INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

SURVISHNO

SURVEILLANCE
VIBRATIONS
SHOCKS
NOISE

Vibes in action

8-10 JULY 2019
INSA LYON

Proceedings

BOOK OF PROCEEDINGS

SURVISHNO 2019

A joint organisation of the conferences
Surveillance, Vibration Shock and Noise, and
Experimental Vibration Analysis

General Chair : J. Antoni, D. Rémond

Table of contents

Wednesday

Angular approaches

Angular vibration on-site measurements and application to torsional analysis on industrial cases

François COMBET, Louis BOISSELEAU

DYNAE, rue Condorcet, 38090 Villefontaine, France
francois.combet@eiffage.com

Abstract

The measurement and analysis of torsional phenomena remains an uncommon and challenging task to perform in the industry. The measurement of the torque via strain gauges provides explicit results but can be difficult to implement on-site and is generally limited to a low rotation speed. The motor current measurement is easier to perform but does not always reflect the torque instantaneous variations. The measurement of the Instantaneous Angular Speed (IAS) presents an interesting alternative as several techniques exist which are relatively easy to install on-site. This is classically performed using optical encoders or magnetic pickup sensors. A lesser known technique is based on Laser Torsional Vibrometry (LTV), using parallel beam laser vibrometers, which has the advantage of being totally non-intrusive. One difficulty however of IAS based diagnosis techniques is then to interpret the IAS amplitude, due to the lack of rules and criteria in this domain, on contrary to translational vibration. Advantages of these different techniques for IAS on-site measurements are discussed, and applications are then presented on industrial cases: the test and certification of the coupling of a fuel injection pump and the torsional analysis of the flexible coupling of a Diesel-generator group.

1 Introduction

The measurement and analysis of torsional phenomena remains an uncommon and challenging task to perform in the industry. The measurement of the transmitted torque via strain gauges provides explicit results but can be difficult to implement on-site and is generally limited to a low rotation speed. The motor current measurement is an easier technique to perform but was shown to not always reflect the torque instantaneous variations, due to the filtering effect of the shaft line inertia on the higher frequency torque & speed variations. The measurement of the instantaneous angular speed (IAS) presents an interesting alternative as several techniques exist which are relatively easy to install on-site. This is classically performed using optical encoders or magnetic pickup sensors. The analysis of the IAS variations was shown to allow the detection of roller bearing faults [3-4]. Another technique is based on Laser Torsional Vibrometer (LTV), using parallel beam laser vibrometers, which has the advantage of being totally non-intrusive [5-6].

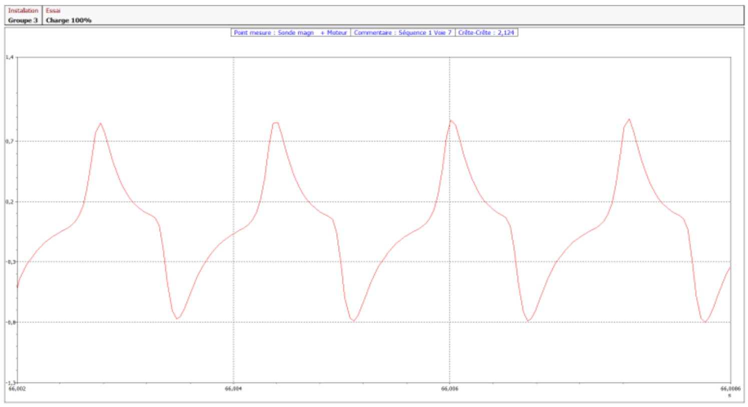
We first discuss the advantages of different techniques for IAS on-site measurements. We then present their application on industrial cases: first the test and certification of the coupling of a fuel injection pump, then the torsional analysis of a flexible coupling of Diesel-generator group.

2 Angular vibration measurement techniques

2.1 Use of impulse sensors (optical encoders & magnetic pickups)

The use of impulse sensors for IAS measurement from strip bands or optical encoders is quite common and well documented. The computation of the IAS can then be performed by the elapsed time method [8].

The IAS can also be measured by a magnetic pickup. In this case the computation is preferably done by frequency demodulation, as the measured signal is non-sinusoidal and modulated in amplitude (depending on the rotation speed).

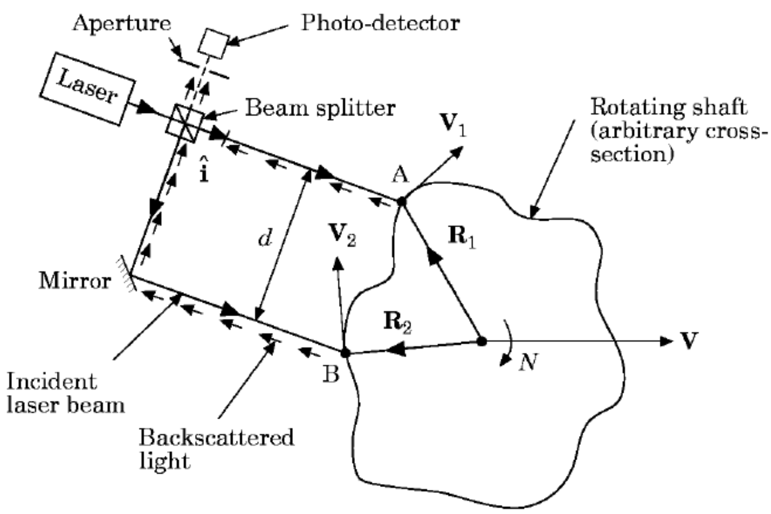


Signal measured by a magnetic pickup

2.2 The Laser Torsional Vibrometer

The Laser Torsional Vibrometer (LTV) is another technique that offers *in situ* measurement, thus avoiding machinery downtime and can function on rotating components of arbitrary shape. The optical geometry used makes the instrument insensitive to solid body oscillation of the target or operator as well as to the cross-sectional shape of the component [5].

The laser beam with wavelength λ is divided into two equal intensity parallel beams separated by distance d:



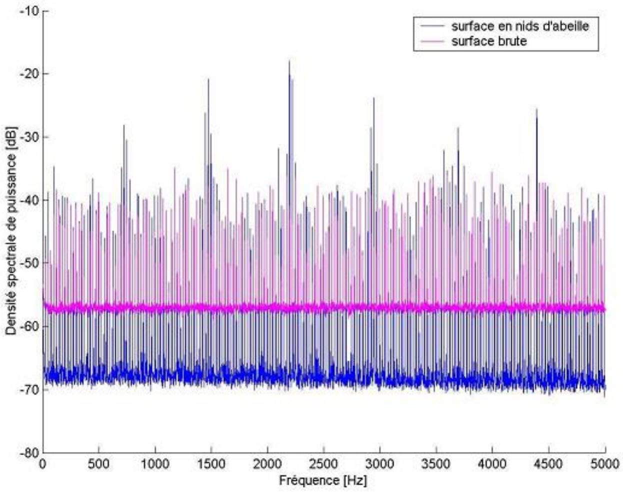
The laser torsional vibrometer optical geometry (picture taken from [5])

The 'beat' frequency between the two backscattered light beams received by the photo-detector corresponds to a Doppler frequency *fD*, which is directly proportional to the speed of rotation Ω of the target component:

$$fD = (2d / \lambda)\,\Omega$$

the fluctuating part of which is the IAS variation. The frequency response of the instrument is dictated by that of the demodulation system used and the usual bandwidth of practical interest is up to 10 kHz.

In practice however this technique is limited by the Signal / Noise ratio of the signal, which is influenced by a run-out phenomenon induced by the target surface roughness. This creates a speckle pattern periodicity at the shaft rotation. For use at very low levels of torsional vibration the speckle pattern periodicity can be attenuated by modulating the spatial position of the incident laser beams from side to side in a random manner.



Train of harmonic components induced by the run-out phenomenon

It should be noted that the setting and calibration of the LTV must be performed *in situ*, i.e. while the shaft is rotating, at least at a low speed.

Note that the LTV also offers the possibility of successive IAS measurement on different sections of the shaft line while the equipment is running. If a phase reference (tachymeter) is used, a torsional Operational Deformed Shape (ODS) of the shaft line can thus be performed.

2.3 In search of criteria and guiding rules for IAS assessment

One difficulty when dealing with angular vibrations is the lack of criterion in the literature in order to assess the maximum allowed amplitude of the measured IAS.

For combustion engines there is an industry regulation for the allowable magnitude of the Peak-Peak twist angle of the crankshaft, which must be below **0.4° PP** [7]. Some constructors have also defined their own criteria, depending on their experience. Note that Pr Nerubenko underlines in [7] that a *torsional vibration* is an angular vibration which implies a *twist* of the shaft, and must not be confounded with the *fluctuation of the rotation speed*, or the RPM non-regularity, which is another dynamical problem. The latter may be defined by the coefficient:

$$C = \Omega\_Peak\text{-}Peak \,/\, \Omega\_mean$$

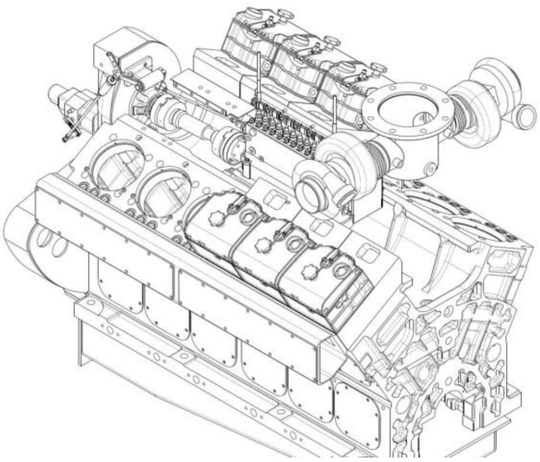
For combustion engines the guide numbers for coefficient *C* are in the range **0.6–1.2%**. Apart from the combustion engine industry we cannot seem to find any other criterion for the angular vibrations.

3 Industrial case studies

We present here applications of IAS measurement techniques to the torsional analysis of industrial equipments on two case studies.

3.1 Certification of a fuel injection pump coupling

This first application deals with a fuel injection pump of a 12V Diesel engine used on a ship. After installation of a new type of injection pump, the coupling of the pump happened to repeatedly break and was changed after only a few hours of service. It was then decided to test other types of coupling and to perform torsional vibration measurements, as a torsional resonance was suspected to be the cause of the damage.

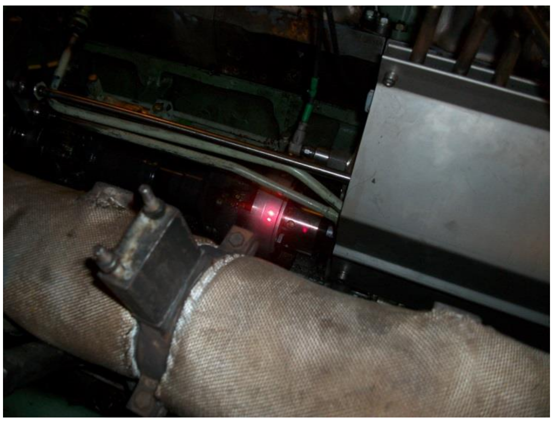


Injection pump installed in the V of the engine

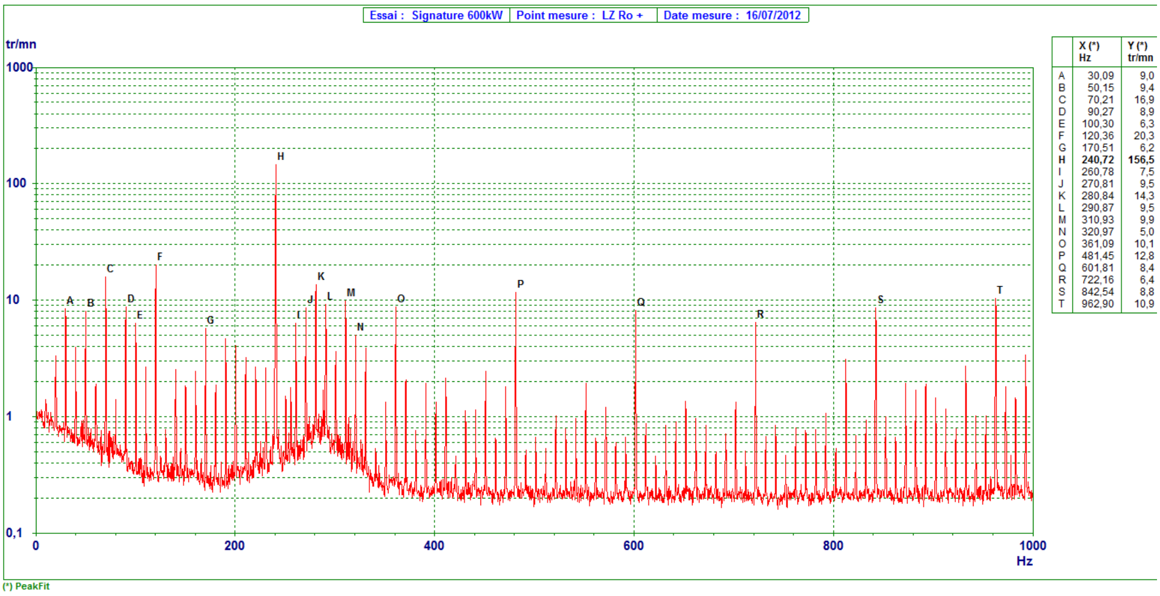
The pump is installed in the 'V" of the engine and is run by a cardan coupling, which is run itself by the pinion cascade of the distribution. The methodology was as follow:

- Measurement with the initial coupling on the Diesel engine in order to qualify the existing system,
- Measurement with different couplings on a test bench and selection of that showing the best results,
- Measurement with the selected coupling on the ship for validation.

