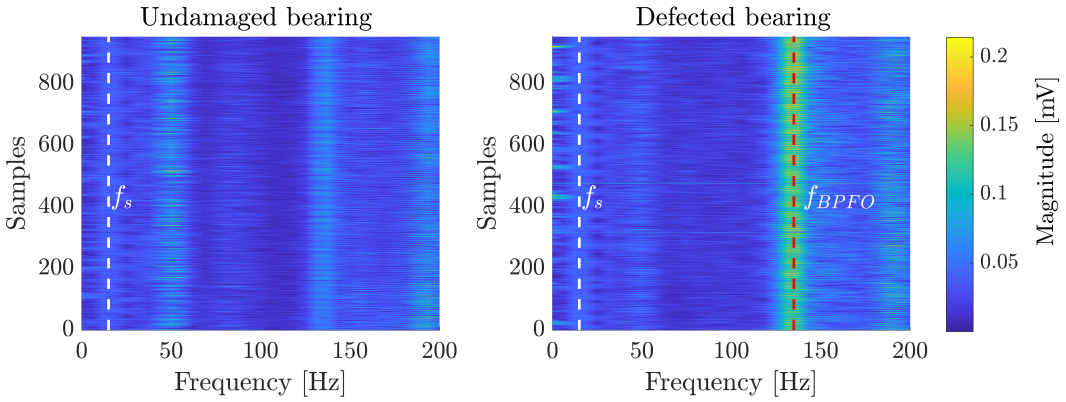


Figure 5.3.5: STFT data within 0 to 200Hz. Image on the left shows spectrum from a clean bearing and on the right from a defected outer raceway.

The dimensions shown in Table 4.1.1 allow the calculation of the Ball Pass Frequency in the inner (BPFI) and outer (BPFO) raceway using Equations 2.1.1. The resulting frequencies are shown in Table 5.3.3.

Table 5.3.3: Bearing damage characteristic frequencies values.

Element	**Symbol**	**Bearings**					
		Speed [RPM]					
		NU 1010 ECP	NU 2244 ECP				
		927	20	40	60	80	100
Inner raceway	f_{BPFI} [Hz]	173.37	2.93	5.86	8.79	11.72	14.66
Outer raceway	f_{BPFO} [Hz]	135.63	2.07	4.14	6.21	8.275	10.34

5.3.4 Run-to-Failure test on scaled-down rig

Until now, the novelty detection was performed by using a geometrically seeded defect on the bearing surface and a priori knowledge of the defect existence was available. In a real situation, this information will not be available, otherwise, a novelty detection stage would not be necessary. The objective of this experiment was to use the vibrational data in order to predict damage onset on an overloaded bearing. The experiments were performed in parallel with the neutron diffraction experiments as they allowed to monitor the test bearing condition during the test.

A neutron diffraction experiment tried to understand the strain evolution after overload events in bearings by recreating an overload defect as previously explained and measure the dynamic strain generated in the contact area of the bearing during operation. A 40kN radial load on the shaft of the bearing subjected the bearing to fatigue by periodically cycling the stresses generated on the contact zone during each roller pass. This value was found to be the minimum load in order to initiate a plastic deformation in the subsurface of the material.

As there was a limited time frame to perform the experiments, a rotational speed of 927 RPM was selected to increase the probability of an early failure. Consequently, for a pristine bearing loaded at the elastic load limit at 40kN, and with a dynamic load rating C of 46.8kN, the L10 model predicts a life estimation of 1.68 million cycles. This represents a life estimation of 30hrs at a rotational speed of 927 RPM. In terms of the speed magnitude and its comparison with a real WTGB, such magnitudes are not uncommon to be found, as healthy gearboxes normally reach up to 1800RPMs in the high-speed shaft connected to the generator [12]. Therefore, this speed can be deemed acceptable as it falls within the expected range of operation in a WTGB. Moreover, this speed was kept fixed for all the test conditions, as it allows consistency in the vibrational response of the bearing.

The measurements were performed in parallel with the neutron diffraction experiment and provided means to estimate a damage onset during the test. For this test the data obtained for the baseline were entirely used as the test set and taken as the reference of the normal condition state. The experiment consisted of continuously measuring the vibrational signature of the bearing and recording AE signals for further damage localisation.

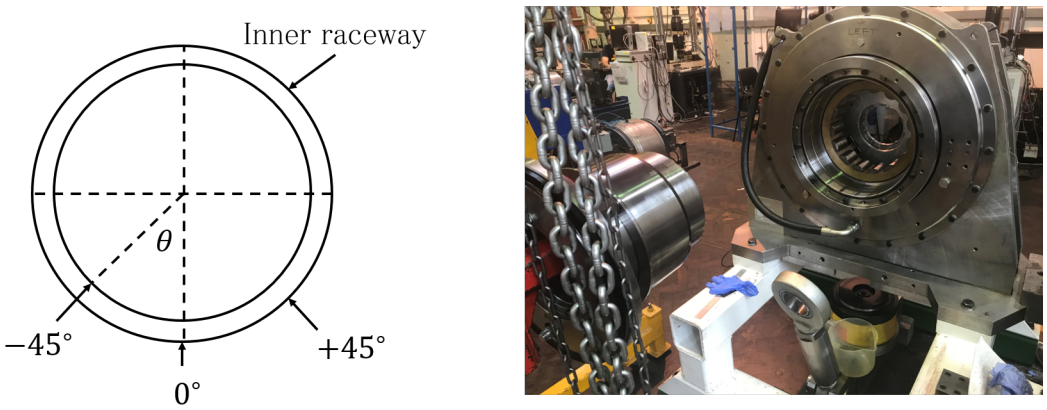
Clearly, this long term measurement experiment led to some restrictions in the amount of data for recording; Especially with AE, as the storage capacity was limited by the sampling frequency of 1MHz in this case. This resulted in storing the vibrational and AE data in intervals of 15 minutes for two consecutive days.

5.3.5 Multilife rig features

In contrast to the scaled-down rig, a defect was machined on the bearing inner raceway of the Multilife rig, as previously described; This is mainly because this component remains stationary and it is constantly subjected to localised fatigue points whereas the outer raceway constantly changes its orientation and thus redistributes the contact points around the loading zone.

Additionally, three different damage angles were selected in order to represent the damage scenarios around the loading zone. The angles selected were along the bearing loading zone at $0°, +45°$ and $-45°$ as shown in Figure 5.3.6(a). The procedure was performed by using the same bearing rotated at the angles mentioned before. Figure 5.3.6 shows the mounting procedure of the inner raceway into the MultiLife rig.

The bearing was oriented taking as a reference a zero degree reference notch machined on the sleeve component. Due to size and weight limitations, the bearing replacement cannot be done without lifting equipment. A 5-tonne crane was used to perform the manoeuvre as shown in Figure 5.3.6(b).



(a) (b)

Figure 5.3.6: (a) EDM notch at $-45°$ and (b) Mounting of the inner raceway into MultiLife rig.

Even though the installation might induce certain randomness in the results, efforts were made to reduce such variations by following an installation procedure and keeping them consistent for each bearing replacement. Additionally, the MultiLife rig has mechanical features that allows the self alignment of different components by means of alignment pins and guides thereby resulting in the reduction of variation during each installation process. The bearing installation process can be described as the following steps:

1. Remove/Install the power supply and sensors around the rig.
2. Remove/Install the guards.
3. Remove/Install hydraulic hoses.
4. Remove/Install the main belts.
5. Remove/Install the loading arms and hydraulic jack.
6. Remove/Install the main shaft using a 5-tonne crane.
7. Remove/Install the cover to access the main bearing.
8. Remove/Install the main bearing using a bearing extractor mounted in a 5-tonne crane.
9. Remove/Rotate bearing in the sleeve.
10. Assemble the rig following the reverse order.

Baseline data was obtained from the clean bearing at 20, 40, 60, 80 and 100 RPMs as shown in Figure 5.3.7. The details shown were mainly due to the shifting of the shaft and ball spin frequency components. A radial load of 10kN, which is 78.5 times less than the dynamic load rating, was maintained constant during the experiments by using a hydraulic jack installed in the rig. This load was maintained at a relatively low value in order to avoid any significant deflection of the inner raceway and to avoid any damage progression.

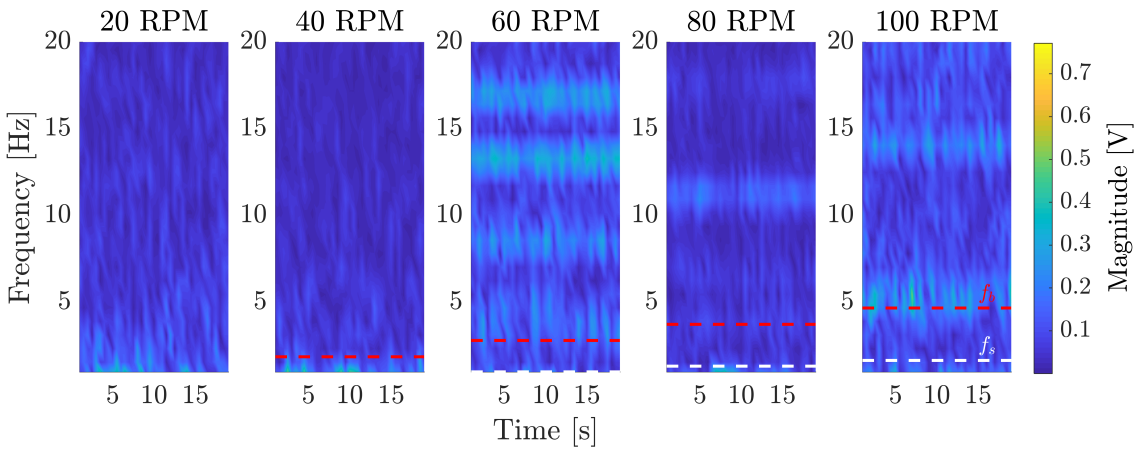


Figure 5.3.7: Baseline data at different rotational speeds. The white and the red dashed line represent the shaft speed frequency and ball pass frequency component f_s and f_b.

The features obtained in the MultiLife rig were rather more difficult to obtain, due to the subtle defect iteration between the inner raceway and the rollers and due to the highly modulated signal as showed in Figure 5.3.8. These features, as mentioned before, were enhanced during the pre-processing by obtaining the envelope of the acceleration signal and finally processed by applying the STFT.

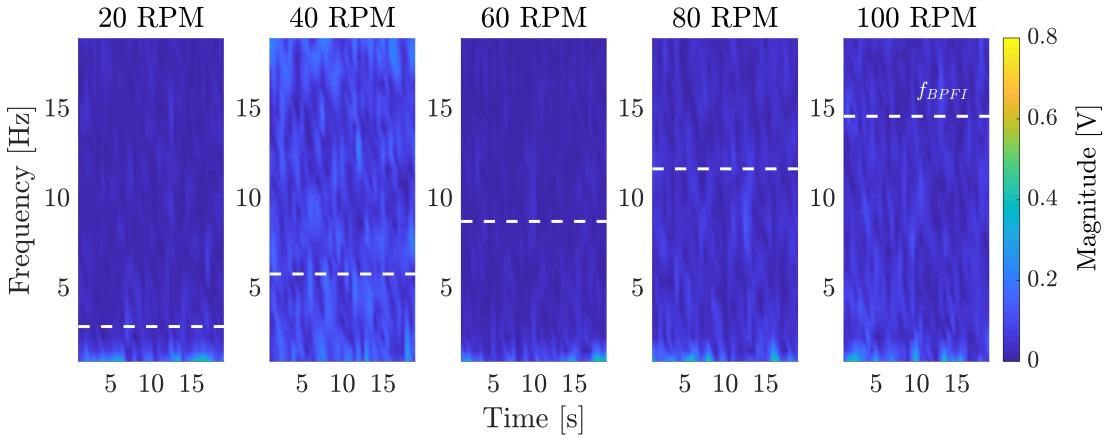


Figure 5.3.8: Weak features obtained after filtering the signal obtained at different rotational speeds.

The damage feature was obtained by replacing the undamaged bearing in the MultiLife rig and taking the acceleration measurements in the vertical axis, collinear with the applied radial load.

The features are shown in Figure 5.3.9 where higher amplitude peaks appear at the corresponding defect frequency in the inner raceway. The white and the red dashed line represents the theoretical shaft and inner raceway frequency calculated from Table 2.1.1.

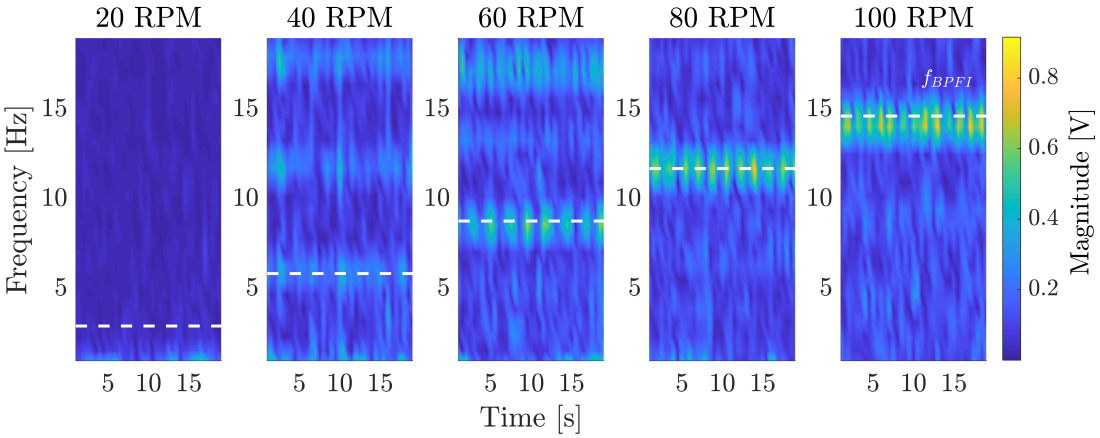


Figure 5.3.9: Damage features generated by EDM notch at 0 degrees and at different rotational speeds. The white dashed line indicates the inner raceway defect frequency f_{BPFI}.

In summary, the features obtained for both rigs have been processed as described in Table 5.3.1. Additionally, to test the localisation capabilities a set of different damage conditions were simulated as shown in Table 5.3.4.

Table 5.3.4: Damage conditions and operational conditions on scaled-down and MultiLife rig.

Scaled-down rig			**MultiLife rig**		
Speed [RPM]	Defect	Angle	Speed [RPM]	Defect	Angle
927	Notch 1	0°	20 to 100 RPMs in 20 RPMs interval	Notch 1	0°
	Notch 2	11°		Notch 2	−45°
	Overload	0°		Notch 3	+45°
Damage conditions	3		Damage conditions	15	
Normal conditions	1		Normal conditions	5	

5.4 Data processing and feature extraction for damage localisation

Similarly to the damage detection stage, it is important to select features that capture the relationship between and input and output data. In this case, TOA features have been shown to provide a useful information for damage localisation using AE signals. A few technical aspects in the generation of such maps will be explained in the following section together with some signal processing methods to process these features. Finally, a method to encode the output variables for the algorithm will be explained.

5.4.1 Time of Arrival determination of AE signals

The time of arrival or onset estimation of AE signals is one of the most common requirements of AE processing. This information is normally used to count the signal burst or use the time of arrival estimation for localisation techniques such as traditional triangulation methods. A few approaches are commonly used for this purpose such as the traditional First Threshold Cross method and the more sophisticated AIC pickers.

5.4.1.1 First Threshold Cross method

This is one of the simplest methods to estimate the onset of wave arrival, and can simply be defined as a threshold value above the noise level. The main disadvantage of this method is related to the selection of the threshold level as it can arbitrarily set or calculated based on the background noise characteristics. Higher threshold levels have the effect of delaying the time of arrival by selecting an onset far away or in the extreme case not selecting any onset as the threshold level is higher than the AE signal.

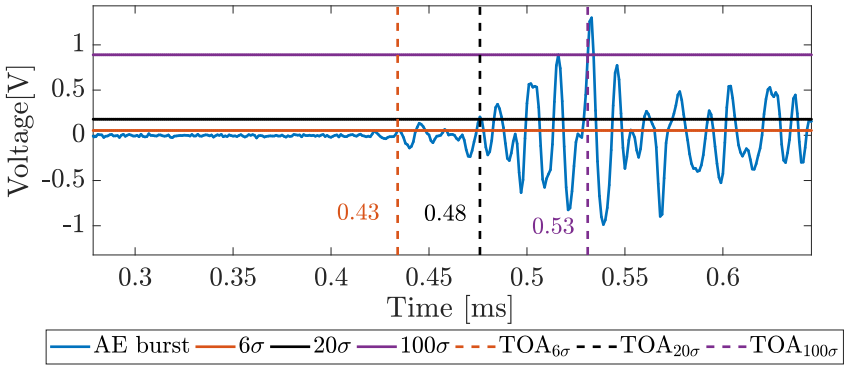


Figure 5.4.1: Three different threshold levels at 6σ, 20σ and 100σ above noise level on AE signal. The TOA estimations are represented as $TOA_{6\sigma}$, $TOA_{20\sigma}$ and $TOA_{100\sigma}$ respectively.

On the other hand, low threshold levels can lead to premature onset times and false burst detection. An example is shown in Figure 5.4.1 where three different threshold levels were calculated based on the background noise standard deviation σ at 6σ, 20σ and 100σ. These three threshold values gave three different onset time around $0.434ms$, $0.476ms$ and $0.531ms$ respectively selecting different wave modes that propagate at different velocities. This problem may be partially overcome by using a parameterless onset estimator method described below.

5.4.1.2 Akaike-Information-Criterion (AIC) picker

The AIC picker is based on the method used for model selection [128] and works on the assumption that a signal can be divided into two different stationary processes separated by a specific time. The stationarity condition implies that the mean and variance do not change over time. This method has been used for detecting the first arrival of seismic data [129] and, in the field of SHM, to detect the wave's first arrival in AE data [90, 130]. The AIC equation for the purpose of AE onset detection is defined as:

$$AIC[t] = t\log_{10}[\sigma\{y[1:t]\}] + (N-t-1)\log_{10}(\sigma\{y[i+1:N]\}) \tag{5.4.1}$$

where t, σ, y, and N are the time, the variance, the signal and the sample size. The time onset is the value where this function is minimised and is estimated as:

$$t_o = \arg\min\{AIC[t]\} \tag{5.4.2}$$

The method separates the signal in two parts; the uncorrelated noise before t_o and the AE signal after. The method applied on a AE signal excited by the Pencil Lead Break (PLB) method estimated its onset at $0.421ms$ as shown in Figure 5.4.2. The advantage of using this method over a traditional threshold method is that an accurate onset of arrival can be calculated without specifying a threshold value. A trade-off in using the AIC method is that it is highly dependent on the time window size. Multiple local minima can be generated due to inappropriate window length but normally this can be solved by providing an estimation of where the onset should be in the time series data.

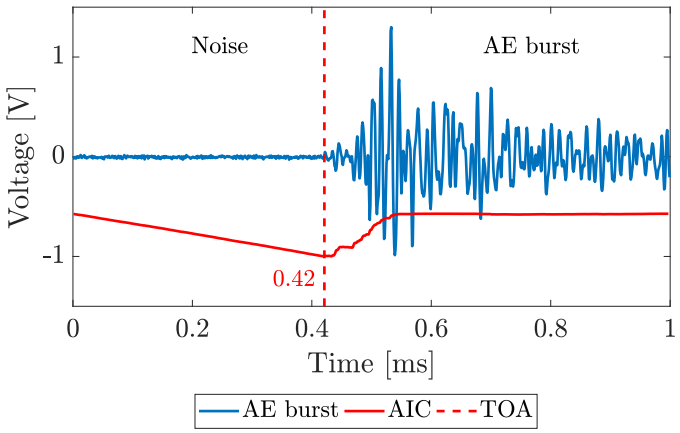


Figure 5.4.2: AIC function on AE signal, the minima obtained from the AIC function represents the AE signal onset.

5.4.2 Considerations for AE sensor placement

Despite it appearing to be theoretically possible to generate a mapping of the form described in Equation 2.4.13, there are a few technical difficulties that impose a constraint on the equipment and the sensor arrangement to ensure that unambiguous and valid mappings may be obtained.

1. **Sampling frequency and sensor spacing**

   The lowest time difference or the highest TOA difference resolution that can

be obtained is the inverse of the sampling frequency, which means that there will exist a minimum distance requirement between sensors. As in the case of a $C = 5000m/s$ and a $T_s = f_s^{-1} = 1 \times 10^{-6}s$ then the minimum distance will be around $5mm$. This also has an impact on the accuracy of estimating a source using this approach as it will limit the location resolution. For the above values, it will limit the ability to locate a defect with a resolution of $5mm$. In the context of roller bearings, this limit will be enough to localise damage, as the extent of the loaded zone where any incipient damage is prone to occur due to RCF normally occurs within a loaded zone of around 40° [117]. Therefore, in terms of damage relocation using the MultiLife rig, this resolution will be enough for this purpose.

2. **Sensor placement**

   There are multiple scenarios where ambiguous mappings can be generated. Figure 5.4.3 shows a scenario where a 2-D localisation is performed by placing the sensors $S1$ and $S2$ on the same axis thus allowing a plane of symmetry. This will lead to two different possible position solutions $P1$ and $P2$. The same occurs for a case where source localisation is performed in a 3-D scenario where sources $P1$ and $P2$ are outside the plane region. This is due to the two AE sources generating the same TOA difference at both positions. As mentioned before, the MLP has the properties of a universal approximator and training it with such a data set will violate the condition of approximating a continuous mapping.

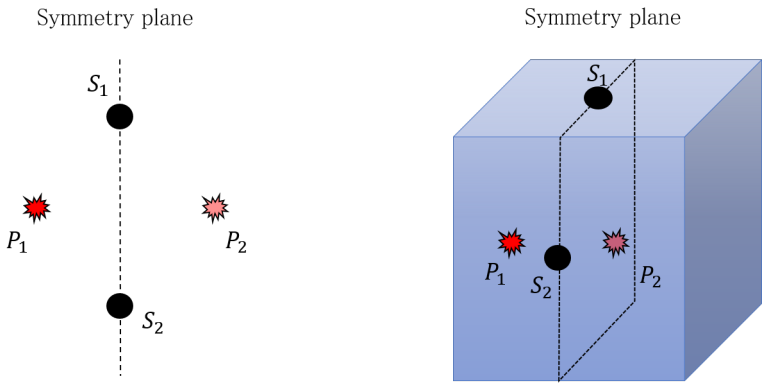


Figure 5.4.3: A few scenarios where a wrong TOAD map can be built, both mappings will lead to multiple solutions in a 2-D and 3-D space.

3. **Sensor condition and bonding condition** Faulty sensors can give wrong and inconsistent measurements and the use of coupling gels can lead to inconsistent and non-repeatable measurements. In practise, it is difficult to attain repeatable bonding conditions. This can be partially solved by using semi-permanent bonding mechanisms such as glues. Several considerations should be taken into account such as partial or complete debonding might occur thus leading to inconsistent measurements.

