Article:

Fuentes, R., Dwyer-Joyce, R.S. orcid.org/0000-0001-8481-2708, Marshall, M.B. orcid.org/0000-0003-3038-4626 et al. (2 more authors) (2020) Detection of sub-surface damage in wind turbine bearings using acoustic emissions and probabilistic modelling. Renewable Energy, 147 (1). pp. 776-797. ISSN 0960-1481

https://doi.org/10.1016/j.renene.2019.08.019

Detection of sub-surface damage in wind turbine bearings using acoustic emissions and probabilistic modelling

R. Fuentes[1], R. Dwyer-Joyce[2], M.B Marshall[2], J. Wheals[3] , E. J. Cross[1]

Dynamics Research Group, Department of Mechanical Engineering, The University of Sheffield[1]

Leonardo Centre for Tribology, Department of Mechanical Engineering, The University of Sheffield[2]

Ricardo Innovations ltd[3]

Abstract

Bearings are the culprit of a large quantity of Wind Turbine (WT) gearbox failures and account for a high percentage of the total of global WT downtime. Damage within rolling element bearings have been shown to initiate beneath the surface which defies detection by conventional vibration monitoring as the geometry of the rolling surface is unaltered. However, once bearing damage reaches the surface, it generates spalling and quickly drives the deterioration of the entire gearbox through the introduction of debris into the oil system. There is a pressing need for performing damage detection before damage reaches the bearing surface. This paper presents a methodology for detecting *sub-surface* damage using Acoustic Emission (AE) measurements. AE measurements are well known for their sensitivity to incipient damage. However, the background noise and operational variations within a bearing necessitate the use of a principled statistical procedure for damage detection. This is addressed here through the use of probabilistic modelling, more specifically Gaussian mixture models. The methodology is validated using a full-scale rig of a WT bearing. The bearings are seeded with sub-surface and early-stage surface defects in order to provide a comparison of the detectability at each level of a fault progression.

Keywords: Acoustic Emission, Condition Monitoring, Bearings, Damage Detection, Probabilistic Modelling, Wind Turbines

1. Introduction

Bearing failures are the leading source of downtime in Wind Turbine (WT) gearboxes and the root cause for this is attributed to Rolling Contact Fatigue (RCF) [1–3]. In the majority of loading conditions, fatigue damage begins its life at the *surface* of materials, where high stresses and imperfections due to manufacturing and surface wear coalesce and lead to crack initiation. The case in bearings is unlike typical fatigue damage. Hertzian contact mechanics dictates that, under the compressive load at the contact between a rolling element and a bearing, the location of maximum stress will lie a small distance under the surface at the point of contact between a roller and the bearing surface. This has some important consequences regarding the damage progression of a bearing. A growing crack will spend most of its time under the surface, where it has minimal impact on the operation of the rest of the system. However, once a crack emerges on the surface, the progression of failure is accelerated through contact with the rolling elements and this will generate spalling. At the point of initiation of spalling, the progression of damage is quick as debris is introduced into the rest of the mechanical system, thus accelerating the overall failure of the gearbox.

Currently, WTs are designed with an overall target lifetime of 20 years [4], a design requirement which extends to all of their subsystems. However, the average service life of wind turbine gearboxes often falls much below the 20 year target [5]. This is a problem; even though gearboxes are not the most unreliable subsystem, they do cause the most downtime [5]. Minimising gearbox failures is thus a key element in increasing overall wind turbine productivity [2]. Because bearing surface damage releases debris into closed-loop oil systems, sampling the oil quality and checking for debris within the oil system is, to date, still used as a reliable technique for diagnosing the overall condition of WT gearboxes [6, 7]. It is also at this point that vibration-based monitoring systems are able to detect the presence of defects. The fact that bearing damage has reached the surface and introduced debris into the oil system motivates the need for detecting fatigue cracks in bearings before they reach this stage, so that preventive maintenance can be carried out and impact to the rest of the gearbox can be minimised. Detecting *subsurface* damage at the incipient stage has been identified as a critical aspect of wind turbine condition monitoring [3]. Rolling contact fatigue is exacerbated in planetary gearboxes, where the bearing raceway is loaded exclusively in the torque direction. This exerts a compressive load on the same point along the circumference of the raceway, as illustrated in Figure 1. In order to avoid the problems associated with the introduction of debris and accelerated failure, it is highly desirable to be able to detect the *incipient* failure of the bearing, at the point where a fatigue crack has just initiated. There are three critical aspects that will determine the outcome of a damage detection system [8]: 1) the physical sensing system, 2) the damage-sensitive features extracted from the data and 3) the damage identification strategy applied to those features. This paper addresses these problems. As for the physical sensing system, Acoustic Emissions (AE) are proposed as a measurement strategy. The damage-sensitive features extracted from AE data play a fundamental role in the ability to identify damage. In this paper, the state of the art of AE features are reviewed and compared and new features are proposed using advanced signal processing tools. Lastly, a rigorous damage identification strategy is proposed that addresses the key challenge of discerning operational and environmental effects from the damage-sensitive features. This is carried from a probabilistic modelling point of view, using Gaussian mixture models in combination with dimensionality reduction tools.

1.1. Subsurface cracks

The interest in subsurface cracks has grown since the realisation that fatigue cracks in gearbox bearings tend to start around non-metallic inclusions [10], introduced during the manufacturing process. The presence of these inclusions, coupled with high stress concentrations under the surface, leads to the development of fatigue cracks, often referred to as White Etching Cracks (WEC), White Structure Flaking (WSF) [11, 12], or simply "butterfly" cracks due to their butterfly shape (with the "wings" following a path from the inclusion, out towards the surface). These cracks tend to grow in the region around 1mm under the contact surface of typical WT bearings [13] and it has been proposed that their formation is driven both by chemical and mechanical processes. Chemically, it is the diffusion and release of hydrogen into bearing steel [14], through lubrication and water ingress that drives the formation of WECs. Mechanically, over-

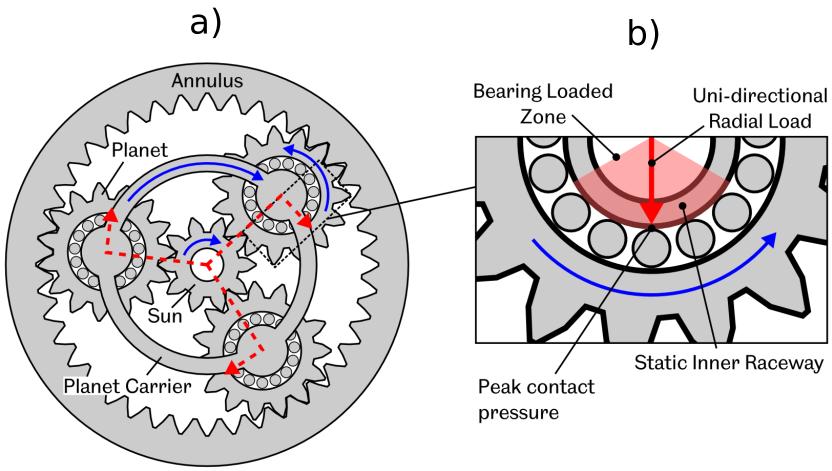

Figure 1: a) Diagram showing gearing setup for a planetary gearbox. Note the planet bearings are constantly loaded in the torque direction, indicated by the red arrows. b) zoom-in to one of the planet bearings, highlighting the loaded zone [9]

load events arising from wind gusts, breaking and torque reversals drive stress concentrations around non-metallic inclusions to yield point and lead to the formation and growth of WECs. Since the realisation that inclusions in bearing steel directly lead to subsurface cracks, the quality control of the manufacturing processes has dramatically improved. However, inclusions will always be present even in today's high standard of steels. In fact, it has been shown that it is typically the smallest inclusions that lead to the greatest stress concentrations and therefore the development of WECs [12]. An example of a WEC at the initial stage of propagation is shown in Figure 2, observed on a WT bearing section [12].

1.2. Damage detection with Acoustic Emissions

When considering the dynamic response of a system, it is a generally well-accepted principle that the physical size of damage is inversely proportional to the frequency at which its effects will be manifested in its dynamic response [15–17]. Furthermore, there is a well established relationship between the AE response of a metal and fatigue crack growth [18]. With this in mind, subsurface damage on bearings represents the smaller end of the scale, requiring relatively high frequency measurements, when com-

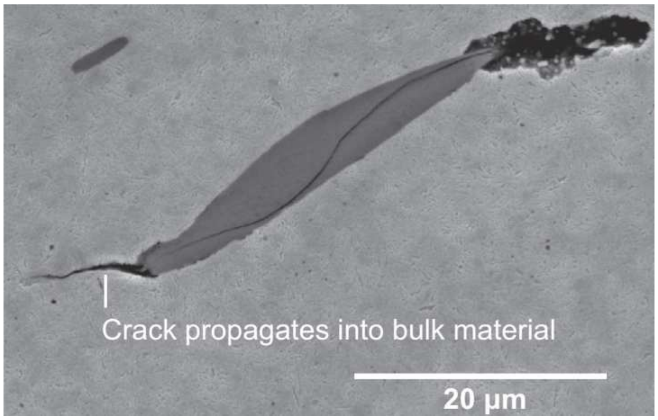

Figure 2: Example of a WEC propagating around a non-metallic inclusion [12]

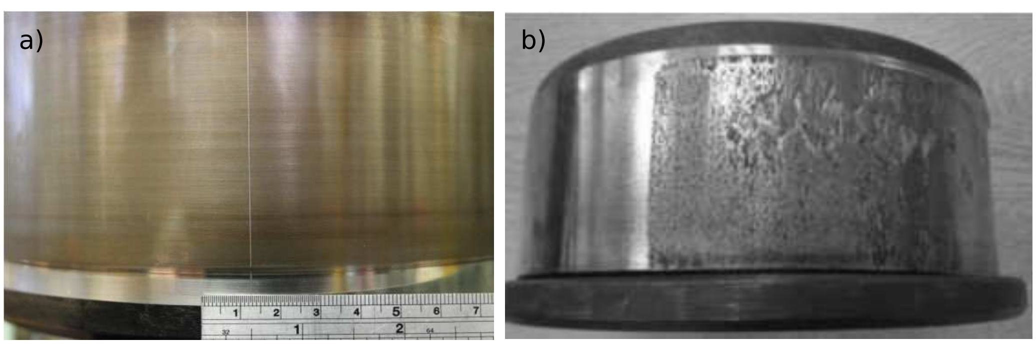

Figure 3: Illustration of damage in representative WT planetary gearbox bearings showing a) a line etch similar, used in [9] which is of similar form to the surface damage introduced in this study (see Figure 6 for the actual profiles) and b) damage arising from real operational conditions, focused on the point at which spalling occurs [12].

pared to traditional vibration-based monitoring, in order stand a reasonable chance of being detected. Following this reasoning, this paper investigates the detectability of subsurface cracks using Acoustic Emission (AE) measurements.

Acoustic Emissions (AE), when used within a Structural Health Monitoring (SHM) or Non-Destructive Testing (NDT) context, are high frequency stress waves that propagate through a material. These waves can be generated by a number of different mechanisms including stress, plastic deformation, friction and corrosion. AE occurs as a result of the release of elastic energy by any one of these mechanisms, which is then propagated through the material as elastic waves. The concept of AE in engineering structures is analogous to the release and propagation of energy that takes place during earthquakes [19], as a result of fractures within fault planes. It is generally accepted that the movement of slip-planes characteristic of micro-cracks during the application of stress and yield [20, 21] leads to the generation of AE. Features extracted from AE measurements have been shown to be successful indicators of the early onset of cracks in various applications [18, 22, 23]. AE testing is a passive method, in the sense that one is listening to the acoustic response of the material when mechanical stress is applied to it. Most materials will have a certain level of AE activity even in an undamaged state when stress is applied to them, the technical term for this is the Kaiser effect [24]. However, when defects such as cracks or spalling are present in the material, the AE response when stress is applied will tend to be more frequent, of higher amplitude, and may have different spectral characteristics depending on the material properties of the medium where the waves propagate. Friction processes also tend to generate AE, as the impact between micro-asperities that occurs during contact of two materials releases elastic energy into the system [25]. There is no clear definition of the frequency range that constitutes an AE measurement. This will depend on the physics of the particular defect; a typical AE stress wave generated from the initiation of a crack in steel can range from 50kHz to 2MHz [21, 26].

AE being now a popular measurement technique, its application to bearing monitoring hasn't gone without attention, but the problem of detecting small subsurface cracks is currently far from solved. One of the barriers to investigating this problem is that subsurface damage is hard to find, validate and measure in an operational bear-

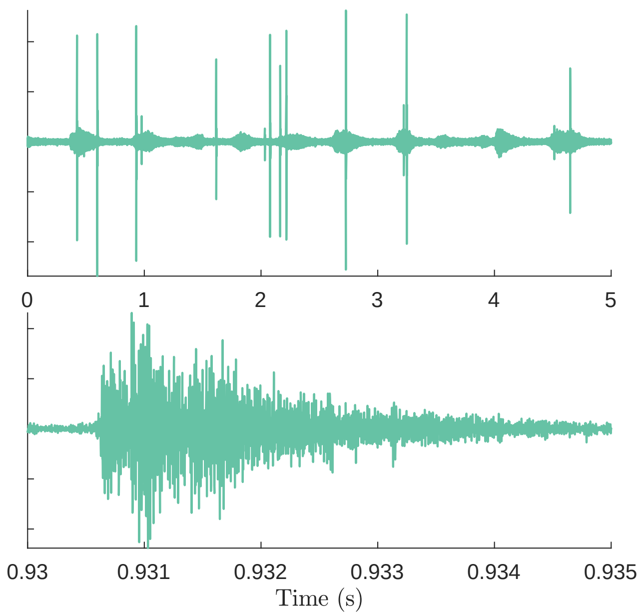

Figure 4: Example of a) an AE signal measured from an undamaged bearing in operation and b) a zoom-in to one of the “hits” characteristic of AE measurements.

ing and much more so within a wind turbine, so all investigations resort to laboratory experiments under controlled environments. The major problem with this lies in the defect. Most investigations (including some by these authors) use artificially seeded defects at the surface to show that a given methodology is able to detect incipient damage [26–36]. An example of such defect is illustrated in Figure 3a. The problem with surface defects (from an investigative point of view) is that they represent the later stages of damage, quickly lead to spalling and are also relatively easy to detect as contact between rollers and surface damage releases large amounts of acoustic energy into the system. An example of a late damage state is shown in Figure 3b, at which point detection through AE or even vibration becomes trivial with the current state of research. A number of studies have attempted to tackle this issue and offer investigations on bearings run from undamaged conditions all the way to failure, usually under an environment that accelerates failure [37–39]. While these investigations significantly advanced the general field of AE-based monitoring, they have not been carried out with statistical rigour. More specifically, the detectability of incipient damage is only viewed qualitatively and no attempt is made at quantifying it.

Detecting damage using data-driven methods can be cast as a *novelty detection* problem. First a statistical model of the system in its undamaged state is characterised and an alarm threshold is established. Further observations are judged based on a *novelty index*; a distance metric that quantifies how far the new observation lies from the baseline undamaged state. The decision of whether an observation belongs to a damaged or undamaged state is then given by whether the novelty index lies above or below the threshold. Under this approach the detectability is judged solely in terms of change from a baseline condition and can be quantified in terms of false positives and false negatives. It also provides ample opportunities for objectively judging different statistical and signal processing methods in terms of their false positive and negative rate performance. A central part of this paper involves performing this comparison using rigorous methods of novelty detection.

A critical aspect of any damage detection system is its ability to separate the effects of damage from those of normal gearbox operation. This is particularly true of AE measurements. The recorded AE signals will not only contain the high frequency stress waves produced by crack growth and plastic strain, they will also contain a significant amount of AE energy that is unrelated with any damage mechanism. A large part of this "benign" energy being released as AE in a gearbox will come from internal stresses and friction. A key challenge is then to separate the AE activity related to normal operation from those that are not. In WT gearboxes, operational variability will naturally arise from varying wind speeds. In variable-speed WTs, changes in speed will result in a wide variety of AE activity, with or without the presence of defects. On a basic level, varying rotational speed will introduce variability into any features derived from a frequency domain or periodicity based analysis of the AE signals, such as those presented in [40, 41]. Varying speed will also introduce changes to the background noise resulting from friction throughout the gearbox and will also lead to changes in oil temperature. Oil viscosity is highly dependent on temperature and this completely changes the lubrication regime of the gearbox, which has a direct effect on the resulting background AE noise signature [42]. AE monitoring of fixed-speed WTs will also suffer from this type of variability. Varying wind speed introduces changes in the internal loads of the gearbox, and the release of energy as AE in materials is directly related to applied loads. Furthermore, load variation also causes variation in AE energy through changing temperatures and the effect that this has on the lubrication regime. Higher temperatures generally lead to lower oil viscosities, which in turn drives an increase in friction through higher asperity contact between the surfaces of roller elements and bearing raceways.