The measurements with the initial coupling were performed with the LTV device aiming the coupling of the injection pump, which is running at 600rpm:



The spectrum of the instantaneous angular speed (IAS) signal is showing an elevation of the background noise around 280Hz, which seems to amplify the spectral component at 2X the injection frequency (*Finj*=120Hz):

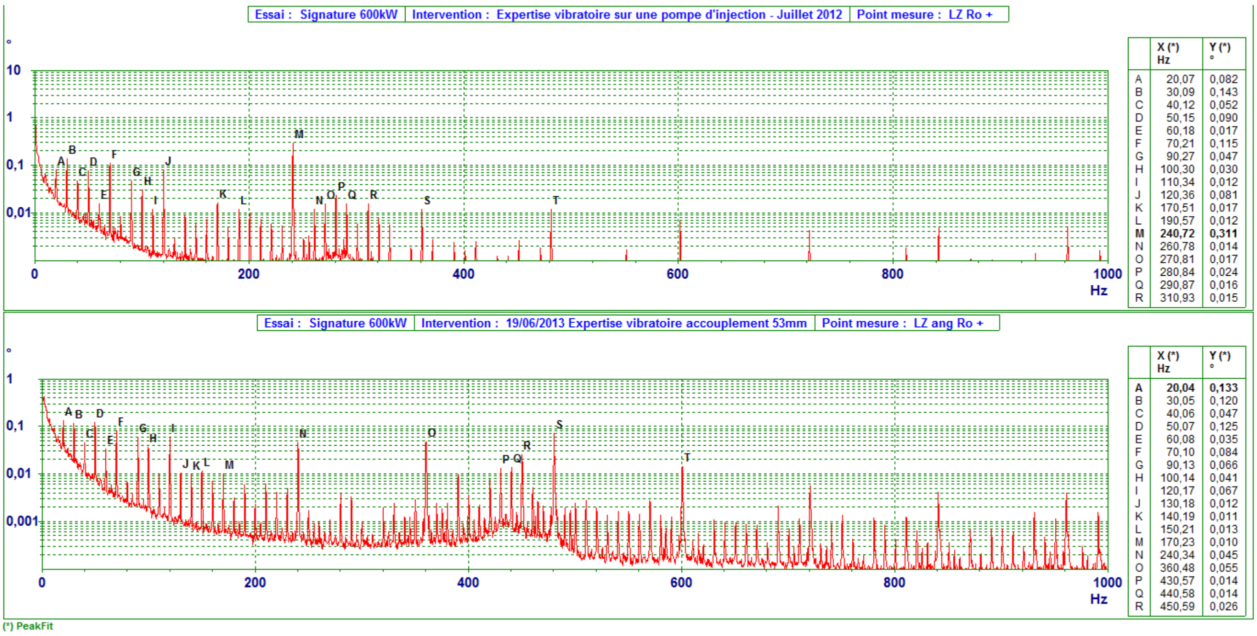


Spectrum of the IAS

The amplitude of the IAS variation at 2X Finj is 156.5rpm Peak-Peak, corresponding to an angular variation of 0.31° 0-P. A limit value was defined empirically for any component of the integrated IAS spectrum, at **0.25° 0-P**.

The injection pump was then tested on a test bench with different types of couplings (1 cardan & 3 disk couplings with different diameters). The best results were obtained with a 53mm diameter disk-coupling.

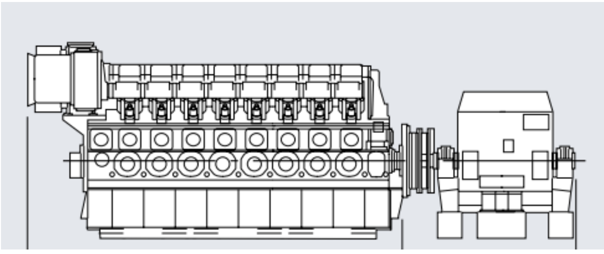
This coupling was then mounted on the engine on board for an endurance test. The amplitudes of the IAS spectral components are significantly lower compared to the original coupling and below the criterion of 0.25° 0-P. They also remain stable after 4 months of service.



Comparison of the angular vibration spectra (in ° 0-P) with the original (top) and new coupling (bottom)

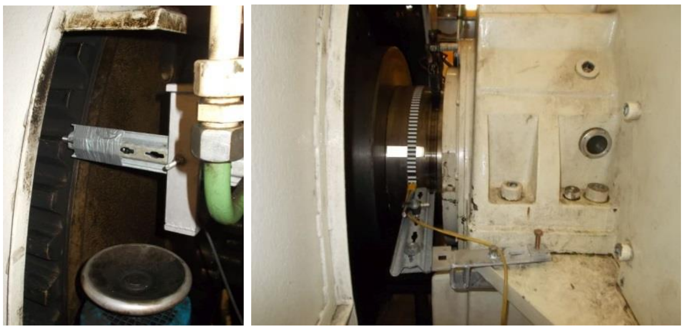
3.2 Torsional analysis of a flexible coupling

This second case study deals with the torsional analysis of a coupling of a high power Diesel group. The equipment is constituted by an 18V Diesel engine driving a generator via a flexible coupling. The shaft line of the group is running at 500rpm.



Diesel engine and generator group

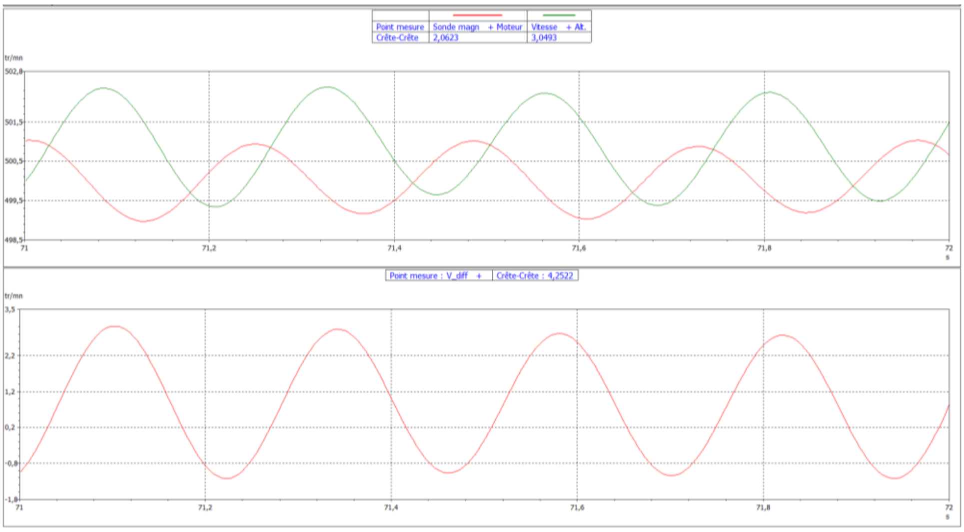
After a sudden and destructive damage of the coupling, it was decided to perform IAS measurement on each side of the coupling in order to analyse its torsional behaviour in service. On the generator side the IAS was measured by using an optical sensor and a strip band with 125 pulses. On the engine side we used a magnetic pickup placed near the toothed wheel of the engine (74 teeth).



Speed measurements on the engine and on the generator

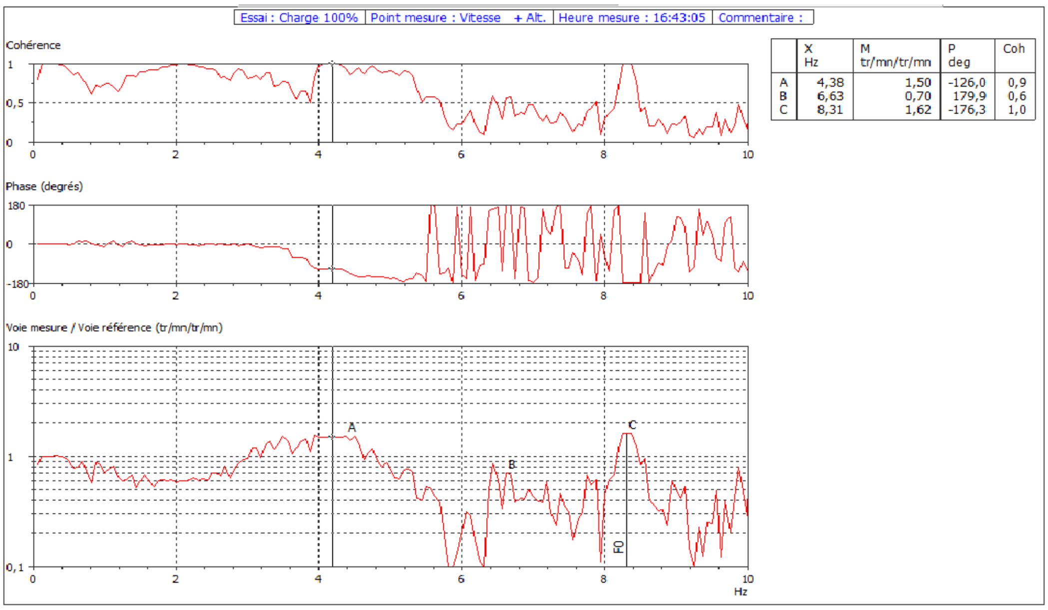
Note that the imperfect junction of the strip band stuck on the generator shaft is inducing periodic spikes at the rotation frequency. These spikes can then be suppressed by a dedicated algorithm. Besides, an advantage of these spikes is to provide a 1X/rev phase reference on the generator shaft.

The analysis of the IAS when the group is running at full load reveals an IAS fluctuation at the combustion cycle frequency of the Diesel engine (i.e. half the rotation frequency) which is present on each side of the coupling. Moreover we observe that this component is phased out and amplified from the motor side to the generator side. The associated speed fluctuation due to this component is about 2rpm PP on the motor side and 3rpm PP on the generator side. This corresponds to a coefficient *C* of speed non-regularity of 0.4% and 0.6% respectively, which seems acceptable. However the fluctuation of the differential speed between the engine & generator is higher (4.25rpm PP), and the corresponding twist of the coupling is about 1° PP.



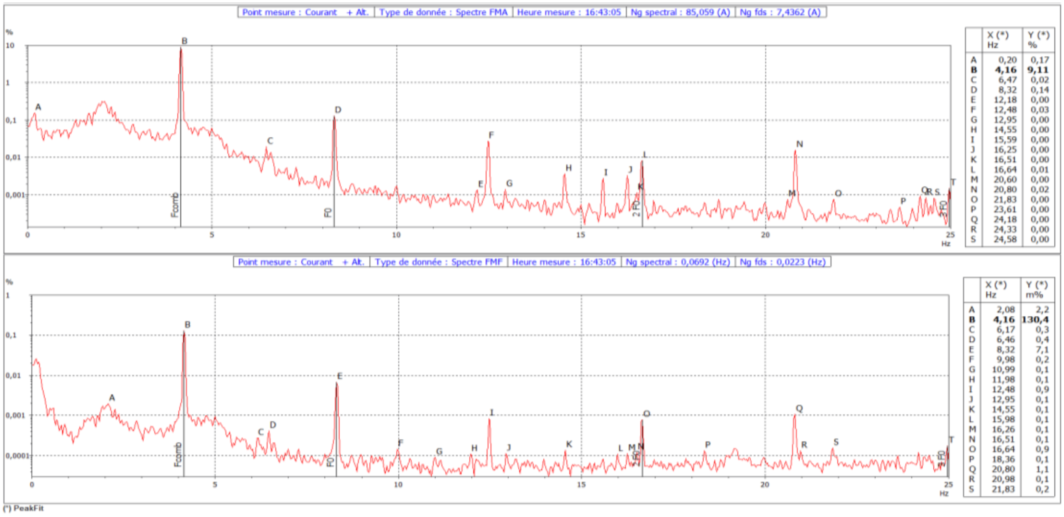
Filtered IAS on the motor & generator sides and differential speed showing a fluctuation at the combustion cycle frequency

The amplification and the phasing of the IAS component at the cycle frequency seem to indicate the proximity of a torsional modal frequency of the shaft line that may amplify the cycle frequency component at 4.17Hz. This seems to be confirmed by the following Frequency Response Function (FRF) calculated between the IAS of the engine and of the generator, which shows a 90° phase lag at a frequency slightly under 4Hz.



FRF calculated between the engine & generator IAS

A measurement of the electrical current was also performed on the generator and shows a strong amplitude modulation at the cycle frequency (9%). Thus the observed IAS fluctuation at the cycle frequency is inducing a torque modulation and so a modulation of the generated electrical power. This modulation is then inducing a torsional fatigue of the elements of the shaft line, especially of the coupling.



Analysis of the generator current amplitude & frequency modulation functions at full load

We also note the presence of an amplification of the background noise around 2Hz on the current amplitude modulation spectrum, which may indicate a torsional modal frequency of the shaft line. This modal frequency is also observed by a transient oscillation on the IAS signals at the moment of the electrical coupling of the generator. Thus it seems to involve the electromagnetic forces of the generator.

We attempted to perform a torsional modelling of the shaft line as in [9], however some parameters still remain unknown: the inertia of the engine and also the torsional stiffness due to the electromagnetic forces of the generator (which seems difficult to estimate from its electrical properties). Further tests are yet to be performed on this group, e.g. testing a flexible coupling with a lower torsional stiffness in order to reduce the incriminated torsional modal frequency close to the combustion cycle frequency.

4 Conclusion

We attempted here to show the interest of different IAS measurement techniques for the torsional analysis on industrial case studies. The use of a specific technique will depend on the application and on the possibilities on site. Advantage of the VRL is to be totally non-intrusive and to avoid machine downtime; however its setting must be performed while the equipment is running.

We also showed the advantage of combining IAS measurement with other measurements, such as the electrical current or the transmitted torque, in order to obtain a full understanding of the torsional behaviour of the equipment.

References

[1] J-L Vasselin, Practical methodologies for on-site measurements of torsional natural frequencies – application to industrial cases, Surveillance 7 international conference (2013).

[2] F Combet, Gear fault diagnosis by motor current analysis – application to industrial cases, Surveillance 7 int. conf. (2013).

[3] L. Renaudin & al, Natural roller bearing fault detection by angular measurement of true instantaneous angular speed, Mechanical Systems and Signal Processing 24 (2010) 1998–2011.

[4] Adeline Bourdon & al, Reconstruction of angular speed variations in the angular domain to diagnose and quantify taper roller bearing outer race fault, Mechanical Systems and Signal Processing 120 (2019) 1–15.