5.4.3 Data input/output encoding

At first glance, it seems that a NN will learn any data mapping relationship by presenting it regardless of any correct treatment of the input features. An important aspect of NN applications is that the way the data are encoded during the learning stage will have an impact on the data generalisation capabilities. At this stage, finding better ways to present data in order to improve generalisation can become a non-trivial task as this step involves additional methods to properly present data to the NN. This step normally involves the treatment of categorical variables, continuous variables and periodic variables.

The importance of choosing a suitable manner to encode input/outputs can be seen as finding better ways to allow the NN to interpolate. Normally this can be seen as finding surfaces that fit the input/output data into a more smooth representation. That means that data with similar characteristics are represented as close as possible in the input/output space, whereas data with different characteristics should be widely separated. A particular case is when a periodic variable such as time, the day of the week or, as in the current situation, the TOA mapping for a circular geometry is needed to be encoded. A common problem with such variables is that the highest and lowest values of the input/output range are conceptually adjacent. This problem occurs when a mapping defined on Equation 5.4.3 is built.

$$\langle \Delta t_{ij} \rangle \rightleftarrows \langle \theta \rangle \quad i \neq j \tag{5.4.3}$$

The problem arises at the lowest and highest values of the angle variable θ. For a physical localisation perspective these two positions are the same but for regression purposes these two values are relatively distant. That means that similar values of Δt_{ij} are generated at values of 0° and 360°. Clearly, from the regression perspective, rather far away values of Δt_{ij} are expected and thus a better way is required to encode this periodic behaviour on the output data. Even though this problem has not been discussed before for AE localisation, a general treatment has been discussed at length in literature [108, 131, 132] for a different type of problems. The advice is for a data transformation defined in Equations 5.4.4 and 5.4.5:

$$\theta_n \rightarrow \langle \sin\theta_n, \cos\theta_n \rangle \tag{5.4.4}$$

$$\langle \sin\theta_n, \cos\theta_n \rangle \rightarrow \arctan\left(\frac{\sin\theta_n}{\cos\theta_n}\right) \tag{5.4.5}$$

This finally leads to restating the mapping in Equation 5.4.3 as:

$$\langle \Delta t_{ij} \rangle \rightleftarrows \langle \sin\theta_n, \cos\theta_n \rangle \quad i \neq j \tag{5.4.6}$$

This encoding allows the representation of a one-dimensional feature into a two-dimensional variable that allows an unambiguous and smooth representation of the angular data. Geometrically, this encoding can be seen as the polar transformation of the angle variable to a Cartesian coordinate system in a unitary circle as shown in Figure 5.4.4. Notice that with this representation each angle value will have a unique pair of $\langle x, y \rangle$ values around the circle circumference.

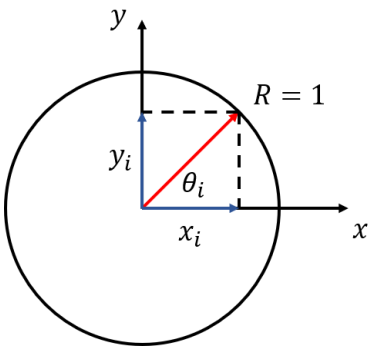


Figure 5.4.4: Data encoding of periodic variables using a unitary circle.

Another problem with these types of variables is the difficulty of determining the statistical parameters such as the mean and standard deviation. This is mainly due to a dependence of a specific reference around the circumference, this can be circumvented by defining the circular mean, variance and standard deviation as given in Equations 5.4.7, 5.4.8 and 5.4.9:

$$\langle \bar{x}, \bar{y} \rangle = \langle \frac{1}{N} \sum_{n}^{N} \cos\theta_n, \frac{1}{N} \sum_{n}^{N} \sin\theta_n \rangle \tag{5.4.7}$$

$$S = 1 - R = 1 - \frac{1}{N} \left\{ \left(\sum_{i}^{N} x_i \right)^2 + \left(\sum_{i}^{N} y_i \right)^2 \right\}^{0.5} \tag{5.4.8}$$

$$s = \sqrt{-2 \log R} \tag{5.4.9}$$

Additionally, from the perspective of a localisation algorithm, it is necessary to quantify the accuracy of the defect localisation. Later on, these expressions will be used to express the localisation accuracy in terms of the circular mean and standard deviations.

5.4.4 Feature extraction for damage localisation using AE

This section will explain the experimental method used to obtain the damage localisation features. The process of generating a map of TOA is shown in Figure 5.4.5. In order to generate a map of TOA, the geometry of interest is discretised and excited using a PLB source. Each point P_i generates an artificial source that represents a potential damage location. The signals measured from each point are

later processed and extracted using the AIC method and finally a feature vector is built for each excitation point.

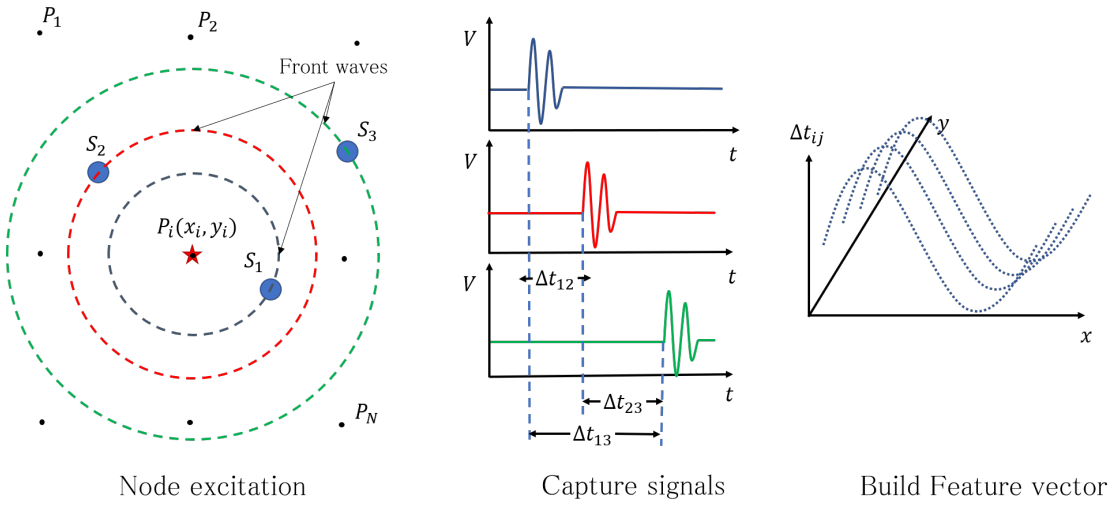


Figure 5.4.5: Feature extraction procedure. The nodes are excited at the corresponding positions, then the TOA values are extracted from the recorded signals and finally a feature vector is constructed.

5.4.5 Time of Arrival extraction for scaled-down rig

In the case of the scaled-down rig, the set of measurements were obtained by dismantling the bearing upper support as shown in Figure 5.4.6. This was performed because the defect initiation was expected to occur in the outer raceway. The maximum contact stresses generated at the outer raceway occurred at fixed positions rather than being periodically redistributed during each bearing rotation as in the case of the inner raceway, thereby subjecting the outer raceway to a more severe fatigue scenario.

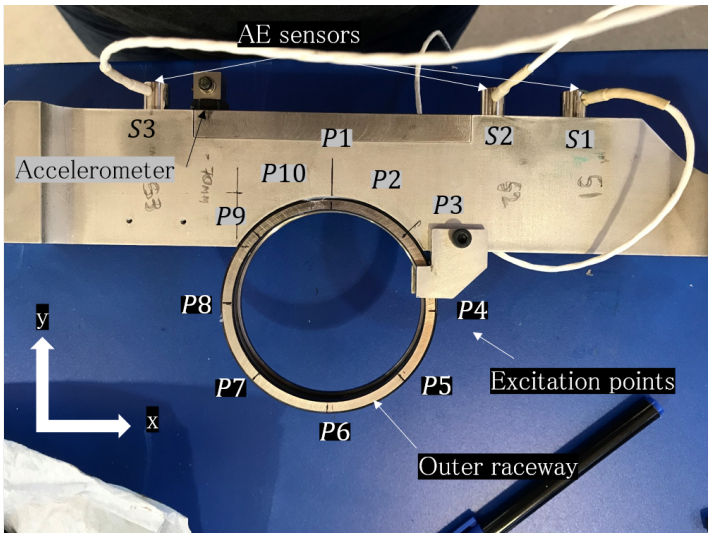


Figure 5.4.6: Excitation points on the scaled-down rig. A set of three AE sensors were installed on the top of the supporting structure.

The internal walls of the outer raceway were discretised into 10 different angles identified as $P1, ..., P10$ where a set of 30 measurements per angle were performed around the internal walls of the bearing. These will simulate the effect of the defect interaction between the rollers and the outer raceway every time the rollers pass through the defect.

The excitations on the inner side of the outer raceway were performed by breaking a 3mm long 0.5mm HB pencil lead (PLB), also known as Hsu-Nielsen source [62]. This source generates a highly repeatable excitation similar to the AE signals and consists of breaking a brittle pencil lead on the surface of the structure [88]. The excitations were performed on the positions identified in Table 5.4.1 with the mean and variance of these measurements contained in Table 5.4.2.

Table 5.4.1: Excited positions around the test bearing.

ID	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Angle [degrees]	0	11	45	90	135	180	225	270	315	326

Table 5.4.2: Mean and standard deviation of the generated map.

ID	Angle $[\theta]$	$\Delta\bar{t}_{12}[\mu s]$	$\Delta\bar{t}_{13}[\mu s]$	$\Delta\bar{t}_{23}[\mu s]$	$\sigma_{12}[\mu s]$	$\sigma_{13}[\mu s]$	$\sigma_{23}[\mu s]$
P1	0°	6.30	5.13	-1.17	0.95	1.04	0.53
P2	11°	5.37	-1.93	-7.30	1.13	1.28	0.98
P3	45°	4.13	-8.73	-12.86	0.77	1.23	1.04
P4	90°	6.53	-7.10	-13.63	1.07	1.52	1.35
P5	135°	6.83	-7.13	-13.97	0.79	1.04	1.09
P6	180°	8.90	5.77	-3.13	1.69	1.52	1.16
P7	225°	7.13	14.60	7.47	0.82	1.25	1.19
P8	270°	7.00	16.13	9.13	1.23	1.16	1.43
P9	315°	8.53	17.00	8.47	1.36	0.87	1.04
P10	326°	8.66	10.07	1.40	1.09	1.01	1.19

The interface between the outer raceway and the upper bearing support used a fluoride-based grease as a couplant. This allowed the reduction of any potential air gaps between the contact interface that tend to increase the reflection coefficient and consequently impact the AE signal transmission. Additionally, a layer of cyanoacrylate was locally applied at the ends of the interface to keep the bearing fixed during the PLB excitations.

Figure 5.4.7 shows the Δt_{ij} values obtained for each pair of sensors. The measurements obtained show a periodic pattern that seems to repeat every 360° degrees and give unambiguous values for each angular position. This is expected as the points are excited in a circular shape. These values can be used to map a single and unique position in the bearing; Even though the previous statement is true for a human operator, further preprocessing must be performed in order to make this data suitable for a machine learning algorithm. This data set was later encoded using Equations 5.4.4 and 5.4.5 in order to obtained a more suitable feature vector. Figure 5.4.7 also shows that a relatively small variance was found in the measurements at each position but a relatively weak feature in between sensors 1 and 2 can impose some difficulties in terms of training an algorithm for regression purposes.

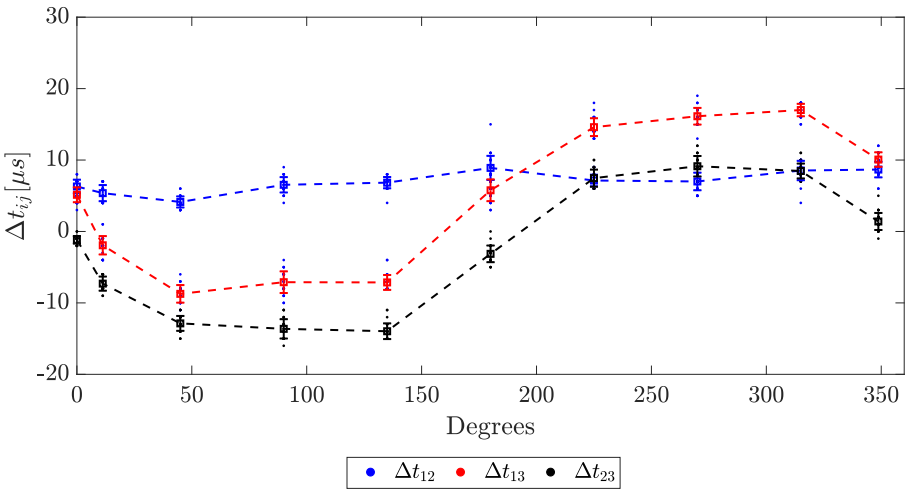


Figure 5.4.7: Time of Arrival Difference in NU1010.

This weak feature arose due to the relatively short distance between sensors 1 and 2 and the variance in the measurements thus generating a relatively noisy feature. Ideally, it is always desirable to obtain a clear feature but, in real applications, this might be constrained due to issues such as the small size of the components being monitored.

A set of 300 observations in total were processed as shown in Figure 5.4.7. In order to provide a balanced dataset, 30 observations at each point were split into 12 points for training, 12 for validation and 6 for testing. This will allow the same number of observations per class (excited points) giving a total of 120 observations for training, 120 for validation and 60 for testing. As in this case the number of training samples is 120 patterns a range of models between 2 to 20 hidden units were defined based using Equations 3.3.17 and 3.3.17 using an observation to weight ratio from 1 to 10. These models were later cross-validated using the 10-Fold method.

5.4.6 Time of Arrival extraction for MultiLife rig

The MultiLife rig runs using NU2244 ECP bearings which are regularly used in WTGB applications. AE sensors were installed inside the bearing inner raceway and separated by 120°. In order to perform the training map, the whole mechanism was dismantled resulting in the inner sleeve and the bearing being isolated from the

main component. This dismantling procedure was performed in the Lea Lab in the University of Sheffield and required at least two skilled persons to operate the crane and removing the security doors, the loading mechanism, the hoses, radiator and covers. This procedure normally took on average one day and was performed every time access was required to change the defect orientation as will be explained later in this chapter. A pristine bearing was placed in the inner sleeve and engaged with the AE sensors inside it.

The bearing was discretised by dividing its circumference in 20 segments of 18° at 3 different circumferences with a separation of 27mm. The measurements at the midsection are defined as $h = 0$ as shown in Figure 5.4.8. This discretisation gave a total of 60 nodes around the bearing surface. The nodes were developed using a drawing software and then printed on a paper. The nodes printed on the paper were carefully cut using a scalpel to leave a small hole for marking by attaching it to the bearing surface. In order to assign a reference frame, the nodes at 0° were aligned with the sleeve reference notch as seen in Figure 5.4.8.

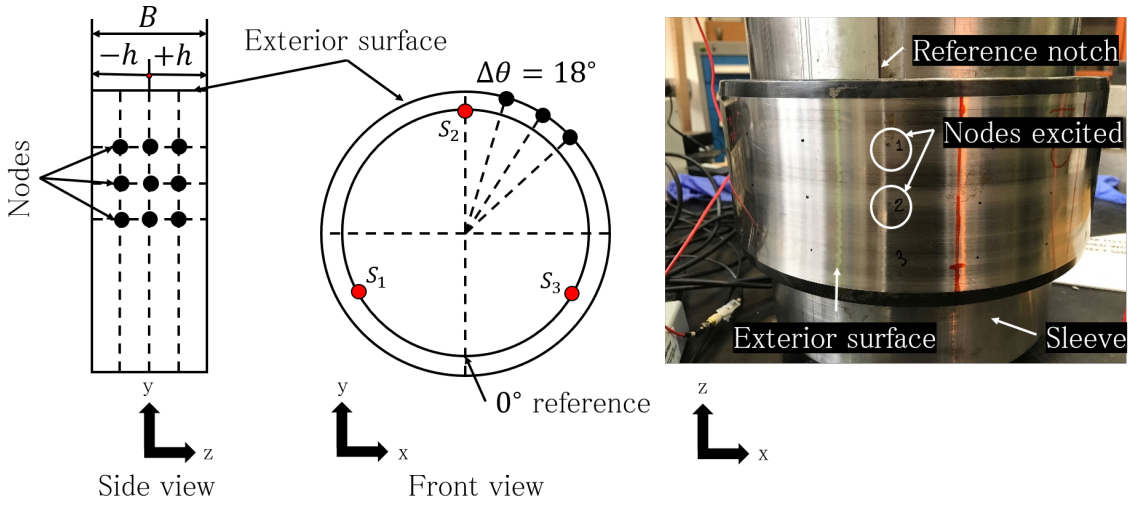


Figure 5.4.8: Excitation nodes at the surface of the NU2244.

The signal from the AE sensors were amplified at 40dB using the Mistras pre-amplifiers connected to a 28VDC power supply as shown in Figure 5.4.9. This allowed the capture of a visible signal within a range of 5V. The data were captured with a PicoScope 6000 at a sampling frequency of 2.5MHz by using a rising edge threshold of 0.5V.

The artificial AE signals were obtained by exciting each node 100 times using a 3mm length 0.5HB pencil lead giving a total of 6000 measurements. The data were previously examined in situ in order to ensure that the bursts captured were consistent. This step allowed the processing of the AE bursts automatically during the time extraction procedure.

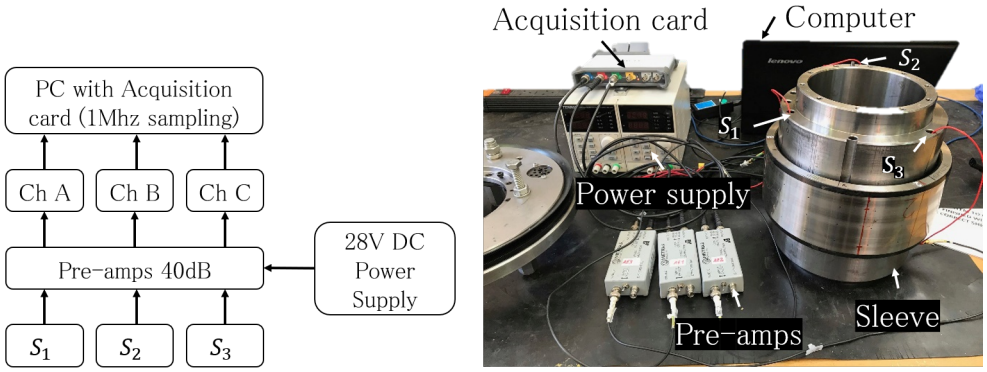


Figure 5.4.9: Experimental set-up for map acquisition procedure. The set-up consisted of an acquisition card, amplifier, AE sensors and a computer.

The signal was processed using a Matlab script that implemented the AIC picker with a fixed window of 500 samples. The data were later identified with each corresponding ID that mapped each position around the surface of the bearing. This time onset information was later used to calculate the TOA difference map by obtaining the time difference between each sensor. As the convention of the Delta T method the features were designated as: Δt_{12},Δt_{13} and Δt_{23} finally resulting in the map shown in Figure 5.4.10.

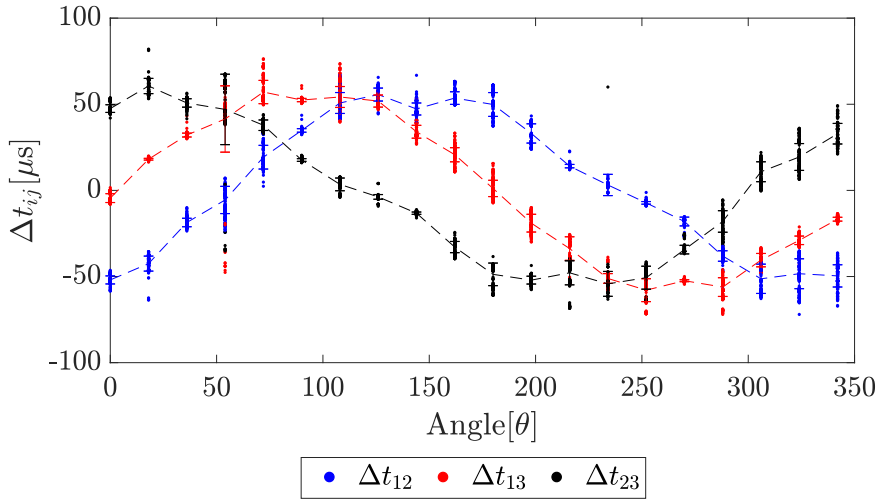


Figure 5.4.10: Time of Arrival Difference observations for each pair of sensors in MultiLife rig.

The data were finally arranged in terms of their corresponding angle and circumference as shown in Figure 5.4.11. This pattern clearly shows a sinusoidal behaviour along the angle values where the 360 degrees position matches with the 0 degrees angular position.

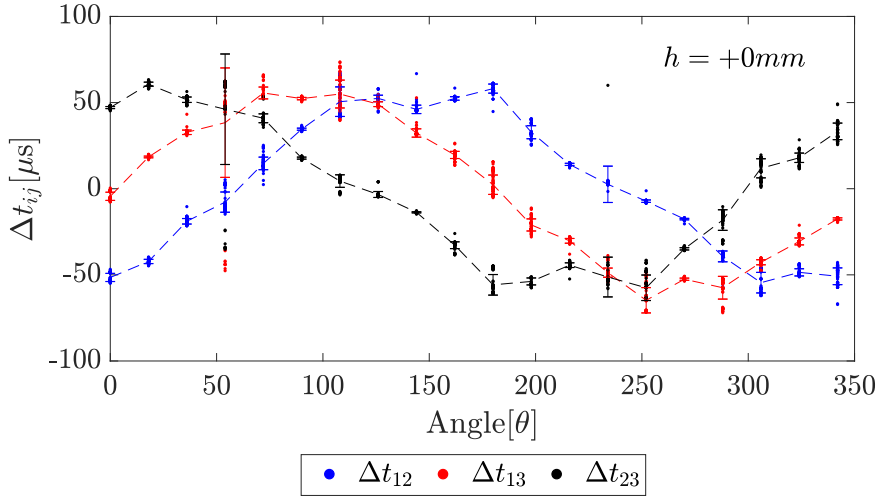


Figure 5.4.11: Time of Arrival Difference in NU2244 at midsection.

The data for the +27mm and -27mm are shown in Figure 5.4.12 and 5.4.13, the time difference measured follow the same behaviour and distribution.