In general, studies investigating bearing monitoring, using either vibration or AE measurements, have not taken into account this operational and environmental variability. Recently steps have been taken in [26] to collect AE measurements on a failing bearing at varying loads and rotational speeds, consequently, the notion that load and speed affect AE both directly, and indirectly, through variations in temperature was confirmed. However, again no attempt was made at quantifying the detectability in terms of comparing a statistical model of the undamaged bearing against the rest of the observations.

A further complicating factor when performing bearing monitoring in practice is detecting incipient faults from practical measurement locations. Since the early days of bearing monitoring it was recognised that detecting even a large seeded fault was much easier if AE is measured in the inner raceway of the bearing compared to the outer raceway [27]. However, in practice it is clearly less invasive to place a sensor on the outside of the gearbox. Placing a sensor inside a gearbox would often require machining and this can lead to overall reduced gearbox reliability.

In summary, the current challenges facing bearing monitoring are clear:

- To detect incipient failure at the subsurface stage using a rigorous statistical methodology.
- To perform such detection under the effects of operational and environmental variability.
- To use practical and non-invasive measurement locations to carry out the monitoring

This paper presents a general methodology for detecting damage in rotational components and validates this using a rig that is representative of the operational environment of a WT epicyclic gearbox. More specifically, this environment involves realistic varying loads and speeds as well as the changing temperature and oil properties that these create. A focus is placed on detecting subsurface as well as small surface cracks. The approach to inducing subsurface damage used here is different from previous studies; instead of taking an undamaged bearing and running it until failure, subsurface damage was carefully seeded in a bearing in order to be able to carry out a robust and conclusive comparison of damaged and undamaged states. Furthermore, three more surface faults were seeded in a different bearing in increasing sizes ranging from 5μm to 50μm width. The reasoning for using seeded defects here is that it is the most robust and conclusive way of validating a damage detection methodology because one knows the exact state of the bearing at every stage. However, a strong focus was placed on using defects, both subsurface and surface-level, that are representative of the very early stages of fatigue damage in bearings. A description of the experimental rig and the defect-seeding procedures is given in Section 2. Particular attention is devoted to the damage-sensitive feature extraction process of AE signals. In broad terms, there are two major and different types of features extracted from AE signals 1) those based on the characteristics of discrete bursts of energy, often termed AE *hits* and 2) those based on analysis of the periodicity of the global AE signal. While in the literature, studies tend to focus on either one approach or the other, here it is of interest to investigate the performance of each type of damage-sensitive feature. A description of the signal processing methods used to derive the different damage-sensitive features is given in Section 3. The third element of the damage detection methodology is the statistical pattern recognition. The methodology used in this paper is to treat the problem as one of novelty detection, where the probability distribution of the damage-sensitive features is modelled for data belonging to an undamaged state. This allows for the computation of a *novelty index* on subsequent observations, and when this exceeds a given alarm threshold this can be indicative of damage. The particular procedure, based on Gaussian mixture modelling is described in detail in Section 5.3. The validation of the entire methodology, including the various damage sensitive features and the novelty detection is presented in Section 5, where results are presented for the data-set collected on the experimental rig with the four different stages of damage.

2. Experimental set-up

The experimental rig used in this study was devised to investigate planetary gearbox bearings, due to their propensity to fail prematurely owing to the fact that the load transferred from the rotating outer raceway to the static inner raceway peaks at a fixed position along the circumference of the inner raceway. This loading condition is illustrated in Figure 1 and leads to short fatigue lifetimes around the loaded section of the bearing, leaving an "un-used" fatigue life outside of this region. This rig has been used in previous investigations into bearing monitoring using various techniques ranging from vibration [43] to AE [35, 36] and ultrasound monitoring [9]. However, this is the first study considering subsurface defects, as well as surface defects with widths under 100μm.

The objective of the rig is to generate a compressive radial load on the inner raceway inside a planetary sun bearing sub-assembly. In order to achieve this, the rig comprises of two bearings: an inner "test bearing", which is housed inside an outer "main bearing". The inner test bearing then houses a stationary shaft, which is connected via two steel lugs to a hydraulic ram, capable of delivering a total compressive load up to 1600kN. In order to apply rotation, the inner raceway of the outer bearing is coupled to the outer raceway of the inner bearing. A tensioned pulley is then used to drive these two raceways together, with power delivered from an electric motor. A cross-sectional diagram of the assembly is shown in Figure 7a, while Figure 7b shows a photograph of the entire rig, highlighting the main components. Figure 8 shows a more schematic view of the main components and applied loads. The inner bearing is coloured in red and the outer bearing is coloured in blue. The rolling elements are shown in light grey. The main interest in this investigation is the inner raceway of the inner bearing. Together with the shaft, this inner raceway remains stationary, with the rolling elements revolving around it, a compressive load being applied at the bottom (via the hydraulic ram). Due to the constant compressive load, it is here where damage normally initiates in planetary sun bearings and so all seeded defects in this study are located so that their position lies exactly at the bottom of the circumference of this inner raceway.

The inner bearing used in this rig is an NU2244, which is typically used in WT gearboxes. Its inner raceway has a bore diameter of 220mm, and the outer raceway has an outer diameter of 400mm. Its maximum dynamic load rating is 1600kN, while its fatigue load limit is of 250kN. Further specifications of this bearing can be found in [44].

2.1. Seeded Defects

This section details the defects seeded into the inner raceway. For the purposes of this study, two types of defects were seeded in order to emulate increasing damage levels: subsurface and surface defects. Overall, a total of six bearing conditions were examined, summarised in Table 2.1. Note that *two* undamaged bearings were used in the experiment, in order to generate robust training and validation datasets for the data-driven damage detection models. Further discussion on the importance of having two undamaged bearings, for validation purposes is given in Section 5

Label	**Condition**	**Severity**	**Bearing**
UD1	Undamaged		A
UD2	Undamaged		B
D1	Subsurface Damage	800kN	C
D2	Subsurface Damage	1000kN	C
D3	Surface Damage	5μm	D
D4	Surface Damage	20μm	D
D5	Surface Damage	50μm	D

Table 1: Summary of bearing conditions

2.1.1. Surface defects

Surface defects represent a damage condition in a relatively advanced stage. The objective here was to generate defects as small as possible in order to emulate the early stages of surface damage. In previous work at the University of Sheffield [36], a spark erosion technique was used to etch the surface of a raceway to emulate a surface crack, which generated surface defects of approximately 200μm width (similar to the damage shown in Figure 3a). In order to achieve a smaller defect, more representative of the early stages of a surface crack, a Cubic Boron Nitrite (CBN) grit was used to scratch the surface using a six-axis Computer Numerical Control (CNC) machine. This was performed at three different pressures, with each one at a different angular position on the raceway. The aim of using three different angular locations along the raceway was to be able to perform a test with three different defect sizes by simply positioning the different defects on the loaded zone of the bearing. The angular positions of the defects were such that only one defect was loaded at any given instant in time, these are shown in Figure 5. The target sizes for the seeded defects were 5μm, 20μm and 50μm, although the actual profiles obtained are shown in Figure 6. These profiles were taken by first filling the scratches with silicon in order to extract a negative of the profile, and then measuring this with a three-dimensional optical profiler. Note that damage conditions D4 and D5 seem very similar to each other. In fact, most of the profile shown in Figure 6a and b is the pattern of material removed from the inside of etch. The profile of the inner parts of the etches were in fact too small to measure with the available profilers. Note that the curvature of background of each profile shown in Figure 6 is due to the bending of the silicon samples and not the curvature of the bearing.

2.1.2. *Subsurface defects*

A key element of this paper is the study of the detectability of defects in a bearing before they propagate to the surface. To achieve this, a subsurface defect was seeded to a second raceway by means of compressing its outer surface with a rolling element. The compression was applied using a hydraulic press capable of applying up to 2000kN. Subsurface yield was estimated to occur at 1000kN for this bearing, using Hertzian contact mechanics relationships. To ensure that subsurface yield occurred, while also preventing the damage propagating to the surface, the yield process was monitored using AE. Some of the observations on AE from this damage seeding are further discussed in [21].

During the seeding of subsurface damage, a large increase in AE energy was ob-

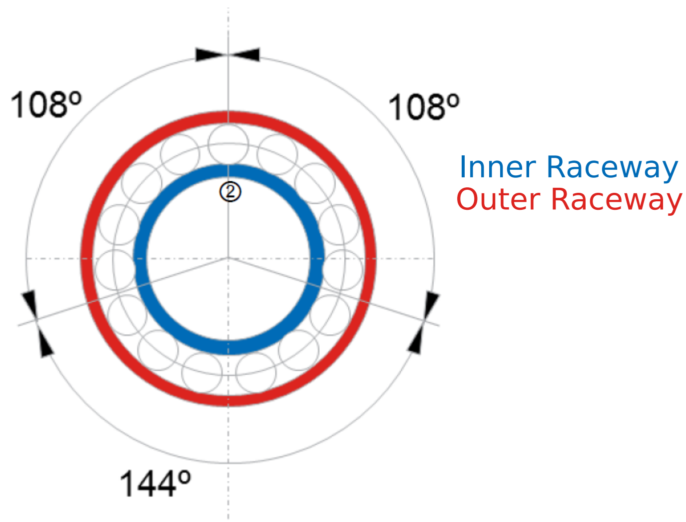

Figure 5: Angular positions of defects along raceway. Note that this angular separation ensures that when one roller passes over one of the defects, none of the rest of the rollers will pass over the other two defects at the same time.

served in the 800kN to 1200kN range. Visible surface damage was only found when a bearing was loaded beyond 1700kN. Although several tests were carried out on numerous raceways, on the final raceway, faults were seeded using two compression levels, at 800kN and 1000kN. These were applied on the same circumferential indices as for the surface damage, so that only one damage site is located within the loaded zone of the raceway.

To summarise, two damaged bearing raceways were used for testing. One raceway contained three surface etches with increasing sizes, to emulate increasing levels of damage. The second raceway contained two seeded subsurface cracks, with increasing levels of maximum compressive load.

2.2. Test conditions

The objective of this study is to perform damage detection of realistic defects in a realistic operational WT gearbox environment. In order to achieve this, a test schedule was designed to capture, for each of the raceway conditions (outlined in Table 2.1), the effects of varying load, speed and temperature. Preliminary tests were conducted,

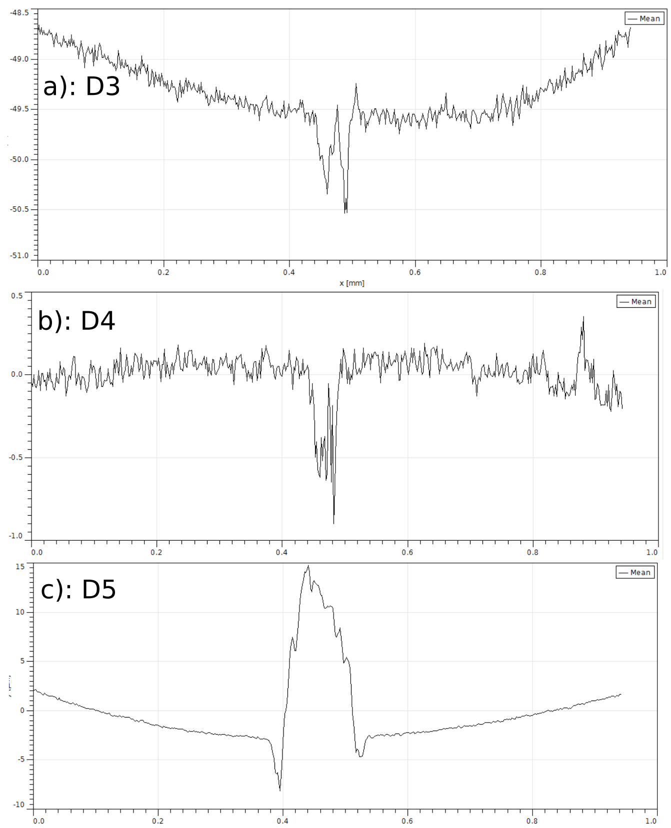

Figure 6: Average (negative) profiles of the three surface damage sites, taken using a 3D optical interferometer profiler, according to Table 2.1. Note that the profiles of the smallest defects only capture te average width, together with excess material on the sides.

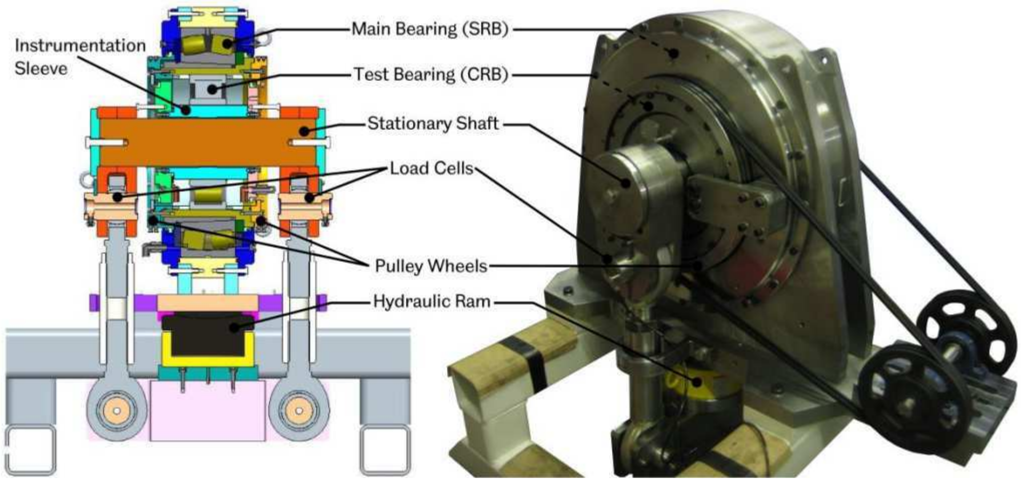

Figure 7: a) Diagram showing fixture, rig bearing and test bearing, outlining the location of AE measurement channels. b) Photograph of rig in the lab.

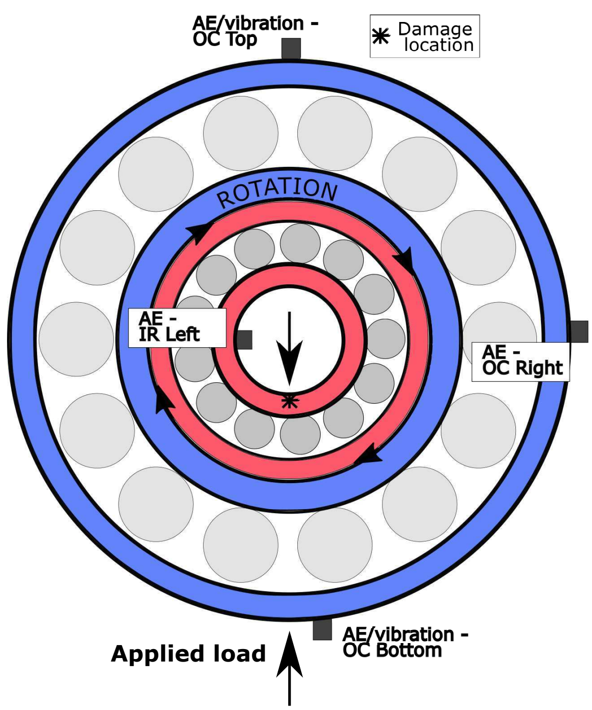

Figure 8: a) Diagram showing fixture, rig bearing and test bearing, outlining the location of AE measurement channels. b) Photograph of rig in the lab.

stepping the compressive load at 100kN steps from 0 to 1200kN. This pointed to three major regimes of AE activity, around the low load (0-400kN), medium load (400kN-800kN) and high loads (800kN-1200kN). For this reason three loads were selected for a test schedule: 200kN, 600kN and 1000kN. The temperature of the rig proved difficult to control precisely. The factors that affect the rig and oil temperatures are the operating load, bearing condition (a failed bearing introduces debris into the system and drives the temperature up through friction), ambient temperature, accrued usage time, whether a heat exchanger is present in the oil system and whether this is aided by a cooling device (such as a fan). In this rig the only means of controlling the temperature directly are via the heat exchanger, the operating load and sequencing of the applied loads. In general it is easier to warm up the rig, than to cool it down, as once it runs and a load is applied, it will quickly warm up and reach a stable temperature. Therefore, it was decided to split the tests into low and high temperatures. This split is reasonable, given that the main effect that temperature introduces (to the AE activity) is an increase in friction at higher temperatures from a reduction of viscosity [45, 46]. In order to keep the temperature down, the low temperature runs were performed early in the morning, testing low loads first and keeping cooling fan on.

Table 2.2 shows the complete schedule of tests carried out. This is also a realistic scenario given that the temperature in a WT gearbox will vary in a similar fashion. When not operational, or at low wind conditions, modern gearboxes will keep circulating the oil through a heat exchanger, to keep it at a stable temperature (and thus the optimal viscosity).