[5] The laser torsional vibrometer: a step forward in rotting machinery diagnostics, N. A. HALLIWELL, Journal of Sound and Vibration (1996) 190(3), 399-418.

[6] B. J. Halkon & S. J. Rothberg, Angular (pitch and yaw) vibration measurements directly from rotors using laser vibrometry, Mech. Systems & Signal Proc. 46 (2014).

[7] G. Nerubenko, Tuned Torsional Vibration Damper, SAE International 2013-01-1897.

[8] H. André & al, Precision of the IAS monitoring system based on the elapsed time method in the spectral domain, Mechanical Systems and Signal Processing 44 (2014) 14–30.

[9] H. S. Han & al, Parametric study to identify the cause of high torsional vibration of the propulsion shaft in the ship, Engineering Failure Analysis 59 (2016) 334–346.

Towards the use of hybrid models for diagnosis and prognosis in turbomachinery health management

P Stephan HEYNS[1], Brian ELLIS[1], David H DIAMOND[1], Ronald G DU TOIT[1], Ronnie SCHEEPERS[2]

[1]Centre for Asset Integrity Management, University of Pretoria, Pretoria, South Africa
stephan.heyns@up.ac.za

[2]Research Testing and Development, Eskom, Rosherville, Johannesburg.

Abstract

Turbomachinery failures are often caused by the dynamic behaviour of rotating blades. The implications of such failures may be serious and practical blade condition monitoring methodologies are therefore becoming important. In this work the idea of hybrid models that rely partly on data driven blade tip timing and partly on physics based finite element models is explored. The work is founded on a Bayesian linear regression based blade tip timing technique which is combined with a stochastic finite element model. This stochastic hybrid approach is applied for the identification and classification of turbomachine blade damage. For the purposes of demonstration, discrete damage is incrementally introduced to a simplified test blade of an experimental rotor setup. The damage identification and classification processes are further used to determine whether a specified damage threshold has been reached, therefore providing sufficient evidence to schedule maintenance.

1. Introduction

Failures of turbomachinery are often caused by the dynamic behaviour of rotating blades. The financial and production implications of such failures may be very significant and appropriate blade condition monitoring methodologies are therefore of critical importance [1,2]. Blade tip timing (BTT) is a non-intrusive measurement technique for online measurement of turbomachine vibration. Essentially it senses when a blade passes a number of proximity probes distributed circumferentially and mounted radially through the turbomachine casing above the row of rotor blades being measured, to determine the blade tip time of arrival. This can be linked to the blade vibration by employing an accurate measure of the once per revolution reference signal. The technique is non-intrusive and online monitoring is possible.

BTT is therefore often regarded as a feasible technique to track the condition of turbomachine blades and prevent unexpected and catastrophic failures. The processing of BTT data to find the associated vibration characteristics is however not trivial. In addition, these vibration characteristics are difficult to validate, therefore resulting in great uncertainty of the reliability of BTT techniques. To deal with the uncertainties of the method, various new concepts have been introduced [2,3,4]. These ideas deal primarily with diagnosis. Techniques for prognosis to assist with maintenance decision making is however becoming more important. Mishra et al. [5] explored a range of techniques of interest to accomplish this through the use of hybrid models that merge physics based and data driven approaches into a unified approach.

In this work the idea of hybrid models is pursued further in the context of turbomachinery blades, by proposing an approach comprising a stochastic finite element model (FEM) based modal analysis and a Bayesian linear regression (BLR) based BTT technique. The use of this stochastic hybrid approach is demonstrated for the identification and classification of turbomachine blade damage. For the purposes of this demonstration, discrete damage is incrementally introduced to a simplified test blade of an experimental rotor setup. The damage identification and classification processes are further used to determine whether a damage threshold has been reached, therefore providing sufficient evidence to schedule maintenance. It is shown that the proposed stochastic hybrid approach may offer benefits for practical implementation. This paper highlights some aspects from of a comprehensive study by Du Toit, Diamond and Heyns [6] as well as some additional results that further corroborates the approach.

2. Hybrid methodology

Figure 1 outlines the proposed hybrid approach for the identification and classification of damage in a turbomachine blade. The process firstly comprises a data driven blade tip timing analysis to track the relative change in natural frequency of a blade to identify and infer the extent of blade damage at a given time. A blade damage threshold is established through finite element analysis. The BTT analysis is repeated continuously until this threshold is reached. Once this threshold is reached, the operation of the turbomachine should be stopped and relevant maintenance be conducted. Subsequently the blade natural frequency and amplitude determined from the blade tip timing results are clustered using predetermined mean values as initial cluster centres. This clustering then allows the severity of the damage to be classified.

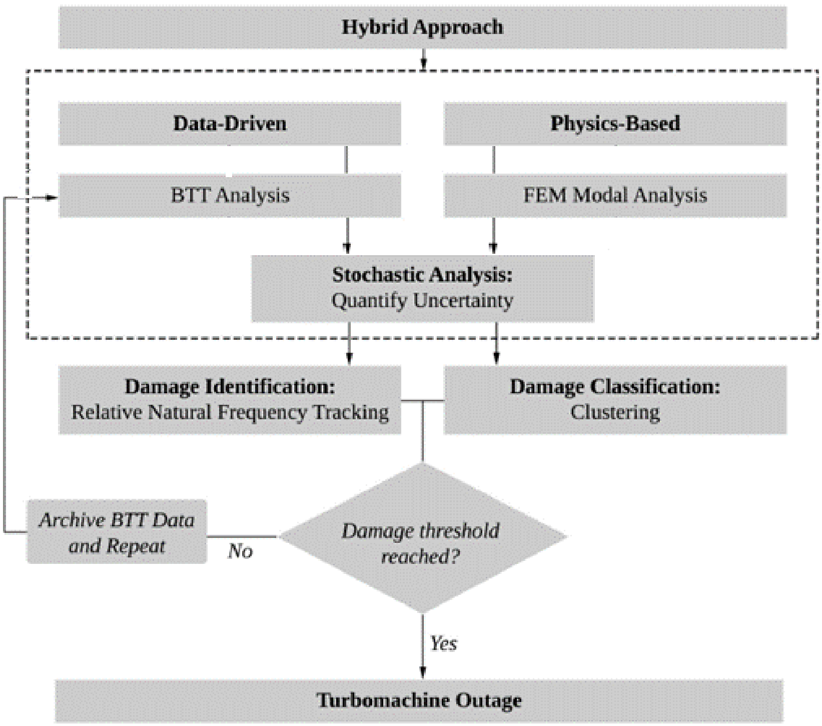


Figure 1: Schematic overview of the proposed hybrid methodology

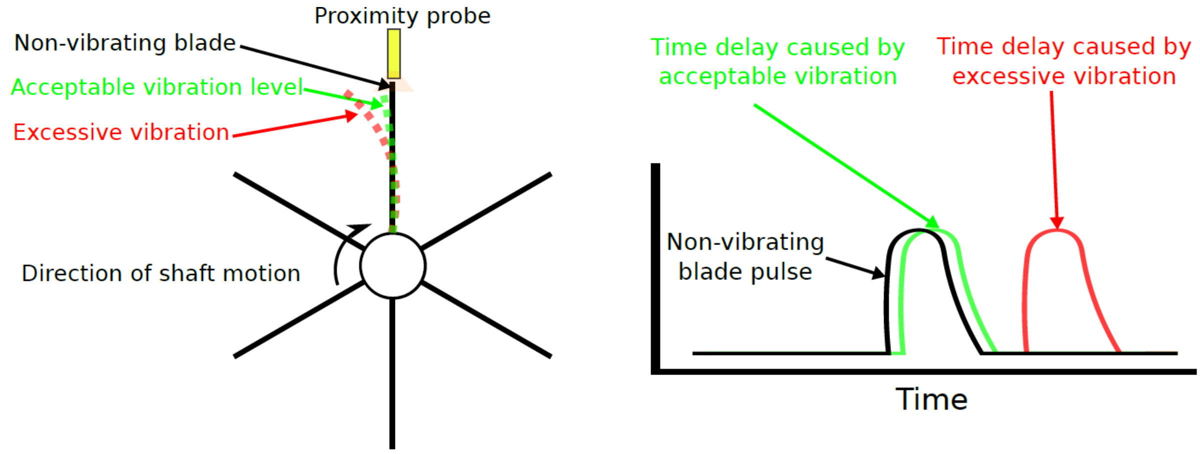
The rationale behind adopting such a hybrid approach is to:

- Use finite element analysis (FEA) to establish a baseline for comparison before BTT tests are performed.
- Project expected blade condition before it becomes possible from the BTT measurements.
- Correct for real blade behaviour and aspects not considered in the FEA.
- Enhance remaining useful life estimation.

The current methodology comprises two analysis types, namely damage identification in which the relative change in natural frequency is tracked to identify and infer the extent of the blade damage, and damage classification in which the natural frequency and blade response amplitudes derived from the BBT results are clustered using predetermined mean values as initial cluster centres to determine the severity of the blade damage. A damage threshold is determined based on FEA.

3. Data-driven investigation

The basic idea behind blade tip timing is to measure the time of arrival (ToA) of blades passing a proximity probe in the casing. This data is then analysed to determine the vibrational state of the blade. The ToA is dependent on the rotational speed of the shaft. However, a vibrating blade will either arrive earlier or later than expected at the proximity probe (see figure 2). Extracting the blade vibration from the ToA measurements can be complex in practice. Many algorithms have been developed for this purpose. Most of these algorithms are based on so-called *indirect* methods in which the maximum amplitude and corresponding frequency at resonance is determined during transient operating conditions and only one or two proximity probes are required, or *direct* methods in which the maximum amplitude at each rotational speed is determined during steady state and at least four probes are normally used.



(a) (b)

Figure 2: Blade tip timing (a) Rotor blade during operation (b) Proximity probe output

In this work we utilise a direct BTT method based on statistical inference [2]. The technique employs Bayesian linear regression (BLR) which offers a number of practical advantages: The tip displacement amplitude and phase with respect to the angular reference signal may be determined at each measured rotational speed. Firstly, this results in a detailed picture of these parameters and their changes over the operating domain. Secondly the processed data considers the whole range of inferred solutions, therefore resulting in a stochastic solution. Lastly, the stochastic nature of the processed data allows one to establish confidence intervals for amplitude and phase and renders the approach noise tolerant.

The BTT technique employed here assumes a single-degree-of-freedom (SDOF) model for the blade vibration. Equation (1) describes the blade tip displacement x at a specific time t, rotor order EO and circular frequency ω:

$$x_i(t) = A_i \cdot \cos(\omega \cdot t_i) + B_i \cdot \sin(\omega \cdot t_i) + C_i \quad \text{where} \quad \omega = EO \cdot \Omega \tag{1}$$

BLR is used to infer the values of the constants A, B and C as probabilistic quantities. The equations are solved for each revolution i at a corresponding shaft speed Ω, with EO also inferred from a probabilistic approach. The parameter set **x** is solved for each revolution and forms part of a multivariate normal distribution with associated mean μ_i and covariance matrix Σ_i as per equation 2.

$$\mathbf{x}_i = \begin{pmatrix} A_i \\ B_i \\ C_i \end{pmatrix} \quad \boldsymbol{\mu}_i = \begin{pmatrix} \mu_{A_i} \\ \mu_{B_i} \\ \mu_{C_i} \end{pmatrix} \quad \boldsymbol{\Sigma}_i = \begin{pmatrix} \Sigma_{AA_i} & \Sigma_{AB_i} & \Sigma_{AC_i} \\ \Sigma_{BA_i} & \Sigma_{BB_i} & \Sigma_{BC_i} \\ \Sigma_{CA_i} & \Sigma_{CB_i} & \Sigma_{CC_i} \end{pmatrix} \tag{2}$$

These random values are substituted in equations (3) to quantify the amplitude and phase:

$$\widehat{A}_i = \sqrt{A_i^2 + B_i^2}$$
$$\phi_i = \arctan\left(\frac{B_i}{A_i}\right) \tag{3}$$

The maximum amplitude and the associated natural frequency is indicated by equation (4)

$$f_{n_{\widehat{A}}} = f(\vartheta) \quad \text{where} \quad \vartheta = \max_{i \in N} \widehat{A}_i \tag{4}$$

while the natural frequency can also be derived from the phase angle

$$f_{n_{\hat{\phi}}} = f(\varrho) \quad \text{where} \quad \varrho = \frac{f_u + f_l}{2} \quad \text{subject to} \quad \Delta\hat{\phi} \geqslant \pi \tag{5}$$

For this work we used an experimental setup (see figures 3 and 4) comprising a rotor assembly, and excitation mechanism, sensors, a data acquisitioning and signal generation system. Labview was used to generate the signals for controlling the motor speed.