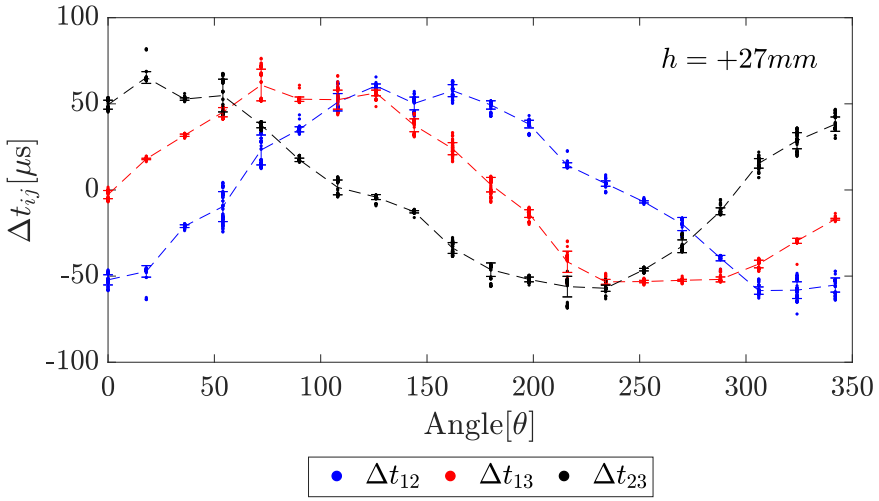


Figure 5.4.12: Time of Arrival Difference in NU2244 Set 1.

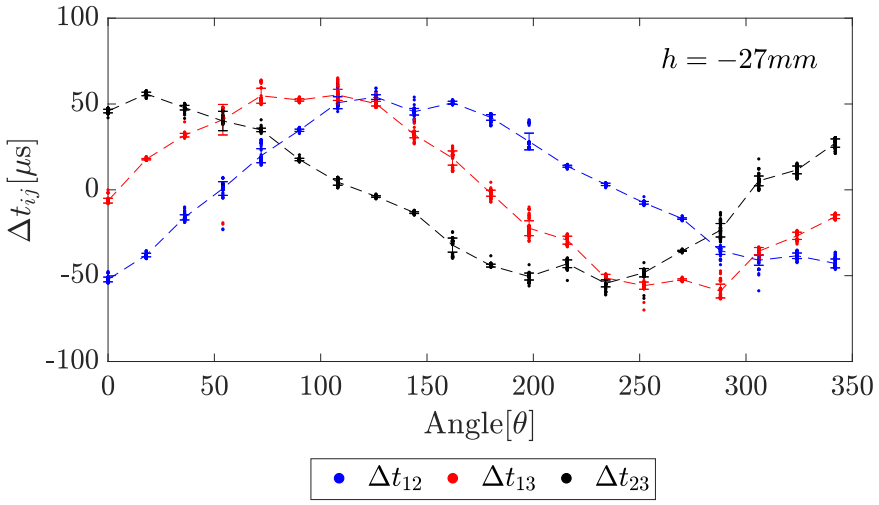


Figure 5.4.13: Time of Arrival Difference in NU2244 set 3.

Even though there is a sufficiently large amount of data obtained from the MultiLife rig with 100 observations (per class) for each excited point, cross-validation of the NN models was performed in order to keep the majority of the points for training purposes. In such cases, the K-Fold cross-validation method is suitable for small test sets around or less than 20% and allows to select a model with the best generalisation performance from the available dataset [108]. The 6000 observations were split into

a set of 2700 observations for training purposes, 2700 for validation purposes and 600 points for testing. This resulted in 45 points per class for training purposes, 45 for validation and 10 for testing purposes. Considering that 2700 points are used for training this allows the assessment of networks from 45 to 450 hidden units by calculating the upper and lower limits using Equations 3.3.17 and 3.3.18.

5.4.7 Defect bursts extraction and signal selection criteria

Unlike the case where the data set used for training was obtained using a simple data acquisition interface and PLB excitations, the data recorded from the EDM and overload defects were processed using an interface designed using MATLAB 2017. This allowed the selection of small portions of the original signal and extraction of the TOA from the three channels which were then saved in a separate file for later processing. This allowed to keep the AE bursts in a relatively small file that contains each single burst captured in a window size of 1000 samples.

In real life applications, the AE signal recorded will be contain by noise. As the aim is to reduce the bursts obtained from secondary sources and obtain a clear signal to process the TOA values the following selection criteria were followed:

- **Periodicity**: Bursts are filtered based on their periodicity as the main interest is to solely capture the bursts obtained from the interaction between the rollers and the defect. This assumes that periodic events that match the defect frequency are generated by the defect in the bearing component. On the other hand, aperiodic events are assumed to be generated from a different source so therefore are not selected for feature extraction.

- **Non-overlapped events**: Even though some sources can show periodicity, some of them are slightly contaminated by noise generated by secondary sources. This can be seen as two different bursts overlapping each other and thus providing an unclear onset.

- **Moderate Signal-Noise ratio**: Background noise sometimes hides the AE burst and, in extreme cases, buries the whole AE burst. In such cases, this makes the wave onset estimation difficult as this generates ambiguous onset estimations.

During the data analysis using the EDM notch, it was possible to categorise the following burst types. Figure 5.4.14 shows three types of burst. The first shows a clipped signal with undistinguishable onset. The second burst is a signal with a low SNR and the third burst with a clear onset. In order to mitigate the error induced in the onset estimation, only bursts such as the last one were selected for extracting TOA values.

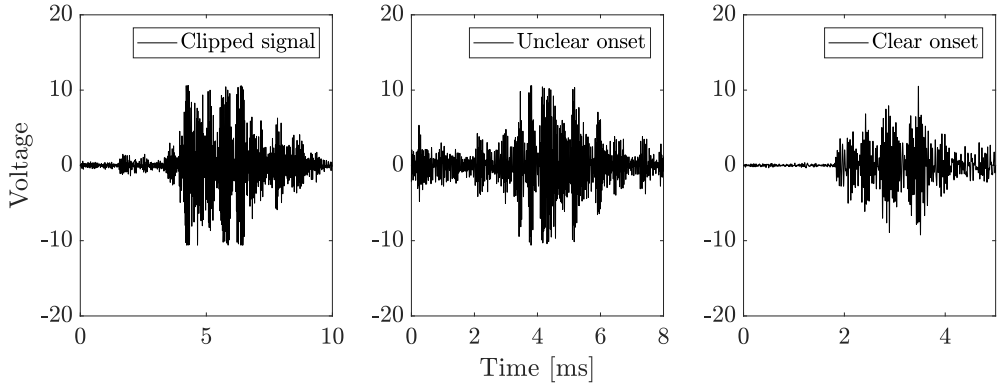


Figure 5.4.14: Bursts examples obtained in the scaled-down rig using an EDM notch.

5.4.8 Post-processing of localisation predictions

Unlike in more simple structures, the localisation of AE sources in bearings becomes more challenging due to secondary sources and the noisy environment. Even though this can be mitigated to some extent under well-controlled conditions, it is almost impossible to avoid completely.

Additionally, severe noise contamination in AE signals can also affect the time onset selection. These effects may be observed by generating an analytical model of the burst and systematically contaminate the signal with Gaussian noise. The signal was modelled using the same approach in some studies [117, 133] where the impact response is modelled as $y(t) = \exp(-Bt)\cos(2\pi f_r t)$ and contaminated with Gaussian

noise. The terms B represent the damping coefficient and f_r the dominant resonant frequency. Figure 5.4.15 shows the variations of onset calculations. Clearly, the onset determination for the signal with a signal to noise ratio of 5 becomes challenging due to the multiple minima obtained in the AIC function.

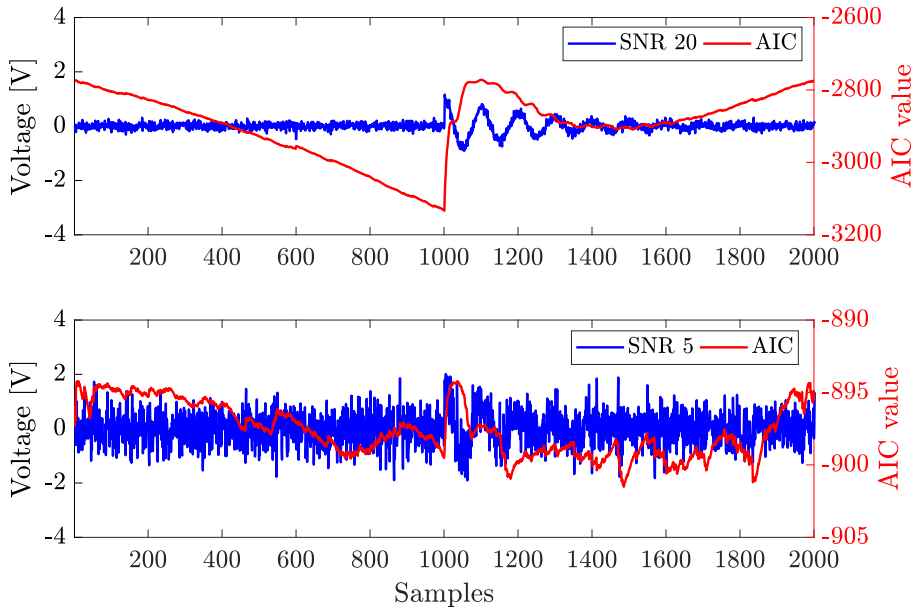


Figure 5.4.15: Effects of noise on AIC picker on AE data contaminated with two different Gaussian noise levels.

As discussed before, noise can lead to variations in the onset calculations using the AIC method. A highly contaminated signal might lead to inaccurate TOA measurements. This high variance in the input data might even lead the NN to extrapolate and thus reducing the reliability of the prediction. In order to overcome this issue, the data of the predicted outputs from the NN will be treated as a distribution.

Figure 5.4.16 shows a NN prediction on an EDM notch at +45 degrees using sets of 25 different measurement for each histogram. As may be seen at each empirical distribution, the parameters such as the mean and the circular standard deviation remained constant independently of the set of measurements. This leads to the assumption that the parameters of the distributions remain constant over time thus leading to stationarity in the data.

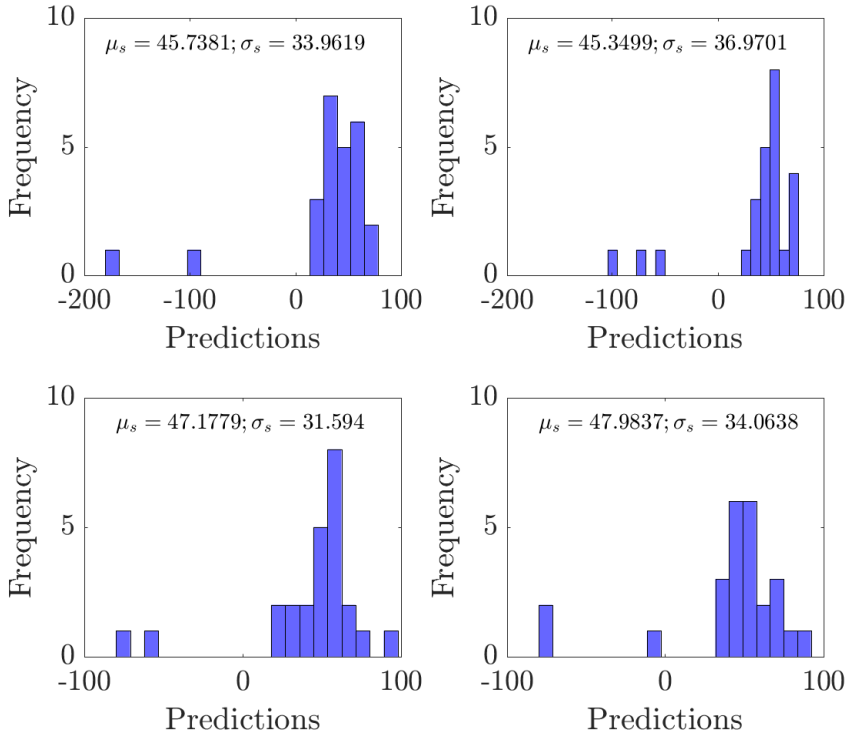


Figure 5.4.16: EDM damage prediction at +45 degrees using a NN with 4 different sets of 25 measurements.

5.5 Time-Delay and Sum Beamforming approach for damage localisation

5.5.1 Scanning at contact interface

The test was performed in the scaled-down rig and considered the problems as a two-dimension localisation problem. The problem was simplified by assuming a homogeneous propagation media with a constant propagation speed at a specific frequency value. The use of this approach requires the information of the sensor positions, the geometry of the component and the propagation speed. Figure 5.5.1 shows the positions of each AE sensors and the reference frame used for this localisation approach.

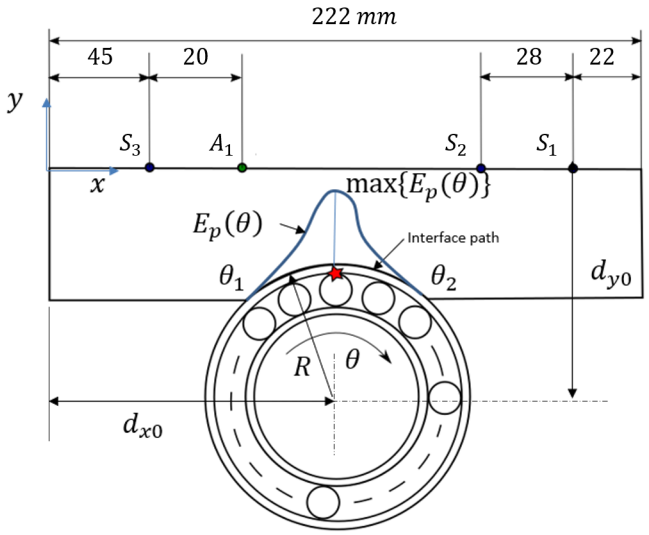


Figure 5.5.1: Bearing interface dimensions.

The localisation region is constrained in the contact interface within the θ_1 and θ_2 as shown in Figure 5.5.1. The maximum value of the beamformer output, defined as $E_p(\theta)$, gives an indication of the main defect source. The angular position was transformed into a Cartesian system using Equations 5.5.1:

$$\begin{aligned} x_k &= +R\cos\theta_k - d_{y_0} \\ y_k &= -R\sin\theta_k - d_{x_0} \end{aligned} \tag{5.5.1}$$

The propagation speed of the zero-order symmetric mode for a 17 mm thick aluminium plate was calculated using the numerical solution of Equation 2.2.3 and by assuming the values of the longitudinal and shear wave propagation speed of 6300 and 3130 m/s respectively [134]. Although the material was known beforehand, this information is not always available, and consequently, it might lead to localisation inaccuracies such as the underestimation or overestimation of the defect position from the sensors. This may result in errors of the same magnitudes as the percentage values of the speed variation.

5.5.2 Time-Delay calculation

In this case, due to the size of the application and wave speed propagation, it is not deemed acceptable to consider a planar wave propagation. In this case, for an

array of 155 mm length, a wave speed propagation of 5500 m/s and a frequency of 30kHz, using Equation 3.4.3 the radial distance will be $r = 26cm$. Considering that at least any hypothetical AE source will be well below that region, the Near-Field assumption will be adequate to describe the wave front behaviour by describing the steering vector as:

$$| \vec{r}_{ki} |= \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \tag{5.5.2}$$

where x_i, y_i,x_k, y_k represent the sensor position and the source location respectively. The delays can be calculated as:

$$\tau_{ki} = \frac{| \vec{r}_{ki} | - | \vec{r_o} |}{C} \tag{5.5.3}$$

Equation 5.5.3 can be understood as the delay that will take a spherical wave to reach a point specified by the vector $\vec{r}_o$, to the source $\vec{r}_{ki}$. The indices k and i represent the nodes in the discretised geometry and the sensors. Figure 3.4.2 shows a planar and spherical wave propagation arriving at an array of sensors separated by a distance d.

5.5.3 Signal processing

In this example, bounded media represents a challenge in terms of the infinite dispersive modes generated by transient waves. This problem leads to multiple propagation speeds to be considered in the frequency domain. The approach in this work was to narrow-filter the signal using a band-pass filter with a central frequency of ten times the bandwidth and calculate its corresponding propagation speed for the zero-order propagation mode $C_{s0}(30kHz) \approx 5500m/s$. The filter specifications can be found in Table 5.5.1.

Table 5.5.1: Narrow filter parameters.

Parameters	**Values**
Central frequency f_c	30 kHz
Bandwidth B	3 kHz
Magnitude A	80 dB

Equations 5.5.2 and 5.5.3 were used into Equation 3.4.1 in order to produce the beamformer output described as Equation 3.4.2. The beamforming power is calculated along the contact area and produces the graphs shown in Figure 5.5.2. This

figure shows interference patterns produced by exciting the outer raceway at -45,-11,0,+11,+45 degrees using the PLB method. The maximum peaks at each pattern represents the location of the AE source defined by:

$$\theta = \text{argmax}\{E_p\} \quad \theta_1 \leq \theta \leq \theta_2 \tag{5.5.4}$$

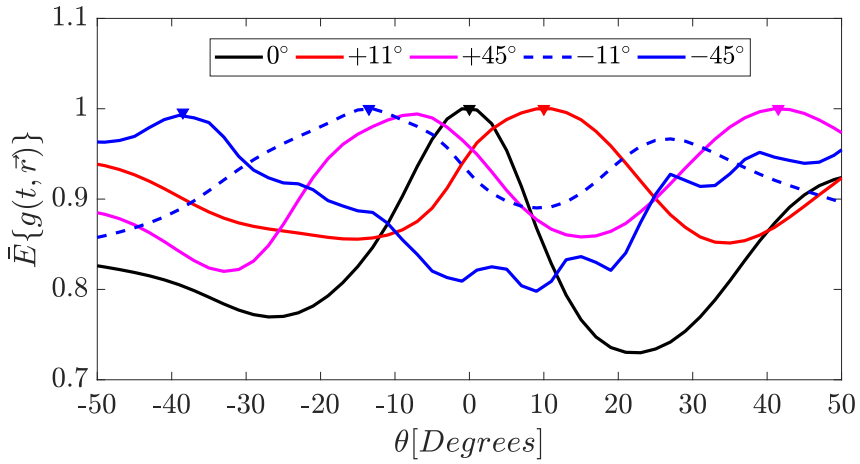


Figure 5.5.2: Beamformer output at contact interface for PLB excitations.

5.6 Data processing and feature extraction for AE sensor self-assessment

5.6.1 Experimental impedance measurement device

The admittance curves were measured using an external device that excited the PZT patch and measured the current passing through it. The resultant electrical admittance $Y(j\omega)$ was obtained by calculating the ratio of the component's voltage $V(j\omega)$ with the current $I(j\omega)$ passing through it as::

$$Y(j\omega) = \frac{I(j\omega)}{V(j\omega)} = Z^{-1}(j\omega) \tag{5.6.1}$$

The inverse of Equation 5.6.1 gives the impedance term $Z(j\omega)$ as it is the reciprocal term for admittance $Y(j\omega)$. Notice that the terms of current and voltage in Equation 5.6.1 are represented in the frequency domain.

As the voltage and current signals are measured in the time domain, the following relationship allows a domain transformation as:

$$Y(j\omega) = \frac{I(j\omega)}{V(j\omega)} = \frac{\mathcal{F}\{I_p(t)\}}{\mathcal{F}\{V_p(t)\}} = -\frac{1}{R_s G}\frac{\mathcal{F}\{V_s(t)\}}{\mathcal{F}\{V_p(t)\}} \tag{5.6.2}$$

Equation 5.6.2 shows that the voltage and current functions are transformed into the frequency domain using the continuous Fourier Transform $\mathcal{F}\{.\}$. The terms R_s, and G represent the PZT resistance and a gain term. Both terms can be seen as a linear amplifier that allows the measurement of the weak current signal from the PZT patch. As in the case for sampled signals the discrete form of Equation 5.6.2 can be represented as:

$$Y[j\omega] = -\frac{1}{R_s G}\frac{FFT[V_s[t]]}{FFT[V_p[t]]} \tag{5.6.3}$$

where $FFT[.]$ refers to the Fast Fourier Transform algorithm. The measurement device built for this purpose was based on the work developed by Peairs *et al.* [136]. They proposed using a low-cost impedance measurement device for impedance monitoring of structures. Despite being relatively simple to build and relatively inexpensive, the device demonstrated a remarkably comparable performance with a commercial HP4194A impedance analyser. A schematic of the impedance measurement device is shown in Figure 5.6.1.

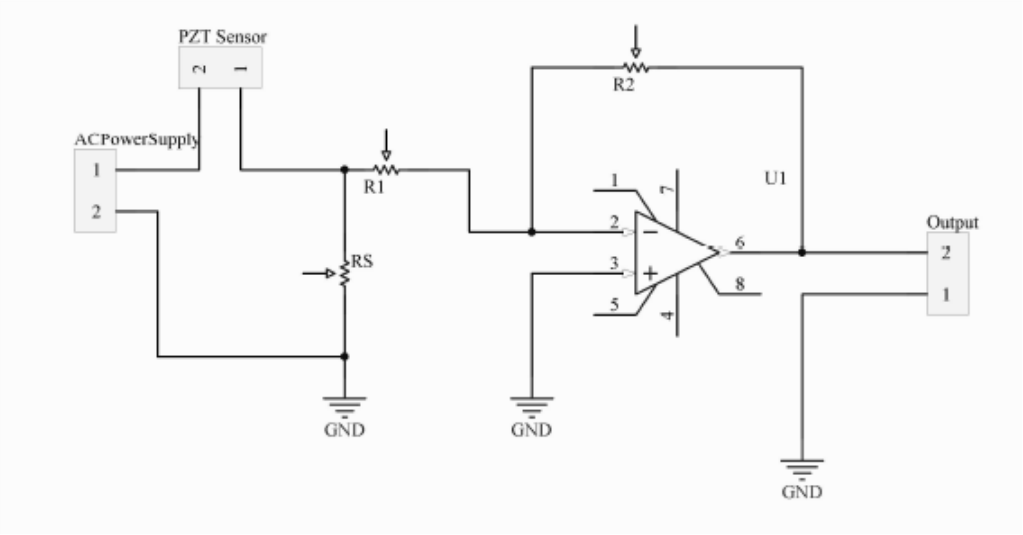


Figure 5.6.1: Impedance measurement circuit used to monitor the impedance values of a PZT patch.

The only modifications performed on the circuit were the addition of two variable resistors to allow a variable gain in the circuit. This was added to amplify the signal's

output at the amplification stage as mentioned before. The gain values are defined by the ratio of two resistance values defined as:

$$G = -\frac{R_2}{R_1} \tag{5.6.4}$$

It is important to mention that the measured impedance curves are still an approximation to the real impedance curves. This is mainly due to the relatively small resistance values in PZT patches. In order to overcome this, the electric current through the PZT patch is measured from a sensing resistance R_s that also works as a limiting resistance to avoid overloading the patch. Thus, for relatively small values of R_s, the patch current $I_p(j\omega)$ should approximate the measured value $I_s(j\omega)$ as:

$$I_p(j\omega) \approx I_s(j\omega) \tag{5.6.5}$$

A resistance value of 200Ω was used in the impedance circuit as a previous work reported that values less than 200Ω can approximate the current flowing through the PZT patch [136]. Thus, the voltage across the patch $V_p(j\omega)$ is relatively close to the input voltage $V_{in}(j\omega)$ as:

$$V_{in}(j\omega) \approx V_p(j\omega) \tag{5.6.6}$$

This card has two ports, an input and an output port. The input port ACPowerSupply was used for the excitation signal and the output port for the measured patch voltage. An additional port was used to connect the PZT patch terminals. The gain values R_1 and R_2 are set using linear variable potentiometers connected on an LM741CN operational amplifier with a bandwidth of 1MHz. The parameters used in the card are shown in Table 5.6.1. The selection of the gain values was performed based on the minimum levels achieved to magnify the measured signal.