2.3. Instrumentation

Four AE sensors and one accelerometer were used to measure the overall dynamic response of the bearing throughout the tests. The positions are illustrated in Figure 8. Three AE sensors were placed outside the bearing, mounted on the Outer Casing (OC) of the outer raceway. One AE sensor was placed inside the Inner Raceway (IR), by machining a small part of the stationary shaft in order to fit the sensor. Furthermore, the AE sensor located in the IR was sprung-loaded into position. The IR AE sensor was located along the circumference of the IR, at angular orientation of 60deg from

Speed (RPM)	**Load (kN)**	**Temperature**
100	200	Low (Fan ON)
100	600	
100	1000	
100	1000	High (Fan OFF)
100	200	
100	600	

Table 2: Bearing test schedule. Each row was performed sequentially, and this schedule was used for every bearing condition.

the damaged zones (the bottom). In these tests, the IR sensor is the closest to the damage, which is the main AE source of interest. Because of the compressive load applied to the inner raceway, the bottom is in direct contact with the rollers, while the top develops a slight clearance. This defines the acoustic path of the AE stress waves to propagate from the source (the damage), down through the rollers, outer bearing and casing and around the circumference of the rig. Therefore, of the three sensors on the Outer Casing, the OC bottom location is closest to the AE source, followed by the OC-right and OC-top (as illustrated in Figure 8). Note that due to symmetry, it was deemed reasonable to not include a measurement position on the OC left side.

The three OC-AE sensors used in this study were Mistras 3MICRO-30D, fitted with a differential cable for noise reduction. The IR sensor was a Mistras NANO-30, which is a non-differential sensor. The Micro30D has a marked resonance at approximately 350kHz, while the Nano30 has a flatter response in the range of 200kHz-500kHz. This is relevant as the sensor frequency response shapes the acquired signals significantly. Compared with vibration sensors, the frequencies of interest are much broader, in the range of 50kHz - 2MHz. It is therefore much harder to achieve a flat frequency response across all frequencies of interest, and so one must accept the significant mechanical filtering that the sensor applies to the "true" underlying AE stress wave. It must also be noted that sensor-to-sensor variability is usually much more significant in AE sensors, compared with vibration instrumentation. One tri-axial accelerometer was mounted next to the AE sensor at the OC-top location.

All data acquisition was recorded with a National Instruments (NI) C-DAQ system. This comprised of several modules, with data recorded at the following sample rates:

- NI-9223, four-channel analogue module sampling at 1MHz with 16-bit Analogue to Digital (ADC) conversion.
- NI-9234, three-channel IEPE module sampling at 51.2kHz.
- NI-9213, thermocouple, for temperature measurements (one sample per test).

Several operational parameters were also acquired in order to assess the influence of each one on the AE response. These parameters were:

- Test Bearing Speed (RPM).
- Left-side and Right-side Load (kN).
- Oil Temperature.
- Casing Temperature.

Data was acquired in "trials" of ten second duration. For each bearing condition in Table 2.1 and each operational condition in Table 2.2, ten different different trials were gathered. The resulting data-set thus comprises approximately 1260 time series records, each taking approximately 322MB of memory, totalling 491GB of data.

3. Damage-sensitive features from AE measurements

The goal of this section is to discuss the signal processing required to derive damage sensitive features from AE data. In broad terms, there are two approaches to this problem and both will be covered here. The fundamental idea behind AE monitoring is to detect the energy release characteristic of the interaction between stress and the plastic deformation around a crack in the form of high frequency stress waves. These short bursts of high frequency data, often referred to as "hits" have been illustrated in Figure 4. The first approach is to identify these hits and to characterise them. On a non-rotating system with low levels of background noise, the simple presence of a hit could be indicative of damage. In this setting, simply characterising the rate of generation of AE hits, or "hit count", would be a sufficient damage-sensitive quantity. However, in rotating systems a certain amount of background AE activity will be expected, as already discussed in Section 1.2. In this setting, it is more desirable to work with damage-sensitive features of individual AE hits. These could be as simple quantities such as the energy, duration and amplitude of hits, but could also be based on more complex models, such as Fourier or autoregressive coefficients. Detecting an AE hit and computing features from these short bursts of energy represents one of the two major different strategies for signal processing. This strategy will be discussed in more detail in Section 3.1. One key advantage of dynamic response data originating from rotating machinery is its tendency to be periodic, and this can be taken advantage of for efficiently deriving damage sensitive features. In terms of AE, one would expect a burst of AE energy to occur every time one roller passes over a damaged area. This information can be encoded efficiently with Fourier coefficients of rectified signal envelopes. This type of periodicity-based analysis represents the second major approach to deriving damage-sensitive features, and is discussed in more detail in Section 3.4.

3.1. Hit-based features

One of the key points of AE data, from a signal processing point of view, is that the bursts captured by the AE acquisition system are very short in comparison to the large amount of time that needs to be spent monitoring. Because of the high sample rates required to capture these high frequency waves, this means that a lot of noise is recorded, in comparison to the amount of useful AE bursts. To put this in context, the bursts recorded from a yielding steel specimen may last in the order of 2000 μs. If one were to monitor at 1MHz for 1 second and expect 15 bursts (which is roughly how many bursts are expected in the rig in this paper at 100 RPM), this would mean that approximately 3% of the data points are informative and the rest is noise. Given the high sample rate, data storage and handling becomes an issue if one wishes to monitor for long periods of time. This has led the AE monitoring community to develop hit-identification strategies, where an AE hit is defined as a burst large enough for it to be likely to be caused by material fracture. Non-rotating systems are naturally "quiet" in their undamaged states and it is normally straightforward to identify hits by setting a threshold on the overall AE signal; the value of the threshold would mostly be determined by the background noise level of the environment and the electrical noise. Rotating systems on the other hand are noisy; there are more sources of acoustic noise, and bursts that are not related to damage but generated by friction, roller impact, arising from minor misalignments and transient loads. These lead to significant and often varied levels of background noise. Having a constantly changing noise level introduced by periodic friction complicates the basic process of identifying an AE hit. Figure 4a illustrates an AE signal with varying background noise, where some AE hits are evident well above the background noise. The problem is that the lower energy hits that may lie close to, or even be buried under, the noise. In order to identify these AE hits, a special adaptive threshold methodology had to be devised.

AE data streams comprise millions of points and the feature extraction process being described here is applied to hundreds of data files. Efficient computation of features is therefore required. In order to achieve efficient compression of AE signals, while preserving the information contained within them, a multi-level Discrete Wavelet Transform (DWT) was applied to all AE signals in this study. Each level of a DWT first splits the signal, using a half-band quadrature mirror filter into its low and high frequency components and decimates the signal by half. Each level of decomposition comprises wavelet coefficients, each representing half of the frequency band of the level above with half the number of points. Multilevel DWT is a popular data compression tool in the general context of signal and image processing [47]. Its application to AE data makes sense given that the information in the signals is contained in a short bandwidth, dictated by the resonance of the sensor. In this study, the sensors used all had resonances in the range of 100-300kHz. For this reason, a two-level DWT was applied that split the signal into two frequency bands, of 0-250kHz and 250-500kHz. Only the lower frequency wavelet coefficients were used effectively reducing the number of data points by half, while keeping all the information of the sensor resonant frequencies. If one were to perform monitoring using a broad-band measurement technique, such as a Laser Doppler Vibrometry (LDV) or fibre brag-grating, this step should be applied carefully so as not to throw away important information. However, piezoelectric transducers will always be resonant around a narrow band and so performing a DWT that retains only that band is bound to be an efficient pre-processing step.

3.1.1. AE hit identification

The objective of hit identification is to establish the existence of an AE hit and its location in time. As discussed above, simple threshold strategies, which work well in non-rotating machinery, fail to correctly identify all of the AE hits in a gearbox setting. The problem is that when non-stationary background noise is present in the system, the appropriate threshold that separates a high energy event from background noise will change with time. If a threshold were to be applied directly to AE data in this setting, it will either capture all high energy hits and leave out the lower energy ones, or be set low enough to capture low energy events but also be triggered constantly by background noise.

In order to correctly identify AE hits, a thresholding strategy is required that identifies the presence of a hit, within a constantly changing noise floor. The methodology developed here makes use of a hit *identification function*, which computes the difference between the local signal energy E_t and a lagged version of itself at E_{t-a}. The difference is then normalised against the local noise level at $t - a$. The resulting hit identification function $H(t)$ captures rapid changes in energy against the local background noise. The local energy can be defined as a moving Root Mean Square (RMS) statistic within a given short time period. The identification function is formally defined as:

$$H(t) = \frac{E(t) - E(t-a)}{E(t-a)} \tag{1}$$

where a represents the lag of the local energy difference. Its value is critical to the success of the identification function, it should represent the expected time over which an AE event will reach its maximum energy. In this study, the lag was tuned empirically and a value of $a = 500\mu s$ was used. The presence of a hit is established when the identification function exceeds a prescribed threshold, T_H. A value of $T_H = 2$ was used in all hit identification procedures presented in this study. This can be interpreted

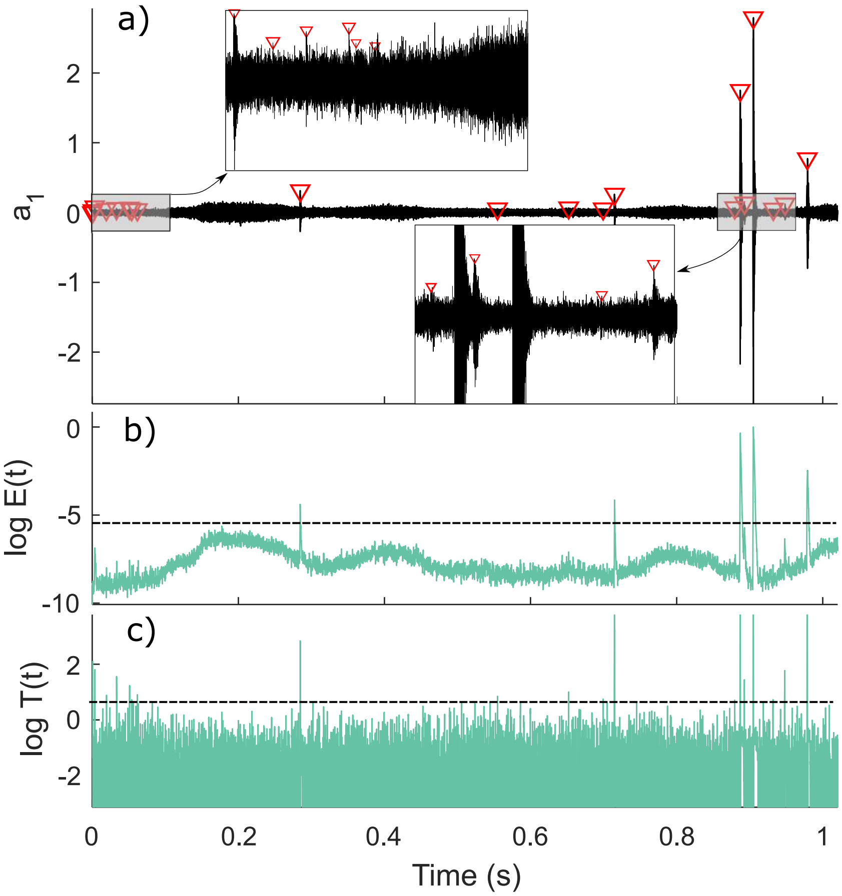

Figure 9: Illustration of effects of setting a threshold either too low or too high, showing a sample of raw AE data with periodically varying noise level

as a hit being defined when the value of the local energy quickly rises to two times its baseline level.

In summary, the steps taken for detecting the presence of a hit, for every AE channel are:

- Decimate signal with a wavelet decomposition
- Take an envelope $E(t)$, of the wavelet coefficients, to capture the amplitude modulation of the process. The envelope could consist of a Hilbert transform or a moving RMS to compute the local signal energy.
- Compute the hit identification function given by equation (1).
- Set a threshold over $H(t)$ and record all instances where this threshold is exceeded.

These steps are illustrated in Figure 9, using a one-second AE recording of an undamaged bearing, taken from the OC-top location (see Figure 8). Figure 9a shows the wavelet coefficients of the 0-250kHz band for this data. Figure 9b shows the local energy function, $E(t)$, while Figure 9c shows the hit identification function derived using Equation (1), including the threshold of 2, the exceedance of which defined an AE hit. The AE hits identified from exceedances of $H(t)$ are shown with triangular markers. Note that using this adaptive thresholding methodology, it is possible to identify AE hits across the entire scale of energies. Figure 9a illustrates this by zooming-in into two regions where low-energy AE hits are present which would have clearly not been identifiable with a simple threshold over the raw data, the wavelet coefficients or the local energy. This procedure is important as it enables the characterisation of individual hits even across the entire range of energy levels.

Once the presence of a hit has been identified by the adaptive thresholding strategy, further steps are required to identify the precise start and end times of each AE hit. For a given hit, a rough start time is already given by the time of exceedance of the threshold over the hit identification function. The end time is defined as either a) the point at which the local energy decays back to within 10% of the local baseline level (before the hit started) or b) the start time of another AE hit occurring before the energy decays back to the local baseline.

Up to this point in the processing, only a rough start time for the hit has been identified, based on the local energy function. However, this will only be accurate to within the given time window used to evaluate $E(t)$. Furthermore, stress waves in a material can propagate in various different modes. The fastest mode will tend to that of longitudinal waves, but this will also carry the least amount of energy. Shear, surface and possibly also Lamb waves (depending on the thickness of the material) may arrive after the first arrival of longitudinal waves, all carrying much more energy. This time delay carries information regarding the total distance a stress wave has travelled, so it is important to capture the precise time of the arrival of the longitudinal mode. Employing a threshold for onset identification, the longitudinal wave will invariably be missed and the arrival of the shear or surface modes is more likely to be captured instead. To overcome this, the methodology proposed by Kurz [48], based on Akaike's Information Criterea (AIC), is used here in order to identify the precise moment of the onset of AE waves. This method computes a cumulative variance of a hit, forwards and backwards and creates an AIC function as the superposition of these two. The point at which this function reaches a minimum indicates the highest change of information (or variance) in the signal and thus the onset of the AE wave can be established by looking for a minimum of this function. An illustration of the AIC function indicating the minimum, where the onset is defined is shown in Figure 10.

3.2. AE hit summary features

Once a table of start and stop positions has been extracted from the AE data for every channel, it is relatively straightforward to go back to the signal and save only the waveforms at those time instances. This is the strategy that has been adopted; it significantly reduces the amount of data stored, and focuses all the post-processing on the data points corresponding to AE hits only, which as discussed before, comprise only a small percentage of the data points in the signal.

There are numerous features that can be extracted once the waveform has been captured. Because an AE waveform is a transient event, there are some key features

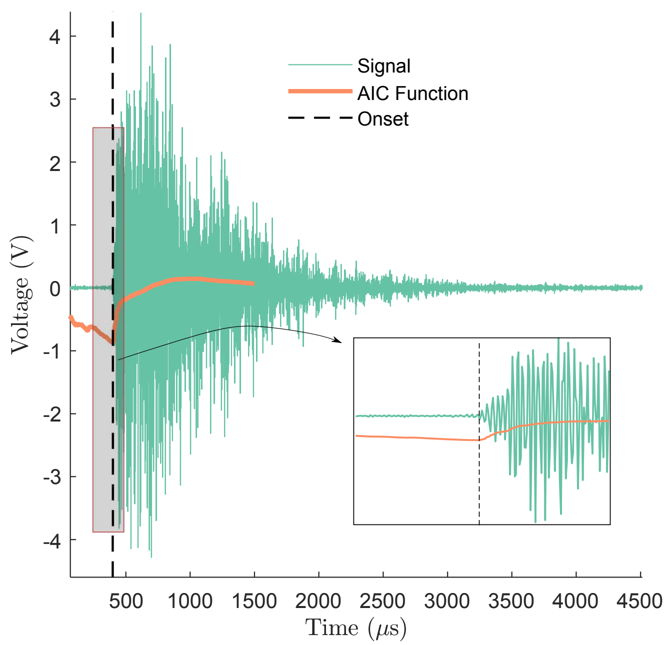

Figure 10: Illustration of AIC onset function on a sample AE hit. The onset is defined where the function reaches its minimum, indicating the greatest change in variance within the signal window.

that can characterise it in general, but simple, terms.

Possibly the most informative feature is the energy contained in the waveform. Different sources of AE will release stress waves at widely different energy levels. The energy is easily computed as the sum of squares of the data points. The power normalises the energy by the duration of the signal. In the case of a transient waveform, such as that of an AE hit, energy, power and RMS will all be related to each other since the duration is a function of the total energy, because of the exponential decay in amplitude. This means that there is some correlation between these variables.