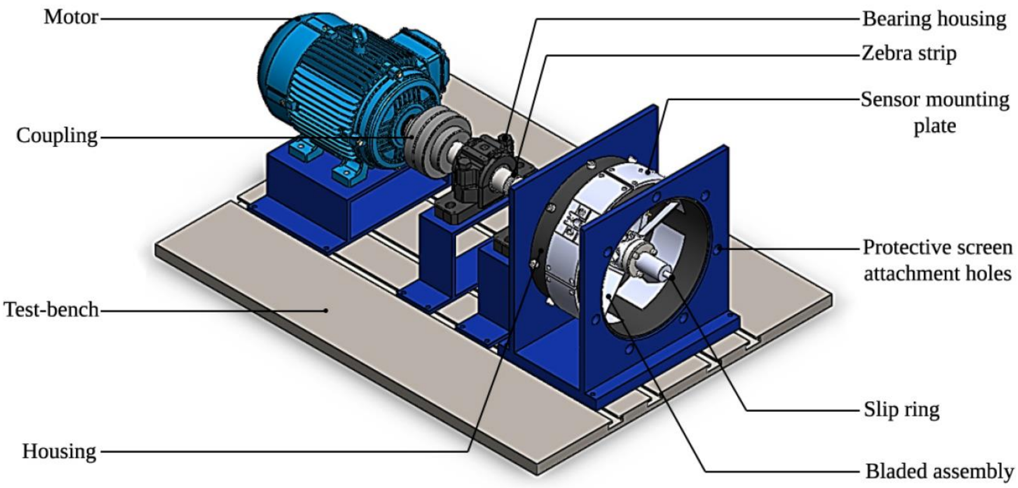


Figure 3: Schematic of the experimental setup

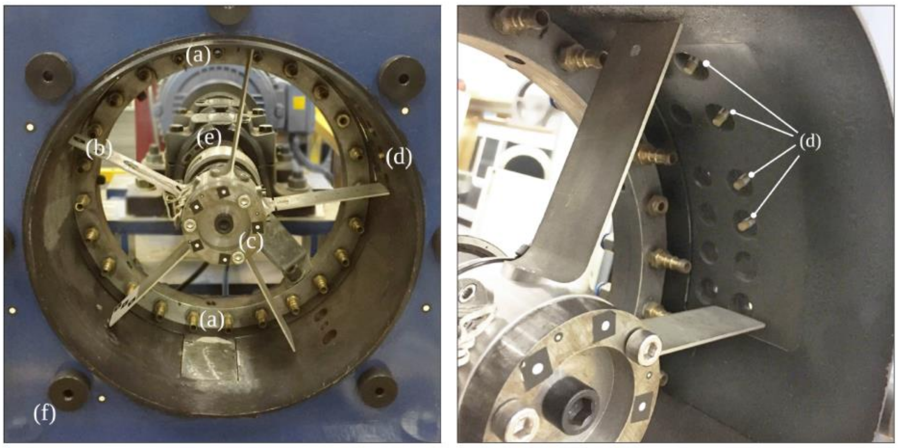


Figure 4: Photos of the rotor and casing

The bladed assembly and casing setup comprises (a) compressed air supply nozzles for blade excitation (b) a bladed assembly with 5 aluminium blades (c) a central hub with a slip ring arrangement for validation measurements using strain gauges (d) four irregularly spaced eddy current probes and (5) a shaft connected to a motor with a variable speed drive.

In this study damage was artificially introduced to a single blade (blade 2) of width 40 mm in twelve discrete steps ranging from 0 mm to 8.61 mm. Damage was classified in three ranges: non-severe (damage increments 1 to 6, ranging from 0 to 1.81 mm), mid-severity (damage increments 7 to 9, ranging from 3.11 mm to 5.60 mm) and severe (damage increments 10 to 12, ranging from 6,97 mm to 8.61 mm).

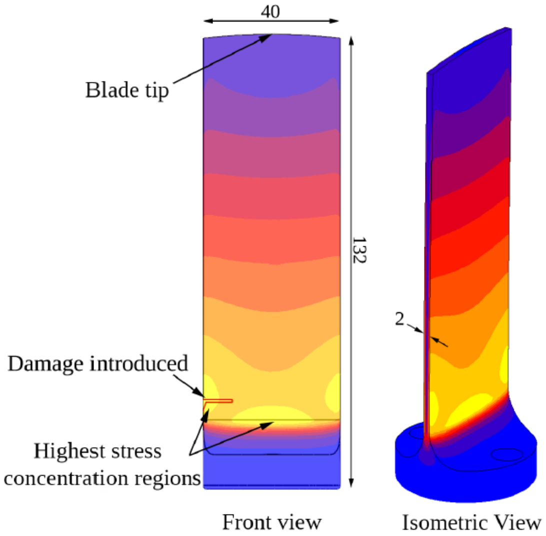


Figure 5: Damage introduced on blade 2, relative to the areas of highest stress concentration for mode 1, as determined from a FEA

The rotor shaft speed was subsequently ramped up from 1195 r/min to 1330 r/min in order to pass through the blade resonant frequencies to allow for the estimation of the amplitude and phase parameters as indicated in equations 4 and 5. BTT tests were repeated six times for each of the 12 damage increments, i.e. a total of 72 tests were conducted. To allow comparison of the extracted blade phase results between the tests, all measurements were synchronised with respect to a specific shaft encoder pulse.

4. Physics based investigation

A 3D FEA was conducted to determine the likely blade resonances at specific operational speeds. Centrifugal loads were accounted for by applying angular velocities to all elements. Thermal loads were accounted for by varying the material properties. The FEA was intended to supplement the BTT results as outlined in sections 5 and 6.

Marc Mentat was used to perform the FEA with the Lanczos algorithm to compute the natural frequencies and amplitudes for the simplified blade. The damage was again introduced in 12 discrete stages corresponding to the damage in the experimental test blade. Slight variations in angle (approximately 10°) and size (length, width and height – about 12%) were introduced for each discrete stage to account for uncertainties in true crack and crack location measurements as well as material properties and centrifugal loads. 89885 ten-noded tetrahedral finite elements were used for the basic geometry. Samples of these parameters were randomly selected from a uniform distribution within the angle and dimension ranges. Aluminium was selected as material with uncertainty again being introduced by varying the material properties within 12% of the reference values.

Variation in these parameters resulted in variation in the natural frequencies for each test within a particular series of tests. This allowed a mean with its associated confidence intervals around the mean, to be determined for each damage condition.

5. Damage identification

The first part of the proposed hybrid procedure considers the tracking of the relative change in the blade natural frequency (see equation 6) to identify and infer the degree of blade damage. Tracking the relative natural frequency change aims to give a more general indication of the blade condition, due to differences in the responses determined from the BTT results and the FEA results, which might not necessarily capture all the important physics associated with the rotating blade response.

The procedure therefore relies on tracking the relative changes of the derived BTT natural frequency results to infer the extent of the blade damage. The FEA allows one to estimate the expected blade conditions (relative change in natural frequency) corresponding to a particular damage size that is considered critical. This process is illustrated in figure 6.

The change in natural frequency from a reference state Δf_{ni} (undamaged in this case) is first quantified. f_{n0} is the original natural frequency and f_{ni} is the current natural frequency.

$$\Delta f_{n_i} = \frac{f_{n_0} - f_{n_i}}{f_{n_0}} \times 100 \tag{6}$$

The aim is to determine if the blade damage threshold has been reached. The challenge however is of course that the actual crack size of the blades can in practice not be determined while the online BTT measurements are made, and that the extent of blade damage must therefore be inferred from the processed BTT results.

Figure 6 shows two iterations of the proposed identification procedure. The figure shows the results from tracking the BTT relative natural frequency based on amplitude and phase based results.

The process follows the following steps:

- Define a blade damage level that would justify physical inspection. This is calculated from the FEM based on the principles of fracture mechanics or fatigue analysis and leads to the identification of a Level 1 damage threshold as seen ion figure 6.
- The FEM modal analysis is performed stochastically at this discrete crack size in order to quantify uncertainty. The mean μ_{FEM1} and the standard deviation σ_{FEM1} corresponding to this damage increment is subsequently recorded.
- A new variable δ_{damage} is then defined as in equation 7. This parameter represents the difference between the relative change in the natural frequency from the BTT results (for a particular batch of tests) and the relative change in natural frequency of the FEM results at a particular discrete damage size, represented by Level 1:

$$\delta_{damage} = \Delta f_{n_{BTT}} - \Delta f_{n_{FEM}}(K) \tag{7}$$

K corresponds to the predetermined crack size in the FEM modal analysis. δ_{damage} is a normal probability distribution with both Δf_{nBTT} and $\Delta f_{nFEM}(K)$ having associated normal distributions.

- The calculation of the mean and the variance of δ_{damage} requires that a number of repetitive BTT tests must be performed and that the mean μ_{BTT} and the standard deviation σ_{BTT} be determined.
- The probability $P(\delta_{damage} \leq 0)$ is determined for a batch of BTT tests and the chosen Δf_{nFEM} (Level 1 in this case). This probability is found from the cumulative distribution function (CDF) of δ_{damage} with the associated mean and variance as shown below:

$$\delta_{damage} \sim \mathcal{N}(\mu_{BTT} - \mu_{FEM}, \sigma_{BTT}^2 + \sigma_{FEM}^2) \tag{8}$$

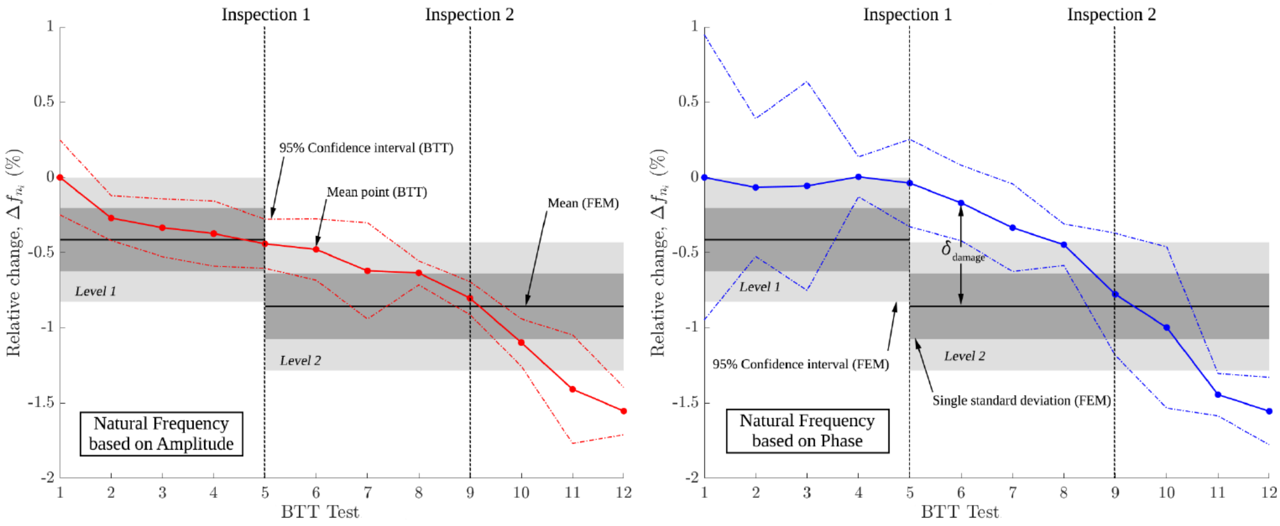


Figure 6: Damage identification at ambient temperature of 22°C

$P(\delta_{damage} \leq 0)$ is the probability that the relative change in natural frequency derived from the BTT measurements Δf_{nBTT} equals or exceeds the permissible relative change in natural frequency derived from the FEA, $\Delta f_{nFEM}(K)$. The damage threshold X_{dt} is based on a selected probability $P(\delta_{damage} \leq 0) > X_{dt}$

The user must decide what an acceptable probability would be to justify maintenance. Repetitive BTT tests are conducted until this probability value (damage threshold) is reached. It is important to note that the damage threshold X_{dt} may be selected to be conservative.

The above steps are repeated after every inspection or maintenance until a blade needs to be replaced. After each inspection a new damage level, based on the FEM results may be prescribed to determine a new blade damage threshold.

Figure 6 demonstrates two iterations of the damage identification process. In this diagram Level 1 corresponds to a relative discrete crack size of 3.95% and Level 2 corresponds to a relative discrete crack size of 9.675%. Both Levels 1 and 2 are arbitrarily selected here for demonstration purposes, but would in practice be based on physical evidence.

A very comprehensive set of experiments were also done at elevated temperatures of 41° and 98°C, using two 2 kW commercial heaters to heat the air stream through the machine. As a further example of the application of the procedure typical results for 98°C are also shown in figure 7. This figure confirms the general behaviour observed at ambient temperature (figure 6) and at 41°C (not shown here).

6. Damage classification

The damage identification process presented in section 5 relies on repetitive measurements to be available. The damage classification process described here however aims to allow the use of a single set of BTT measurements to determine the blade condition. A high level of confidence must therefore be attached to this measurement. To accomplish this a damage classification procedure that is based on the clustering of BTT natural frequency and amplitude values is investigated. Here we use the well-known K-means clustering technique, which assigns the observations with the nearest means to a certain cluster of data points. Physically this means that the vibrational characteristics of the blade from a specific BTT measurement, are assigned to an existing cluster of vibrational characteristics with the nearest mean. The aim of this cluster is to classify the severity of the blade damage according to which group the vibrational characteristics are assigned to. This means that the damage severity may be established from these clusters.