Table 5.6.1: Impedance measurement card parameters.

Parameters	**Values**
Gain: G	5.34
$R_1[\Omega]$	14.56×10^3
$R_2[\Omega]$	77.74×10^3
$R_s[\Omega]$	200

5.6.2 Experimental set-up

A 5 mm Aluminium plate was used as a fixed structure. A set of sensors were installed using different bonding conditions as early described. A Picoscope 6000 was used to excite the PZT using an external output. The measurements were recorded using the analogue input channels at 2 MHz. The PZT sensors were connected using the differential configuration so only the voltage changes and current were measured in between the patch terminals. Figure 5.6.2 shows the measurement set up used in this experiment together with the connections of the impedance card to the patch and the acquisition system.

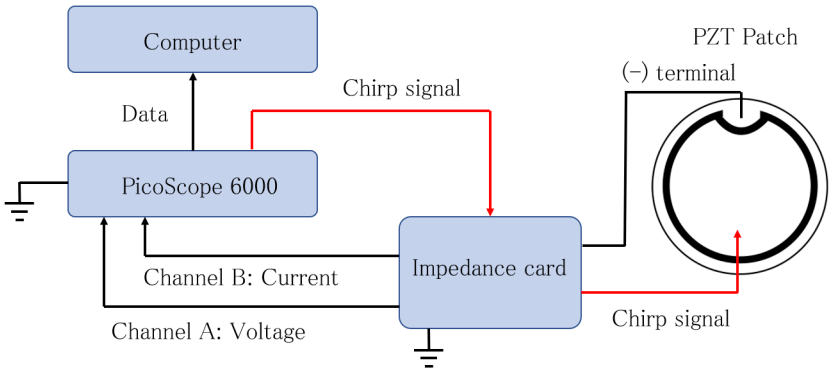


Figure 5.6.2: Acquisition device and set-up for impedance measurements.

5.6.3 Impedance measurement

The Electrical Impedance (EI) was actively measured by exciting the PZT patch with a linear chirp function as shown in Equation 5.6.7, 5.6.8 and Figure 5.6.3. This will allow a fast estimation of the frequency response of the system by sweeping a sine wave within a specific frequency range. Table 5.6.2 shows the parameters used in Equation 5.6.7 to excite the PZT element.

$$V_{in}(t) = V_0 \sin\left(2\pi\left(0.5ct^2 + f_1 t\right)\right) \tag{5.6.7}$$

$$c = \frac{f_2 - f_1}{T} \tag{5.6.8}$$

The maximum excitation frequency was based on the hardware design limitations as the signal generator and the operational amplifier were only allowed to work around 1 Mhz. Figure 5.6.3 and 5.6.4 show a time domain and a time-frequency representation on the chirp function respectively. Notice the linear frequency component variation along time of the chirp signal with respect to time.

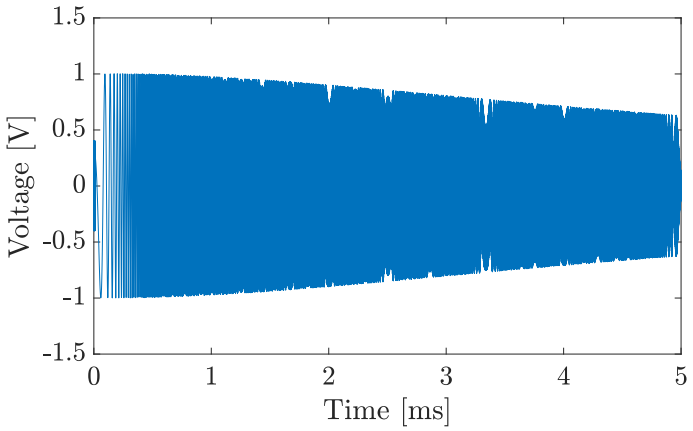


Figure 5.6.3: Chirp signal used to excite the PZT elements.

Table 5.6.2: Chirp excitation signal parameters.

Parameters	**Symbol**	**Values**
Amplitude	V_0	1 V
Wave type	$V_{in}(t)$	Sine
Start frequency	f_1	1×10^3 Hz
Stop frequency	f_2	1×10^6 Hz
Frequency Increment	Δf	100 Hz
Time increment	Δt	0.5×10^{-6} s
Sweep time	T	5×10^{-3} s

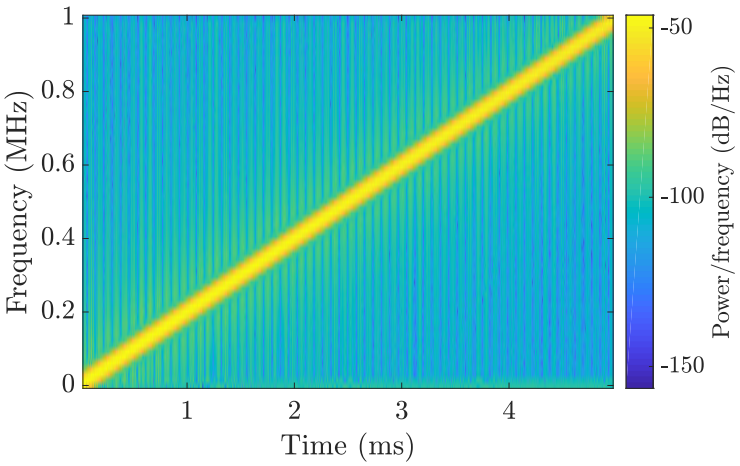


Figure 5.6.4: SFTF of the excitation signal using a Chirp function, sweeping from 0 to 1 MHz.

Different debonding levels were simulated by controlling the amount of bonding area. This was achieved by controlling the amount of area bonded between the sensor and the PZT surface with a 0.2 mm thin paper mask in between as shown in Figure 5.6.5. In a practical application, this value depends on the viscosity of the adhesive and in the case of cyanoacrylate, a layer thickness of 0.2 mm is the maximum value achievable [137]. Therefore, providing a consistent thickness for a fully bonded and partially bonded scenario. Additionally, an assumption that debonding will be generated symmetrically as shown in Figure 5.6.6 was made. Even though this condition is difficult to achieve in a real scenario, the analytical impedance model described on Equation 2.5.2 is defined in terms of an effective force at the bonded interface, therefore deeming this assumption acceptable for arbitrary bonding areas.

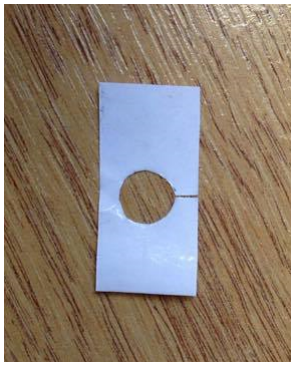


Figure 5.6.5: Paper mask used to simulate different debonding conditions.

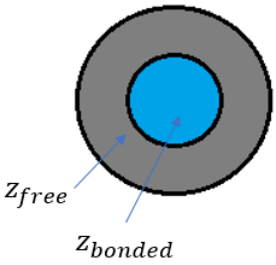


Figure 5.6.6: PZT bonding scenarios. The blue and grey area represent the bonded and debonded area respetively.

The blue and grey areas represent the bonded and unbonded areas respectively. The bonding was achieved by using cyanoacrylate over the bonded area and allowing it to cure for 15 minutes to avoid any changes in the impedance measurements over time. In order to quantify the amount of debonding, the damage metrics were defined in terms of percentage of detached area A_d as:

$$DS[\%] = \frac{A_d}{A_b} * 100 = \frac{\pi(R^2 - {r_b}^2)}{\pi {r_b}^2} * 100 = \frac{(R^2 - {r_b}^2)}{{r_b}^2} * 100 \tag{5.6.9}$$

where r_b represents the radius of the bonded section A_b and R represents the PZT patch radius. Four different damage sizes were defined as shown in Table 5.6.3 by gradually increasing the radius of the hole in the paper mask.

Table 5.6.3: Debonding at different levels.

Damage size [%]	**Area** [mm^2]	**Radius** [mm]
0	78.54	5
25	58.90	4.33
50	39.15	3.53
75	19.63	2.50

5.6.4 Admittance features

As expected from the analytical expressions, the linear part of the experimental admittance curves behave similarly to the behaviour described with the 1-dimensional model discussed early. Figure 5.6.7 shows a shifting on the slope towards the maximum limit expected in the free condition. Even though it does not seem to monotonically increase with the damage progression, the feature shows two distinguishable features whether a completely detached and bonded scenario occurs.

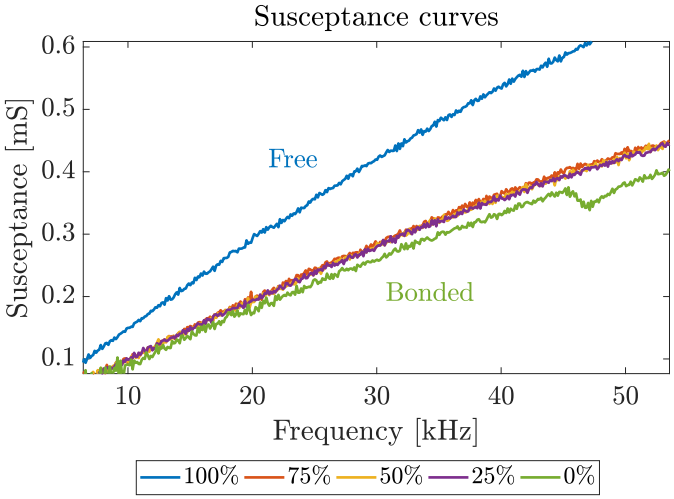


Figure 5.6.7: Linear admittance curves at different damage progression.

Interesting features have been seen in the absolute value of the impedance curves as shown in Figure 5.6.8. An apparent shift on the resonance frequencies occurs, as the curves show a tendency to shift the resonance frequency to the left as the debonding progresses. Under a completely free condition the experimental resonance frequency of the radial mode is $228kHz$ which coincides with the data provided by the PZT manufacturer. Similarly this occurs for the conductance and susceptance features shown Figure 5.6.9 and 5.6.10.

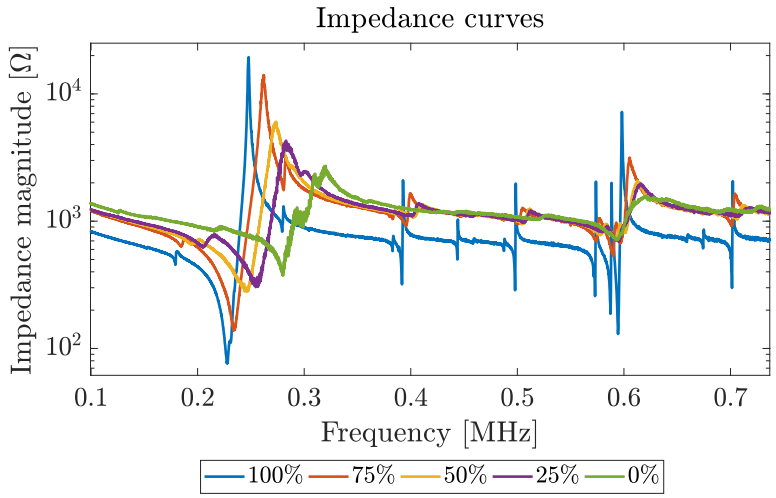


Figure 5.6.8: Absolute value of impedance curves on semi-logarithmic scale.

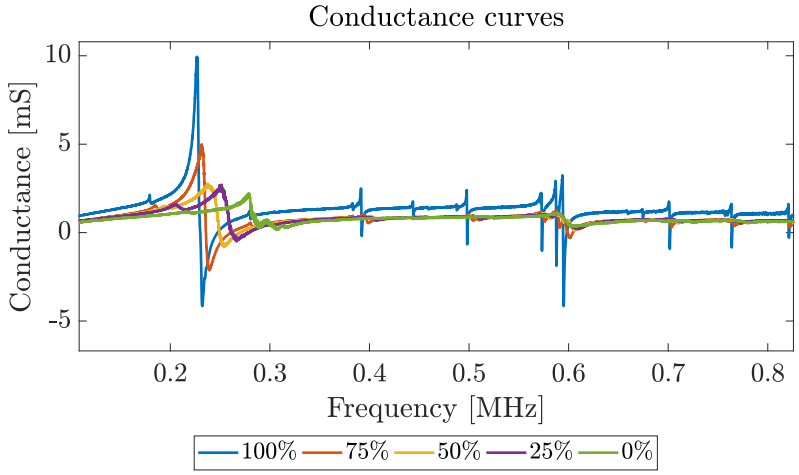


Figure 5.6.9: Conductance curves at different damage progression.

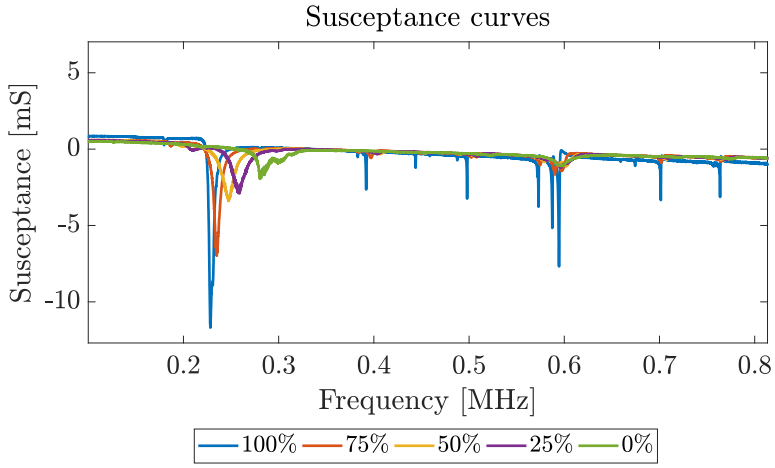


Figure 5.6.10: Susceptance curves at different damage progression.

A set of 32 measurements were acquired of each of the bonding scenarios. The measurements were later reduced to a four-dimensional feature vector that contained the susceptance values at 10, 20, 30 and 40 kHz. These values were selected as they were contained in the linear part of the impedance curve. No advantage of increasing the dimension of the feature vector in order to track the slope variations is expected with higher dimensions.

5.7 Chapter summary

- A procedure to simulate a late onset of damage and overload damage was described in this chapter. This section provided a methodology to simulate an overload defect using a FEM, the results obtained from the plastic area generated during the overload showed a great agreement with the experiments. Moreover, a significant reduction in the remaining life was possible, this agrees with the fact that overload events have a detrimental effect on WTGB.

- A process to extract the features and assemble the feature vector using vibrational data was explained. Although the features were selected and processed similarly for the scale-down, further signal enhancement on the MultiLife rig was required. This resulted in unambiguous features that are suitable for damage detection purposes.

- A feature extraction process using AE signals was performed in order to obtained a TOA map. Similarly to the unambiguity requirement for novelty detection, a unique map for TOA was desired. Additionally, a data encoding procedure was used to represent periodic variables as a means to avoid the introduction of ambiguous mappings. This stems from the requirement that smooth functions are required for NNs to learn the data relationship.

- An approach using beamforming exploited a different AE feature, a delay between waves is obtained and used to generate an interference pattern from the AE waves. Although this is not a ML method, the use of this feature might be useful where relative noise levels are encountered, and an accurate TOA estimation might not be deemed possible. The procedure required to use analytical methods to calculate the time-delays associated with an area of interest and also required to make assumptions on the behaviour of the wave propagation.

- A sensor self-assessment section described the use of impedance measurements to provide information about bonding degradation in AE sensors. These features were extracted by exciting a PZT with an external device and pre-processing the signals using the FFT. It was shown that this approach led to distinguishable features between a healthy and damaged conditions. Moreover, it was shown that this behaviour occurred as predicted by the linear impedance model described in this section. This ultimately, explaining the physical relation between bonding and slope variation.

CHAPTER 6

Results and discussions

6.1 Novelty detection

In this section the results of the damage detection stage using outlier analysis and LDA are shown for the scaled-down and MultiLife rig.

6.1.1 Outlier detection results for scaled-down rig

The results of the outlier detection in the fatigue rig can be seen in Figure 6.1.1. A threshold level of 99% was implemented and this gave a total of 0% of false positives for both training and test set whereas a significant amount of true negatives were generated. A general observation arises from observing the mean values of each damage condition, despite showing a high degree of overlap between the normal and faulty condition it is possible to discern a deviation from the normal condition.

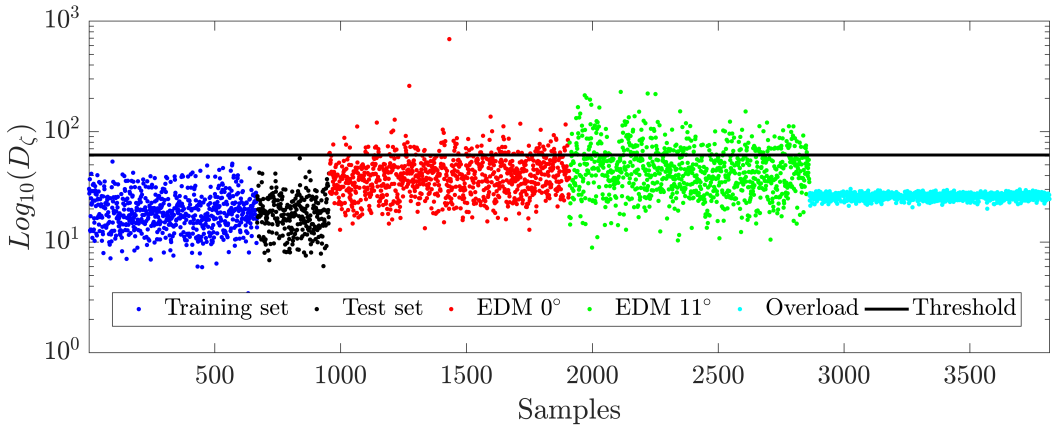


Figure 6.1.1: Mahalanobis distance for training set (Blue), test set (Black), EDM notch at 0° (Red), EDM at 11° (Green) and overload failure (Cyan) at a 99% threshold value.

This overlapping will impose difficulties to estimate a threshold level that efficiently separates the undamaged and the damage state. Any decrease in the threshold level will lead to eventually increase the number of false positives from the normal condition. The percentage of False Positives (FP) and False Negatives (FN) are shown in Table 6.1.1 where the two different threshold levels were estimated using the Monte Carlo method.

Table 6.1.1: Outlier false positives and negatives at different threshold levels for the scaled-down rig.

Threshold level [%]	**Training** FP [%]	**Test** FP [%]	**Notch 1** FN [%]	**Notch 2** FN [%]	**Overload** FN [%]
99	0	0	87.94	76.10	100
90	0	0.34	80.50	66.98	100

6.1.2 Outlier detection in a run-to-failure test

Figure 6.1.2 shows the Mahalanobis distance during the time the experiment lasted. As shown with the EDM defects, it was possible to distinguish between two different states by analysing the trend in the Mahalanobis distance.

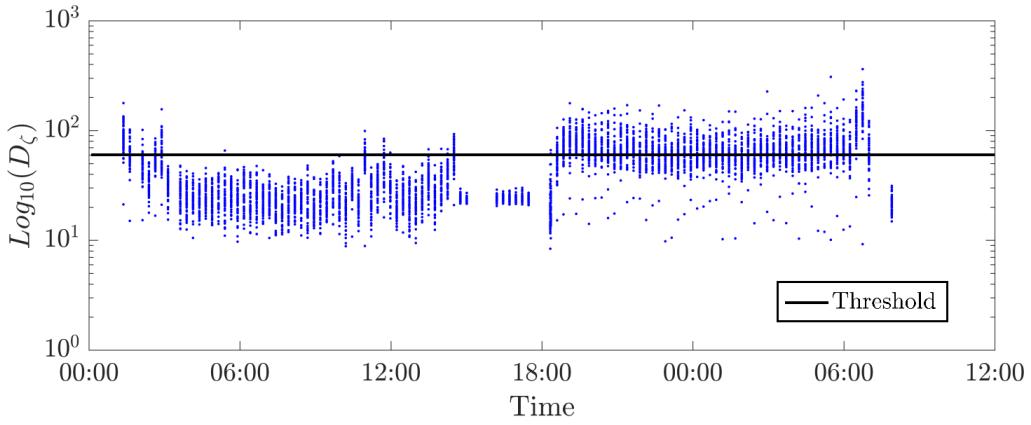
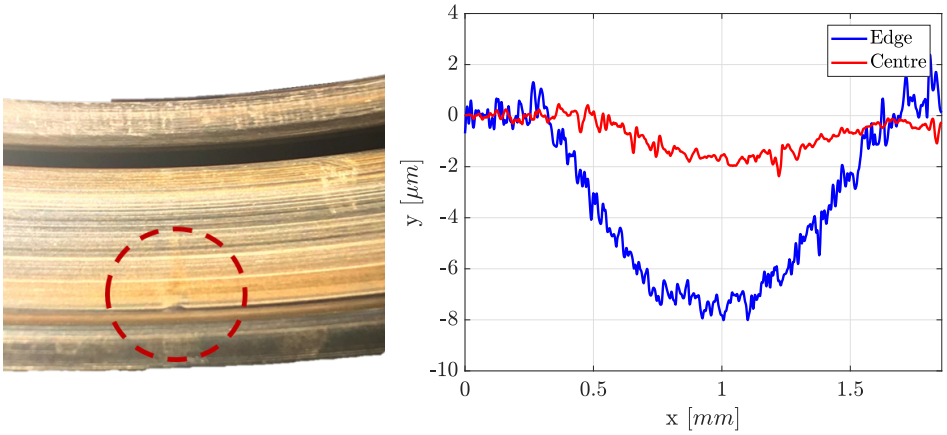


Figure 6.1.2: Run-to-failure Mahalanobis distance.

A slight change in the mean of the Mahalanobis distance at around 18:00 of the second day of the experiment shown a potential deviation from the normal condition. This was also supported by a small amount of surface damage from the edge of the outer raceway as shown in Figure 6.1.3a and 6.1.3b. Additionally, it is possible to notice a gap at 15:00 and 17:45, these gaps occurred due to the rig stalling and consequently requiring to change the support bearings and the electric motor. This resulted from the long term usage and excessive demand for these components during the test.



(a) Surface Damage (b) Profilometer measurements

Figure 6.1.3: (a) Surface damage at one of edges the outer raceway inside the red dashed circle and (b) Profilometer measurements at the contact zone.

6.1.3 Outlier detection results for MultiLife rig

An estimation of the covariance matrix was obtained from 70% of the data, thus 1113 samples were used for training and the remaining 477 for testing purposes. An individual covariance matrix was calculated for 5 different speed conditions. The results obtained are shown in Figure 6.1.2 with a 99% threshold.