The rise time is defined as the time difference between the waveform onset and its maximum amplitude. The information this carries is valuable because, due to the difference in speeds of different wave modes, some will arrive first and some later, thus giving a rough indication of how far the source is from the sensor. In practice, in a steel structure, waves will propagate as longitudinal, transversal, surface, and possibly Lamb waves. These waves will all travel at different speeds and will carry different proportions of the total energy of the wave-front. The Lamb wave modes may or may not be excited, as their existence requires that the wavelength of the AE be of the same order of magnitude to the thickness of the material it is travelling through. An investigation of Lamb waves is outside the scope of this paper, but their use should not be discarded and is marked as future work. In steel, longitudinal, shear and surface waves arrive in that order. The amount of energy they carry is also given in that order. Therefore the first arrival will always be from a longitudinal wave, and the maximum amplitude will tend to be recorded at the arrival of a surface wave. The usefulness of this is that the rise-time of an AE hit is a useful feature as it gives an indication of how far the wave has travelled. Waves that come from far away will have high rise-time (separation between longitudinal and surface waves) while the opposite is true for short rise times.

Other features that are collected are the peak amplitude of the signal, the total duration and the decay time. The duration is defined, during the hit extraction process, as a decay after the peak amplitude to a level within a specified tolerance of the baseline noise, immediately before the hit. The duration will tend to be a function of the energy in the waveform, but also of the physical mechanism exciting the wave, and it is therefore a useful feature. Once all of these features for each hit are computed, they are assembled so that inference with a Bayesian network can be performed with them.

In summary, six summary statistics are taken from every AE hit:

1. Maximum Amplitude
2. Power
3. Energy
4. Rise-time
5. Decay time
6. Duration

These features ae commonly used in the detection of damage from AE measurements [22, 36, 41]

3.3. AE hit Auto Regressive coefficients

Whilst the hit summary statistics may provide sufficient information for detection, their general drawback is that they provide a simplistic representation of the signal, they also require a significant amount of pre-processing (such as the identification of an accurate onset), which can be prone to error. An alternative is to represent the signal in terms of a model that automatically captures the main characteristics of the signal. In this paper, Auto Regressive (AR) models are used as a damage-sensitive feature that provides a more detailed representation of the individual AE hits. In this paper, AR model weights, $\mathbf{w}$, are fit via linear regression to every single AE hit extracted using the procedure outlined above. As with the summary statistics, this AR model is fit to the single-level Discrete-Wavelet-Transformed signal, thus halving the number of points to compute and focusing on the frequency range of interest (250kHz to 500kHz in this case).

3.4. Modulated signal envelope features

If one signal processing paradigm were to be singled out as having enabled large-scale fault identification in rotation machinery, taking frequency decompositions of rectified signal envelopes would easily win. The idea is simple; every time a periodic load is applied to a gearbox component (roller-bearing, gear-tooth contact for example) which contains a sizeable defect, this will generate a high frequency burst of energy, which can often be sensed at other points in the gearbox. These high-frequency burst will not necessarily be evident behind a frequency spectrum of the dynamic response, because the actual frequency content of these burst will be dictated by various resonant frequencies at which these bursts are transmitted. These resonant frequencies are characteristic of the gearbox assembly, not of the bursts and so will be generally excited during normal gearbox operation. What *is* characteristic of these bursts is that they happen at periodic intervals and this period can single out the particular component that is generating them. The result of this is that the amplitude modulation of the dynamic response signal contains more information about damage processes than does the signal itself. One simple and well-established way of extracting this amplitude modulation is via the use the Hilbert-Huang transform. A simple frequency analysis, via a Discrete Fourier Transform (DFT), of the signal envelope has been shown to highlight well defects in many different types of rotating machinery [49]. Whilst this idea was originally applied to vibration signals which measure dynamic response in a much lower frequency range, this technique has been applied to AE measurements with a good degree of success [28, 33, 40].

In this paper, the DFT coefficients of AE signal envelopes are used as a damage-sensitive feature. This provides a useful point of comparison, given that there so far, this type of feature have been widely used in the majority of papers investigating damage detection using vibration and AE in rotating systems [28, 33, 40, 41], and more specifically, detection of sub-surface damage.

As discussed in the previous sub-sections, the AE signals are originally sampled at relatively high sample rates (1MHz in this case), in order to capture the high frequencies at which the stress waves characteristic of AE travel (250kHz-500kHz in this case). By contrast, the frequencies at which one would expect to find evidence of damage in the amplitude modulation of these signals is much lower, belonging in the range between zero and the low hundreds of Hertz. In the specific case of this bearing rig, the ball pass frequency between the rolling elements and the bearing is estimated at 15Hz, hence, there is a large disparity between the bandwidth of the original envelopes and the frequency content of interest. A DFT at this original sample rate would yield very poor frequency resolution and would also be computationally expensive. In order to remedy this, in this paper, the Hilbert transform is applied to the one-level Discrete-Wavelet-Transformed signal. This halves the number of points used for computation of the envelope and also only takes the envelope over the frequency bandwidth of interest, thus eliminating the potential for noise to be introduced here. After this envelope is derived from the DWT, these wavelet coefficients are further low-pass filtered and down-sampled down to an effective sample rate of 100kHz, one-tenth of the original sample rate. At this point, a Short Time Fourier Transform (STFT) is applied to down-sampled envelopes, with a window length of 250000 points (2.5 seconds). This is enough to capture a potential of 45 cycles of damage-related AE bursts in every window, with frequency resolution of 0.2 Hz. This frequency resolution is appropriate to capture the damage process at the expected ball pass frequency of 15Hz. Note that for further processing (the probabilistic modelling detailed in Section 4), this damage-sensitive feature was truncated to 2502 spectral lines, which yields an effective analysis bandwidth of 12.5kHz.

4. Probabilistic Modelling

The process of detecting damage from the observed AE damage-sensitive features is a problem of searching for outliers in statistical data. An outlier can be defined as an observation that is different enough from the rest of the observations that it is likely to have been generated by a different mechanism [50]. There are two fundamental elements of outlier analysis in data. The first is a statistical model of the reference (undamaged) condition data. This model is often assembled as a *probability density*. The second element is a statistical distance that measures how far any given observation is to the centre of the data mass, relative to the reference probability density. For the purposes of this paper, an "observation" shall be defined as a multivariate vector of damage-sensitive features, evaluated at one instance in time. In the context of AE data (as discussed in Section 3) this could comprise, for example, a vector of AE hit statistics, autoregressive model coefficients, or rectified signal envelope spectra.

In a large number of application domains [50], including SHM and condition monitoring [16, 51], it is common to assume that the underlying probability density of the reference data can be safely modelled as a Gaussian distribution. Under the Gaussian assumption, the *Mahalanobis Squared Distance* (MSD) is a good measure of the relative closeness between observations and the reference set. The approach of modelling damage-sensitive features with single-Gaussian distributions and using an MSD as a novelty index is now in wide-spread use in the field of statistical outlier analysis [50] as well as in the field of Structural Health Monitoring (SHM) [16, 51, 52]. However, a major drawback of the single-Gaussian distribution approach is that it is unsuitable for modelling the probability density of data that has been generated by multiple regimes. In monitoring contexts, multiple regimes often arise from changing environments and operation. In the case of bearings, varying loads, speeds and temperatures will generate differing characteristic responses in the AE features. These will manifest themselves as multiple modes in the probability density of AE features. One way of modelling these complex probability densities is through the use of mixture distributions, discussed in the following section.

4.1. Dimensionality reduction

One characteristic of some of the damage-sensitive features being used in this study is that they are high-dimensional. The AR coefficients comprise vectors of 150 dimensions while AE envelope spectra contain 2500 dimensions. Generally speaking, most novelty detection schemes rely on computing distance metrics between feature spaces of new observations against feature spaces representing normal conditions. In this setting, it is a well recognised result and an effect of the "curse of dimensionality", that the contributions of individual dimensions to a distance metric tend to get masked in high dimensional feature spaces [53]. This presents a problem - if damage will only introduce a change to a handful of dimensions within a high dimensional feature, this may not show up when computing a novelty index. Furthermore, high dimensional features also present a problem when computing covariance matrices of statistical models. It is a generally well-accepted rule that one needs at least twice as many observations as there are dimensions in a feature space in order to begin to accurately capture the covariance structure of the data.

In order to remedy these problems, the authors turn to the use of standard dimensionality reduction techniques. Here Principal Component Analysis (PCA) is used as a dimensionality reduction strategy. Its use is wide-spread within the field of statistical learning [54, 55] as a data visualisation and pre-processing tool. PCA can be viewed as a class of linear Gaussian models [56] and is represented by a linear transformation from a d-dimensional data/feature space $\mathbf{y}$ into a lower-dimensional space $\mathbf{x}$, called the principal component scores,

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \varepsilon \tag{2}$$

where $\mathbf{C}$ is a $d \times d$ PCA rotation matrix and ε is an additive Gaussian noise term. It is an orthogonal transform designed so that $\mathbf{x}$ contains a rotated version of $\mathbf{y}$ aligned in the directions of greatest variance. The resulting PCA scores, $\mathbf{x}$, therefore contain most of the information (in terms of variance of the original data-set) within the first few dimensions. It is therefore common practice to use only a handful of the dimensions of $\mathbf{x}$, this could be determined by observing how much variance is lost as one throws away dimensions. In this paper, however, the dimensionality of the PCA scores is fixed to five, so as to enable a comparison between different damage-sensitive features without introducing the effects of differing dimensionality of the feature space into the outlier analysis process.

The PCA rotation matrix, $\mathbf{C}$ must be estimated. By definition, $\mathbf{C}$ is the eigen-decomposition of the covariance of the data/feature set $\mathbf{y}$, so it can be estimated through an eigenvalue analysis. However, this may not scale well to very high dimensions. An alternative approach is an iterative Expectation Maximisation (EM) algorithm, which leads to the notion of probabilistic PCA [57, 58]. This paper uses the EM approach described in [57] to learn matrix $\mathbf{C}$

In the scenario being investigated, one does not have access to damaged condition data sets. The PCA rotation matrix, $\mathbf{C}$ is learned using the same data used for the novelty detector (described below, and so it is only representative of the undamaged class. Note that detection would be *easier* if $\mathbf{C}$ could be learned using both damaged and undamaged data sets, as the main differences in both of these sets would likely represent the greatest variance, leading to two different columns in $\mathbf{C}$ describing each condition. This would lead to a significant separation of the two different conditions in the PCA scores, $\mathbf{x}$. However, what is of interest here is the performance of a novelty detector on a projection of the undamaged class only. The discussions that follow will discuss the problem of novelty detection. This novelty detection is carried ot in the lower-dimensional domain of PCA scores, which is denoted throughout the paper as $\mathbf{x}$.

4.2. Gaussian Mixture Models

A natural way of dealing with probability densities that arise from multi-regime processes is to partition the space of features into the different regimes and to fit a density model to each region of the feature space. Ideally, one would have a label of the regime type associated with each observation. However, in practice this is difficult to attain and there may be more naturally occurring clusters than one has labels for. It is therefore more desirable to work with algorithms that automatically partition the space into different regions. This task is generally referred to as *clustering* and there is a wide choice of algorithms available for carrying this out [55]. Here, the focus is on novelty detection, so whatever clustering scheme is used should also define a probability density over the feature space. The Gaussian Mixture Model (GMM) is used here because 1) it is a flexible density estimator for multi-modal data, 2) there are efficient algorithms for clustering, or data partitioning with GMMs and 3) it is possible and straight-forward to derive a novelty index in order to perform damage detection with this model. This subsection discusses these three important points.

The concept of using a GMM for novelty detection within an SHM and condition monitoring context has been investigated in some recent studies[59–61], including some by these authors[62]. The approach to using a GMM to achieve novelty detection taken here follows that of [62], where a novelty index is derived using the probability density of the GMM. The parameters of a GMM model with K components comprise of the set of K means, covariances and mixing proportions. For notational simplicity, these can be encoded in the vector $\boldsymbol{\theta} = \{(\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k), (\mathbf{S}_1, ..., \mathbf{S}_k), (\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_k)\}$. Learning the appropriate parameters of a GMM involves choosing a parameter set, $\boldsymbol{\theta}$ that maximises an objective function. Damage detection then involves evaluation of the likelihood of new observations, given the reference parameter set $\boldsymbol{\theta}$.

4.2.1. The GMM likelihood function

The likelihood of the model plays a central role in the damage detection approach of this paper: it is the objective function used for parameter estimation, it defines the probability density and is also therefore a useful novelty index. The Gaussian mixture is defined as a weighted sum of Gaussians, so its probability density can be written as,

$$p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y}|\mu_k, \mathbf{S}_k) \tag{3}$$

where, $\mathcal{N}(\mathbf{y}|\mu_k, \mathbf{S}_k)$ represents the normal Gaussian density of the k^{th} component, with mean vector μ_k and covariance matrix $\mathbf{S}_k$. The term π_k defines the relative contribution of component k to the total density, also known as responsibilities. Equation (3) defines the *likelihood* of the model, given observed data $\mathbf{y}$. This quantity plays an important role in determining the model parameters as well as deriving appropriate novelty indices. In practice, it is easier and more computationally stable to work with the *log-likelihood* of the model. This is true both for parameter learning, as well as for evaluation of novelty indices for damage detection. Written explicitly as a sum over all observations, the log-likelihood for the GMM is,

$$\ln \mathcal{L}(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln\{\pi_k p(\mathbf{y}_i|\mathbf{z}_k)\} \tag{4}$$

where z_{nk} denotes the posterior responsibility of component k for generating the i^{th} observation. This is a convenient form for the log-likelihood function as it is given as a sum of logarithms This results from a formulation in terms of hidden variables (see [56, 58]). The model parameters are estimated using maximum-likelihood learning, which involves maximising the log-likelihood of Equation (4). Unfortunately, evaluating this quantity involves knowledge of the partitioning of the data-set into the different clusters. This cluster-assignment is not known a-priori which makes this an optimisation task with missing data, or hidden variables. The formulation of the GMM log-likelihood function in terms of these hidden variables has already been discussed in Section 4.2.1. The Expectation Maximisation (EM) algorithm was derived to deal with exactly this type of problem; it is a general framework for performing maximum-likelihood parameter optimisation of models with hidden data [63]. Furthermore, it has been shown that EM is a suitable learning algorithm for a wide class of linear Gaussian models with latent variables [56], to which GMMs belong to. EM is used in this paper as a learning strategy for all GMM models of damage-sensitive features.

4.2.2. GMM model selection through cross-validation

A common issue is that of choosing an appropriate model order; the number of components, K, of the GMM. Too many components will result in the GMM over-fitting the density estimate (assign very high density in regions of data where observations exists, and no density elsewhere). On the other hand, using too few components would fail to correctly capture the complexities of the true underlying distribution. Ideally, the number of components should be determined directly from the training data. In this paper, the authors turn to the use of cross-validation in order to perform this task. Cross validation procedures divide the available training data into different training and testing subsets, so that the model error is evaluated on previously un-seen data and so the generalisation performance of the algorithm can be evaluated. For this paper, k-fold cross-validation is used, as this is a well-known robust and efficient way of performing cross-validation in a wide variety of statistical models [54]. K-fold cross-validation divides the available training data into K different partition, or folds. These are selected at random. Then, the GMM model is trained using EM, on $K - 1$ folds, thus leaving one fold out of the training set. The model error is then computed for the left-out fold, and the process until every fold is held-out at least once as a test set. The benefit of this is that the model error can be quantified entirely in terms of its performance on previously unseen data, and so this gives a good insight into the generalisation performance of the model.

As for the model error, two different quantities will be used in order to assess the quality of the model fit. The first is the Bayesian Information Criterion (BIC) [64], a quantity that is derived from the model likelihood, but penalises models of higher complexity (more components). The BIC considers the intrinsic function of the GMM as a density estimator. However, it does not consider its function as a classifier, more specifically in the novelty detection case, a one-class classifier. Therefore, the second error metric considered in the cross-validation procedure in this paper is the misclassification rate. However, in order to treat the GMM model as a classifier, one must also consider that the overall model comprises both the density estimation *and* the novelty threshold together. How this is determined plays a central role in the success or otherwise of the damage detection scheme. The threshold estimation scheme used in this paper is discussed below. The error metric considered is the exceedance rate of the estimated threshold over the GMM density model observed on the held-out fold of the cross validation procedure.