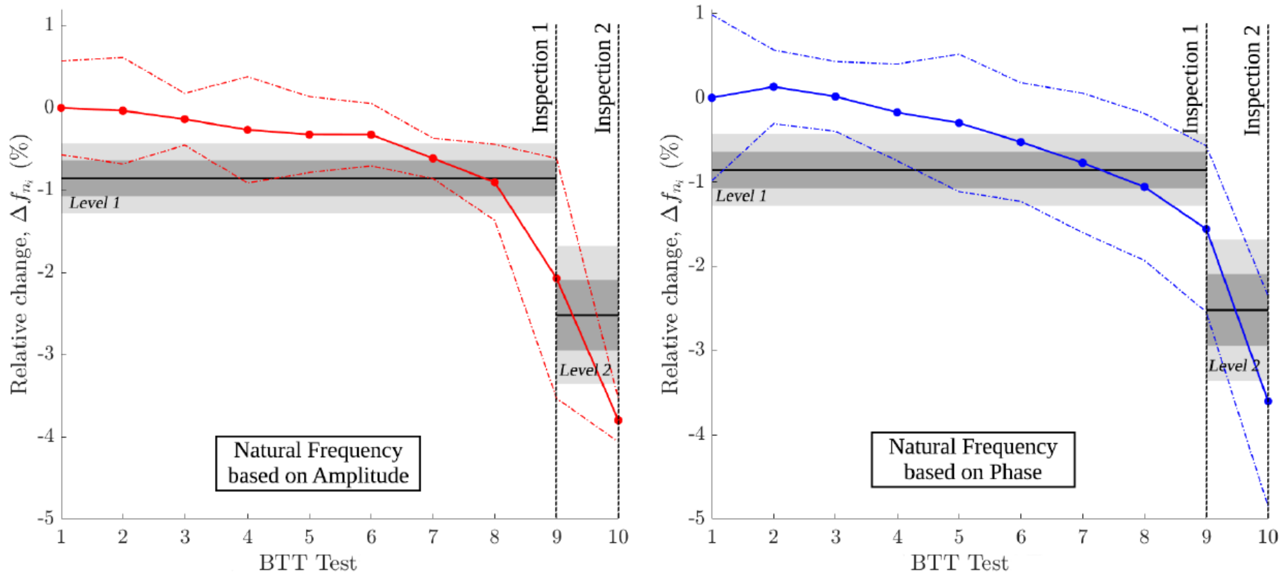


Figure 7: Damage identification at 98°C

The process follows the following steps:

- The amplitude and associated natural frequency of a specific blade are extracted from the BTT measurements.
- We propose clustering of the natural frequency and amplitude results using predetermined mean values. The predetermined values correspond to the natural frequency results from the FEA for mode 1. The blade amplitudes are however not computed from the FEA since this would require a computational fluid dynamic analysis. This is however a complex process and infeasible in practice, and is avoided here by simply assigning zero amplitudes to the initial clusters.
- In this work the data is partitioned in three clusters, namely non-severe damage, mid-severity damage and severe damage (see section 3). These partitions are not enforced on the BTT data before the K-means clustering commences. However after clustering is completed the individual natural frequency points and associated amplitudes are used to determine the accuracy of the final classification.
- The initial cluster centres correspond to zero amplitude values and the mean natural frequency values of the FEA results partitioned using the scheme mentioned above. This results in three amplitude and natural frequency combinations to be used as starting points for the centroids. These initial cluster centres enable individual BTT measurements to be classified to the closest partition.
- The BTT amplitude and natural frequency values are both normalised over the range 0 to 1 for use in the K-means clustering. The initial cluster centroids based on the FEA are also scaled using this scheme.
- The point-to-cluster-centroid distances are computed for all of the individual BTT points. These points are considered with no indication of which damage increment the points belong to. The overall averages of the points are calculated and the new centroid locations are allocated. As a result the BTT points are classified into the associated group of most likely Range of Damage (RoD) that the amplitude and natural frequency would represent. In figure 8 RoD I, RoD II and RoD III represent the new averaged clusters for the undamaged, middle damaged and largest damage increments respectively. The incorrectly classified points are also shown in figure 8.

Figure 8 shows the results of the K-means clustering implementation as part of the damage identification process. RoD I, RoD II and RoD III represent the new averaged clusters for the non-severe, mid-severity and severe damage classes respectively. Incorrectly classified points are indicated by a Roman numeral above the marker.

The overall classification accuracy is 78%. If only the RoD III case is considered, the accuracy would be 94%. However in one case an actual RoD III was incorrectly classified as a RoD I point. This is obviously dangerous and points to further research being required.

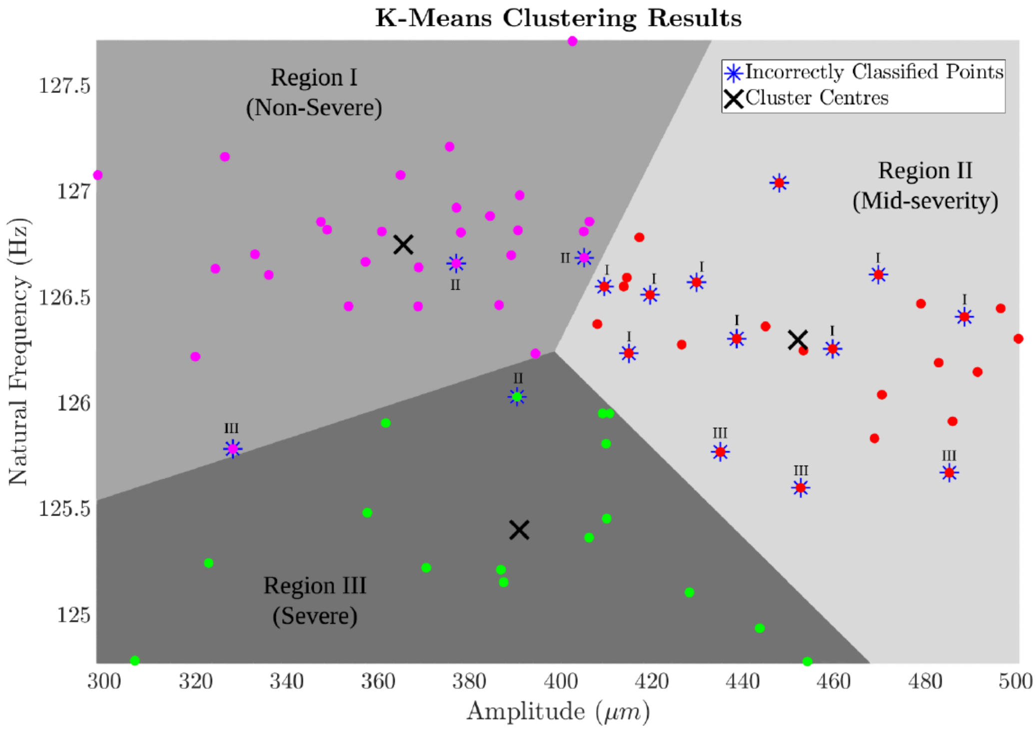


Figure 8: K-means clustering results

7. Conclusion

This work represents a first attempt to develop a hybrid methodology encompassing data-driven BTT and physics based finite element analysis for turbomachinery blade diagnostics and prognostics. This hybrid method uses the outputs of a BTT technique that is based on Bayesian linear regression and stochastic finite element analysis.

An experimental study was conducted on a simplified test rotor, with discretely introduced damage on a test blade.

The work led to a damage identification procedure based on the probability that the relative change in natural frequency of the BTT results is as large as that determined by the finite element modal analysis (at a chosen discrete damage size) projects it to be. This probabilistic damage identification process demonstrates the ability to infer the extent of blade damage.

A damage classification process is introduced to determine the blade condition using a single set of BTT measurements. To simplify the process, K-means clustering is used to classify the derived BTT amplitude and natural frequency values. The predetermined finite element analysis natural frequency results are used to initiate clusters and cluster centroids. The clustering of the derived BTT vibrational characteristics to the nearest cluster centroid enables the severity of the blade damage to be classified. While 78% classification is achieved, it is however shown that the possibility exists of classifying a severely damaged blade as non-severely damaged. This is of course potentially dangerous in practice. Future research into the performance of alternative classification procedures is therefore required.

We do however believe that a useful step has been taken towards the use of hybrid models for diagnosis and prognosis in turbomachinery health management.

Acknowledgments

The authors gratefully acknowledge the Eskom Power Plant Engineering Institute for their support of the development of the test facilities that were used for this investigation.

References

[1] Gubran, A.A. & Sinha, J.K. Shaft instantaneous angular speed for blade vibration in rotating machine. Mechanical Systems and Signal Processing, 44(2014), pp.47-59.

[2] Diamond, D.H., Heyns, P.S. and Oberholster, A.J. A comparison between three blade tip timing algorithms for estimating synchronous blade vibration. 9th World Congress on Engineering Asset Management, WCEAM 2014, 28-31 October 2014.

[3] Diamond, D.H., Heyns, P.S. and Oberholster, A.J. Improved blade tip timing measurements during transient conditions in a state space model, Mechanical Systems and Signal Processing, 122(2019), pp.555-579.

[4] Diamond, D.H. and Heyns, P.S. A novel method for the design of proximity sensor configuration for rotor blade tip timing. Journal of Vibration and Acoustics, 140(2018), pp. 061003-1-8.

[5] Mishra, M., Saarti, J. Galar, D. and Leturiondo, U. Hybrid models for rotating machinery diagnosis and prognosis. Estimation of remaining useful life. Luleå University of Technology, 2014.

[6] Du Toit, R.G., Diamond, D.H. and Heyns, P.S. A stochastic hybrid blade tip timing approach for the identification and classification of turbomachine blade damage, Mechanical Systems and Signal Processing, 121(2019), pp.389-411.

Condition Monitoring

CM Base, a universal gateway to condition monitoring datasets

Cécile Capdessus[1], Frédéric Bonnardot[2]

[1]PRISME, Univ. Orléans, 21 Rue Loigny-la-Bataille 28000 Chartres, France
Cecile.capdessus@univ-orleans.fr
[2]Univ Lyon, UJM St-Etienne, LASPI, EA3059, F-42023 Saint-Etienne, France
frederic.bonnardot@univ-st-etienne.fr

Abstract

Condition monitoring is a lively research domain, with hundreds of researchers sharing new techniques coping with complex problems, such as variability of the operating conditions, complex machinery monitoring, diagnostics, prognosis... Considering the many methods that have been proposed these last decades, it is striking how few of them have reached the industrial application. One of the reasons is that the intensive validation of these methods over real-life data required to ensure their reliability is often difficult to achieve. The results obtained over one dataset may not be easily reproduced over another one. Furthermore, can the comparison between two techniques be objective, when applied to different datasets?

Some laboratories and companies generously share the datasets recorded on their test benches or industrial devices. Unhappily, it is often difficult for them to know for certain who has worked on their data, which is both frustrating and a real problem when they report to their authority about the diffusion of the data, in order to get the subsidies required for the test bench exploitation. From the user's point of view, there can also be difficulties in finding the right datasets for a specific study among the jungle of all that is proposed on the web. How to find the appropriate dataset for a prognosis study, or a fatigue study, or the study of any specific kind of damage? Another problem can be met by those who wish to share their datasets with the scientific community but do not have the technical skills or staff to do so.

CM Base is a web portal that aims at facilitating the sharing of the data, by offering many a functionality, such as a list of all the existing datasets and test benches with all related papers, searching facility allowing to extract from the base all datasets related to a specific problem or papers related to a specific dataset.

1 Introduction

These last decades have seen a wide expansion of on line condition monitoring, due to various factors. First, the evolution of sensors, computing and communication technologies allow setting up condition monitoring systems more easily and at lower cost. Second, the need of such systems has increased. Indeed, in the present competitive international context, production companies cannot bear the consequences of unplanned production lines shutdown. Furthermore in many domains, products are lent rather than bought by the clients, or bought with some service package, including monitoring. The development of machine learning and Artificial Intelligence (AI) has also opened new fields for automated surveillance.

In this context, the need for reliable monitoring techniques has led industry as well as academic researchers to explore many possibilities, often based on the use of a set of sensors and the elaboration of relevant health indicators through the application of signal processing techniques. The elaboration and the testing of these methods require the production of experimental data.

Many laboratories have set test benches and recorded measure signals datasets and some of them chose to share these datasets with the scientific community through a web site. But it reveals difficult for them to get information about the diffusion of their dataset: for instance who it has been used by, the conferences and papers the obtained results have been published in. For their part, the users have some difficulties in finding the right dataset to test their technique or even to know about the available ones. Furthermore, some laboratories wish to share their datasets but lack computer knowledge or facility to do so.

CM Base is a database meant both to gather all the information about existing on line datasets and to help laboratories putting online their datasets. It can be accessed through a web site that allows inserting a new dataset, inserting a paper related to an existing dataset, and browsing the basis in order to find some datasets relevant to a specific problematic with all related papers. In order for the basis to be as widely used as possible, the interface has been designed to simplify all procedures, which makes CM Base a very easy to use application.

In section 2, the need is analysed and a solution is designed, in section 3 the chosen technical solution is presented, in section 4 a user guide for the use of CM Base is given, and in section 5 a conclusion and some perspective are proposed.

2 Analysis of need

2.1 Reliability and reproducibility

Reliablility and reproducibility of the techniques have become one of the main topics in all scientific domains [1], peculiarly in all that is related to data processing and statistical analysis. Indeed, for a specific technique to be validated, it is not enough that it has been tested on one set of data. The quality of the obtained results can depend of some parameters related to the specific experimental setup, the location of the sensors, all things that are different from a test bench to another. Furthermore, real-life signals are yet another challenge. Techniques that have been tested on measure signals collected from a lab test bench can lack robustness when it comes to measures collected from an industrial device [2, 3]. It is thus of utmost importance that researchers might access to a wide set of different measures, in order to test the robustness of their proposed techniques and to specify in which conditions these can be applied. Researchers should thus be given access to a wide range of measure signals, collected on both lab test benches and real-life devices.

2.2 Comparison of techniques and measure signals

No indicator can be considered perfect for the monitoring of a system. Depending of its stage, the presence of some damage may produce on the measure signal alterations of a different nature. For instance, a localised defect on a roller bearing can be detected with techniques well adapted to transient signals, whereas once the damage has spread, such techniques will fail to put it at light. Furthermore, different kind of measure signals can be used to monitor a system: current, vibrations, temperature… For these reasons, most condition monitoring AI systems use both different kinds of measure signals and different kind of indicators. It is thus interesting to compare either the results obtained through one specific method on different datasets, or through different methods applied to a specific dataset. This can help defining how to combine these techniques and merge different indicators in order to ensure a reliable monitoring system. It would thus be interesting for the condition monitoring community to have access not only to different kind of measure signals, but also to the related bibliography.

2.3 Choice of an appropriate dataset

When addressing some specific topic that may be bearings, gearboxes, diagnosis, prognosis, wear, variable load, variable speed or any other relevant topic for condition monitoring, accessing to a measure dataset well adapted to this topic is not an easy task. All labs are not equipped so as to set up an experiment and while many datasets are now available on the web, it takes time to check whether they have the right profile. It would be convenient that all these datasets can be accessed through a browsing system that would allow getting straight access to the well-fitted ones. And then, for each dataset, it would be interesting to have also access to the related state of the art. On the top of it, it would ensure to the researchers who have already published on this dataset that any newcomer is informed of their work.

2.4 Storage and feedback

Many labs and even companies have chosen to share their datasets through their lab webpages or at the occasion of a conference contest [4]. In order to justify the maintaining of their test benches, they must prove to their organisation that sharing the data with the scientific community is profitable. This profit can be expressed in terms of international influence or appropriation of new techniques. The problem is that they often get no more than the number of times their data have been downloaded, or at best who asked for them. They rarely get information about what has really been done on their data. They sometimes discover by chance that their team is thanked in a paper for having shared the data. It would be interesting for them to get a list of all papers that have been published about techniques that were applied to their dataset. In other words, it would be useful to ty to each dataset the related bibliography. Furthermore, some researchers would like to share their data with the community but have no technical means to do so. It would be interesting to provide the community with a storage facility to share their dataset.