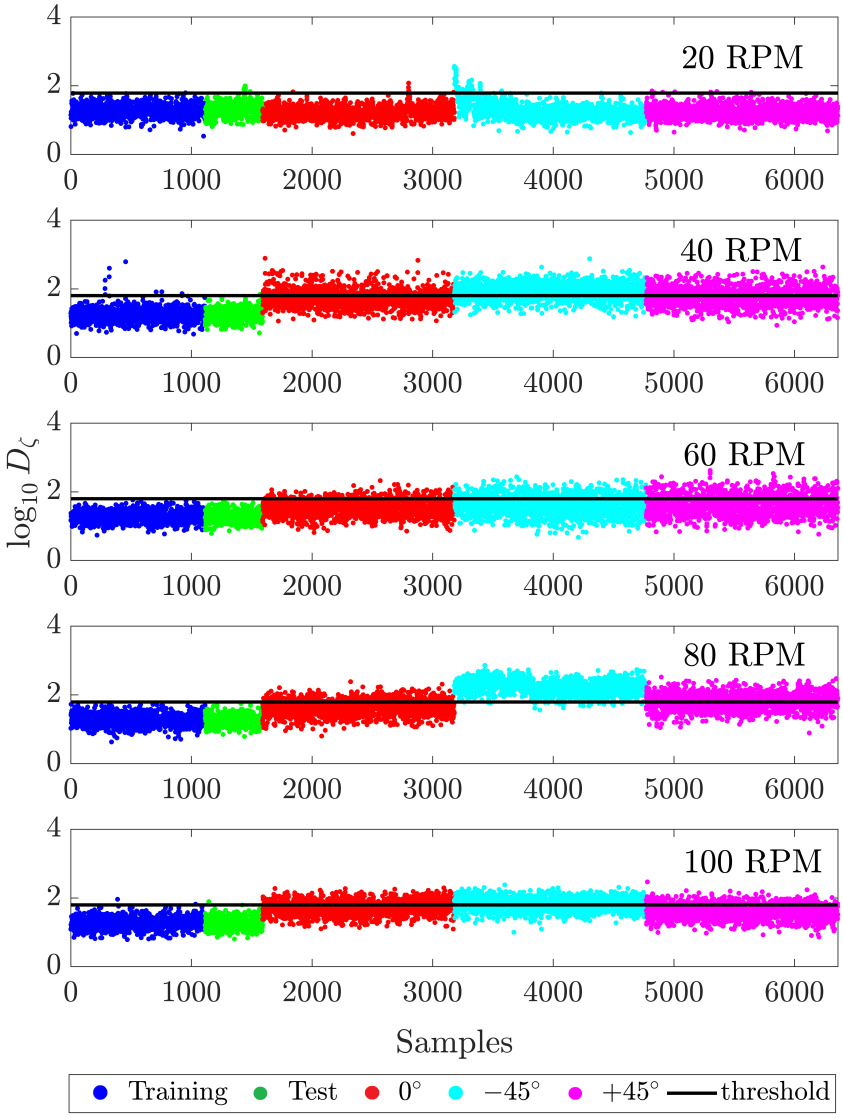


Figure 6.1.4: Outlier detection in MultiLife rig at a 99% threshold level.

Table 6.1.2 shows the number of False Positives and False Negatives at a 99% threshold level at different shaft speeds and defects.

Table 6.1.2: Percentage of False Positives (FP) and False Negative (FN) from the test set at different damage scenarios in MultiLife rig using the Mahalanobis Distance and a 99% threshold level.

RPMs	**Notch 1**		**Notch 2**		**Notch 3**	
	FP [%]	FN [%]	FP [%]	FN [%]	FP [%]	FN [%]
20	2.1	99.7	2.1	96.2	2.1	99.62
40	0.21	65.03	0.21	25.16	0.21	56.79
60	0	87.23	0	72.51	0	69.31
80	0	75.35	0	1.7	0	49.37
100	0.21	62.77	0.21	42.89	0.21	81.76

The use of the Squared Mahalanobis Distance (SMD) shows to learn the normal condition data as the percentage of False Positives values stay relatively low at 2.1%, this is consistent with different rotational speeds. A difficulty in classifying the majority of the damage data have been encounter as in general, a value above 25.16% of False Negatives has been found. The worst prediction occurred at 20 RPM where a total of 99.62% of the damage data was classified as undamaged data. In a practical sense, an underperformance was expected at low rotational speeds as at these ranges the vibrational amplitudes at the characteristic defect frequency will remain relatively low. This ultimately leading to data that shows similar characteristics to the baseline data.

6.1.4 LDA results for scaled-down rig

As a means to understand the classification capabilities of the method, a standard procedure of dividing the data set into a training and test set was performed. The set of weights were determined from the training set which compromised 70% of the data corresponding to the normal and damaged state for both rigs. The data were projected onto a lower-dimensional space as shown in Figure 6.1.5. The data were

separated by the linear boundary corresponding to zero, i.e. where the discriminant function $y(\mathbf{x}) = 0$.

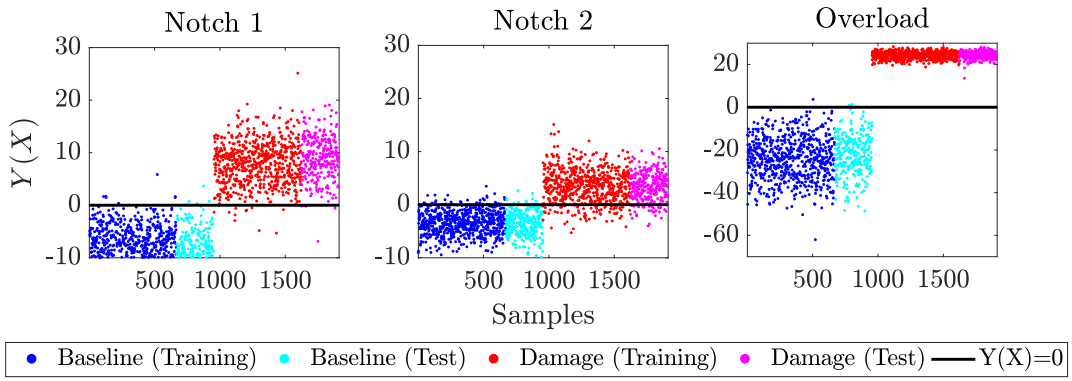


Figure 6.1.5: Class separation between damaged and undamaged states in the scaled-down rig at 927RPMs.

The faulty condition in the scaled-down rig was successfully separated for the different damage conditions as shown in Figure 6.1.5. This shows that there is a significant separation between the baseline and the different damage conditions. Additionally, these separations seem to occur regardless of the defect orientation as there is no significant difference between the defect localised at 0° and 11° degrees. Comparing the data with the outlier detection method shown that the percentage of misclassification in significantly reduced as shown in Table 6.1.3. Moreover, the damage obtained in the run-to-failure test has shown a great separation between the normal and damaged state. This could potentially be due to the high dissimilarity between the baseline data and the damage features in the reduced dimensional space.

Table 6.1.3: Percentage of False Positives (FP) and False Negative (FN) from the test set at different damage scenarios in scaled-down rig using LDA.

RPMs	**Notch 1**		**Notch 2**		**Overload**	
	FP [%]	FN [%]	FP [%]	FN [%]	FP [%]	FN [%]
927	1.75	1.40	4.55	10.84	1.05	0.00

6.1.5 LDA results for MultiLife rig

The data obtained in the MultiLife rig shown a similar behaviour at lower shaft speeds as seen in Figure 6.1.6. Clearly the data from the faulty conditions and the normal condition have similar distribution parameters at a lower-dimensional projection, thereby making it difficult to differentiate between a normal and damaged state at this speed. This has to do with the lower vibration levels obtained at the defect frequency, as the defect interaction has a low excitation level at that speed.

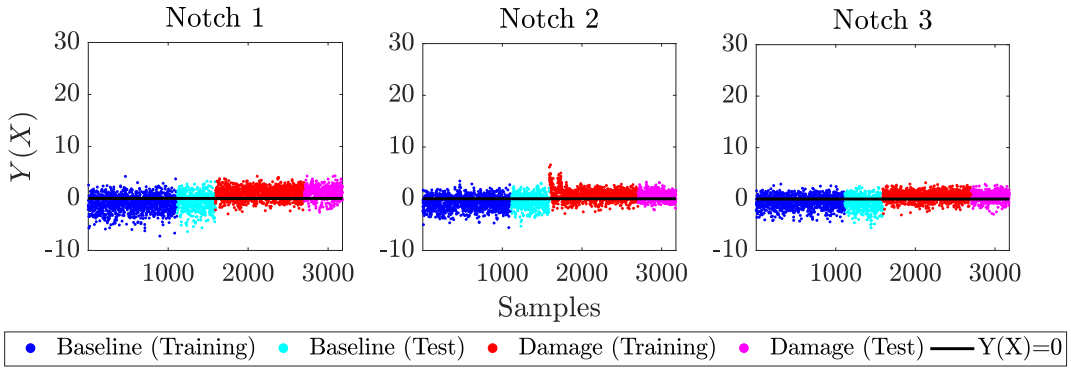


Figure 6.1.6: Class separation between damage and undamaged states at 20 RPM in MultiLife rig.

This is clearly shown after a threshold of 40RPMs is reached as shown in Figure 6.1.7. Even though both conditions seem to slightly overlap, a distinction between normal and damage conditions is evident. This behaviour also occurs at 60, 80 and 100RPMs as shown in Figure 6.1.8, 6.1.9 and 6.1.10. These conditions showed a trend in reducing the number of false positives and false negatives at higher shaft speeds as shown in Table 6.1.4.

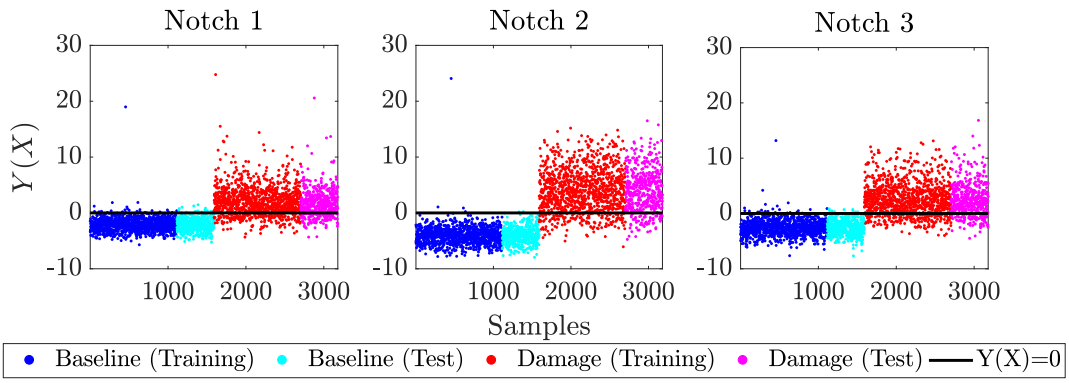


Figure 6.1.7: Class separation between damage and undamaged states at 40 RPM in MultiLife rig.

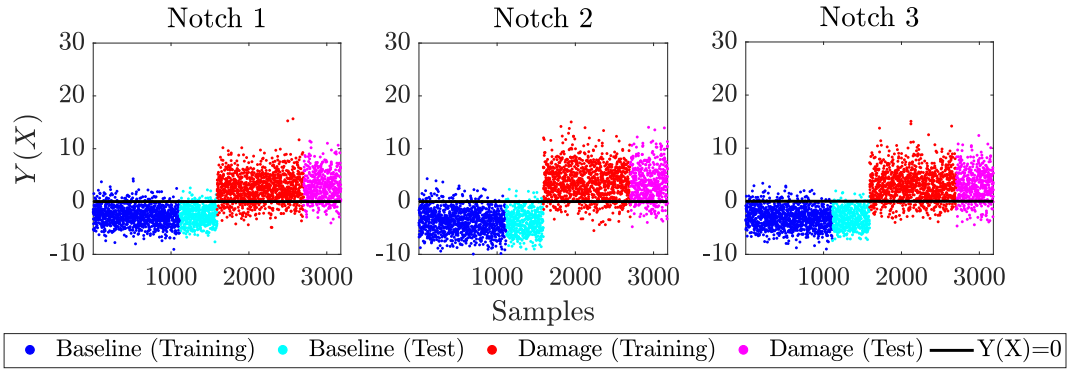


Figure 6.1.8: LDA results in MultiLife rig at 60 RPM.

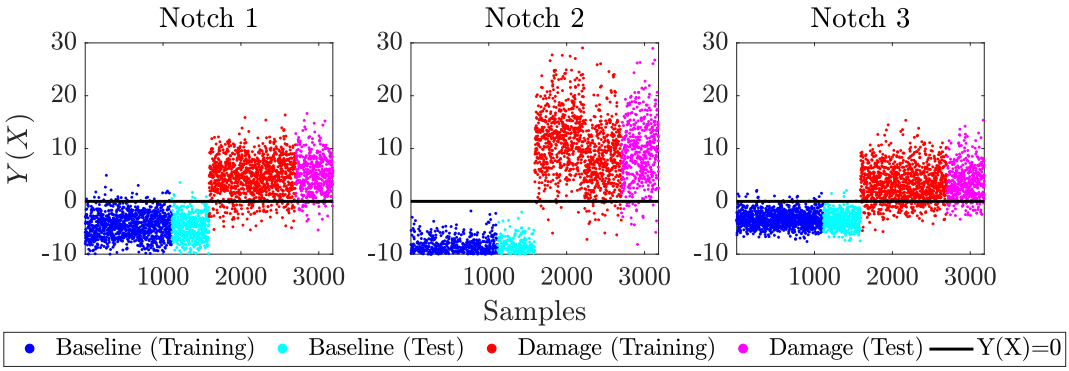


Figure 6.1.9: LDA results in MultiLife rig at 80 RPM.

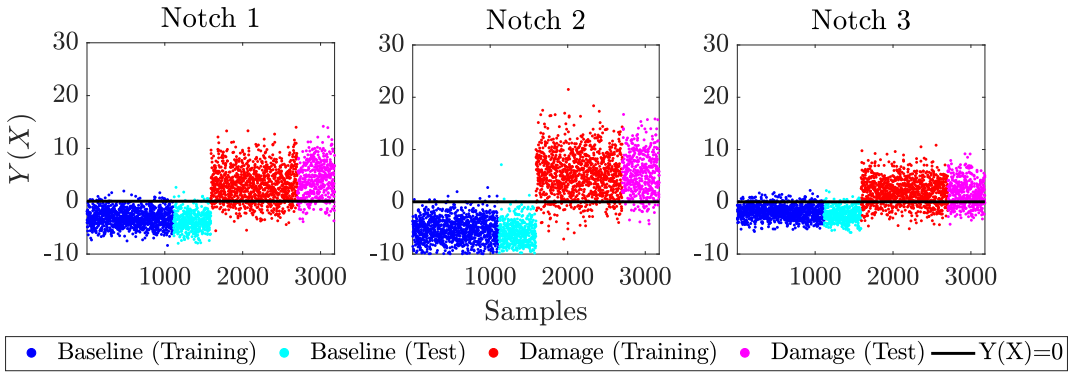


Figure 6.1.10: LDA results in MultiLife rig at 100 RPM.

Table 6.1.4: Percentage of False Positives (FP) and False Negative (FN) from the test set at different damage scenarios in MultiLife rig using LDA.

RPMs	**Notch 1**		**Notch 2**		**Notch 3**	
	FP [%]	FN [%]	FP [%]	FN [%]	FP [%]	FN [%]
20	38.99	16.35	39.62	32.08	29.14	27.46
40	4.19	22.64	0.42	13.63	3.14	18.87
60	6.29	12.37	3.35	13.21	3.56	14.26
80	1.89	7.34	0.00	4.40	1.47	9.64
100	2.52	9.85	0.84	6.08	3.35	28.51

6.2 Damage localisation

The previous section addressed whether the damage was present in bearings and allowed the discrimination between damaged and undamaged states. In this section, the results for the damage localisation stage for the scaled-down and MultiLife rig using a NN approach will be shown. Finally, the results for the beamforming approach will be shown for the scaled-down rig.

6.2.1 Damage localisation in scaled-down rig

The validation errors from different NN models were calculated based on the sum-of-squares error expression defined in Equation 3.3.8, which calculates the difference between the target value and the value predicted by the neural network. These errors are associated with a specific dataset, such as the training, validation and test dataset. In this case, for model selection purpose, a validation error is calculated from a validation dataset which has not been previously seen by the network during the training phase.

A model for each fold was selected based on the lowest validation error obtained, this can be seen as the point where an inflexion point occurs and no benefit occurs through increasing the complexity of the model. Figure 6.2.1 shows the validation error for the 4th fold where the minimum validation error occurred in a NN with 8 hidden units.

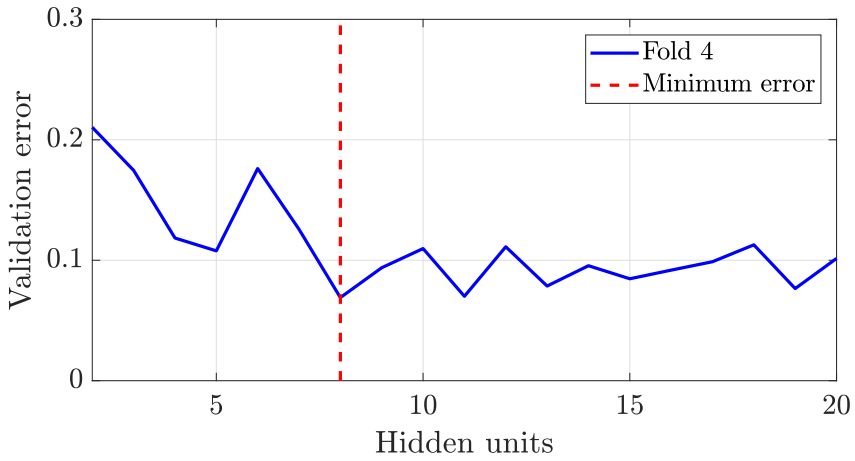


Figure 6.2.1: Cross-validation error results for the scaled-down rig for the 4th fold.

As shown in Figure 6.2.2, the optimal networks were found to be the ones with a range of 7 to 19 hidden units. These networks were then assessed using the test set showing that the best performance was obtained with a network of 17 hidden units giving a test error of 0.127. This resulted in the following selected models for each fold as shown in Figure 6.2.2. The lowest model was the NN with 8 units giving a test error and validation error of 0.106 and 0.135 respectively.

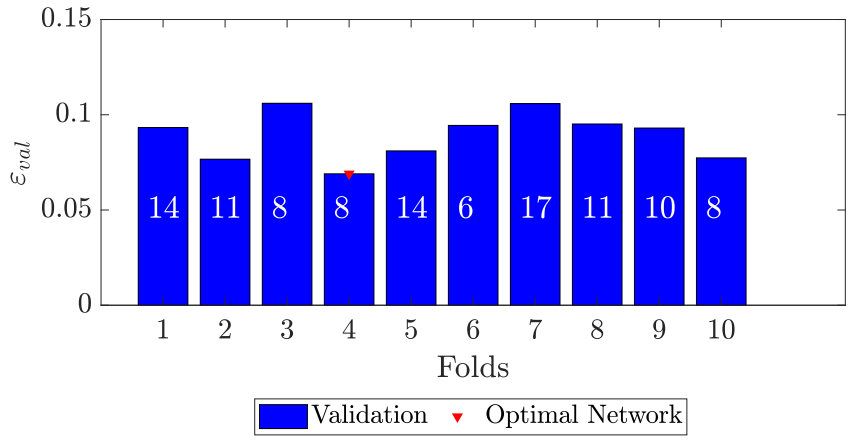


Figure 6.2.2: Validation error for candidate models for each fold, where the model with 8 units showed the best validations performance.

The resulting optimal network predictions using the test set obtained from the PLB are shown in Figure 6.2.3. The resulting predictions using the test set which has not been previously seen by the NN during the training or validation stage show a linear relationship between the predicted and target values. Therefore, this shows that the predictions are correlated with the target values.

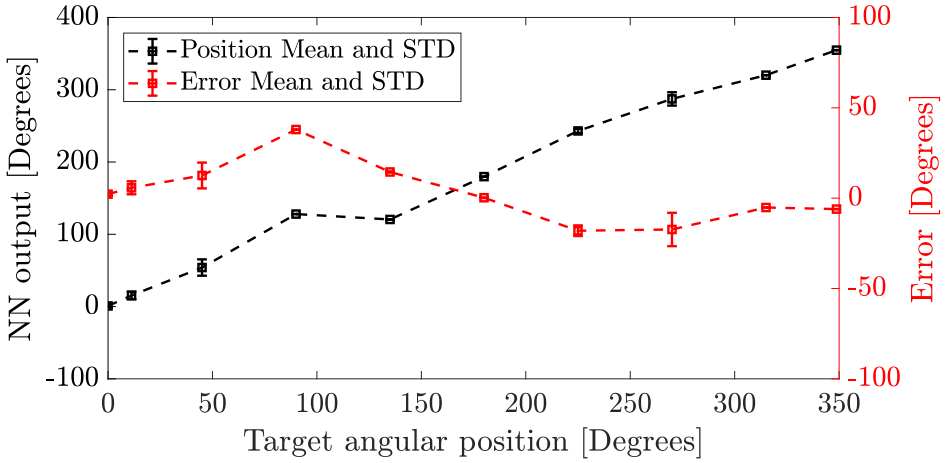


Figure 6.2.3: Test set predictions around bearing circumference.

Figure 6.2.4 shows the results from introducing the EDM notch at 0 and 11 degrees and the overload generated on the bearing outer raceway. These results were obtained from 100 observations from the notch and overload experiment. These bursts were processed similarly as the bursts used for training purposes aided with the interface

built-in Matlab as previously shown. Although, there is a high variance in the angle prediction, the post-processing of the 100 observations led to the expected defect positions by using the average of the NN predictions.

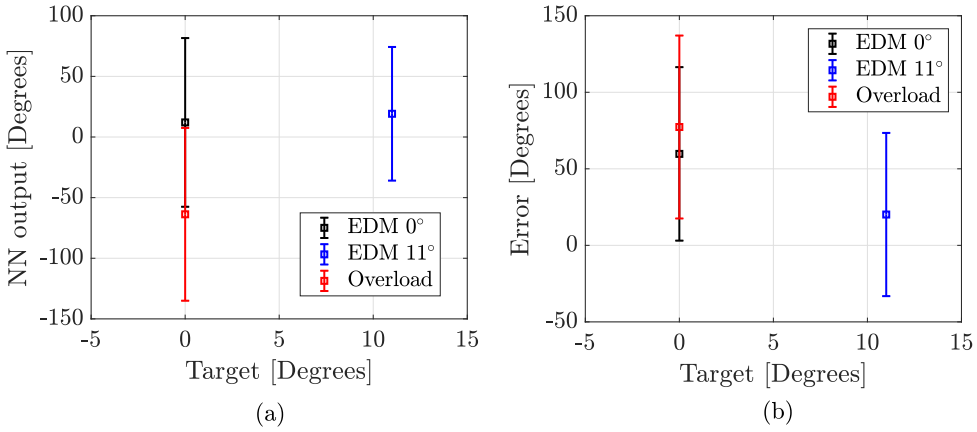


Figure 6.2.4: (a) Predictions using EDM defects at 0 and 11 degrees and overload defect at 0 degrees. (b) Error distribution from NN predictions.