4.3. Detection thresholds

An important aspect of any novelty detection scheme is the definition of the detection threshold. The detection threshold defines the decision boundary between which observations are classed as normal or abnormal. In this case, because the negative log-likelihood of a Gaussian mixture is being used as a novelty index, the detection threshold will be given in terms of this quantity. There is no ultimate gold standard method for estimating an appropriate novelty threshold, and there is a wide variety of approaches to this problem in the literature [65]. In the problem at hand in this paper, the primary requirement is that the threshold minimise the number of false positive and false negatives. The other consideration of practical importance in this work is whether the training data, from the undamaged condition, contains outliers. Outliers in a training set will manifest themselves as observations with high novelty indices. If these are taken into consideration for the threshold determination for example, through the use of maxima or percentiles of the training novelty indices, the resulting threshold will be biased by these outliers. This would in turn mean that observations from potential damaged classes may fall under the threshold. On the other hand, placing the threshold too close to the majority of training novelty indices will result in an over-sensitive classifier that produces a high number of false positives. In this work, a robust approach to threshold estimation was used, following ideas from Monte Carlo sampling and Extreme Value Statistics (EVS), which have been used in the past as robust means of threshold estimation [51, 66, 67]. When deciding on an appropriate novelty threshold, one is interested in the distribution of maximum values of the novelty indices. The distributions of the exponential family (to which the Gaussian belongs) are generally poor at predicting extreme values, due to the fact that the tails of the distribution decay exponentially to infinity. This is not a particularly good representation of the maxima (and minima) of physical events. Extreme value theory offers an alternative for modelling the tails of statistical distributions. It dictates that the extremes will be governed by either one of three distributions: Gumbel, Frechet or Weibull. The Gumbel distribution is of particular interest since it is the limiting distribution of the maxima of Gaussian random variables, so it would be a suitable distribution for modelling the tails of a Mahalanobis (un-squared) distance. However, when faced with more complex probability densities of the undamaged class data (as is the case in this application of bearing monitoring), other novelty indices might be used which may not conform to the Gaussian assumption. Such is the case of the negative log-likelihood of the GMM, being used here as a novelty index. In this case, the Generalised Extreme Value (GEV) distribution offers a solution to modelling the tails of arbitrary probability distributions [68]. Its use has previously been investigated in the context of SHM [69].

Here, the novelty threshold is defined using a GEV distribution fitted to random samples of negative log-likelihoods drawn from the estimated GMM density. The threshold estimation procedure follows the Monte Carlo approach outlined in [51], except that here, a GEV is used instead of an empirical cumulative distribution function when assessing confidence levels. The threshold estimation procedure is outlined in Algorithm 1. The idea of it is to draw N_s random samples from a GMM with parameter set $\boldsymbol{\theta}$ representing the normal condition, and to evaluate their novelty index (the negative log-likelihood). This process is repeated for N_t different trials, in which the maxima of each trial is recorded and stored in a vector $\mathbf{z}$. A GEV distribution is then fitted to this vector of maxima, from which a threshold, T, can be estimated using the GEV Cumulative Distribution Function (CDF),

$$GEV_{cdf}(z) = \exp\left\{-\left(1+\xi\frac{z-\mu}{\psi}\right)^{-1/\xi}\right\}, 1+\xi\frac{z-\mu}{\psi} > 0 \tag{5}$$

where μ and ψ are location and scale parameters respectively and ξ is an additional parameter which determines the type of distribution the GEV fit belongs to (from the family of Gumbel, Frechet or Weibull).

Algorithm 1 Threshold Estimation

---

procedure MC-GEV THRESHOLD($\boldsymbol{\theta}, N_s, N_t, C_{onf}$)

  for $i = 1 : N_t$ **do**

    $\mathbf{y} \leftarrow$ draw N_s samples from GMM with parameter $\boldsymbol{\theta}$

    $\mathbf{L} = -\log p(\mathbf{y}|\boldsymbol{\theta})$

    $z_i \leftarrow \max(\mathbf{L})$

  end for

  Fit a GEV distribution to vector of maxima, $\mathbf{z}$

  $T \leftarrow \text{argmin}||GEV_{cdf}(\mathbf{z}) - C_{onf}||_2^2$

  return T

end procedure

4.4. Procedure Summary

To summarise, the procedure for using a Gaussian mixture for damage detection suggested here is as follows:

1. Select a (damage sensitive) feature vector $\mathbf{y}$ to represent the data from a healthy condition. Any operational and environmental changes should be captured in $\mathbf{y}$. In this paper, three features are considered from the AE data: hit summary statistics, hit AR coefficients and signal envelopes.
2. Select a subset of $\mathbf{y}$ to use for training the model, $\mathbf{y}_{train}$. Select another subset to test the model predictions on: $\mathbf{y}_{test}$.
3. Project high dimensional features, $\mathbf{y}$ onto a suitably lower-dimensional domain $\mathbf{x}$.
4. Use cross-validation to decide on an appropriate number of clusters, K, for a GMM model, the training data set.
5. Train a GMM, using the EM algorithm, on the entire $\mathbf{x}_{train}$ set, using the number of clusters K determined from the cross-validation step.

6. Evaluate negative log-likelihood function on $\mathbf{y}_{train}$, point-by-point, using equation (3), and set the detection threshold $\mathcal{T}$ using the procedure described in Algorithm 1.
7. Evaluate $-\log \mathcal{L}$ on any new observations, and check whether this falls above or below the detection threshold.
8. If the model correctly captures the variability of the healthy condition, exceedances of $\mathcal{T}$ indicate damage, or other previously unseen (and possibly benign) changes.

5. Experimental results

This section details the results of applying the damage detection framework described in Section 4 to the AE damage sensitive features computed using the methods outlined in Section 3 to the wind turbine bearing experimental set-up, described in Section 2.

Recall that damage detection is performed by fitting a Gaussian mixture model to the PCA projection of the damage-sensitive AE features. The results of interest are the quantification of detection performance of the identified GMM model on observations from bearings with increasing levels of damage. The damage levels are outlined in Table 2.1.

All GMM models were trained using data from condition UD1, from the first undamaged bearing, tested on the other half of the UD1 set and then validated on the data set from UD2. Each AE channel is being considered separately, therefore each will generate a different number of hits at different time indices. The envelope spectrum-based features also generate a much less dense quantity of features per trial. In order to keep a consistent indexing and to enable a quantification of false positives and negatives across the three different features and four different AE channels, the decision threshold was applied to the maximum novelty index of each 10 second recording. If this falls above the detection threshold, then the trial is classified as normal, otherwise it is classed as outlying. This makes it possible to quantify false positives and negatives, because the ground truth of the state of each trial is available. On the other hand, it would be impossible (at this stage) to establish the ground truth as to whether an individual AE feature was generated by a damage or a benign mechanism, within the stream of hundreds of thousands of AE hits. Bearing this in mind, the random reshuffling to divide the UD1 data set into training and testing sets was carried out by reshuffling the trial indices, rather than individual feature indices. This ensures that the same training and testing data is used for all AE channels and features. The splitting of available undamaged condition data into training and testing helps to understand the generalisation performance of the GMM as a density estimator. However, validating the detection performance on a second undamaged bearing measures the robustness of the entire damage detector - including the choice of damage-sensitive features.

Whilst the aim was to keep the environmental variations consistent across the different bearing conditions in order to carry out a fair comparison, the reality was that the temperatures inevitably fluctuated slightly between the different conditions. Most critical is the oil temperature, as this relates directly to the viscosity and therefore the lubrication regime. All tests were carried out at a low and a high temperature regime, but the precise temperature of each regime varied slightly. In general, operating the gearbox at higher temperatures leads to increased levels of AE activity due to the higher level of asperity contact. Of the two data sets collected on undamaged bearings, UD1 was collected at a slightly lower temperature than UD2. Training a novelty detector on the lower temperature data set and validating its predictions on a data set from a higher temperature provides a more robust validation than would be by testing against data from similar temperatures. For this reason, the data set from UD1 was used for training, and UD2 for validation. This presents a harder problem than would be by training on UD2 and validating on UD1.

The temperature distributions for UD1 and UD2 are illustrated in Figure 11a, using a kernel density estimate (with a band-width of 4). This shows the density of trials at the different temperatures, split by training, testing and validation sets. Note that this shows the training and validation distributions after applying the random shuffling to select training and testing sets from UD1. Two things are clear. The first is that there are two clear regimes of temperature, shown by the two different modes of the density curves. The second is that, for the validation set, the temperature distribution is shifted towards higher temperatures. This means that UD1 contains a major regime around the 25°C range which is not completely captured in UD2. Conversely, UD2 contains a number of trials above 50°C that are not present in either the training and testing sets of UD1. Figure 11b shows the temperature distributions of the damage condition data sets. Note that the data for the lower damage levels (D1, D2 and D3) were collected at slightly higher temperatures than the training and testing sets, but overall lower than the validation set. The modes of the D1 and D2 - the two subsurface damage sets lie at 50°C, which is still within the reach of the density of the training set. The data for the highest level of damage (D5) was collected at relatively low temperatures, with the mode for the lower temperature regime having a significant level of density in the range between 20°C and 25°C.

5.1. Damage sensitive features

Three damage-sensitive features were computed from the raw AE data in order to carry out a comparison of damage detection performance, as described in Section 3. Two of these were based on individual AE hits: summary statistics and Auto Regressive (AR) model coefficients. One of the features was designed to capture the periodic nature of the amplitude modulation of the AE signals, so the frequency spectra of AE envelopes was used. The hit summary statistics have a relatively low dimensionality of $d = 5$. The model order for the AR coefficients was chosen to be 150, so in this case $d = 150$. The envelope spectra originally had a much higher dimensionality of $d = 2500$, although this was truncated to 150 as it was found that the higher dimensions (belonging to higher frequencies) only contributed in terms of added noise and did not add a significant amount of information.

Recall that all features were computed from a discrete-wavelet-transformed domain not directly from the raw data. Only the high frequency component of the single level DWT is used, in order to leave out the low frequency bands where the AE sensors are not resonant. The features are, therefore, only representative of the 250kHz-500kHz frequency band. These features are illustrated in Figure 12 (for the OC top sensor), which shows the median and inter-quantile ranges of the three features (along each column), grouped by different levels of damage. The shaded area represents the percentile regions of $\pm 25\%$ around the median, while the dashed lines represent the extrema; the

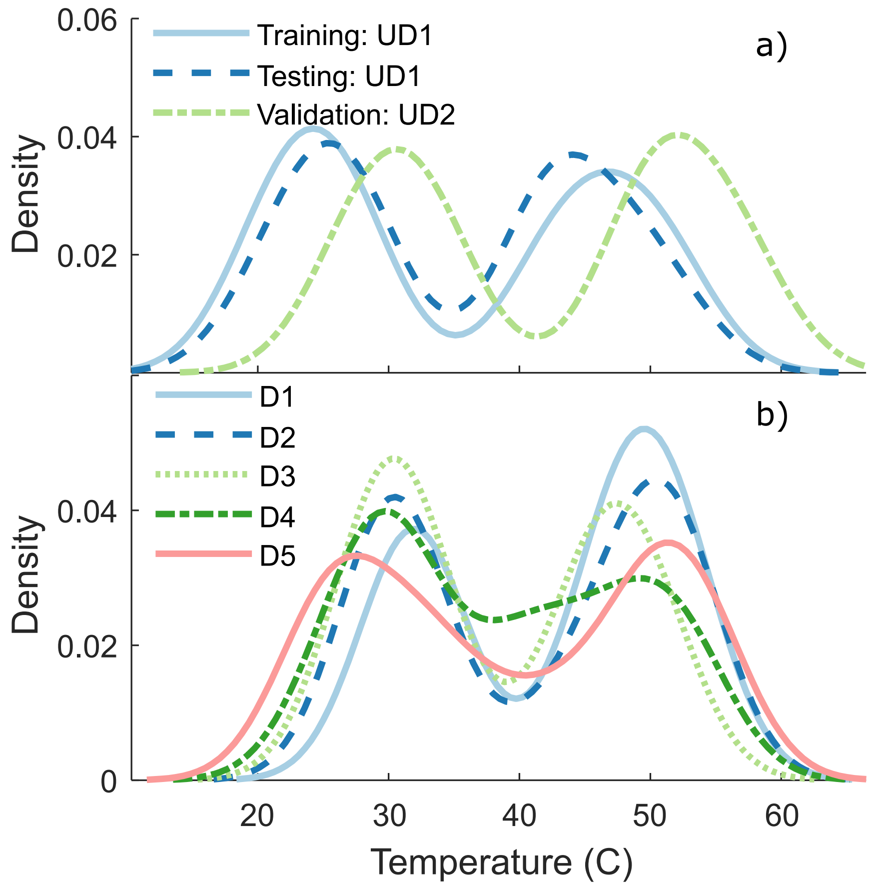

Figure 11: Kernel density of temperature ranges for a) the training, testing and validation trials and b) the damage condition trials.

1% and 99% percentiles. The AE features observed in this gearbox are characterised by a large number "extreme" events, which completely mask non-robust measures of location and scatter (a regular mean and standard deviation). The visualisation in terms of inter-quantile ranges of Figure 12 allows a qualitative appreciation of the shape, scatter and extremes of these features. Note that the scale has been adjusted so that the upper 99^{th} percentile of the undamaged feature vectors is visible on the plots. This makes it harder to examine the details of the features, but enables an appreciation for the large difference between the average process and the extreme AE events. The development from non-damaged to damaged is visually clear across all features in Figure 12. A common factor between all three features is an increase in variance of the feature vector as damage progresses, and this is most evident towards state D5.

These robust measures of location and spread of the features make it possible to visualise how in this case, damage does not manifest itself as a sharp change to the average feature vector. Instead, the baseline characteristic AE activity remains largely the same, with the addition of extra AE activity that is characteristic to the damage process. This is true for both hit-based features, but not so for the envelope spectra, which is capturing information across relatively large time-scales. It is evident from Figure 12c that the envelope spectra completely shifts its median as damage is introduced and progresses, along with an increase in variability. While this may be a desirable property for this damage-sensitive feature, it should also be noted that its variability in the undamaged condition is much larger, and this will result in greater variance in any density model fitted to it. This adversely affects detectability, especially at the lower damage levels. Of the three features compared in Figure 12, AE hit statistics, which have the lowest dimension (6), also have the lowest variability in the undamaged state.

In order to reduce the dimensionality, PCA was applied to all three damage sensitive features, using only the training data (from the undamaged condition) to derive the PCA rotation matrix $\mathbf{C}$ (see Equation (2)). The dimension of the PCA scores was chosen to be $m = 5$ as this captures most of the variance contained in the original full-dimensional features. Figure 14 provides an illustration of the first 5 PCA scores for each damage-sensitive feature. Because of the large quantity of individual feature vectors, which greatly vary in scale, a visualisation of the PCA scores of individual feature vectors is not effective. Instead, Figure 14 shows the standard deviation of each PCA score for individual trials (10-second recordings). In order to see how environmental and operational changes have an effect on the PCA scores, the bottom row of Figure 14 shows the applied load to the bearing and casing temperature for each trial. Focusing on the two undamaged conditions, it is clear that temperature drives the variance of the first principal component on all three features. In particular, note that the higher temperature range of UD2 has the greatest variance of all UD1 and UD2 trials. The PCA scores of AE hit statistics and envelope spectra show a clear increase in overall variance as damage is introduced, which increases as damage progresses. This effect is marked by large increases in the first score (this represents the direction of greatest variance). The situation is different for the hit-based AR coefficients, where the effect of damage is more markedly seen from the second score onwards. The reason for this is that an AR model is insensitive to scale, so a simple change in the overall energy of the waveform will not yield a different coefficient set. A change in the shape of the waveform, on the other hand will lead to a change in the coefficient set. The fact that damage is more evident from the third score onwards indicates that the AE waveforms generated by a damage mechanism "look" largely the same in this case, but have subtle differences. This is consistent with the visualisation of the median and inter-quantile ranges of AR coefficient vectors provided in Figure 12b. Damage does not completely change the shape of the AR coefficient set, it subtly changes and generates more variability in some of the dimensions.

Another way to provide a qualitative view of how the damage process affects the three different damage sensitive features is to visualise how the *probability density* of these features changes with different classes of damage. Figure 13 provides a visualisation of a two-dimensional kernel density estimate evaluated on all three damage sensitive features for conditions UD1, D1,D2,D3,D4 and D5[1]. The kernel density was evaluated on the first and second PCA components of each feature for the OC Top sensor location. Note that this is helpful even for the lower-dimensional features such as AE hit summary statistics, as it allows for visualisations of a decomposition of the data

[1]UD2 was omitted since it is visually similar to UD1

in the two directions of greatest variance. The contours of Figure 13 indicate regions of equal probability density, and only four contours have been drawn, so as to divide the probability density into four quantile regions. Two major regimes are clear across all three features for the undamaged state. These represent the high and low temperature regimes. The first damage level is not strongly visually evident, but the progression of damage is clear across all three features, albeit in different ways. As damage appears, there is a clear change in the *shape* of the density. In D2, the second mode that was evident in UD1 is now masked by a much higher density closer to the core density. This is the case for all three features. While this is indicative of a change, this will be harder to detect given the relative closeness of the change to the majority of the data mass. As surface damage appears, all three features develop regions with high density away from the core of the data mass. Given this difference in the shape of the probability density of the features as damage progresses, these observations are bound to generate high negative log-likelihoods with respect to a model trained on UD1.

5.2. Cross validation analysis of GMM

In order to gain an insight into the output variance and generalisation performance of the novelty detectors, cross validation was used as described in Section 4.2.2. Two metrics were considered: the exceedance of the detection threshold as an error metric and the Bayesian Information Criterion. It is important that this is carried out using only the data available for training, as this is the scenario one is faced with when performing a realistic monitoring task. A 10-fold cross validation procedure was used, considering cluster sizes in the range of $K = 1, ..., 15$. Using any further than 10 folds on this data set tends to result in ill-conditioned covariance matrices for some of the GMM components. The 10-fold cross validation results using misclassification rate are shown in Figure 15, for all three damage sensitive features. The curves in Figure 15 show the median of the cross-validated output as a solid line, while the greyed-out area encloses the regions between the 5^{th} and 95^{th} percentiles.