2.5 Specifications

The application should allow:

- Sharing a dataset ;
- Sharing the reference of a publication related to a dataset ;
- Browsing the base in order to get a list of all datasets relevant to:
    - A specific topic (diagnosis, prognosis…);
    - A specific device (gearbox, bearing, engine…);
    - A specific context (lab test bench, industrial recordings);
    - Specific operating conditions (constant or variable load or speed for instance);
- Getting the bibliography related to any dataset.

The access should be authorized only to registered visitors, so that the application must include some access management part. The interface must be easy to use, so that contributors do not lose time whenever they add a dataset or a paper, or browse the base.

3 Technical solution

3.1 Strategy of development

CM Base is accessed through a web site. This web site can be used in two ways: as a data navigator to search data by keywords and obtain associated bibliography, or as an editor to declare existing data and databases, associated articles and documents.

The catalogue of data, database, associated links, keywords and bibliography are stored in a relational database. The associated documents and pictures are directly stored on the server. CM Base is not a data storage server (like dropbox) that manages hundreds of MegaBytes of data, but a gate allowing access to the data. Only links to data and small document are stored into the database. The user can use a file-sharing solution such as Mendeley [5] to store his measure datasets, and then declare these data in CM Base to give both acces to the data and information about it: keywords, little document, link to files, …

The website was developed by using standard LAMP tools: Apache, MySql and Php and a Linux Debian web server hosted by Université Jean Monnet of St Etienne on a Virtual Machine.

In order to provide a long-term solution not sensitive to update we avoided using a web Content Management System (CMS) like Wordpress, Drupal, Joomla, … The solution was directly coded in php [6]. In the same way, instead of using css toolkits like bootstrap that require extra files and regular update, a more classical css formatting was used , based on flexbox.

A Model-View-Controller architecture [7,8] was used to organise the project. The database was designed through a Merise analysis [9] and is managed by sql requests in Php.

3.2 Database structure

The organisation of the database is illustrated by figure 1. Each box corresponds to a table, which is equivalent to a sheet from a spreadsheet. The name at the top is the name of the table/entity. The following elements are the fields, which are the equivalent to the column names in a spreadsheet. The interest of relational database is that it allows establishing some links between the tables. These links are symbolised by rounded boxes. For example, all fields related to a member, i.e. his member_ID, last name, e-mail, and so on, are stored in a Members Table/Entity. A link is established with the Papers Table in order to associate to each member the publications he introduced into CM Base. This link is created by adding a (table_ID, member_ID) pair to the IntroducedBy Table. More technical details about this representation could be found by searching literature on Merise method for Database Analysis.

The main tables are:

- The Members table, to store a list of people registered in CM Base,
- The Papers table that contains a list of papers stored in the database,
- The Datasets table, to store information about each test bench or data collected in the database,
- The Ressources table, in which are collected a list of files, links, pictures.

The links between tables enable us to find articles associated with datasets (About relation) or to make a link between type of Sensors and Datasets. The use of a database facilitates data management since all creation, updates of data are made by requests to MySql server, so that all the storage of the information is managed by Mysql.

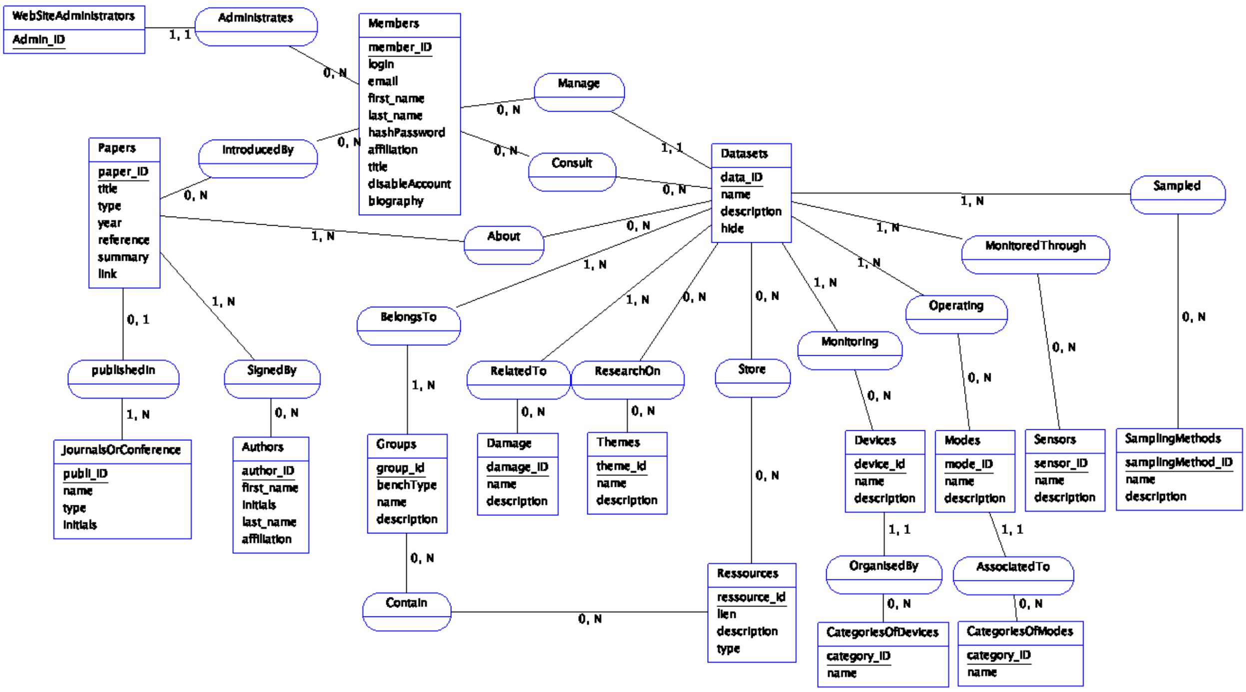


Figure 1: Database Conceptual Model

3.3 Organisation of the Web site

Access to the data is provided through a web site. A menu enables accessing to the following pages:

- "Search the base" to browse inside the collection of data,
- "Add my data" that enables the user to insert its own data into the database or made some links with other databases,
- "Add my papers" to declare papers linked with a dataset,
- "My profile" to change or give personal details,
- "My contributions" to list papers and data inserted by the logged user.

The next section will present some of the web pages as an example.

4 User guide

In this user guide are presented some of the functionality of CM Base. A more exhaustive user guide is available on CM Base web site.

4.1 First connection

Figure 2 shows the home of the database. An account must be created for the first connection by using request account link in the menu. The request account page asks some personal details, an e-mail, a biography and a password.
After connection, more options are available on the menu as shown in figure 3.

4.2 Search the base

The “search the base” web page lists the content of the dataset and enables to make search by using tags associated with the dataset (sensors, device, type of sampling, …). These tags are the same as those used for the dataset insertion (see “Add my data” section).

Figure 2: Home page

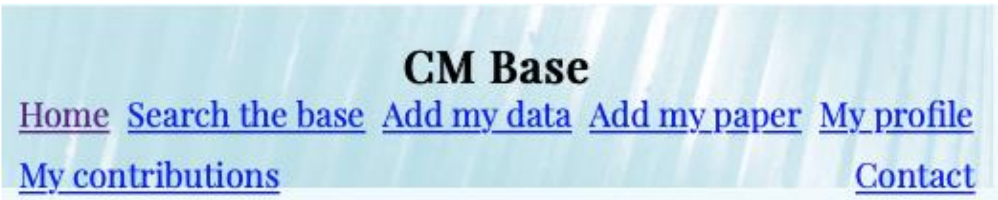
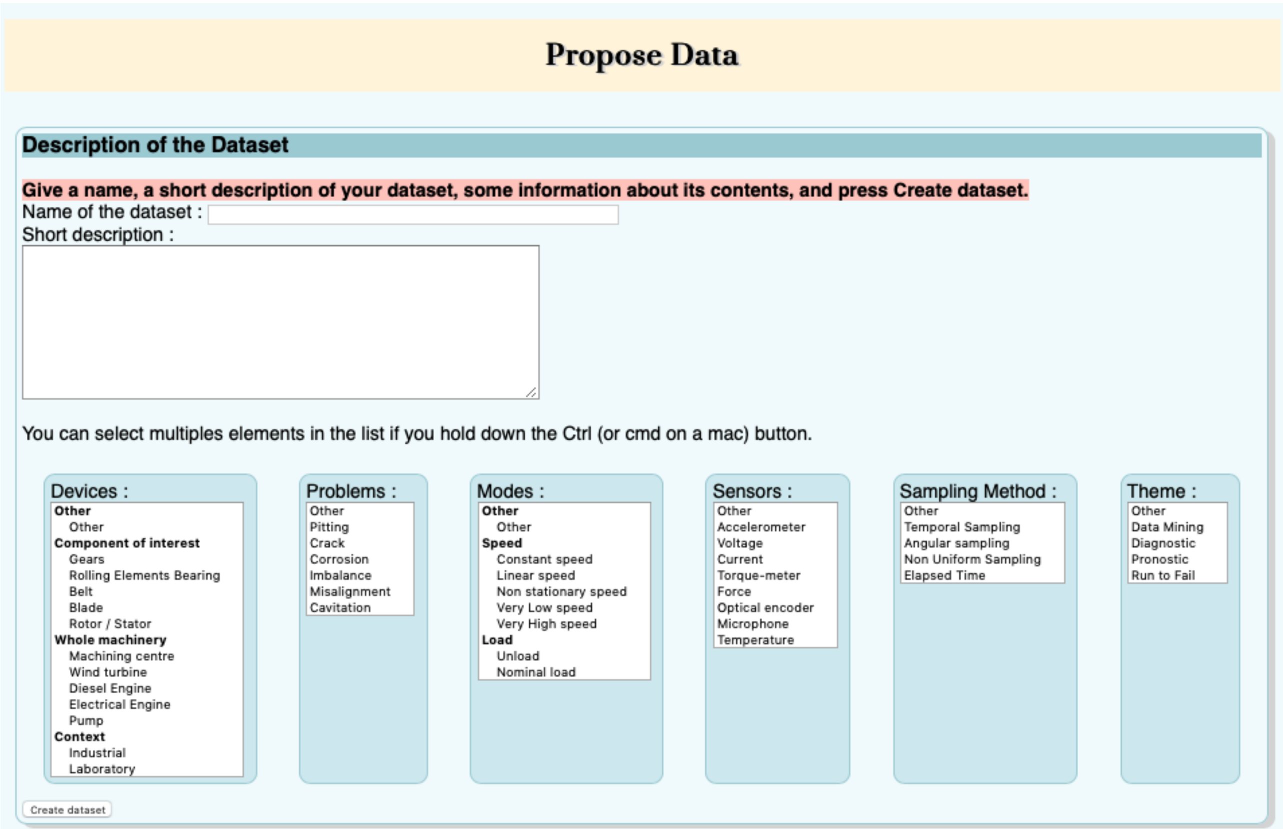


Figure 3: CM Base menu

4.3 Add my data

The menu item "add my data" allows inserting your data into CM Base and sharing them. The associated page is shown in Figure 4. At first, only the "Description of the dataset" part is displayed. Once this part is completed and validated by clicking on the "Create Dataset" button, the next elements appear.
The description of the dataset and main information sections display the same elements as those displayed in a dataset list or dataset search answer. Each element can be modified after submission if necessary.
The user should provide a dataset name, a short description (short enough to be displayed in a table of answer). A longer description can be provided as a pdf file in the main information file.
The six thematic lists are very important and should be filled with care, since the selected items will be used to perform sharper browsing through the base. One or more elements can be selected.
The "Devices" theme enables identifying the component of interest (gear, rolling elements bearing, …) within the dataset and the associated machinery. It is also necessary to give information about the damage, the sensors and sampling methods. The "Mode" section allows giving extra information on running conditions: speed, load, … Three different acquisition modes are available: diagnostic when it is necessary to find the type of defect, prognostic when it is necessary to estimate a lifetime duration, and, data mining when a large set of data with different states (health, damage, …) are provided.
The "Access to data" section gives three alternatives to provide data (it is possible to combine these options): first, the data can be obtained through an e-mail request to the associated contact, second, one or more links to a file on a server can be provided , and third, if a web page associated with the data already exists, it is possible to provide some web links.
Additional files like calibration information, document of the test bench, scripts to read the data can also be provided.