6.2.2 Damage localisation in MultiLife rig

The cross-validation process required the assessment of 4600 different NN models in total (460 models × 10 Folds) and thus being relatively computational expensive. The results from the cross validations are shown in Figure 6.2.5.

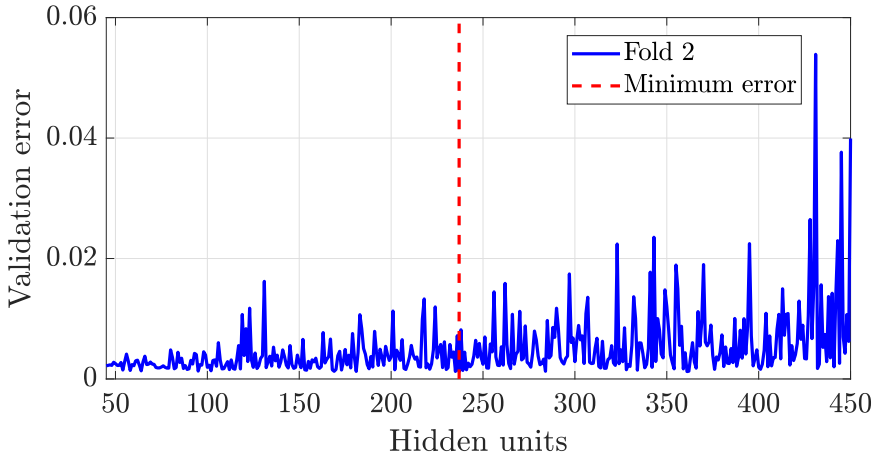


Figure 6.2.5: Cross-validation error results for the MultiLife rig for the 2nd fold.

The results from the cross-validation shown that there is not a significant error reduction after using a network of 237 units as it shown a test and a validation error of 1.22×10^{-3} and 1.54×10^{-2} respectively in Figure 6.2.6. Moreover after that value, the validation error tends to increase as it shows to be an inflection point where more and less complex models showed to underperform.

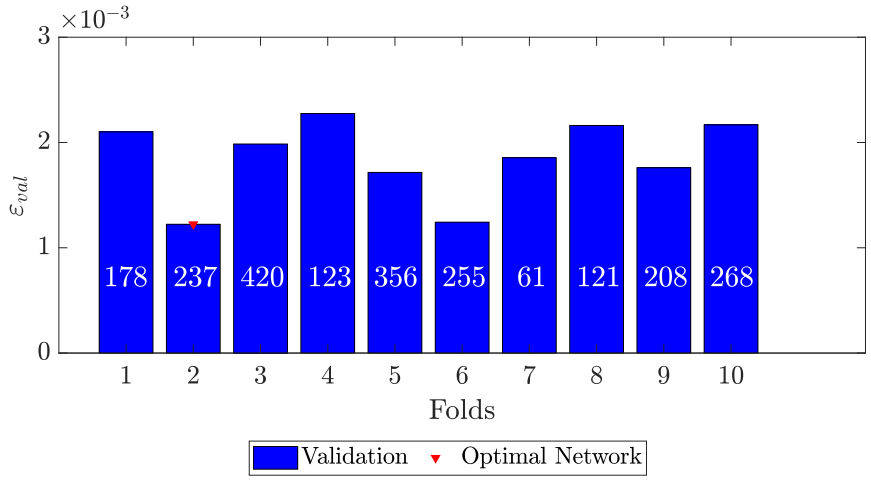


Figure 6.2.6: Validation error for candidate models for each fold, where the model with 237 units showed the best test performance.

The optimised network with 237 hidden units has been evaluated using a test set that has not been previously seen by the NN for the training and validation stage. The result in Figure 6.2.7 shows a source localisation error of 3% where a linear relationship between the predictions and the target values can be observed.

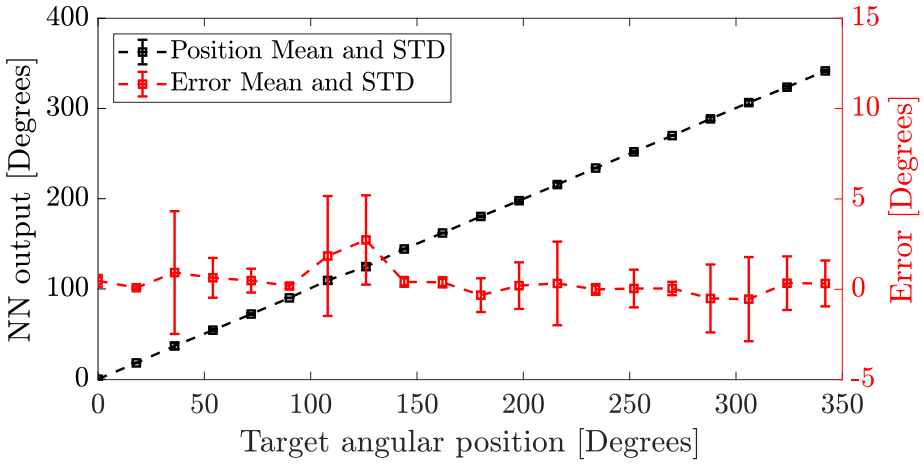


Figure 6.2.7: Test set predictions around bearing circumference.

The optimal network among the candidate networks was identified and used to process the bursts extracted from the EDM notch. Figure 6.2.8 shows the bursts predictions for each damage position at 0, -45 and +45 degrees.

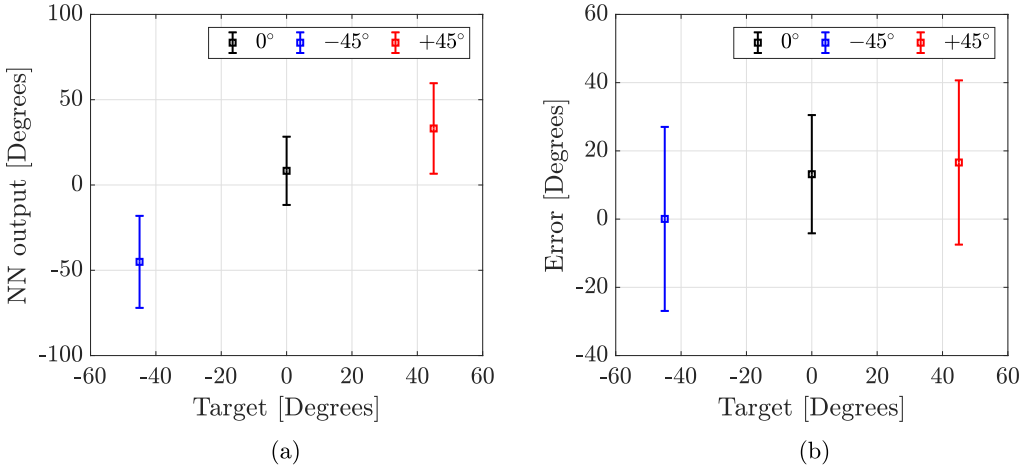


Figure 6.2.8: (a) Prediction angular position and error distribution. (b) Error distribution from NN predictions.

6.2.3 Damage localisation results for scaled-down rig using time-delay and sum beamforming approach

The controlled PLB excitations at 0° showed the best accuracy attained with the method with a relatively small variance. A trend can be seen in the variance when the localisation is attempted at zones far away from the contact zone of the bearing. Burst localised at +45° and −45° shown sometimes a degradation on the constructive patterns allowing the formation of multiple maxima points. This effect might be attributed to the small delay variations measured between sensors 1 and 2. As explained before, the direction of the steering vector depends on the magnitude of the delay calculated between sensors. In this case, these small variations might lead to abrupt changes in the vector orientation which ultimately affects the beamformer constructive patterns.

The localisation of an EDM notch defect was also possible along the contact zone, even though the achieved accuracy was not comparable to the one obtained using the PLB test, having a resulted in a minimum error of 9.4° degrees.

Table 6.2.1: Beamforming localisation results.

Test	Angle [θ]	Mean [$\bar{\theta}$]	STD[θ]	Error [$\Delta\theta$]
PLB	$-45°$	$-0.47°$	$-11.65°$	$44.53°$
	$-11.25°$	$-9.33°$	$+1.4°$	$1.92°$
	$+0°$	$-1.07°$	$+0.83°$	$0.83°$
	$+11.25°$	$+2.67°$	$+2.68°$	$8.58°$
	$+45°$	$+32.6°$	$+16.62°$	$12.4°$
EDM	$+0°$	$+15.13°$	$+21.60°$	$15.13°$
	$+11.25°$	$+9.4°$	$+26.3°$	$1.85°$
Overload	$+0°$	$+20.47°$	$+45.2°$	$20.47°$

6.3 AE sensors self-assessment

A similar approach using the spectral data for performing the novelty detection in bearings was applied here. The features obtained were relatively clean in terms of noise. The baseline dataset was taken to be the condition where the sensors are completely attached to the structure whereas the test dataset consisted of all the different debonding levels at 25, 50, 75 and 100. The last data set consisted of a completely free PZT patch.

An Outlier detection approach using the Squared Mahalanobis Distance (SMD) was used to calculate any onset of failure on the PZT sensors patch. A covariance matrix was built based on the unfaulty sensors data and then used to calculate the distance of the entire data set (non-faulty and faulty features). Figure 6.3.1 shows the results of the Mahalanobis Distance for each set.

A Monte Carlo simulation with 1000 iterations was performed to obtain the threshold value of 99% for 26 training observations, which shown to separate the faulty sensors from the healthy ones.

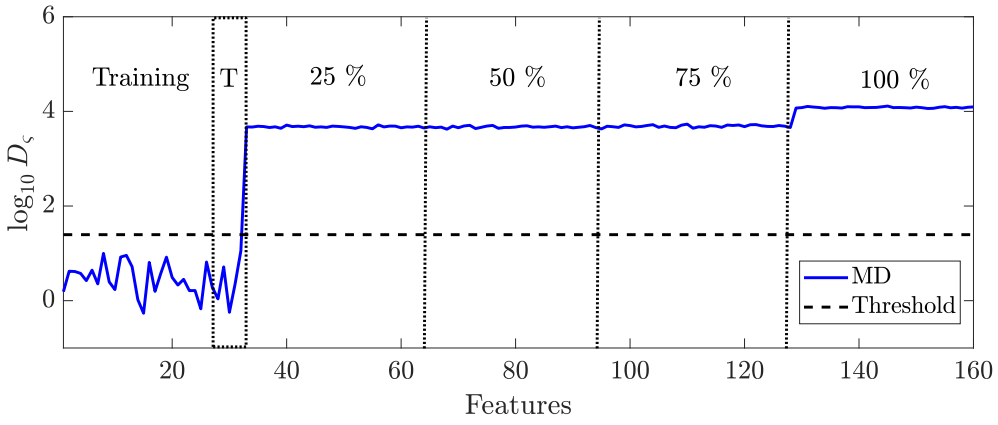


Figure 6.3.1: Outlier detection results for different debonding conditions at a 99% threshold. A 100% of the test set T was correctly classified whereas different debonding leves where detected as a novelty.

Figure 6.3.2 shows the results of applying LDA to the previously described dataset. As was expected, a high degree of separation was attained using this method. It may be seen that no false positives or false negatives occurred in this case. The method was able to classify the debonding condition at 25% as a damaged condition and seems to correctly classify further debonding progression as demonstrated for the completely detached condition.

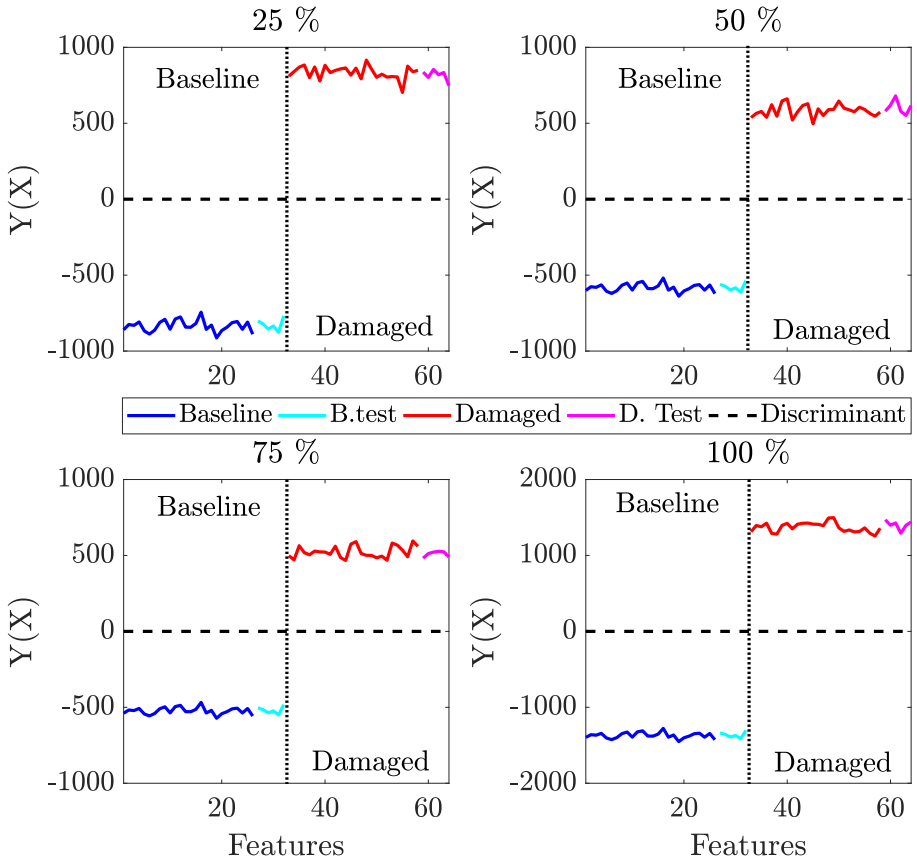


Figure 6.3.2: LDA results for different debonding conditions.

6.4 Discussions

6.4.1 Novelty detection

Two methods have been implemented in this chapter in order to detect damage in roller bearings. A set of time-frequency features have been shown useful as a pattern for differentiating between normal and damage condition states. Both methods were tested using the same feature extraction techniques in order to make a fair comparison and have been proven useful in terms of scalability on a real sized bearing. Considering that the real scale bearing had a higher ratio of bearing-to-defect size and lower shaft speeds compared with the scaled-down bearing, the selected features have been proven useful for damage detection at an early damage onset.

Outlier detection has been shown relatively useful for damage detection using spectral features. Even though a severe overlapping in the Mahalanobis distance occurred, it allowed the distinction between different states. On the other hand, Linear Discriminant Analysis showed great accuracy as it reached the lowest rate of 1.75% and 0% of false positives in the scaled-down and MultiLife rig respectively. LDA showed a higher percentage of detecting novelty on the overloaded bearing as 100% of the damage condition was correctly classified, whereas outlier detection obtained a 100% misclassification rate.

A run-to-failure test showed that it was possible to perform outlier detection using the SMD as a deviation measurement from an undamaged condition. This experiment required to simulate an overload defect using a FEM approach to reduce the bearing life expectancy, although this required a very sophisticated technique for validation, the results can be obtained within reasonable conditions in terms of the number of measurements and material conditions. These issues are acknowledged by the main author in [2].

It is important to mention that the results obtained from the outlier analysis are highly dependant on the threshold level used and for instance on the method used for estimation. This leaves an open criterion of where the threshold level must be in order to reduce the false positives and negatives. In the case of LDA the criterion is slightly more straightforward as it comes more naturally to estimate a boundary for both states from the training data.

Additionally, outlier detection is designed to classify a single point as being an outlier or not, without considering the other test data points. This is achieved by reducing the multidimensional feature vector into a one-dimensional scalar and calculating a Novelty Index without taking into account the structure of the whole observation points. It is important to mention that so far there is an assumption on the data distribution as both methods mainly rely on that the data follows a Gaussian distribution. In practise this assumption might fail, more specifically in the case of LDA as there is no necessity why each class should have the same distribution. As a consequence any deviation from this assumption might lead to underperformance during the classification.

A different approach might be considering a few aspects of the distributions analysed by assuming where the most information content is contained. In this case, the assumption was in the distance between classes such as the case of the LDA approach but other valid assumptions might consider that the information is contained in data variance such as the case of PCA.

The application of LDA has great potential for separating features when there are relatively small variations in the data such as in this case. This method, in contrast with the SMD, managed to find the best representation that maximises the distribution separation between the damage and undamaged state.

LDA also suffers from overfitting issues. This can be seen when optimising the set of weights in order to determine the hyper-plane that best separates the data. Normally this happens when there is a relatively small number of data points. In such cases, when optimising the set of weights, they may become highly dependant on a specific training set. That is not the case when the number of patterns N is higher than the number of dimensions D used in the feature vector as in this problem. If this is an issue a regularised version of LDA might be advisable [138].

Outlier detection can be useful in a situation where the data can be described using a Gaussian Normal distribution. In the case of the scaled-down experiment the feature vector was obtained from one operational conditional where the rotational speed remained relatively constant. This is generally not the case in the practical scenario as normally gearboxes are expected to work at different ranges of speeds that lead to multiple baseline features as shown in Figure 5.3.7.

This can be partially simulated using the MultiLife rig as it allowed the operational speed to be changed between five different values. Intuitively, it is expected that the data from the pristine condition vary accordingly to the rotational speed of the shaft f_s and the ball-spin frequency f_b. The data would no longer follow the Gaussian distribution assumption required to perform outlier detection as the covariance matrix calculated from the training data does not fairly represent the data obtained at different operational conditions. One commonly expected behaviour of performing outlier analysis in this situation is the prevalence of triggering false positives alarms even on data obtained from the baseline condition as shown in Figure 6.4.1.

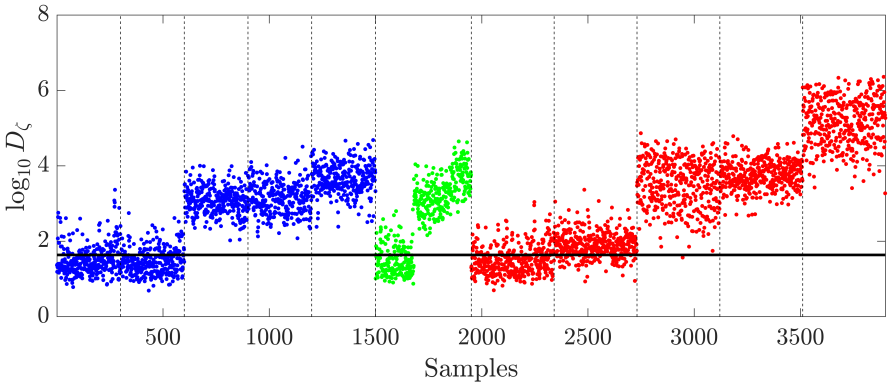


Figure 6.4.1: Outlier detection using each baseline feature, blue dots represent the training sample, green dots represent the test samples and the red dots the damage condition at each rotational speed.

The covariance matrix was estimated from 300 samples for each normal condition giving a total of 1500 samples, a total of 450 samples were reserved for the testing set which contains a total of 90 samples for each case of the normal condition. The threshold level was estimated using the Monte Carlo approach for the total number of observations. The results of Figure 6.4.1 show the Mahalanobis distance obtained for the training, testing and faulty data condition. The faulty data condition was obtained from a defected bearing at a 0° at a rotational shaft speed of 20, 40, 60, 80, 100 RPMs. Even though a vast majority of the 20 and 40 RPMs baseline test data was classified as being undamaged, high-speed ranges were completely triggered as false positives due to data being completely different in terms of their statistical parameters. These problem has been known as variation of environmental and operational conditions and has been extensively studied in the field of SHM. Several approaches have been used to tackle this issue by normalising the data for different conditions and desensitise features to these variations [40]. Clearly, special precautions on data assumptions have to be made in order to improve the robustness of the novelty detection stage.

6.4.2 Damage localisation

The machine learning approach using NN for localisation of AE sources was successfully implemented in both scaled and real size bearings. The principal aims of this method allowed the localisation of defects on complex geometries regardless of propagation speed. These effects were rather captured from the characteristics of the data obtained from a training map and allowed the localisation of simulated defects on bearings.

This chapter allowed the two main problems in the CM localisation stage to be overcome. First, it allowed the localisation of damage in complex structures and second, it allowed localisation without a priori knowledge of the propagation speed in media. These two problems are the most challenging in SHM/CM and, so far, they seemed to be better dealt with a machine learning approach. Additionally, the implementation of this method is not only limited to WTGB, as it can potentially target more sophisticated applications such as wind turbine blades, bridges, aircraft and any other structure/machine that comprises complex geometrical features and unknown material composition.

It was found that the MLP was able to perform under certain conditions. Mostly attributed to the noise dependence of the input data. The scaled-down rig initially showed the possibility of implementing the method in real scale components but a trade-off between the data variance and separability was shown to be the main problem for relatively small components. This situation did not occur with the real scale component which showed greater separation between features.

This chapter used the NN approach and treated the localisation task as a regression problem in terms of a machine learning approach which overcomes many of the analytical and standard triangulation techniques. Nonetheless, it is important to mention that there is a variety of regression methods available that have not been implemented yet for this type of problems. Gaussian Processes is a potential method that might be useful for this applications as it was well implemented for AE localisation before in structures [90].

So far, only time-domain features such as the TOA, have been used in this chapter for the feature selection stage. Unfortunately, they have shown to be greatly affected by noise in certain situations where there is excessive background noise. However, the AIC picker used in this work dealt with this problem to some extent as a great effort in selecting the AE signals was spent during pre-processing. In certain situations where a low signal-to-noise ratio is unavoidable, onset estimation might be challenged as the method might not discriminate between the background noise and the actual AE burst. This situation allows the exploration of several pre-processing techniques that might help to improve this situation. Filtering techniques and the use of time-frequency features to determine the time onset such as the CWT might be a further topic for research.

It was shown that it is possible to localise a late onset of damage by using an EDM notch. Alternatively, it was not completely obvious for an early damage onset such as seen using an overloaded bearing in the scaled-down rig. The main reason for this is because features such as TOA are highly sensitive to noise in the weak bursts that were generated by an early interaction between the defect and the rollers. Despite this, it is still possible to obtain a prediction but its reliability may be questionable as the NN will be required to extrapolate. This is clearly a problem and the reliability of the prediction will be dependent upon how great the extrapolation is. This is certainly not deemed acceptable as NN are poor at extrapolation.