In general, the variance of the misclassification rate decreases as the number of components of the GMM is increased. The AR model coefficients yield the best performance in terms of variance. The envelope spectra yield the best performance in

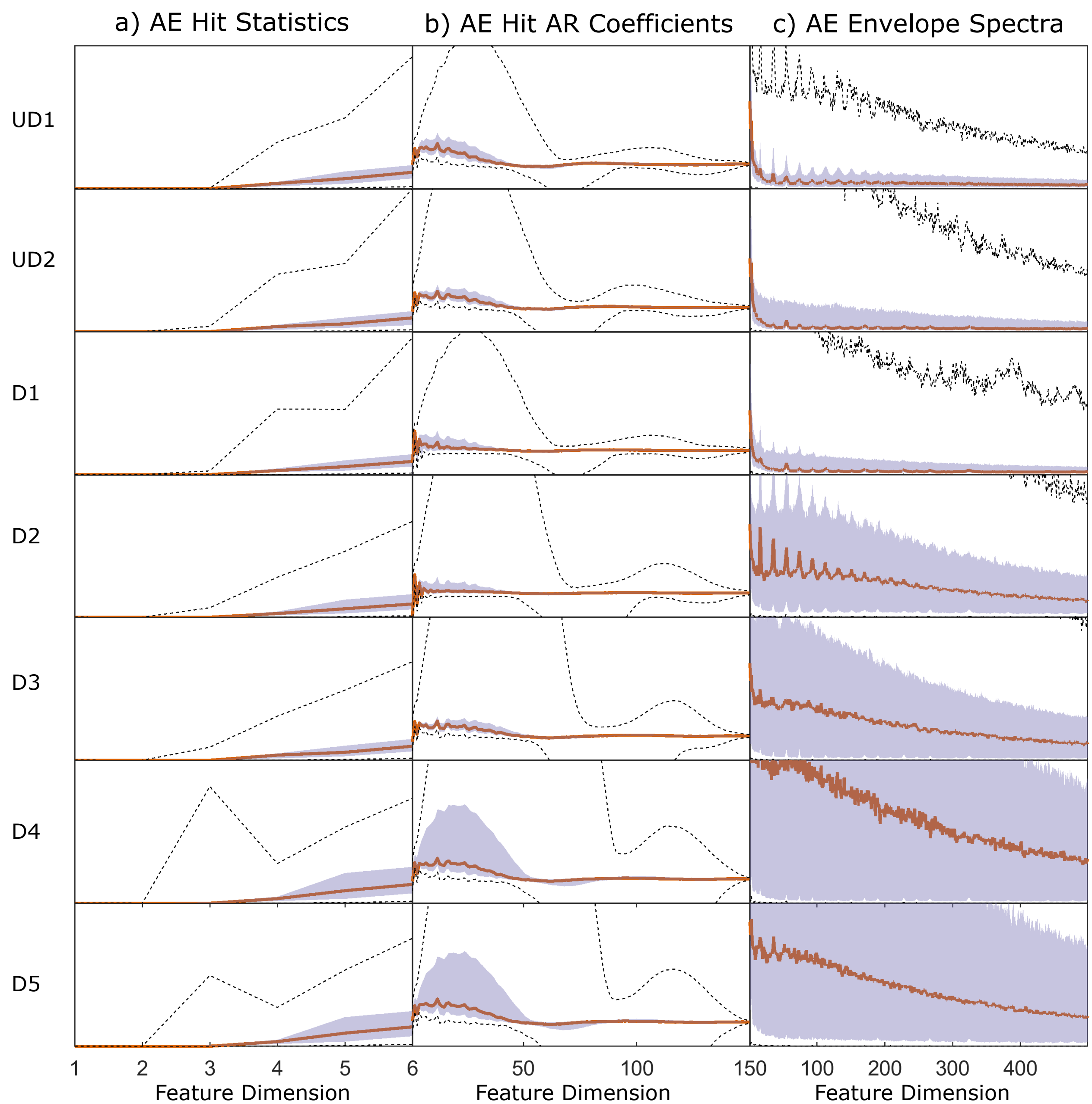

Figure 12: Median and inter-quantile ranges of the three features: a) AE hit statistics, b) AE AR coefficients and c) AE envelope spectra, grouped by different levels of damage. The shaded area represents the percentile regions of $\pm 25\%$ around the median, while the dashed lines represent the extrema; the 1% and 99% percentiles. Note that the vertical axes are not labelled as they correspond to normalised features, but they show the same scale for each feature type.

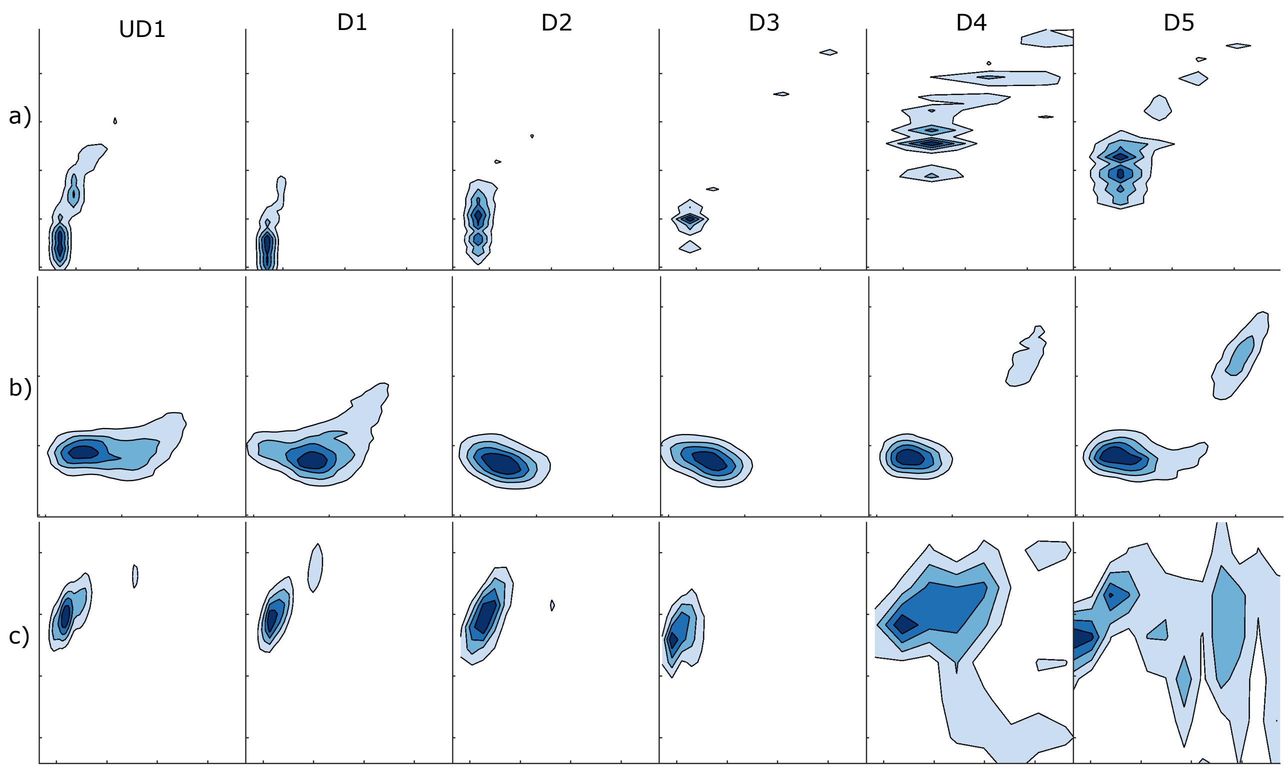

Figure 13: Two-dimensional empirical density estimates of the first and second PCA scores for the three damage sensitive features used:a) hit-based features, b) AR coefficients and c) envelope spectra. These are shown for increasing levels of damage as outlined in Table 2.1

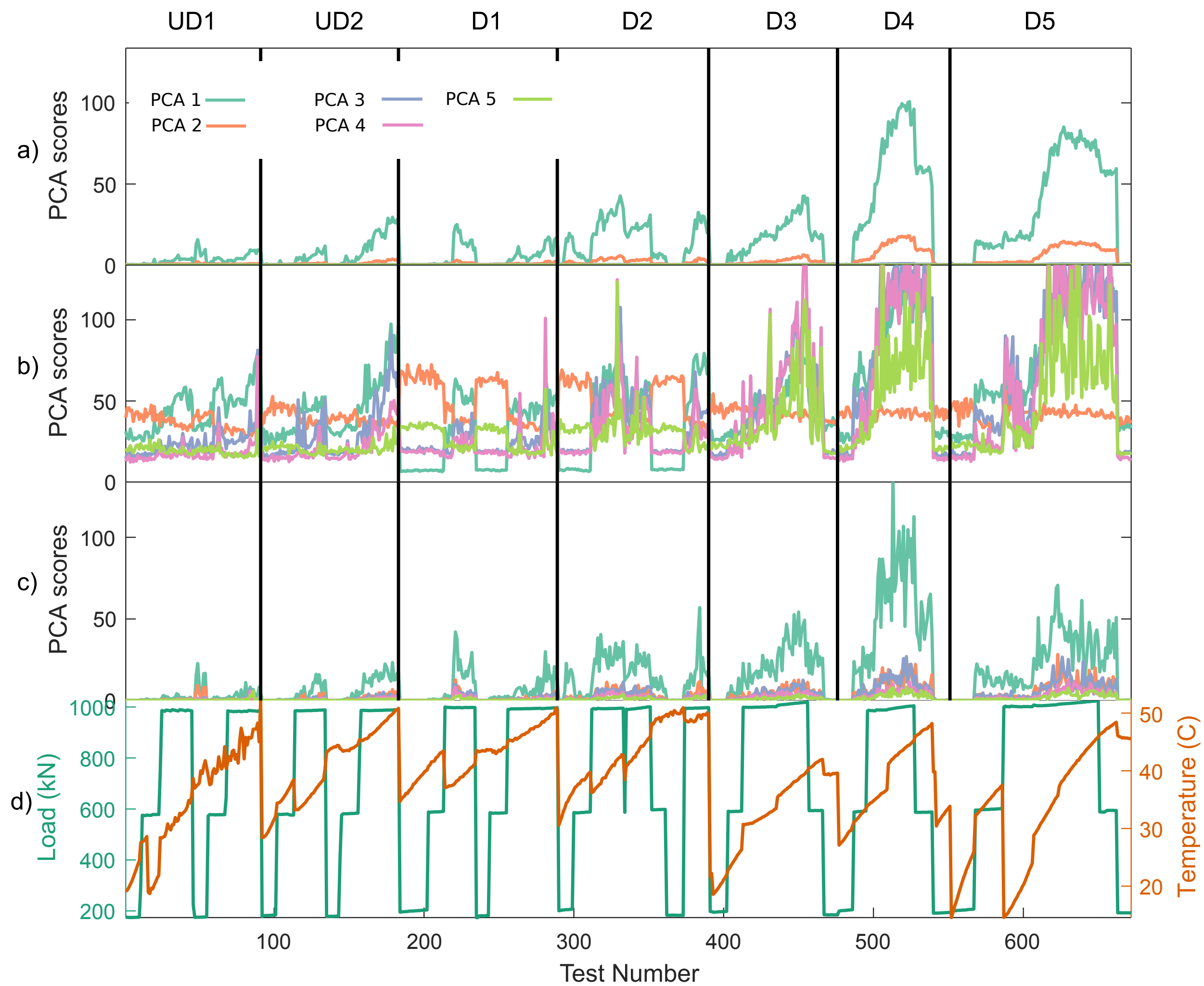

Figure 14: Standard deviation of the first five PCA scores computed per 10-second test for all bearing states. The three damage-sensitive features are shown: a) AE hit statistics, b) AE AR coefficients and c) AE envelope spectra. d) shows the variation in load and temperature throughout these tests.

terms of median misclassification rate, reaching 100% correct classification but with very high variance due to a few stray folds.

Figure 16 shows the cross validation results, using the BIC as an error metric. Note that in this figure, the vertical axes, representing the BIC are not shown as each subfigure has a different scale. However, it is the trend and the variance that are important. In general, as the number of clusters increase, the BIC increases, indicating a better fit, but so does the variance. A high BIC with high variance is indicative of model over-fit. The selection of the appropriate model order has to balance high BIC scores, low BIC variance as well as low misclassification rate and low variance around this.

Several observations can be made from examination of Figures 15 and 16. The first is that while the average BIC increases with increasing cluster numbers (as would be expected), the misclassification rate does not have such drastic improvements and tends to converge quickly. The second observation is that all four AE sensor locations behave in a similar fashion for each of the damage-sensitive features considered. This is more evident in the misclassification rate. Furthermore, each of the three different features has a markedly different optimal number of clusters. The misclassification rate provides greater insight than the BIC on this point. For each feature and AE sensor location, the model order for the GMM was selected as the first to generate a misclassification where 95% of the folds (the upper bounds in Figure 15) have a misclassification under 0.01%. These are marked with vertical lines in Figure 15. It is interesting to note that the AE envelope spectra generate consistently low *median* misclassification rates, but with high 95^{th} percentiles. This behaviour is due to one or two outlying observations, the source of which is likely to be contaminating ambient noise, which sometimes overpowers the envelope spectrum if this is at much higher amplitudes than regular AE activity. It is also worth considering that hit-based features generate in the range of 500 to 1000 observations per trial. Hence outliers tend to hide well above the 95^{th} percentile. On the other hand, the envelope spectra generate only 8 observations per 10-second trial, which means that a small number of outlying observations make is seem like the spread of this error metric is large. Furthermore, this also means that there is less resolution on the misclassification rate. Considering this, the criteria for setting the GMM model order on AE envelope spectra features was that the median reached a 0% misclassification rate.

5.3. Damage detection results

This section presents the results of the damage detection process. The objective is to quantify the performance of each detector. A *detector*, in this context is a GMM of a damage-sensitive feature at a sensor node. Each detector has a different model order, established during the cross-validation procedure described above, and has its own threshold, established using the GEV procedure of algorithm 1.

After making a decision on the GMM model orders for each sensor location and damage-sensitive feature, a GMM was trained using the entire training data-set. The novelty of subsequent observations is assessed by evaluating the Negative Log-Likelihood (NLL) of each feature vector against the reference GMM. This is described by Equation (4). Even though the NLL already represents a logarithmic scale of the original Euclidean distance between the GMM centres relative to their variance, the resulting NLLs evaluated over the entire range of bearing conditions still results in orders-of-magnitude difference in scale. For the purposes of visualisation, $\log$ NLL is used here, noting that this is just a practical transformation for visualising results. Figure 17 shows a kernel density of the $\log$ NLL, for all four sensor nodes and three damage-sensitive features. Each sub-plot in Figure 17 represents a sensor-feature combination, and densities are shown for each bearing condition. Note that all sub-figures in Figure 17 have been zoomed-in on the vertical axes, to focus on the low-density regions, where damage is most evident. The thresholds identified using the GEV approach are shown as vertical dashed lines (note the same log-transformation has been applied to the threshold). In this setting, changes to the baseline AE sound-scape should be evident as regions of higher density of the (log) NLL above the threshold.