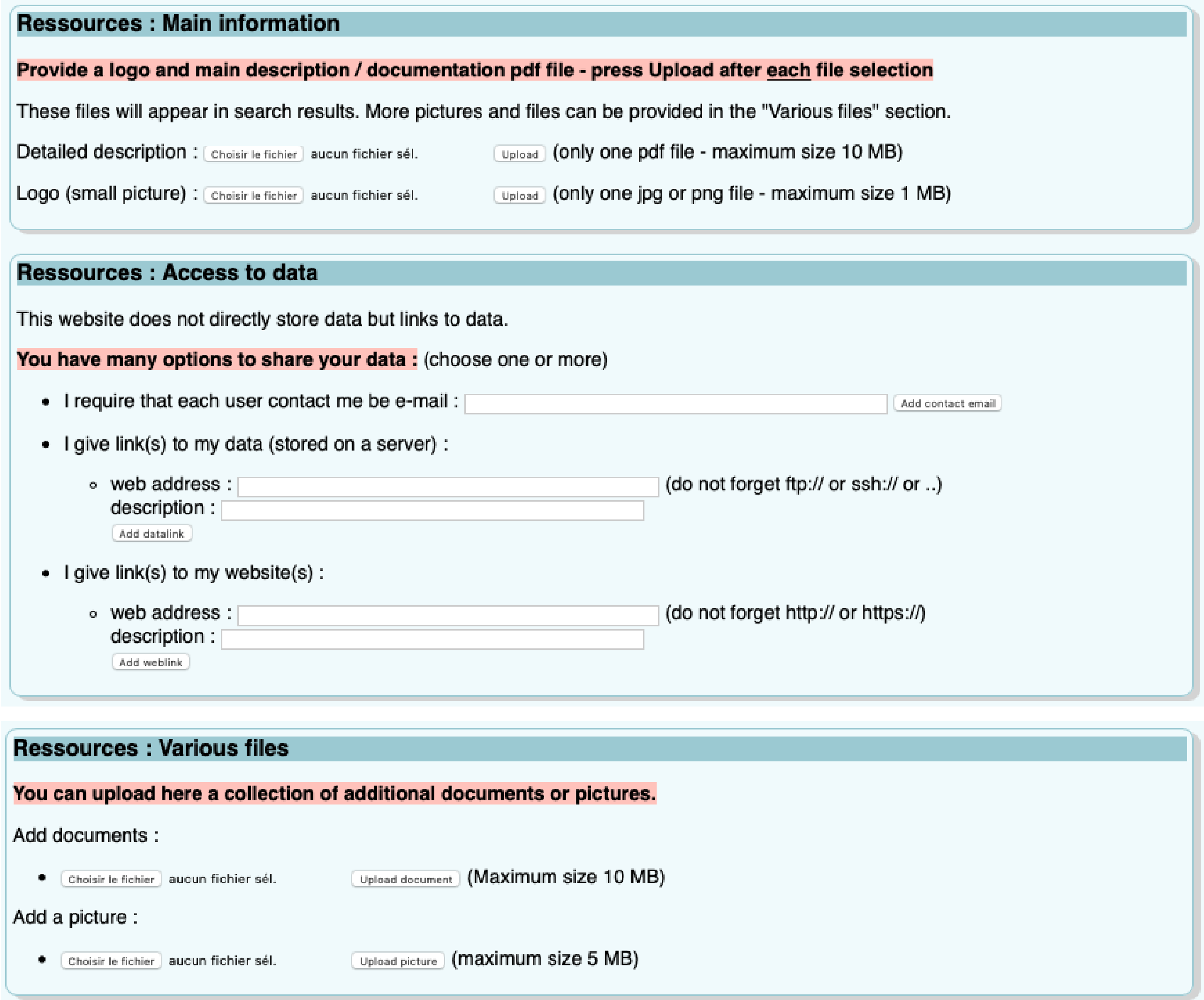


Figure 4: Add my data

4.4 Add my paper

Every user can declare paper associated to a dataset by using the “Add my paper” page. To do so, simply paste the citation of your paper in a format similar to bibliography style and click next. CM Base will analyse your entry and try to identify the author, the title, the journal and so on. The result of this analysis will be presented in a web page and you will be proposed to correct the values and to link your paper to a dataset.

4.5 My Profile

My profile web page allows entering or modifying your personal details, biography, password or delete your account.

4.6 My Contributions

This web page shows a list of submitted datasets as well as associated bibliography. It also gives the possibility to edit or delete the dataset. Since this database is moderated, the dataset should be accepted by the webmaster before being viewable on the web site.

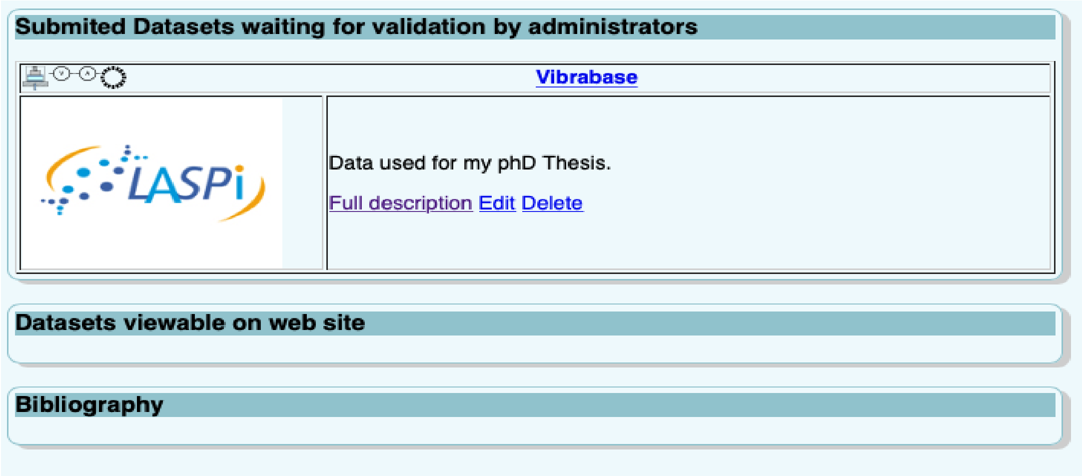


Figure 5: Add my data

5 Conclusion

By proposing a Condition Monitoring data base, CM Base, we aim at filling some gap in our research community. The proposed application will allow finding datasets well-fitted to a specific study, related publications, proposing new datasets and sharing publications. It was designed to be easy to use, and easy to maintain. The structure is such that some new functionality can be added if requested by the users. Some suggestions have already been made: auto alert messages, forums to debate about datasets and methods, challenges proposed by companies on hot topics, groups of datasets used for contest or teaching purpose, overview publications and so on. The interest of the application will depend on its success. The more datasets and papers will be referred to in the base the wider its use will spread. We thus count on your contributions and suggestions to improve the application.

References

[1] M. Baker, Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the crisis rocking science and what they think will help, Nature, vol. 533, no. 7604, 2016, p. 452+.

[2] T. Barszcz, How to bridge academia and industry ? A practitioner's perspective, Keynote lecture, Surveillance 8, Roanne, France, October 20-21, 2015.

[3] R. Klein, *Automatic diagnostics of complex machinery*, Keynote lecture, Surveillance 8, Roanne, France, October 20-21, 2015.

[4] J. Antoni et al, Feedback on the Surveillance 8 challenge: Vibration-based diagnosis of a *Safran aircraft engine*, MSSP, Volume 97, December 2017, Pages 112-144.

[5] *https://data.mendeley.com/*

[6] PHP Notes for Professionals, Free Programming Books, goalKicker.com

[7] G. E. Krasner, S.T. Pope, A Cookbook for Using the Model-View-Controller User Interface Paradigm in Smalltalk-80, JOOP, September 1988.

[8] T. Reenskauag, *The original MVC reports*, thing-model-view-editor, 1979.

[9] J. Gabay, *Merise et ULM pour la modélisation des systèmes d'information*, Dunod, 5ème édition, 2004, ISBN 2 10 007205 6.

Experimental investigation of sensor mounting positions for localized faults detection of epicyclic gear sets

Yu GUO, Zhen LIU, Xing WU, Yinxin YU

Faculty of Mechanical and Electrical Engineering
Kunming University of Science and Technology
Kunming City, Yunnan, China 650500
kmgary@163.com

Abstract

In the literature reports on the vibration based localized faults detection of epicyclic gear sets, the vibration sensor is often mounted on the ring gear or the housing adjacent to the ring gear. However, this sensor mounting position is often too ideal to be utilized in applications. There are different structures of epicyclic gear sets widely used. It is a challenge for selecting a suitable sensor mounting position. In this paper, the sensor mounted on the input/output shaft bearing house for picking up the vibration is experimental investigated, and compared with that from the conventional sensor mounting position. Analysis results shown that the bearing house position can also be employed to expose the fault features of planet gears.

1 Introduction

Planetary gearboxes as important rotating mechanical transmission units are widely used in the wind turbines, helicopter, automobiles, and marine vehicles. However, due to rigorous working conditions in applications, and the wear and crack failures of planetary gear sets occurred frequently [1]. Therefore, the condition monitoring and faults diagnose of planetary gear sets become more and more important.

Compared with the vibration picked up from a fixed-axis gearbox, the vibration picked up from a planetary gearbox are more complicated. In a conventional single-stage planetary gearbox, several planet gears rotate around the sun gear and rotate around their own centers simultaneously, and each planet gear meshes with the sun gear and the ring gear at the same time. On the other hand, the epicyclic motions of planet gears make the mesh positions changed from time to time, which cause the vibration transfer paths between the gear mesh positions and the fix-mounted sensor position are time-varying. As a result, rich modulation phenomena can be observed in the picked up vibration from a planetary gearbox.

As well known that picking vibration is one of important steps in the condition monitoring and faults diagnosis. Generally, it is expected that the picked vibrations contain more significant fault features, which can reduce the requirement to the following signal processing procedures. Then, the sensor mounting position should be considered carefully at first. However, the sensor mounting position of the planetary gearbox is few discussed by literature. In most theoretical literature experimental studies, vibration sensors are often mounted on the ring gear or the housing adjacent to the ring gear [1-3]. However, this sensor mounting position is too ideal to be utilized in applications. Due to different structures of epicyclic gear sets. To some structures, e.g. the ring gear is rotating, the ring gear position is impossible to mount the sensor at all. It is a challenge for the vibration analysis based condition monitoring and faults diagnosis. In addition, part of the research is based on mounting the sensor inside the planetary gearbox for fault diagnosis. For example, mount the sensor on the planet carrier [4,5]. For the built-in sensor installation method, the sensor installation space in a planetary gearbox should be considered firstly, which will put forward new requirements on the design, manufacture and disassembly for the planetary gearbox. In other word, this method has no universal applicability. On the other hand, similar to the sensor installation method for the fixed-axis gearbox picking up the vibration, which mounted the sensor on the input/output shaft bearing housing, should be a feasible sensor mounting position for planetary gearboxes. However, this mounting position has a problem that the vibration transfer path between the mesh point of the ring-plant gear pairs and the sensor is much longer comparing with the conventional ring gear related sensor mounting position. It is worth carrying out an investigation on the different sensor mounting positions for applications of the condition monitoring and faults diagnosis of planetary gearboxes. To address this issue, an experimental investigation has been carried out on a planetary gearbox test rig for the vibration based tooth-root crack faults detection. The well-known vibration separation technique and the synchronous averaging are utilized to extract the faults characteristics through the vibration picked up from different sensor mounting positions for a comparison. Experimental results show that the fault features contained in the observed vibration from bearing housing are weaker than that obtained from ring gear position. However, the bearing housing sensor mounting can also be utilized for the vibration based tooth faults detection by using the vibration separation and the synchronous averaging.

2 Planetary Gearbox Transmission

As shown in Fig. 1, a single-stage planetary gearbox is generally composed of a sun gear, several planet gears, a ring gear and a planet carrier [6]. Different from the conventional fixed-axis gearbox, the vibration transfer paths between the sensor and the meshing points are time-varying under the running condition of the planetary gearbox. Three structures of epicyclic gear sets commonly used are shown in Figs. 2(a)-(c), which correspond to the planetary gear set with the standstill ring gear, with the standstill sun gear, with the rotating sun gear and the rotating ring gear, respectively [7].

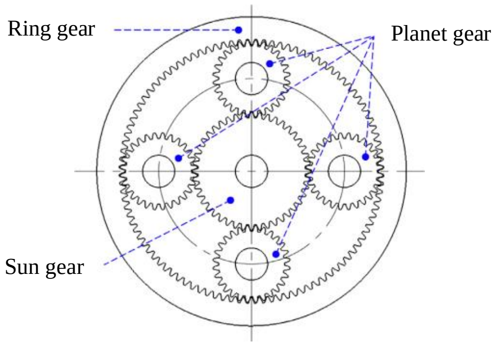


Figure 1: A planetary gear set with 4 planet gears.

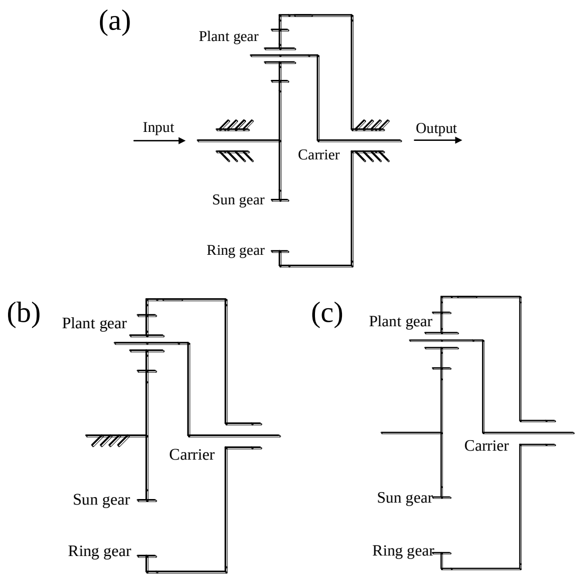


Figure 2 : Structures of planetary gear sets with: (a) standstill ring gear, (b) standstill sun gear, (c) rotating sun and ring gears

In literature reports of the vibration analysis of planetary gear sets, the sensor mounting position is often based on the structure with a standstill ring gear shown in Fig. 2(a). It is noted that the ring gears can also be rotating as the structures shown in Figs. 2(b) and (c), which rise challenges for selecting a suitable sensor mounting position. As well-known that the bearing housing utilized as the sensor mounting position for a fixed-axis gearbox is widely adopted in applications. However, the bearing housing sensor mounting for the vibration analysis of planetary gear sets is still an issue. It has two obviously drawbacks. Firstly, the vibration transfer path is much longer than that of the sensor mounted on the ring gear. Secondly, the interferences from adjacent bearings can lead the picked vibration much noisy. Then, the bearing housing sensor mounting is few reported in the literature. However, the bearing housing sensor mounting can be implemented in most applications of epicyclic gear sets. Then, it is worth investigating whether the bearing housing sensor mounting position is suit for the faults detection of epicyclic gear sets.

In this paper, experimental studies and vibration analysis for the planetary gearbox test rig with a tooth-crack of a planet gear have been carried out based on the vibration picked up by sensor 2 and 3 at the sensor mounting positions shown in Fig. 3, respectively.