Additionally, it was still plausible to generate a map by performing excitations around the component of interest, thus leading to highly accurate predictions using the test measurements for each bearing by using data with the same characteristics. In reality the TOA generated from the defects studied here had a slightly different characteristic in terms of their distribution mainly due to noise levels. A possible solution could be to train the NN with corrupted data adding noise systematically. This has been previously been shown to give an improvement in NN generalisation capabilities [131]. In this case, TOA information across the bearing of the full-sized rig has been included, as a means for accounting for these variations in the TOA features.

An interesting method recently used in AE such as the Beamforming method seemed to allow the localisation of defects on bearings to some extent. The main advantage

of this method is its capability to localise sources by using a different time-domain feature such as the wave delay. This approach will normally require less sophisticated methods to pre-process the signal such as the AIC approach. By using the signal interference patterns, the localisation is achieved by analysing the peaks formed from the constructive/destructive interference patterns. A drawback of this method is that it relies on a constant wave speed propagation and it needs a priori-knowledge of the structure. Therefore, this requires consideration of the geometry of the propagation media and its propagation speed in the material. This method is also constrained by the assumption of an interrupted wave propagation path resulting in it only being able to localise sources in simple geometries. Moreover, this does not account for media that is composed of multiple materials. In this case, the bearing support was manufactured using only one material thus simplifying this localisation approach.

6.4.3 AE sensors self-assessment

The novelty detection methods used in this chapter have shown to correctly distinguish between healthy and faulty sensors. Moreover, it seems that using the linear part of the susceptance features has been shown to be highly sensitive to different debonding extents. Both the SMD and LDA method have shown a clear separation between healthy and faulty data.

Even though the method has been applied successfully to identify faulty sensors, several issues might arise. A baseline dataset must be known a priori and this also assumes that the bonding was performed consistently for the entire sensor network. Moreover, the type of bonding used can also influence the response features.

Whilst both methods seemed to be useful for a novelty detection application they were less successful at representing the debonding progression. The Mahalanobis distance seemed to be unaffected by the increasing debonding level even though a slight increase in the completely free condition was seen in the results. This does not suggest any relationship between the amount of debonding induced as they do not increase monotonically as the debonding progresses.

A similar observation occurred with LDA, as there was no clear relation between the different debonding levels. The best separation occurred at the completely free

scenario as occurred with the SMD. A better representation was observed on the non-linear part of the impedance curves showing a correlation for different debonding levels. This could be further exploited for the prediction of damage progression.

Additionally, this debonding effect seems to have affected the susceptance values, but it is important to state that other conditions might give a similar deviation from the baseline data. Sensor breakage has been shown to have a similar effect on the susceptance curves [101].

Additionally, the quality of the soldering of these components might affect the admittance curves. Figure 6.4.2 shows consistent measurements for completely bonded sensors whereas the sensor with 50% debonding shows multiple peaks and a highly distorted curve. The sensor with 50% debonding was damaged due to a low-quality soldering technique. This behaviour might occur in situations were soldering operations were performed close to the depolarisation temperature and thereby damaging the sensor during installation. This clearly suggests that special precautions are needed to be taken into account during the installation.

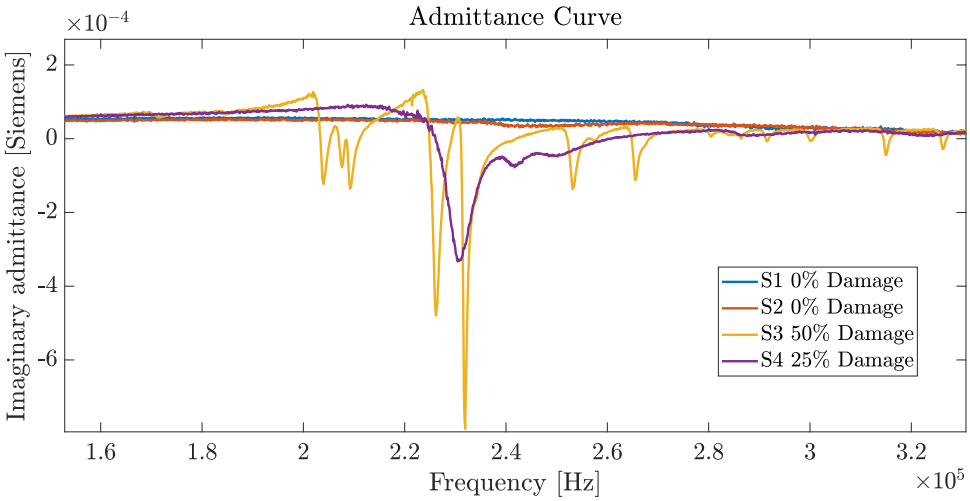


Figure 6.4.2: Depolarisation effect during soldering. A sensor with 50% debonding showed a distorted admittance curve due to depolarisation effects.

CHAPTER 7

Conclusions and further work

This work presented a machine learning approach for CM of WTGB by using vibrational and AE data. The main aspect of this work dealt with a well-established hierarchical approach for SHM/CM where each level of diagnosis was sequentially targeted. Moreover, features were selected from the relevant signals, where vibrational data was more relevant for of incipient damage detection whilst AE allowed the subsequent targeting of the damage location stage.

Several other topics were also considered in this work. The structural integrity of sensors in particular AE sensors was investigated and a means of assessing healthy and faulty sensors was suggested. Finally, an exploratory investigation into a methodology for quantifying overload damage on bearings was conducted. In summary, the following main conclusions can be drawn from the following chapters:

In Chapter 2, a review of current techniques suggested a vast number of signal processing methods for extracting features from vibration signals. Vibrational methods have been proven to be robust for monitoring bearings, in particular. Traditional methods have been proven as a standard for detecting damage via a qualified operator with previous expertise. As mentioned in Chapter 2, sometimes these features are beyond the normal operator intuition and better tools must be developed for this purpose, thus leading towards Machine Learning methods. In literature, novelty detection has been an extensively discussed topic, more specifically in terms of roller bearing applications. On the other hand, damage localisation has had less attention, as detection of damage normally leads to an immediate bearing

replacement. The low cost of small bearings is the main justifying reason for only one level of diagnostics. Moreover, the higher the level of diagnostics required, the more complex the problem becomes. This situation can be potentially useful for other types of applications such as WTGB where the costs justify the implementation of the second level of diagnostics. This stage conjointly with some recent advancements in the WTGB field such as the MultiLife concept can be potentially be integrated in the near feature to reduce the downtime and unexpected breakdowns. Moreover, this can potentially increase the life-span of such components.

In Chapter 3, an introduction to novelty detection using the Outlier detection and LDA method has been explained. The advantage of using Outlier detection resides that only a healthy condition of the structure/machine is needed to represent a healthy condition of the structure/machine. This allows calculating the deviation of the measured conditions from the healthy data distribution. Although this approach is practical in many situations where the data can be easily separated, LDA takes to maximise this separation by exploiting features from the healthy and damaged state. This has an advantage over Outlier detection as a threshold is set by a discriminant function whereas an additional technique is needed for outlier analysis.

Additionally, a background for NN has been included in Chapter 3. The main reason for using this approach was to exploit the localisation by means of understanding the data structure rather than implementing a physics-based model. As is well-known, this is the data-driven approach in SHM/CM. This approach overcomes two difficult problems that analytical models might not solve: Localisation under (1) an unknown propagation speed and (2) in a complex geometry. This chapter provided the main steps to implement the method in this work for regression problems. The background provided information on the MLP architecture, how to define the NN size in terms of the number of hidden units and a method to avoid overfitting by using the K-fold cross-validation method. Finally, the beamforming localisation method was introduced. Although not a ML method, beamforming uses a different set of wave features that were worth exploring.

A methodology for simulating the damage produced by an overload event was explored in Chapter 5. A Finite Element model was developed to investigate the damage progression at different loading events and showed a high level of agreement

with the experimental measurements using Neutron Bragg imaging. A fatigue test performed showed a reduction in the remaining life of the bearing, even though it was a clear fracture in this controlled experiment. The run-to-failure test performed on Chapter 5 showed a severed surface failure. This was also seen on the SMD as it showed a potential deviation from a normal state during the bearing operation. Despite providing a means for predicting the plastic deformation in bearings, it provides a systematic method for inducing controlled defects. This in practise allows the premature reduction in the bearing life in specific locations of the bearing and reduces the test times. Additionally, the use of Neutron Bragg imaging has been shown useful in detecting plastic deformation in bearings. This allowed a non-destructive approach to qualitatively describe the damage due to overload events. Even though the capabilities for predicting the damage extent have been shown, further investigation with more experimental data needs to be performed in the future. This will provide more reliable correlations between damage size and the remaining life in roller bearings and consequently lead to the damage prognosis stage in CM.

Chapter 5 presented a ML approach using vibrational signals, this chapter showed the signal processing steps used to produce unambiguous features. These were enhanced by using a low-pass filter or a Hilbert Transform approach and then extracted using the STFT method. Additionally, Chapter 5 presented the features extracted for the localisation stage, this features represented as TOA values were preprocessed and encoded in order to provide a suitable and unique data representation to the NN.

Additionally, an integral approach was implemented in Chapter 5 where the AE sensors were considered as part of the CM scheme. Debonding not only distort the AE signal in the time domain but also the frequency content of the signal, ultimately leading to the deterioration of the information captured by the AE sensors. A similar methodology used for novelty detection was implemented to detect faulty PZT patches. The use of susceptance curves was shown to be sensitive to bonding degradation.

The features obtained for different bonding conditions were shown to overlap considerably when using the linear part of the susceptance curve. Further analysis of the experimental data showed that the resonance peaks seem to be correlated

with the damage level. Considering that novelty indicating possible debonding will immediately lead to a sensor replacement/re-installation, a damage localisation assessment stage is deemed unnecessary. Additionally, this can be solved by exploiting the redundancy of the sensor network in such situation.

In Chapter 6, the SMD was used as a measure of any deviations from a normal state of the rotary machine. Even though it was possible to identify cases where the bearing behaviour deviated from the normal condition, several overlaps occurred. It showed that using this method to detect a novelty might lead to several false positives when the outlier data is not significantly distant from the baseline distribution. Two aspects are important to notice in this method: (1) It does not exploit any other characteristic of the multi-dimensional feature as it reduces the features to a one dimensional regardless of its data separability and (2) it does not consider the measurements as a whole data distribution.

These questions lead to the exploration of a method that can not only exploit the novelty detection capability but also consider the entire data structure. LDA was shown to answer both questions by maximising the data separability and treating the data as a whole distribution. A significant separation was obtained using this method and allowed clear diagnostic differences between a normal state and a damaged one.

One limitation that seemed to affect both methods was that neither was able to detect novelty at lower rotational speeds. In practise, this would not be an issue, as gearboxes will normally deal with speeds at least 10 times higher than 20 RPMs. Beyond that point novelty detection was capable with both methods.

AE technology was exploited for obtaining meaningful feature maps and, from the experiments, was shown to provide unambiguous features for damage localisation. Even though the generated maps can be used directly as in the case of the Delta-T method to localise damage by manually interpreting the Time-of-Arrival (TOA) data. The use of a Neural Network NN removes the operator's influence and naturally interpolates the data. It is important to note that this point depends on how cautiously the NN has been trained and the ambiguity of the features.

Even though it was initially expected that greater accuracy would be obtained on the scaled-down rig, the TOA features obtained from the defects were highly

contaminated with noise. This led to the overlapping of the features thereby reducing the accuracy during the predictions. The effect of localising defects on smaller parts becomes a challenge due to the requirement of better time resolution for such applications. The localisation on the full-sized bearing did not suffer from this issue as the noise variance was relatively lower. This allowed a clear distinction between each observation as less overlapping occurred between features.

In practise, secondary AE sources become a problem when implemented in such types of mechanisms where effects of dropping rollers and secondary components occur, thus pre-selection of the AE must be conducted. in this case, a criterion was implemented based on the periodicity of the generated bursts. This led to select the AE bursts that were generated at the characteristic defect frequency of the bearing and filter those generated at random times. In a real scenario there is no such behaviour, transient wind speeds might induce variations and thus making it difficult to filter such events.

The use of a NN normally requires a large number of data features in order to generalise over a training sample. This also represents a limitation on the model size, as discussed in Chapter 3. Using such techniques requires not only the consideration of sample size but also the consideration of additional steps such as cross-validation.

Special attention to the features should be taken into account by attempting to avoid ambiguous features that might lead to spurious results. These maps can be shown to provide a unique mapping for each location point. Moreover this can be shown even after the data transformation. The reason why this step is important is that a continuous map must be used to train a NN. The initial features arranged in terms of each angular position were unambiguous but they were not a continuous map. This led to post-process these features and transform them in order to allow them to remove these singularities.

An additional approach was implemented in this thesis as it exploited other types of features on the AE signal. As seen before, the use of TOA maps can be susceptible to noise contamination and can lead to ambiguous results. An alternative approach used the shape of the waveform rather than the arrival onset, thereby resulting in reasonable predictions. The negative aspect is the number of assumptions needed to

be made. Propagation speed must be assumed and, additionally, localisation cannot be performed on complex geometries.

The use of the SMD and LDA have been shown to be useful for novelty detection and capable of diagnosing sensors with at least 25% debonding. This shows the sensitivity of the susceptance curves as features for debonding.

The linear part of the susceptance curves was shown to be useful for novelty detection as they successfully allowed the provision of unambiguous features for bonding degradation. Further tracking can potentially be performed by analysing the resonance peaks at the impedance curve; but this in practise seems less feasible. The main reason is that, after any onset of bonding degradation the first level of detection might be sufficient to highlight the issue and to take corrective action.

This work highlights the need to take great care during AE sensor installation. If in practise any of these methods are expected to be implemented, further steps such as ensuring that the sensors are completely attached to the surface in a consistent manner using the same type and amount of adhesive should be performed.

Bonding effects seem to be more noticeable at the resonance frequencies and the amplitudes of the admittance curve. Even though they are shown to be unambiguous features for damage, their use is not expected to improve the novelty detection performance.

7.1 Further work

A few challenges were encountered in this thesis leaving a few issues to be further addressed. The novelty detection performed in Chapter 6 answered whether or not damage occurred on bearings but under a few specific conditions. Vibrational data required pre-processing of the signal, in the form of signal enhancement via Hilbert Transform to identify defect frequencies. The main issue occurred due to severe overlapping of the features but this was partially solved by using LDA which managed to separate both states.

The main issue could be due to the time-frequency resolution limitations of the Short-Time Fourier Transform (STFT) as it is well-known in signal processing. Further exploration of the Continuous Wavelet Transform (CWT) might be useful as it overcomes this time-frequency resolution problem. A feature vector using the coefficients of the CWT of the vibrational signal might solve this problem as this method does not have the time-frequency limitations observed in the STFT. This allows the exploration of different signal processing techniques to allow unambiguous features for novelty detection.

Novelty detection using SMD was performed through a priori knowledge of the rotational speed of the defect frequencies beforehand. In a real scenario this rotational speed might not even be constant due to the transient behaviour of the wind. An exploration of data normalisation methods might allow the building of more robust baseline conditions in order to perform novelty detection. These methods have been successfully used before to tackle the Environmental and Operational Variations in the SHM context.

The damage localisation chapter showed a way to use NN by systematically following best practises. This required the use of a relatively large data set for training, validating and testing which in this case took several days to obtain. This leads to two different issues for further exploration, either regression techniques that are able to use smaller datasets could be explored or automatic methods for generating the TOA maps could be investigated. The latter can have a great impact on reducing the time spent in building the maps. Even though PLB have shown to generate very repetitive excitations other methods such as laser-based methods have been used to extract these maps automatically by exploiting the thermo-elastic effect [90].

A few aspects of using TOA maps can be considered for further exploration. Even though the features were processed using an accurate technique for determining the AE wave onset, they are sensitive to noise to some extent. As seen in the prediction it is rather difficult to rely on only one measurement and thus the whole distribution might be more useful in this case. Due to this issue, a few questions regarding other types of features that might be less sensitive to noise might arise such as the shape of the AE burst in time and frequency domain.

Additionally, this might lead to assess the use of narrow-filtering techniques as in the time-delay beamformer section to isolate a specific propagation mode.

It is important to notice that the TOA features were obtained from the raw time-domain signal by the assumption that the fastest propagation mode will be captured first. This information was clearly obtained from the time-domain but it does not restricts to build a map based on the arrival of each of the frequency components of the AE signal. A time-frequency approach using the CWT coefficients has been thought by building a map based on the first arrival at a specific frequency.

The use of beamforming allowed the exploration of a different set of features that exploited the signal waveform by forming constructive interference patterns around hypothetical locations. As this method was implemented in a fairly simple structure, it lacks the flexibility to be implemented for more complex structures. Further exploration into methods of overcoming this issue would be of interest.

The novelty detection using the SMD and LDA was shown to be successful in the detecting debonding on PZT patches. A particular problem with that approach is its inability to perform the task in an online manner. The ideal situation would allow the on-demand measurement of an AE signal signature and monitoring any signature changes. This might be performed by using a repetitive excitation such as a PLB on a reference point and monitoring of the signal features.

The model used to predict the damage progression was shown to generally agree with the experimental data. This does not predict remaining life by itself but, rather allows it to be systematically used for controlling the overload size. This certainly opens a new set of investigations where more experimental data must be obtained. Due to limitations of the measurement time and due to the rather probabilistic failure rates enough experiments were not able to be performed in this work. By allowing further experimental testing it may be possible to correlate the damage sizes to the failure rates in bearings.

Even though the techniques presented in this work have shown damage detection and location capabilities, it is important to mention that their implementation in a real WTGB requires further improvements in their robustness. It should be clarified that so far all the diagnosis stages were performed under well-controlled circumstances. In

reality, a priori knowledge of the defect is available, multiple operational conditions might be encountered, moreover, a unique baseline condition might be unavailable. Additionally, the techniques showed here assumed a unique defect instance. In practise, this is rarely the case as there is nothing that can prevent multiple random defect generation around the bearing component and thereby detection and location should allow this possibility. As seen in Chapter 5 damage location was performed based on an a priori knowledge of that certain bursts were generated at a given damage characteristic frequency, thereby filtered from the random pulses generated from the AE signals. Despite being a relatively well-controlled environment it was almost impossible to eliminate these random events and required a certain level of guessing during the filtering. Comparing this situation with what can be potentially be encountered in a real scenario can lead to a difficult task and might trouble the interpretation of the localisation prediction.

So far these methods have been tested in one specific component and no knowledge of whether or not the data obtained for training the damage detection and localisation will work in different WTGB. In reality, multiple WTGB will be encounter in a Wind Turbine Farm and it might need to be considered that variations might be encounter. This again deals with the robustness of the damage detection and location methods and it is an issue that needs to be studied before implementing it. Despite these challenges, it has been shown that the approaches followed in this thesis allowed to target the first and second levels of diagnostics together with an integrated approach to monitor the AE sensors. This work performed a systematic approach using machine learning and allowed to understand the challenges during the implementation of these techniques in WTGB.

As shown in the results in Chapter 6, it is possible to perform damage detection on PZT sensors using susceptance curves as features. So, damage can be detected and further actions such as replacing the sensor or bonding it again can be taken. Further questions such as the extent of damage might arise. From the features obtained from the linear part of the susceptance curve seems difficult to answer. An analysis of the resonance frequency seems to show a consistent feature to understand damage progression and also seems to be an unambiguous feature for novelty detection.

Another possible area for investigation is to study the effects of temperature. It is well known that each piezo-electric component has a temperature limit (Curie Temperature). This temperature limits the operational temperature of the PZT sensor as the sudden depolarisation of the PZT patch will lead to distorted impedance curves leading to inaccurate measurements. The depolarisation effect seems to have an effect on the admittance curves and thus allowing to detect damage in such situations.

Contributions

Journals

1. A. Reid, I. Martinez, M. Marshall, T. Minniti, S. Kabra, W. Kockelmann, T. Connolley, and M. Mostafavi. Mapping of axial plastic zone for roller bearing overloads using neutron transmission imaging. *Materials & Design*, 156:103 – 112, 2018.

2. A. Reid, C. Simpson, I. Martinez, S. Kabra, T. Connolley, O. Magdysyuk, C. Charlesworth, M. Marshall and M. Mostafavi. Measurement of strain evolution in overloaded roller bearings using energy dispersive X-ray diffraction. *Tribology International*, 140:105893, 2019.

3. A. Reid, M. Marshall, I. Martinez, S. Moorby, T. Connolley, M. Mostafavi and S. Kabra. Measurement of strain evolution in overloaded roller bearings using time-of-flight neutron diffraction. *Materials & Design*, 2020.

Conferences

1. I. Martinez, M. Marshall and I. Antoniadou. PZT sensors benchmarking and a wavelet transform approach for acoustic emission source location. *24th International Congress on Sound and Vibration*, ICSV 2017, 2017.

2. I. Martinez, A.G.P. Reid, M. Marshall and I. Antoniadou. A wavelet transform approach for acoustic emission localisation and an examination of PZT sensor self-diagnostics. *BSSM 12th International Conference on Advances in Experimental Mechanics*, BSSM 2017, 2017.

3. A. Reid, I. Martinez, M. Mostafavi, I. Antoniadou and M. Marshall. Overload and failure of wind turbine gearbox bearings. *BSSM 12th International Conference on Advances in Experimental Mechanics*, BSSM 2017, 2017.

Bibliography

[1] SKF. Premature bearing failures in wind gearboxes and white etching cracks. http://evolution.skf.com/premature-bearing-failures-in-wind-gearboxes-and-white-etching-cracks-wec/. Accessed: 21-4-2020.

[2] A. Reid, I. Martinez, M. Marshall, T. Minniti, S. Kabra, W. Kockelmann, T. Connolley, and M. Mostafavi. Mapping of axial plastic zone for roller bearing overloads using neutron transmission imaging. *Materials & Design*, 156:103 – 112, 2018.

[3] Wind energy statistics: 2018. https://www.renewableuk.com/page/UKWEDhome. Accessed: 24-9-2018.

[4] Statistics show bearing problems cause the majority of wind turbine gearbox failures. http://energy.gov/eere/wind/articles/statistics-show-bearing-problems-cause-majority-wind-turbine-gearbox-failures. Accessed: 9-5-2016.

[5] M-H Evans. White structure flaking (wsf) in wind turbine gearbox bearings: effects of 'butterflies' and white etching cracks (wecs). *Materials Science and Technology*, 28(1):3–22, 2012.

[6] B. Johannes, E. Paul, H. Ludwing, and W. Karl. *Ball and Roller Bearings: Theory, Design and Application*. Wiley, 1999.

[7] D. McVittie. Wind turbine gearbox reliability. www.sandia.gov/wind/2006reliability/tuesday/14-brianmcniff.pdf. Accessed: 9-5-2016.

[8] T. Hamilton. A more durable wind turbine. https://www.technologyreview.com/s/416518/a-more-durable-wind-turbine/. Accessed: 9-5-2016.

[9] D. Milborrow. Operation and maintenance costs compared and revealed. 19, 01 2006.

[10] W. Vachon. Long-term o&m costs of wind turbines based on failure rates and repair costs. pages 2–5, 01 2002.