As it has already been illustrated in the previous sections, the effect of damage is different across all three damage-sensitive features being considered. In the case of hit-based features, when damage is present, the original density of the features is preserved and additional bursts of energy related to the damage process are generated. This is evident from Figure 17. In the case of the envelope spectra-based features, there is a marked shift in overall mass of the density of the NLL toward the right. The

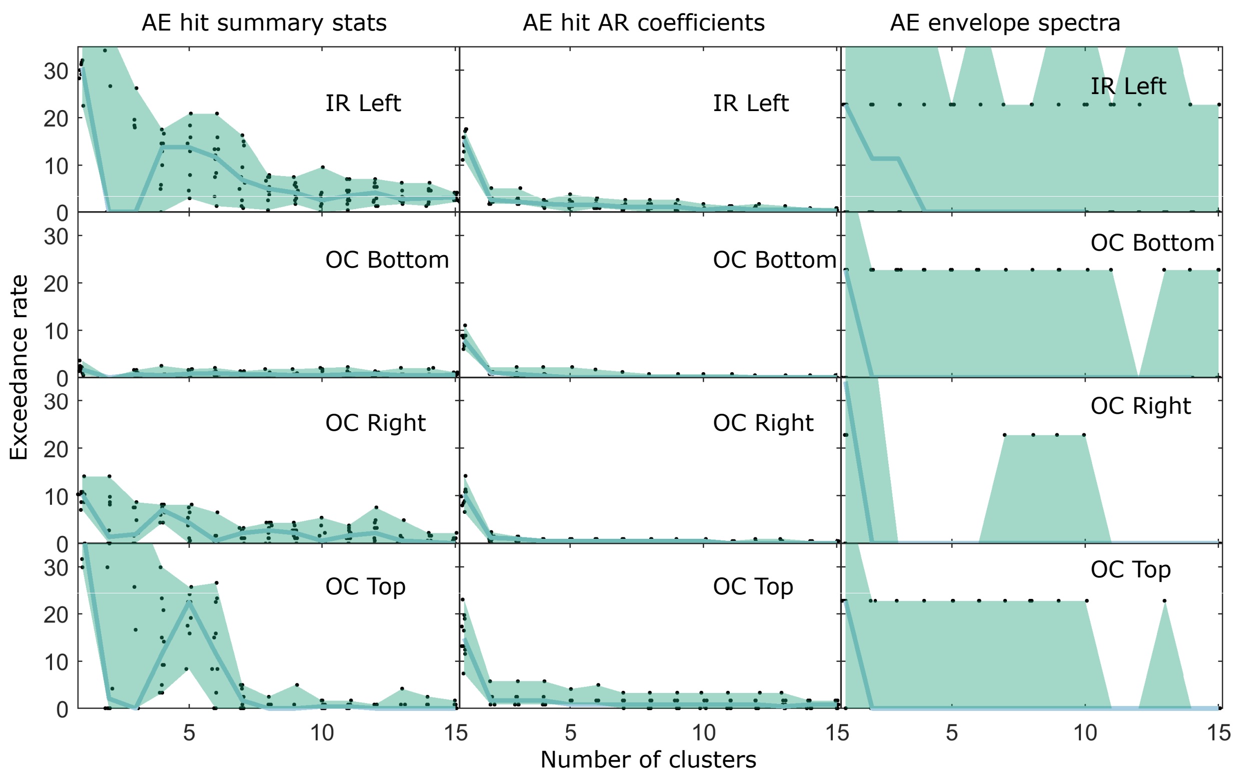

Figure 15: 10-fold cross-validated output of GMM misclassification rate with increasing number of clusters, showing results for the three damage sensitive features at the four sensor locations.

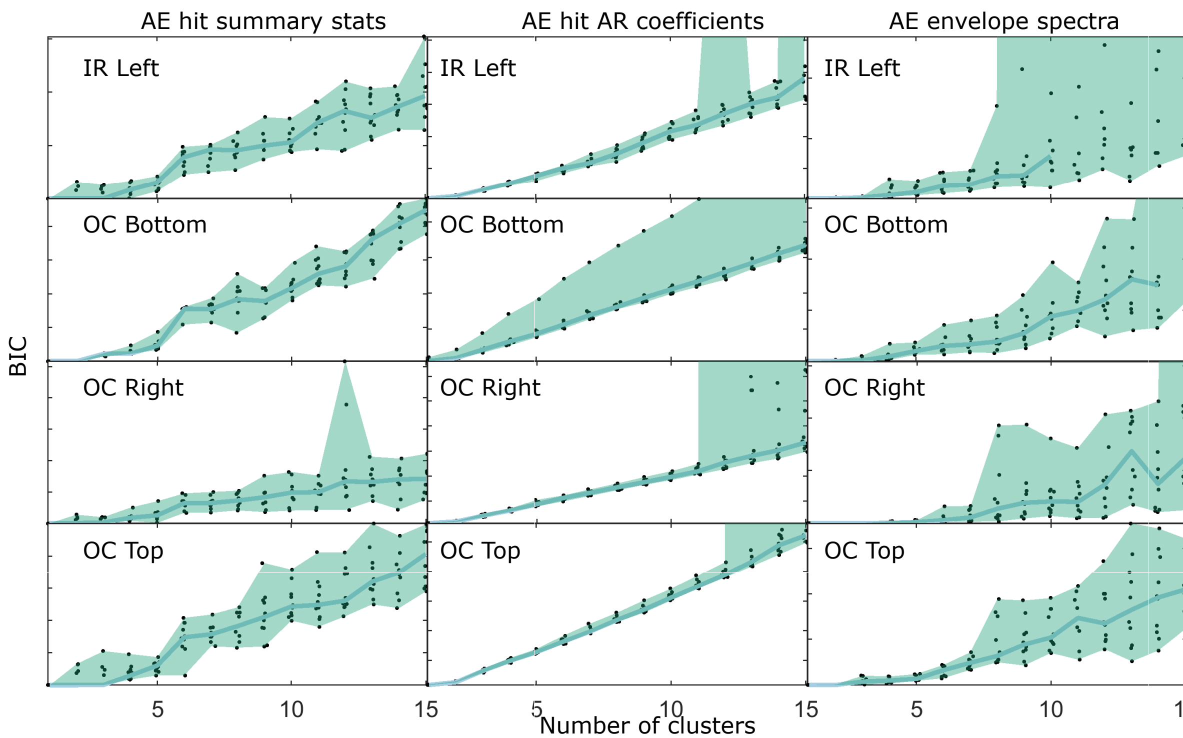

Figure 16: 10-fold cross-validated output of GMM Bayesian Information Criterion (BIC) with increasing number of clusters, showing results for the three damage sensitive features at the four sensor locations.

importance of an appropriate threshold is highlighted in here. Note that there clearly an appreciable shift in probability mass early on in the development of damage. However, the detection threshold represents denotes the point beyond which probability mass of the training features will be negligibly small. If the training features have high variance, this will lead to larger threshold, hence lower detectability, even when this may be visible by a visual comparison of the densities of the novelty indices. This is the case for the AE envelope spectra, where even though there is a considerable shift in probability mass of the NLL, the original high variance of the features places the threshold at a relatively high position. This highlights the difference and difficulty of detecting damage using no prior information of the damage process, as opposed to a retrospective analysis with knowledge of damaged states.

It is impossible to accurately quantify the detection performance based on individual hit results, simply because one does not have access to the ground truth of whether a specific AE burst of energy was generated due to damage or due to a "benign" process within the gearbox. In this case, the available ground truth is the bearing condition of each 10-second trial. For this reason, summarising features of the NLL of each individual trial are used to quantify false positive and negative rates. Two such summarising features are presented here. Noting that it is the extremes of the NLL that flag novelty, Figure 18 shows the maxima of the (log) NLL for the OC top sensor location, for each individual trial and for the three features. The vertical divisions in Figure 18 mark the different bearing conditions, while the horizontal dashed lines indicate the detection threshold. As before, the bottom row shows the average applied compressive load and casing temperature.

For the three features considered, the maxima (log) NLL clearly capture the behaviour of the undamaged process, as the majority of the trials of UD1 and UD2 fall under the detection threshold. This is a robust validation of the overall detection methodology, given there are minimal false positives in the validation set (UD2). In this case, the only false positives come from the AR coefficient features, and at the highest temperature observed in UD2. All three detectors fail to identify the presence of the lowest level of subsurface damage (D1). The second level of subsurface damage, D2, is detectable by the three features, albeit only at high loads. It is reasonable to conclude that it is the applied load that drives the detection, since tests with high temperature but low load have low detection rates. Moving upwards in the damage scale, all three surface damage conditions, D3,D4 and D5 are detectable with the three features. However, note that they all have different degrees of success at this. In general, AE hit statistics and envelope spectra do not detect well under low loads. The AR coefficients on the other hand begin to detect the damage at the lower loads, although only for the most severe of the surface damage conditions.

Considering only the maxima (log) NLL of each trial is still prone to an increased rate of false positives seeing it is likely that even in an undamaged state, rare AE events will be generated that will drive one of the feature vectors to have a high novelty index. This has the potential to judge an entire observation set based on one erratic event while ignoring the information contained in the thousands of other feature vectors contained in that time window. A more robust way of quantifying detection performance would be through the exceedance rate of feature vectors above the detection threshold. This, in effect, quantifies the probability mass of the NLL that falls above the detection threshold, as was illustrated using Figure 17. A positive trial is defined as one where the exceedance rate above the detection threshold falls above the exceedance rate observed on the training set. Using this definition, the detection rate for all sensor locations and feature vectors is given in Figure 19. A positive detection rate on UD1 and UD2 implies a false positive, while the same implies a true positive on the damaged conditions. Overall, it is possible to conclude that all three feature vectors are capable of detecting from the second level of subsurface damage (D2) onwards, with varying degrees of success depending on the sensor location. The AE envelope spectra is overall the worst performing, missing D2 altogether on the OC right location, and with overall low true positive rates. This is attributed to the high variability of the feature vectors, as seen in Figure 12. The AR coefficients, while having overall the highest true positive rate across all locations, also have the highest false positive rate on the validation set (condition UD2), which is undesirable.

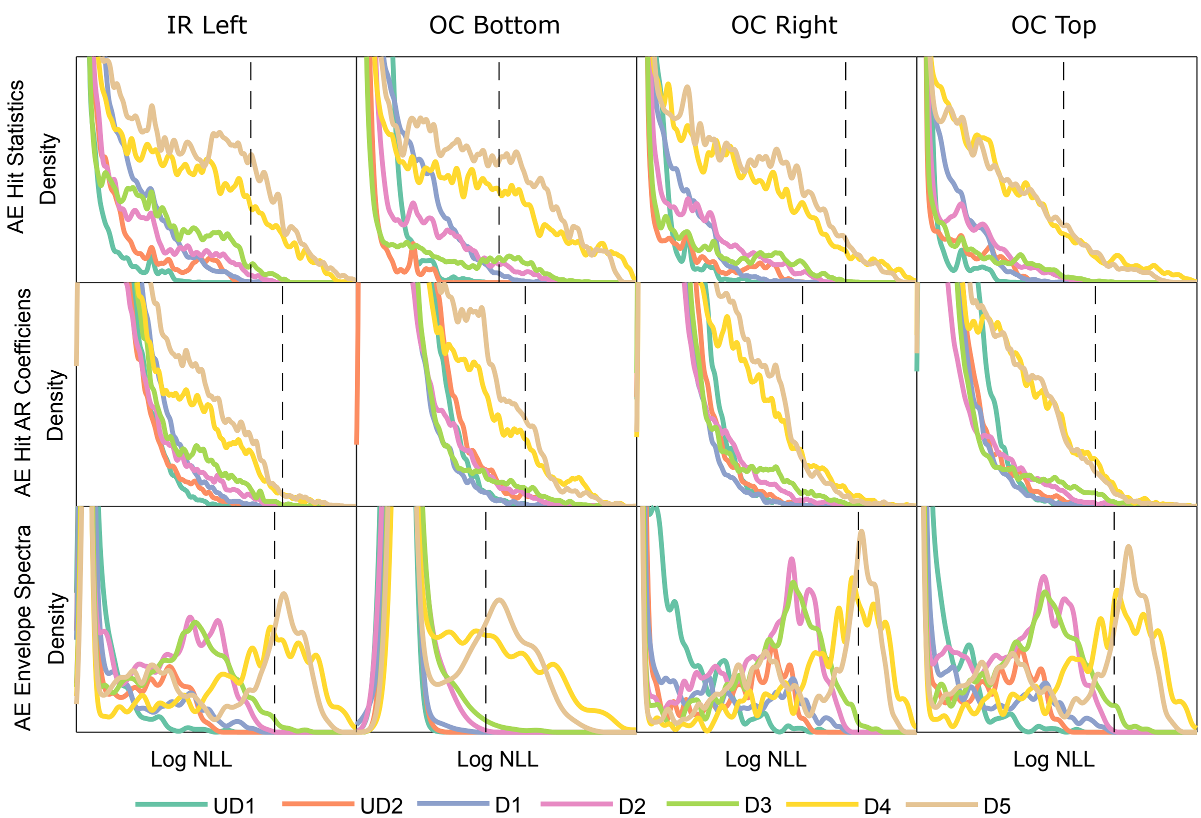

Figure 17: Kernel density of log negative-log-likelihood of GMM model, evaluated on the three different AE features and on the four sensor nodes, for different damage states. The scales have been normalised and adjusted to highlight the tail of the distributions. The vertical dashed lines represent the detection threshold.

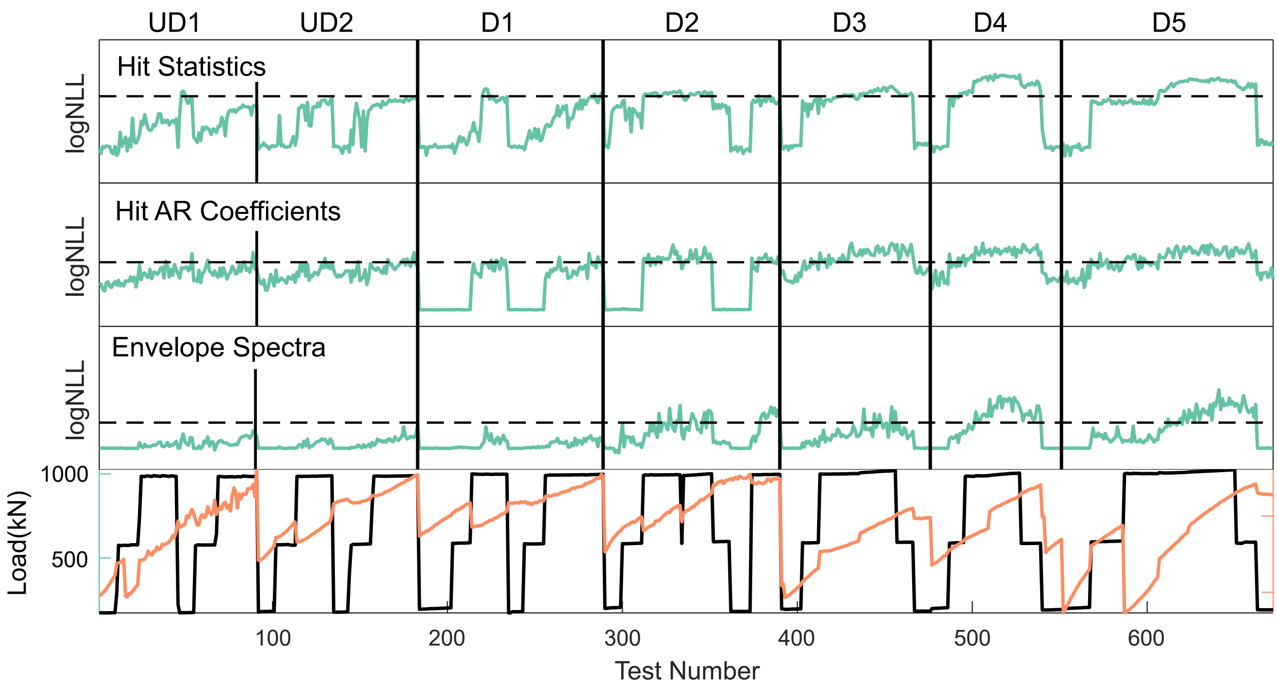

Figure 18: Maxima of the (Log) NLL for each trial, for different bearing conditions and the three different damage-sensitive features (along rows). The horizontal dashed lines represent the detection threshold. Bottom row shows the applied compressive load and casing temperature.

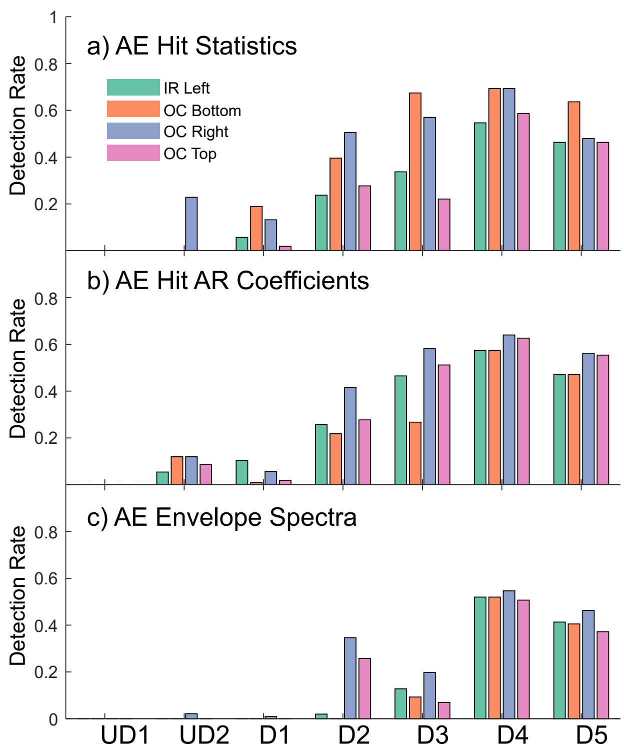

Figure 19: Detection rate for the four different AE sensor locations and the three damage-sensitive feature considered.

6. Conclusions

This paper has investigated the problem of detection of sub-surface damage in Wind Turbine bearings, applying probabilistic modelling to features extracted from Acoustic Emission measurements. Detecting sub-surface damage on an operational bearing is not a trivial problem. It is however, an important problem to tackle. Currently, the majority of gearbox failures can be attributed to bearing failures and this, in turn, accounts for most of the down-time of WTs globally. Damage in bearings starts under the surface, as a result of Hertzian contact mechanics. While a lot of effort has been placed in investigating and developing monitoring strategies for surface damage, sub-surface damage is harder to detect and has received much less attention. This paper focused on this problem. An emphasis has been placed in using measurements that can be used in practice, so AE measurements have been taken from practical sensing locations, at the outer casing of a bearing. An experimental rig was devised in order to replicate the operational environment of a planetary bearing inside an epicyclic gearbox, as these suffer the most from early failure before the end of their prescribed fatigue life.

Even though it is sub-surface damage that was of primary interest in this study, early-stage surface defects were also investigated. A total of five levels of damage were used, two sub-surface and three early-stage surface defects. This allows to investigate the detectability of damage throughout its progression. It also helps to build confidence in the detection scheme by observing that detection rates are higher for larger or more severe defects.

Measurements were taken at four different locations; one sensing location at the inner raceway bearing, close to the location of seeded faults and three around the circumference of the outer casing of the bearing rig. AE data were collected from each sensor and processed separately in order to evaluate the detectability of the different levels of defects at each location. Detection of damage was carried out by first extracting three different damage-sensitive features from the raw AE data, and then fitting a probabilistic model to perform novelty detection. The features chosen in this investigation were hit summary statistics, Auto Regressive (AR) coefficients of the individual AE hit time histories, and the envelope spectra of the raw AE signals. These features all capture a different type of information contained in the AE data. The hit-based summary statistics contain information about the average energy, duration and the distance each individual stress wave has travelled. The hit AR coefficients on the other hand provide a greater level of detail as to the spectral characteristics of the waveforms. Finally, the envelope spectra capture the amplitude modulation of the signal, so any periodic bursts of energy, which is a characteristic manifestation of damage in dynamic response data, should be evident in this feature.

The progression of early bearing failure has been illustrated qualitatively, using a Principal Component Analysis (PCA) projection of the three damage sensitive features used here. This is shown in Figure 13. In the case of the three different damage-sensitive features investigated, it is clear that the larger surface-level defects are evident through a clear change in the probability distribution of the features. This is easy to spot qualitatively. On the other hand, the change in all damage-sensitive features arising from sub-surface damage is not necessarily clear from visual examination. Furthermore, in the presence of high levels of noise common in rotating machinery, it is difficult to establish whether a specific burst of AE energy has been generated by a damage process or belongs to the background noise. This motivates the need for a principled statistical approach to the detection problem. The problem is therefore treated as one of inference under a probabilistic model. In this particular case, because the data were gathered under a range of operational conditions, a Gaussian Mixture Model was used for this task. A GMM was fitted to the damage-sensitive features from an undamaged bearing and the Negative Log Likelihood (NLL) of the model was used as a novelty index.

They key result of this investigation is that it is clearly possible to identify sub-surface damage from a practical measurement location, at the casing of a planetary gearbox bearing. Using the probabilistic framework presented in this paper, it is possible to perform such detection under changing environmental and operational conditions. One of the key aspects of AE activity within a bearing environment is that it is highly dependent on applied load and temperature, as this affects the lubrication properties. The methods developed here have proved to be robust against these challenges. This paper validates the approach under an experimental rig where the environment can be carefully controlled; This is important as temperature, load and lubrication affect the background AE response and hence the detectability of defects. A clear next step in this research would be to validate the detectability and the probabilistic approach on an operational wind turbine, where although there may be less control over the operational parameters, background noise from gearboxes and other source would be more realistic.

7. Acknowledgements

The authors would like to acknowledge the financial support of Ricardo Innovations thorough the sponsorship of several PhD programmes as well as the bearing test rig. R. Dwyer-Joyce would like to acknowledge the EPSRC for funding part of this research through the fellowship on Tribo-Acoustic Sensors EP/N016483/1. Support for E.J. Cross through grant number EP/S001565/1 is also acknowledged.

8. References

[1] F. Oyague, “NREL/TP-500-41160 gearbox modeling andlLoad simulation of a baseline 750-kW wind turbine using state-of-the-art simulation codes,” tech. rep., National Renewable Energy Laboratory, 2009.

[2] W. Yang, P. J. Tavner, C. J. Crabtree, Y. Feng, and Y. Qiu, “Wind turbine condition monitoring: technical and commercial challenges,” *Wind Energy*, vol. 17, pp. 673–693, may 2014.

[3] W. Qiao and D. Lu, “A survey on wind turbine condition monitoring and fault diagnosisPart I: components and subsystems,” *IEEE Transactions on Industrial Electronics*, vol. 62, pp. 6536–6545, oct 2015.

[4] “IEC 61400-1 Wind turbines - Part 1: Design requirements,” 2005.

[5] J. Ribrant and L. M. Bertling, “Survey of sailures in wind power systems with focus on Swedish wind power plants during 19972005,” *IEEE Transactions on Energy Conversion*, vol. 22, pp. 167–173, mar 2007.

[6] D. Coronado and J. Wenske, “Monitoring the oil of wind-turbine gearboxes: main degradation indicators and detection methods,” *Machines*, vol. 6, no. 2, p. 25, 2018.

[7] M. P. Barrett and J. Stover, “Understanding oil analysis: how can it improve reliability of wind turbine gearboxes,” *Gear Technology*, 2013.

[8] C. R. Farrar and K. Worden, *Structural health monitoring: a machine learning perspective*. John Wiley & Sons.

[9] T. Howard, *Development of a novel bearing concept for improved wind turbine gearbox reliability*. PhD thesis, The Univerisity of Sheffield, 2015.

[10] K. Tamada and H. Tanaka, “Occurrence of brittle flaking on bearings used for automotive electrical instruments and auxiliary devices,” *Wear*, vol. 199, pp. 245–252, nov 1996.

[11] M. H. Evans, “White structure flaking (WSF) in wind turbine gearbox bearings: effects of ’butterflies’ and white etching cracks (WECs),” *Materials Science and Technology*, vol. 28, no. 1, pp. 3–22, 2012.

[12] T. Bruce, E. Rounding, H. Long, and R. Dwyer-Joyce, “Characterisation of white etching crack damage in wind turbine gearbox bearings,” *Wear*, vol. 338, pp. 164–177, 2015.

[13] M.-H. Evans, A. Richardson, L. Wang, R. Wood, and W. Anderson, “Confirming subsurface initiation at non-metallic inclusions as one mechanism for white etching crack (WEC) formation,” *Tribology International*, vol. 75, pp. 87–97, jul 2014.

[14] M.-H. Evans, A. Richardson, L. Wang, and R. Wood, “Effect of hydrogen on butterfly and white etching crack (WEC) formation under rolling contact fatigue (RCF),” *Wear*, vol. 306, pp. 226–241, aug 2013.

[15] D. Alleyne and P. Cawley, “The interaction of Lamb waves with defects,” *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, vol. 39, pp. 381–397, may 1992.

[16] S. Doebling, C. R. Farrar, B. Prime, M, and D. Shevitz, “Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: A literature review,” 1996.

[17] K. Worden, C. R. Farrar, G. Manson, and G. Park, “The fundamental axioms of structural health monitoring,” *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 463, pp. 1639–1664, 2007.

[18] J. Baram and M. Rosen, “Fatigue life prediction by distribution analysis of acoustic emission signals,” *Materials Science and Engineering*, vol. 41, no. 1, pp. 25–30, 1979.

[19] C. U. Grosse and M. Ohtsu, *Acoustic Emission Testing.* Gardners Books, 2010.

[20] C. R. Heiple and R. O. Adams, “Acoustic Emission produced by deformation of metals and alloys - Ar review: Part I,” *Journal of Acoustic Emission*, vol. 6, no. 6, pp. 177–204, 1987.

[21] R. Fuentes, T. P. Howard, M. B. Marshall, E. J. Cross, and R. S. Dwyer-Joyce, “Observations on Acoustic emissions from a line contact compressed into the plastic region,” *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, vol. 230, no. 11, pp. 1371–1376, 2016.

[22] K. M. Holford, R. Pullin, S. L. Evans, M. J. Eaton, J. Hensman, and K. Worden, “Acoustic emission for monitoring aircraft structures,” *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 223, pp. 525–532, aug 2009.

[23] J. Hensman, K. Worden, M. Eaton, R. Pullin, K. Holford, and S. Evans, “Spatial scanning for anomaly detection in acoustic emission testing of an aerospace structure,” *Mechanical Systems and Signal Processing*, vol. 25, no. 7, pp. 2462–2474, 2011.

[24] J. Kaiser, *Untersuchungen über das auftreten von geräuschen beim zugversuch*. PhD thesis, Technical University of Munich (TUM), 1950.

[25] V. M. V. M. Baranov, *Acoustic emission in friction*. Elsevier Science, 2007.

[26] A. Cockerill, A. Clarke, R. Pullin, T. Bradshaw, P. Cole, and K. M. Holford, “Determination of rolling element bearing condition via acoustic emission,” *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, vol. 230, no. 11, pp. 1377–1388, 2016.

[27] T. Yoshioka, “Detection of rolling contact sub-surface fatigue cracks using acoustic emission technique,” *Lubrication Engineering*, vol. 49, no. 4, pp. 303–308, 1993.

[28] J. Shiroishi, Y. Li, S. Liang, T. Kurfess, and S. Danyluk, “Bearing condition diagnostics via vibration and acoustic emission measurements,” *Mechanical Systems and Signal Processing*, vol. 11, pp. 693–705, sep 1997.

[29] N. Jamaludin, D. Mba, and R. H. Bannister, “Condition monitoring of slow-speed rolling element bearings using stress waves,” *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering*, vol. 215, pp. 245–271, nov 2001.

[30] A. Morhain and D. Mba, “Bearing defect diagnosis and acoustic emission,” *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, vol. 217, pp. 257–272, apr 2003.

[31] K. R. Al-Balushi, A. Addali, B. Charnley, and D. Mba, “Energy index technique for detection of Acoustic Emissions associated with incipient bearing failures,” *Applied Acoustics*, vol. 71, pp. 812–821, sep 2010.

[32] B. Eftekharnejad, M. Carrasco, B. Charnley, and D. Mba, “The application of spectral kurtosis on Acoustic Emission and vibrations from a defective bearing,” *Mechanical Systems and Signal Processing*, vol. 25, pp. 266–284, jan 2011.

[33] D. Pandya, S. Upadhyay, and S. Harsha, “Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN,” *Expert Systems with Applications*, vol. 40, pp. 4137–4145, aug 2013.

[34] P. Kankar, S. C. Sharma, and S. Harsha, “Fault diagnosis of rolling element bearing using cyclic autocorrelation and wavelet transform,” *Neurocomputing*, vol. 110, pp. 9–17, jun 2013.

[35] R. Fuentes, T. Howard, E. J. Cross, R. Harald-Hestmo, T. Huntley, B. Marshall, Mathew, and R. Dwyer-Joyce, “Detecting damage in wind turbine bearings using acoustic emissions and Gaussian process latent variable models,” in *Proceedings of the 10th International Worklshop in Structural Health Monitoring*, (Stanford University, Palo Alto, CA), 2015.

[36] J. R. Naumann, *Acoustic emission monitoring of wind turbine bearings*. PhD thesis, The University of Sheffield, 2015.

[37] M. Elforjani and D. Mba, "Assessment of natural crack initiation and its propagation in slow speed bearings," *Nondestructive Testing and Evaluation*, vol. 24, pp. 261–275, sep 2009.

[38] Z. Rahman, H. Ohba, T. Yoshioka, and T. Yamamoto, "Incipient damage detection and its propagation monitoring of rolling contact fatigue by acoustic emission," *Tribology International*, vol. 42, pp. 807–815, jun 2009.

[39] Z. Zhi-qiang, L. Guo-lu, W. Hai-dou, X. Bin-shi, P. Zhong-yu, and Z. Li-na, "Investigation of rolling contact fatigue damage process of the coating by acoustics emission and vibration signals," *Tribology International*, vol. 47, pp. 25–31, mar 2012.

[40] B. Kilundu, X. Chiementin, J. Duez, and D. Mba, "Cyclostationarity of Acoustic Emissions (AE) for monitoring bearing defects," *Mechanical Systems and Signal Processing*, vol. 25, pp. 2061–2072, aug 2011.

[41] A. Rai and S. Upadhyay, "A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings," *Tribology International*, vol. 96, pp. 289–306, apr 2016.

[42] C. Jiaa and D. Dornfeld, "Experimental studies of sliding friction and wear via acoustic emission signal analysis," *Wear*, vol. 139, pp. 403–424, aug 1990.

[43] I. Antoniadou, T. Howard, R. Dwyer-Joyce, M. Marshall, J. Naumann, N. Dervilis, and K. Worden, "Envelope analysis using the Teager-Kaiser energy operator for condition monitoring of a wind turbine bearing," in *Applied Mechanics and Materials*, vol. 564, pp. 170–175, 2014.

[44] SKF, "NU2244 Cylindrical Roller Bearing," 2018.

[45] O. Reynolds, "On the theory of lubrication and Its application to Mr. Beauchamp tower's experiments, including an experimental determination of the viscosity of olive oil," *Proceedings of the Royal Society, London*, vol. 40, pp. 191–203, 1886.

[46] D. P. Hess and A. Soom, "Friction at a lubricated line Contact operating at oscillating sliding velocities," *Journal of Tribology*, vol. 112, no. 1, p. 147, 1990.

[47] S. Mallat, *A Wavelet Tour of Signal Processing*. 2009.

[48] J. H. Kurz, C. U. Grosse, and H. W. Reinhardt, "Strategies for reliable automatic onset time picking of acoustic emissions and of ultrasound signals in concrete," *Ultrasonics*, vol. 43, no. 7, pp. 538–546, 2005.

[49] R. B. Randall, *Vibration-based Condition Monitoring*. Chichester, UK: John Wiley & Sons, Ltd, jan 2011.

[50] C. C. Aggarwal, *Outlier Analysis*. New York, NY: Springer New York, 2013.

[51] K. Worden, G. Manson, and N. R. J. Fieller, "Damage detection using outlier analysis," *Journal of Sound and Vibration*, vol. 229, no. 3, pp. 647–667, 2000.

[52] N. Dervilis, E. Cross, R. Barthorpe, and K. Worden, "Robust methods of inclusive outlier analysis for structural health monitoring," *Journal of Sound and Vibration*, vol. 333, pp. 5181–5195, sep 2014.

[53] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, p. 46, 2001.

[54] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, vol. 2nded. Springer series in statistics, 2001.

[55] C. M. Bishop, *Pattern recognition and machine learning*. Springer-Verlag New York, 2006.

[56] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models.," *Neural computation*, vol. 11, no. 2, pp. 305–345, 1999.

[57] S. Roweis, "EM Algorithms for PCA and SPCA," *Computing*, vol. 10, no. 13, pp. 626–632, 1997.

[58] C. M. Bishop, “Latent variable models,” *Published in Learning in Graphical Models*, pp. 371–403, 1999.

[59] E. Figueiredo and E. Cross, “Linear approaches to modeling nonlinearities in long-term monitoring of bridges,” *Journal of Civil Structural Health Monitoring*, vol. 3, no. 3, pp. 187–194, 2013.

[60] E. Figueiredo, L. Radu, K. Worden, and C. R. Farrar, “A Bayesian approach based on a Markov-chain Monte Carlo method for damage detection under unknown sources of variability,” *Engineering Structures*, vol. 80, pp. 1–10, 2014.

[61] J. Kullaa, “Structural health monitoring under nonlinear environmental or operational influences,” *Shock and Vibration*, vol. 2014, 2014.

[62] R. Fuentes, *On Bayesian Networks for Structural Health and Condition Monitoring*. PhD thesis, 2017.

[63] A. Dempster, N. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society Series B Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[64] R. E. Kass and L. Wasserman, “A reference bayesian test for nested hypotheses and its relationship to the Schwarz criterion,” *Journal of the American Statistical Association*, vol. 90, p. 928, sep 1995.

[65] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, “A review of novelty detection,” *Signal Processing*, vol. 99, pp. 215–249, jun 2014.

[66] H. Sohn, D. W. Allen, K. Worden, and C. R. Farrar, “Structural damage classification using extreme value statistics,” *Journal of Dynamic Systems, Measurement, and Control*, vol. 127, p. 125, mar 2005.

[67] D. A. Clifton, S. Hugueny, and L. Tarassenko, “Novelty detection with multivariate extreme value statistics,” *Journal of Signal Processing Systems*, vol. 65, pp. 371–389, dec 2011.

[68] M. Aitkin and D. Clayton, “The fitting of exponential, Weibull and Extreme Value Distributions to complex censored survival data Using GLIM,” *Applied Statistics*, vol. 29, no. 2, p. 156, 1980.

[69] H. W. Park and H. Sohn, “Parameter estimation of the generalized extreme value distribution for structural health monitoring,” *Probabilistic Engineering Mechanics*, vol. 21, no. 4, pp. 366–376, 2006.