[11] T. Howard. *Development of a Novel Bearing Concept for Improved Wind Turbine Gearbox Reliability.* PhD thesis, University of Sheffield, 2016.

[12] T. Bruce, H. Long, and R. Dwyer-Joyce. Dynamic modelling of wind turbine gearbox bearing loading during transient events. *IET Renewable Power Generation*, 9:821–830, 05 2015.

[13] M. Whittle, J. Trevelyan, and P. J. Tavner. Bearing currents in wind turbine generators. *Journal of Renewable and Sustainable Energy*, 5(5):053128, 2013.

[14] ISO 15243-15243:2017. *Rolling bearings – Damage and failures – Terms, characteristics and causes.* ISO, Geneva, Switzerland.

[15] A.T. Harris and M.N. Kotzalas. *Advanced Concepts of Bearing Technology,: Rolling Bearing Analysis, Fifth Edition.* CRC Press, 2006.

[16] G. Lundberg and A. Palmgren. *Dynamic capacity of rolling bearings*, volume 196. Generalstabens Litografiska Anstalts Förl, 01 1947.

[17] ISO 281-281:2007. *Rolling bearings – Dynamic load ratings and rating life.* ISO, Geneva, Switzerland.

[18] H. Schlicht, E. Schreiber, and O. Zwirlein. Effects of material properties on bearing steel fatigue strength. *Effect of Steel Manufacturing Processes on the Quality of Bearing Steels*, pages 81–101, 1988.

[19] D. McVittie. Wind turbine gearbox reliability. `https://windpower.sandia.gov/2006reliability/tuesday/14-brianmcniff.pdf`. Accessed: 29-11-2018.

[20] F. Rasmussen, K. Thomsen, and T. Larsen. The gearbox problem revisited. 01 2004.

[21] A.D. Richardson, M.H. Evans, L. Wang, R.J.K. Wood, M. Ingram, and B. Meuth. The evolution of white etching cracks (wecs) in rolling contact fatigue-tested 100cr6 steel. *Tribology letters*, 66(1):1–23, March 2018.

[22] R. Errichello, R. Budny, and R. Eckert. Investigations of bearing failures associated with white etching areas (weas) in wind turbine gearboxes. *Tribology Transactions*, 56(6):1069–1076, 2013.

[23] R. Errichello, S. Sheng, J. Keller, and A. Greco. Wind turbine tribology seminar - a recap. Technical report, 11 2011.

[24] Ricardo plc. First wind farm trials commence of ricardo multilife(tm) bearing life extension technology. `https://ricardo.com/news-and-media/press-releases/first-wind-farm-trials-commence-of-ricardo-multili`. Accessed: 20-11-2018.

[25] Charles R. Farrar and K. Worden. An introduction to structural health monitoring. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 365(1851):303–315, 2007.

[26] C. Harris M. and A.G. Piersol. *Vibration and shock handbook*. McGraw-Hill, 2002.

[27] R. Bogue. Sensors for condition monitoring: a review of technologies and applications. *Sensor Review*, 33, 2013.

[28] R. Schulz, S. Verstockt, M. Loccufier, K. Stockman, and S. Van Hoecke. Thermal imaging for monitoring rolling element bearings. In *12th International conference on Quantitative InfraRed Thermography, Proceedings*, page 9. http://www.ndt.net, 2014.

[29] A.D. Nembhard, J.K. Sinha, A.J. Pinkerton, and K. Elbhbah. Condition monitoring of rotating machines using vibration and bearing temperature measurements. *Advances in Condition Monitoring of Machinery in Non-Stationary Operations*, 2014.

[30] L. Wang and R.X. Gao. *Condition Monitoring and Control for Intelligent Manufacturing*. Springer, 2006.

[31] R. Barron. *Engineering Condition Monitoring: Practice, Methods and Applications*. Pearson Education, 1996.

[32] B.J. Roylance. Ferrography—then and now. *Tribology International*, 38(10):857 – 862, 2005. Ferrography and Friends - Pioneering Developments in Wear Debris Analysis.

[33] A. Choudhury and N. Tandon. A review of vibration and acoustic measurment methods for the detection of defects in rolling element bearings. *Tribology International*, 32, 1999.

[34] A. Rytter. *Vibration based inspection of Civil Engineering structures*. PhD thesis, University of Aalborg, 1993.

[35] Y. Takeo and F. Takashi. A new acoustic emission source locating system for the study of rolling contact fatigue. *Wear*, 81(1):183 – 186, 1982.

[36] M. Elforjani and D. Mba. Condition monitoring of slow-speed shafts and bearings with acoustic emission. *Strain*, 47(s2):350–363.

[37] B. Eftekharnejad and D. Mba. Monitoring natural pitting progress on helical gear mesh using acoustic emission and vibration. *Strain*, 47:299 – 310, 07 2010.

[38] F.P. García Márquez, A.M. Tobias, J.M. Pinar Pérez, and P. Mayorkinos. Condition monitoring of wind turbines: Techniques and methods. *Renewable Energy*, 46:169 – 178, 2012.

[39] N. Braun and B. Datner. Analysis of roller/ball bearing vibrations. *Transactions of the ASME*, 1979.

[40] C. Farrar and K. Worden. *Structural Health Monitoring: A Machine Learning Perspective*. John Wiley & Sons, 2012.

[41] E. Figueiredo, G. Park, J. Figueiras, C. Farrar, and K. Worden. Structural health monitoring algorithm comparisons using standard data sets. *Los Alamos National Laboratory*, 2009.

[42] H. Ahmadi and P. Salami. Using of power spectral density for condition monitoring of fan. *Modern Applied Science*, 2010.

[43] P. Tse, Y. Peng, and R. Yam. Wavelet analysis and envelope detection for rolling element bearing fault diagnosis – their effectiveness and flexibilities. *Journal of Vibration and Acoustics*, 2001.

[44] I. Antoniadou, Manson G., W.J. Staszewski, T. Barszcz, and K. Worden. A time-frequency analysis approach for condition monitoring of a wind turbine gearbox under varying load conditions. *Mechanical Systems and Signal Processing*, 2015.

[45] A. Ziaja, I. Antoniadou, T. Barszcz, W.J. Staszewski, and K. Worden. Fault detection in rolling element bearings using wavelet-based variance analysis and novelty detection. *Journal of Vibration and Control*, 2016.

[46] W.J. Staszewski and A. Robertson. Time-frequency and time-scale analyses for structural health monitoring. *Phil. Trans. R. Soc. A*, 2007.

[47] B. D. Forrester and Aeronautical Research Laboratory (Australia). *Time-frequency domain analysis of helicopter transmission vibration.* Dept. of Defence, Defence Science and Technology Organisation, Aeronautical Research Laboratory Fishermans Bend, Vic, 1991.

[48] W.J. Staszewski, K. Worden, and G.R Tomlinson. Time-frequency analysis in gearbox fault detection using the wigner-ville distribution and pattern recognition. *Mechanical Systems and Signal Processing*, 1997.

[49] T.H. Park. *Introduction to digital signal processing – Computer Musically Speaking.* 01 2009.

[50] M. Feldman. Hilbert transform in vibration analysis. *Mechanical Systems and Signal Processing*, 25(3):735 – 802, 2011.

[51] C. R. Heiple and S. H. Carpenter. Acoustic emission produced by deformation of metals and alloys - A review. *Journal of Acoustic Emission*, 6:177–204, sep 1987.

[52] J. Kaiser. *Untersuchung über das Auftreten von Geräuschen beim Zugversuch.* PhD thesis, Technische Universität München München, 1950.

[53] A.W. Warren and Y.B. Guo. Acoustic emission monitoring for rolling contact fatigue of superfinished ground surfaces. *International Journal of Fatigue*, 29(4):603 – 614, 2007.

[54] A. Morhain and D. Mba. Bearing defect diagnosis and acoustic emission. *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, 217(4):257–272, 2003.

[55] C. James Li and S.Y. Li. Acoustic emission analysis for bearing condition monitoring. *Wear*, 185(1):67 – 74, 1995.

[56] B. Eftekharnejad, A. Addali, and D. Mba. Defect source location of a natural defect on a high speed-rolling element bearing with acoustic emission. *Annual Conference of the Prognostics and Health Management Society 2011*, 2(026), 2011.

[57] L.M. Rogers. The application of vibration signature analysis and acoustic emission source location to on-line condition monitoring of anti-friction bearings. *Tribology International*, 12(2):51 – 58, 1979.

[58] J.L. Rose. *Ultrasonic Guided Waves in Solid Media*. Cambridge University Press, 2014.

[59] C.B. Schaefer. Ultrasonics and their scientific and technical applications. 1938.

[60] I. Hirotsugu, k. Kikuo, N. Tomoaki, and S. Toshikazu. Time-frequency analysis of dispersive stress waves by means of wavelet transform : Identification of group velocity and application to ultrasonic materials evaluation. *Transactions of the Japan Society of Mechanical Engineers Series A*, 61(581):153–160, 1995.

[61] J. Jingpin, W. Bin, and H. Cunfu. Acoustic emission source location methods using mode and frequency analysis. *Structural Control and Health Monitoring*, 15(4):642–651, 2007.

[62] ASTM E976-15. *Standard Guide for Determining the Reproducibility of Acoustic Emission Sensor Response*. ASTM International, West Conshohocken, PA, 2015.

[63] M. W. Hawman and W. S. Galinaitis. Acoustic emission monitoring of rolling element bearings. In *IEEE 1988 Ultrasonics Symposium Proceedings.*, pages 885–889 vol.2, Oct 1988.

[64] C.C. Tan. Application of acoustic emission to the detection of bearing failures. *National Conference Publication - Institution of Engineers, Australia*, pages 110–114, 12 1990.

[65] C. James Li and S.Y. Li. Acoustic emission analysis for bearing condition monitoring. *Wear*, 185(1):67 – 74, 1995.

[66] F. Hemmati, M. Alqaradawi, and M.S. Gadala. Rolling element bearing fault diagnostics using acoustic emission technique and advanced signal processing. *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, 230(1):64–77, 2016.

[67] C. James Li. On-line detection of localized defects in bearings by pattern recognition analysis. *Journal of Engineering for Industry*, 111(4):331–336, 1989.

[68] A. Rojas and A. K. Nandi. Detection and classification of rolling-element bearing faults using support vector machines. In *2005 IEEE Workshop on Machine Learning for Signal Processing*, pages 153–158, Sep. 2005.

[69] H. Qingbo, Y. Ruqiang, K. Fanrang, and D. Ruxu. Machine condition monitoring using principal component representations. *Mechanical Systems and Signal Processing*, 23(2):446 – 466, 2009.

[70] M. Pirra, E. Gandino, A. Torri, L. Garibaldi, and J.M. Machorro-López. PCA algorithm for detection, localisation and evolution of damages in gearbox bearings. *Journal of Physics: Conference Series*, 305:012019, jul 2011.

[71] D. Kateris, D. Moshou, X-E. Pantazi, I. Gravalos, N. Sawalhi, and S. Loutridis. A machine learning approach for the condition monitoring of rotating machinery. *Journal of Mechanical Science and Technology*, 28(1):61–71, Jan 2014.

[72] A. Ziaja, I. Antoniadou, T. Barszcz, W.J Staszewski, and K. Worden. Fault detection in rolling element bearings using wavelet-based variance analysis and novelty detection. *Journal of Vibration and Control*, 22(2):396–411, 2016.

[73] K. Worden and W.J. Staszewski and J.J. Hensman. Natural computing for mechanical systems research: A tutorial overview. *Mechanical Systems and Signal Processing*, 25:4–111, 1 2011.

[74] A. Timofiejczuk, F. Chaari, R. Zimroz, W. Bartelmus, and M. Haddar. *Advances in Condition Monitoring of Machinery in Non-Stationary Operations: Proceedings of the 5th International Conference on Condition Monitoring of ... Poland*. Springer Publishing Company, Incorporated, 1st edition, 2017.

[75] NDT. AE source location techniques. `https://www.nde-ed.org/`. Accessed: 29-11-2018.

[76] R.K. Miller, E.K. Hill, and P.O. Moore. *Nondestructive Testing Handbook: Acoustic Emission Testing*. Nondestructive testing handbook. American Society for Nondestructive Testing, ASNT, 2005.

[77] A. Tobias. Acoustic-emission source location in two dimensions by an array of three sensors. *Non-Destructive Testing*, 9(1):9 – 12, 1976.

[78] M. Asty. Acoustic emission source location on a spherical or plane surface. *NDT International*, 11:223–226, 10 1978.

[79] P. Barat, P. Kalyanasundaram, and Baldev Raj. Acoustic emission source location on a cylindrical surface. *NDT & E International*, 26(6):295 – 297, 1993.

[80] A. Dragoş Axinte, D.R. Natarajan, and N.N.Z. Gindy. An approach to use an array of three acoustic emission sensors to locate uneven events in machining–part 1: method and validation. *International Journal of Machine Tools and Manufacture*, 45(14):1605 – 1613, 2005.

[81] T. Kundu, S. Das, and K.V. Jata. Point of impact prediction in isotropic and anisotropic plates from the acoustic emission data. *The Journal of the Acoustical Society of America*, 122(4):2057–2066, 2007.

[82] T. Kundu, S. Das, S.A. Martin, and K.V. Jata. Locating point of impact in anisotropic fiber reinforced composite plates. *Ultrasonics*, 48(3):193 – 201, 2008.

[83] T. Kundu, S. Das, and K.V. Jata. Detection of the point of impact on a stiffened plate by the acoustic emission technique. *Smart Materials and Structures*, 18(3):035006, 2009.

[84] T. Kundu, H. Nakatani, and N. Takeda. Acoustic source localization in anisotropic plates. *Ultrasonics*, 52(6):740 – 746, 2012.

[85] D. Aljets, A. Chong, and S. Wilcox. Acoustic emission source location in plate-like structures using a closely arranged triangular sensor array. 28, 01 2010.

[86] G.C. McLaskey, S.D. Glaser, and C.U. Grosse. Beamforming array techniques for acoustic emission monitoring of large concrete structures. *Journal of Sound and Vibration*, 329(12):2384 – 2394, 2010. Structural Health Monitoring Theory Meets Practice.

[87] M.G. Baxter, R. Pullin, K.M. Holford, and S.L. Evans. Delta t source location for acoustic emission. *Mechanical Systems and Signal Processing*, 21(3):1512 – 1520, 2007.

[88] M.J. Eaton, R. Pullin, and K.M. Holford. Acoustic emission source location in composite materials using delta t mapping. *Composites Part A: Applied Science and Manufacturing*, 43(6):856 – 863, 2012.

[89] L.M. Rogers. The application of vibration signature analysis and acoustic emission source location to on-line condition monitoring of anti-friction bearings. *Tribology International*, 12(2):51 – 58, 1979.

[90] J. Hensman, R. Mills, S.G. Pierce, K. Worden, and M. Eaton. Locating acoustic emission sources in complex structures using gaussian processes. *Mechanical Systems and Signal Processing*, 24(1):211 – 223, 2010.

[91] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[92] N.A. Barricelli. *Symbiogenetic Evolution Processes Realized by Artificial Methods*. 1957.

[93] S.K. Al-Jumaili, M.R. Pearson, K.M. Holford, M.J. Eaton, and R. Pullin. Acoustic emission source location in complex structures using full automatic

delta t mapping technique. *Mechanical Systems and Signal Processing*, 72-73:513 – 524, 2016.

[94] T. G. Overly, Park G., K. M Farinholt, and C. R. Farrar. Piezoelectric active-sensor diagnostics and validation using instantaneous baseline data. *IEEE Sensors Journal*, 9:1414–1421, 11 2009.

[95] M. I. Friswell and D. J. Inman. Sensor validation for smart structures. *Journal of Intelligent Material Systems and Structures.*, 10:973–982, 1999.

[96] G. Kerschen, P. De Boe, J.C. Golinval, and K. Worden. Sensor validation using principal component analysis. *Smart Mater. Struct*, 14:36–42, 02 2005.

[97] S. Bhalla and C.K. Soh. Structural health monitoring by piezo-impedance transducers. i: Modeling. *Journal of Aerospace Engineering*, 17(4):154–165, 2004.

[98] N. Saint-Pierre, Y. Jayet, I. Perrissin-Fabert, and J.C. Baboux. The influence of bonding defects on the electric impedance of piezoelectric embedded element. *Journal of Physics D: Applied Physics*, 29:2976, 01 1999.

[99] W.P. Mason. *Electromechanical Transducers and Wave Filters,*. Van Nostrand Company, 1948.

[100] V. Giurgiutiu and A.N. Zagrai. Embedded self-sensing piezoelectric active sensors for on-line structural identification. *ASME Journal of Vibration and Acoustics*, 124:116–125, 2002.

[101] T. Overly, G. Park, K. Farinholt, and C. Farrar. Piezoelectric active-sensor diagnostics and validation using instantaneous baseline data. *Sensors Journal, IEEE*, 9:1414–1421, 12 2009.

[102] G. Lanzara, Y. Yoon, Y. Kim, and F. Chang. Influence of interface degradation on the performance of piezoelectric actuators. *Journal of Intelligent Material Systems and Structures*, 20(14):1699–1710, 2009.

[103] G. Park, C.R. Farrar, F. Lanza di Scalea, and S. Coccia. Performance assessment and validation of piezoelectric active-sensors in structural health monitoring. *Smart Materials and Structures*, 15(6):1673, 2006.

[104] C.N. Sathyanarayana, A. Umesh, and S. Raja. Effect of sensor debonding on lamb wave propagation in plate structure. *ARPN Journal of Engineering and Applied Sciences*, 9:1358–1366, 01 2014.

[105] K. Worden, G. Manson, and D. Allman. Experimental validation of a structural health monitoring mothodology: Part i. novelty detection on a laboratory structure. *Journal of Sound and Vibration*, 259(2):323 – 343, 2003.

[106] K. Worden, G. Manson, and N.R.J. Fieller. Damage detection using outlier analysis. *Journal of Sound and Vibration*, 229(3):647 – 667, 2000.

[107] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[108] C.M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[109] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA, 1995.

[110] I. Nabney, C.M. Bishop, and C. Legleye. Modelling conditional probability distributions for periodic variables. pages 177 – 182, 07 1995.

[111] E. Baum and D. Haussler. What size net gives valid generalization? volume 1, pages 81–90, 11 1988.

[112] D.H. Johnson and D.E. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Simon & Schuster, Inc., New York, NY, USA, 1992.

[113] R.J. Mailloux. *Phased Array Antenna Handbook*. Artech House Publishers, Boston, MA, 1994.

[114] M. Wevers and M. Surgeon. 5.14 - acoustic emission and composites. In Anthony Kelly and Carl Zweben, editors, *Comprehensive Composite Materials*, pages 345 – 357. Pergamon, Oxford, 2000.

[115] B. Jalalahmadi, T. Slack, N. Raje, and N. Arakere. A review of rolling contact fatigue. *Journal of Tribology*, 131, 10 2009.

[116] B. Dolenc, P. Boskoski, and D. Juricic. Distributed bearing fault diagnosis based on vibration analysis. *Mechanical Systems and Signal Processing*, 66-67:521 – 532, 2016.

[117] J. Naumann. *Acoustic Emission Monitoring of Wind Turbine Bearings*. PhD thesis, University of Sheffield, 2016.

[118] Z. Budrovic, H. Van Swygenhoven, P.M. Derlet, S. Van Petegem, and B. Schmitt. Plastic deformation with reversible peak broadening in nanocrystalline nickel. 304(5668):273–276, 2004.

[119] Y. Sun, H. Choo, P.K. Liaw, Y. Lu, B. Yang, D.W. Brown, and M.A.M. Bourke. Neutron diffraction studies on lattice strain evolution around a crack-tip during tensile loading and unloading cycles. *Scripta Materialia*, 53(8):971 – 975, 2005.

[120] E-W. Huang, R.I. Barabash, Y. Wang, B. Clausen, L. Li, P.K. Liaw, G.E. Ice, Y. Ren, H. Choo, L.M. Pike, and D.L. Klarstrom. Plastic behavior of a nickel-based alloy under monotonic-tension and low-cycle-fatigue loading. *International Journal of Plasticity*, 24(8):1440 – 1456, 2008.

[121] ASM International. *Atlas of Stress-strain Curves*. ASM International, 2002.

[122] M. Smith. *ABAQUS/Standard User's Manual, Version 6.14*. Dassault Systèmes Simulia Corp, United States, 2014.

[123] B. Bhushan and P. Lim Ko. Introduction to tribology. *Applied Mechanics Reviews*, 56:B6, 01 2003.

[124] S.R. Agnew, D.W. Brown, and C.N. Tomé. Validating a polycrystal model for the elastoplastic response of magnesium alloy AZ31 using in situ neutron diffraction. *Acta Materialia*, 54(18):4841 – 4852, 2006.

[125] B. Jacobson. The Stribeck memorial lecture. *Tribology International*, 36:781–789, 11 2003.

[126] N. Tandon and B.C. Nakra. Comparison of vibration and acoustic measurement techniques for the condition monitoring of rolling element bearings. *Tribology International*, 25(3):205 – 212, 1992.

[127] P. Wickramarachi. Effects of windowing on the spectral content of a signal. *Sound and Vibration*, 37:10–11, 01 2003.

[128] H. Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1998.

[129] R. Sleeman and Torild van Eck. *Single station real-time P and S phase pickers for seismic observatories*, pages 173–194. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[130] K.M. Holford, M.J. Eaton, J.J. Hensman, R. Pullin, S.L. Evans, N. Dervilis, and K. Worden. A new methodology for automating acoustic emission detection of metallic fatigue fractures in highly demanding aerospace environments: An overview. *Progress in Aerospace Sciences*, 90:1 – 11, 2017.

[131] L. Tarassenko. 7 - design, training and testing of the prototype. In L. Tarassenko, editor, *Guide to Neural Computing Applications*, pages 77 – 97. Butterworth-Heinemann, New York, 1998.

[132] C.M. Bishop. Neural Networks: A Pattern Recognition Perspective. Technical report, Aston University, Birmingham, UK, 1996.

[133] R.B. Randall, J. Antoni, and S. Chobsaard. The relationship between spectral correlation and envelope analysis in the diagnostics of bearing faults and other cyclostationary machine signals. *Mechanical Systems and Signal Processing*, 15(5):945 – 962, 2001.

[134] A. Birks, R. Green, and P. McIntire. *Nondestructive Testing Handbook, Third Edition: Volume 7, Ultrasonic Testing (UT)*, volume 7. ASNT, 1991.

[135] C. Liang, F.P. Sun, and C.A. Rogers. Coupled electro-mechanical analysis of adaptive material systems – determination of the actuator power consumption and system energy transfer. *Journal of Intelligent Material Systems and Structures*, 5(1):12–20, 1994.

[136] D.M. Peairs, G. Park, and D.J. Inman. Improving accessibility of the impedance-based structural health monitoring method. *Journal of Intelligent Material Systems and Structures*, 15(2):129–139, 2004.

[137] G. Habenicht. *Important Reactive Adhesives*, chapter 4, pages 23–43. John Wiley & Sons, Ltd, 2008.

[138] J.H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.