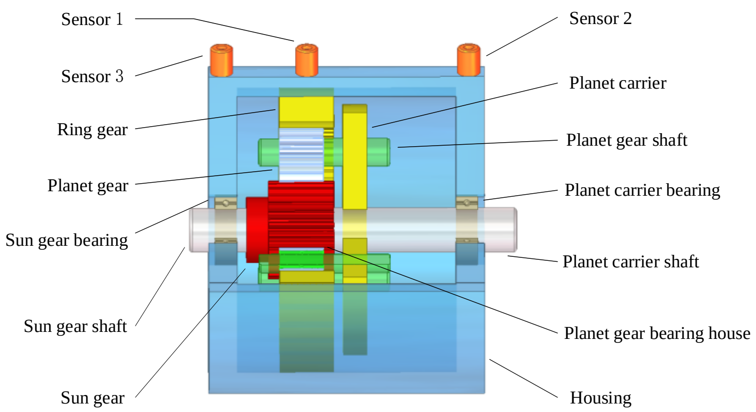


Figure 3: Installation location of sensors.

3 Briefs on Vibration Separation and the Synchronous Averaging

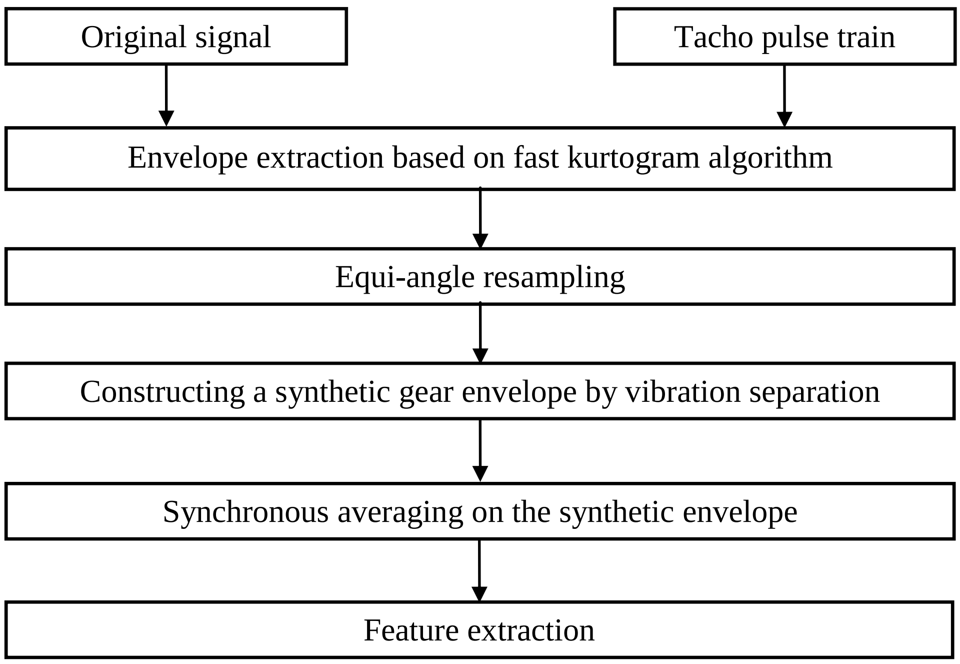


Figure 4 : Brief description of the method flow

As mentioned above, the vibration transfer path between the tooth mesh position and the sensor is time-varying under the condition that the sensor is mounted on the ring gear or the housing adjacent to the ring gear. It is worth mentioning that even though the length of vibration transfer paths seem to be a constant under the condition that the sensor is mounted on the input/output shaft bearing house, the vibration picked up by the sensor is also time-varying. The reason is that the changes of the meshing positions of planet gears make the picked vibration on the bearing house acting as a rotating vector. Therefore, the vibration separation [8,9] and the synchronous averaging [8] should be employed to eliminate the effects from the time-varying vibration transfer path. To extract the weak fault features of planetary gear sets, a combination scheme of envelope extraction, vibration separation and synchronous averaging has been proposed in [6] for the fault diagnosis of planetary gear sets recently. In this scheme, the envelope is demodulated to make the weak impulsive localized fault feature of the planetary gear sets prominent, the well-known vibration separation is utilized to eliminate the influence of time-varying transfer path, and the angular-domain synchronous averaging is used to reduce the influence of speed fluctuation and improve the signal-noise ratio (SNR). The schematic of the envelope-windowed vibration separation is shown in Fig. 4. The main steps are listed as follows [6].

(1) *Envelope extraction*. The fast kurtogram algorithm [10] is employed to extract the complex envelope from a maximum spectral kurtosis value determined demodulation frequency band.

(2) *Equi-angle resampling on the envelope*. Performing the equi-angle resampling scheme on the imaginary and the real parts of the complex envelope in the time domain, the complex envelope in the angular domain can be obtained.

(3) *Constructing a synthetic gear envelope by vibration separation*. The vibration separation technique is applied on the envelope in the angular domain. Then, a synthetic envelope is constructed according to the teeth mesh sequence of the interesting gear.

(4) *Synchronous averaging on the separated envelope.* The synchronous averaging is utilized on the synthetic envelope to remove the non-synchronous components.

(5) *Feature extraction*. The order envelope spectrum is calculated by the Fast Fourier Transform (FFT). Then, the fault feature can be exposed.

More details on the scheme of envelope windowed vibration separation and synchronous averaging can be found in [6].

4 Feature Frequencies of Planetary Gear Sets

Understanding the feature frequencies or orders of planetary gear sets is the core in the condition monitoring and faults diagnosis, by which we can explain the frequency lines in a spectrum reasonably and make a decision on the health status of the planetary gear set. Assuming f_c denotes the rotation frequency of the carrier, the meshing frequency f_m can be given as in [9] by

$$f_m = N_r f_c = N_p(f_p + f_c) = N_s(f_s - f_c) \tag{1}$$

where N_r is the teeth number of the ring gear, N_p is the teeth number of the planet gear, N_s is the teeth number of the sun gear, f_p denotes the absolute rotation frequency of the planet gear, and f_s represents the absolute rotation frequency of the sun gear. Using the carrier as the reference, the corresponding meshing order l_m can be calculated by

$$l_m = \frac{60 f_m}{n_c} = \frac{60 N_r f_c}{60 f_c} = N_r \tag{2}$$

where n_c is the carrier speed. The feature frequency of the planet gear with a tooth-root crack f_p^r is the rotation frequency of planet gear relative to the planet carrier, which is expressed as in [9] by

$$f_p^r = f_p + f_c = f_c(N_r / N_p) \tag{3}$$

And the feature order of the planet gear with a tooth-root crack can be given by

$$l_p^r = \frac{60 f_p^r}{n_c} = \frac{60 f_c (N_r / N_p)}{60 f_c} = N_r / N_p \tag{4}$$

5 Experimental verification

5.1 Experimental description

In the investigation, the experiments have been carried out on a planetary gearbox test rig for the vibration based tooth-root crack faults detection. The test rig is a 75 kW transmission system, which is driven by an AC drive motor with an adjustable speed range from 0 to 2500 rpm. The loading unit is an AC induction generator in tandem with a digital AC drive to regenerate the power back into the system. The test unit is a single-stage planetary gearbox (type: NGW 2K-H) as shown in Fig. 5 with the specific parameters listed in Table 1. The DH904 eddy current sensor is mounted on position 1 shown in Fig. 5 for picking up the tacho pulse train. Three DH112 acceleration sensors are mounted at positions 2, 3, and 4 respectively for picking up the vibrations from the planetary gearbox. A NI 9234 4-channel card is used for the data acquisition with sampling rate 51.2 kHz. It is worth mentioning that the planet carrier shaft is the input and the sun gear shaft is the output. The position 2 is on the sun gear bearing housing, the position 3 is on the housing adjacent to the ring gear, and the position 4 is on the planet carrier bearing housing. The data series used in this study are the vibrations picked up at positions 2 and 3 respectively. And the tacho pulse train is obtained by the eddy current sensor at position 1. In order to simulate the tooth-root crack fault of a planetary gear, a crack of about 4 mm is machined at the tooth root of a planetary gear by the wire cutting method as shown in Fig. 6.

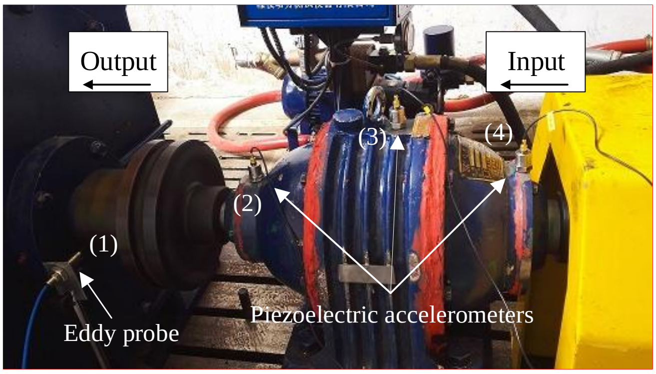


Figure 5 : Test rig of planetary gear transmission

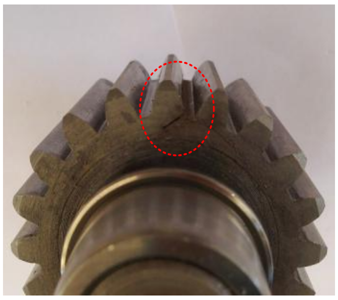


Figure 6 : Planet gear with tooth root crack

Gear	Sun gear	Planet gear	Ring gear
Number of teeth	28	20	71

Table 1: Parameters of planetary gearbox

In the experiment, the rotation speed of the input shaft is at about 1000 rpm. The characteristic orders of the planetary gear in the planetary gearbox can be calculated by Eqs. (2) and (4) in theory, which are listed in Table 2 by using the planet carrier as the reference.

Planetary gearbox meshing order l_m	71.00 ×
Planetary fault order l_p^r	3.55 ×
Planetary carrier frequency order l_c	1.00 ×
Planetary frequency frequency order l_p	2.55 ×

Table 2 : Characteristic orders of the planetary gearbox

5.2 Experimental data analysis

The waveforms in of the raw vibration picked up by the sensors mounted at positions 2 and 3 are shown in Figs. 7 and 8, respectively. The tacho pulse train obtained by the eddy probe at position 1 and the corresponding speed curve are shown in Figs. 9 and 10.

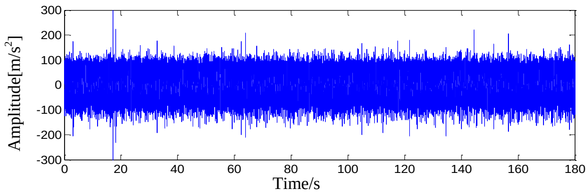


Figure 7 : Vibration waveform observed at position 2

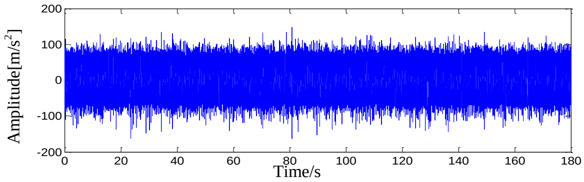


Figure 8: Vibration waveform observed at position 3

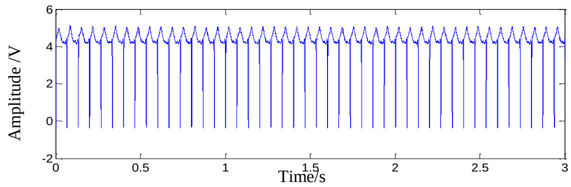


Figure 9: Speed pulse

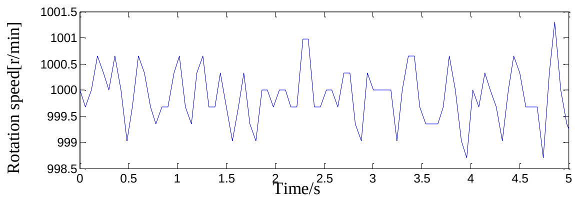


Figure 10 : Speed curve

Using the vibration separation and the synchronous averaging techniques, the order spectra of the vibrations picked up at the positions 2 and 3 are shown in Fig. 11, where the characteristic order (3.55×) and its harmonics related to the tooth-root crack of a planet gear order are exposed clearly. Experimental results show that the fault feature contained in the observed vibration from the bearing housing (position 2) is weaker than that obtained from the ring gear position (position 3). However, it is worth noting that the bearing housing sensor mounting position can also be utilized for the vibration based tooth faults detection by using the vibration separation and the synchronous averaging techniques.

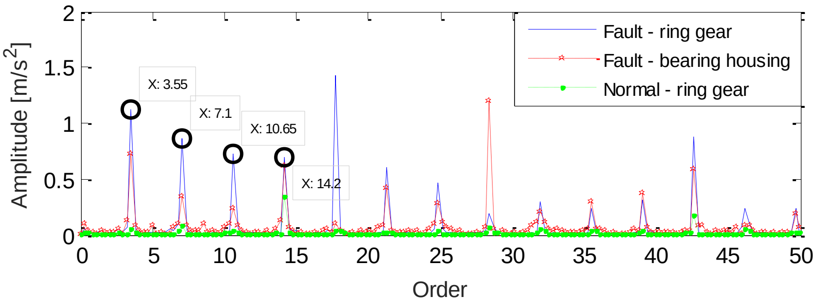


Figure 11 : Fault characteristic order of planet gear fault

6 Conclusion

In practical applications, different structures of epicyclical gear sets widely used. According to a specific structure of a planetary gear set, the sensors should be mounted on suitable positions to pick up vibration. The experimental investigation results show that the sensors mounted on the position of the bearing housing and mounted on the housing adjacent to the ring gear both can be utilized for the vibration based tooth faults detection by using the vibration separation and the synchronous averaging techniques. Moreover, the experimental results also show that the bearing housing sensor mounting position is also can be employed for the faults detection of planetary gear sets, but the fault feature is weaker than that obtained from ring gear position.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 51675251).

References

[1] Y. Guo, L. Zhao, X. Wu, et al., *Vibration separation technique based localized tooth fault detection of planetary gear sets: A tutorial*, Mechanical Systems and Signal Processing, 2019, 129: 130-147.

[2] Y. Lei, D. Han, J. Lin, et al., *Planetary gearbox fault diagnosis using an adaptive stochastic resonance method*, Mechanical Systems and Signal Processing, 2013, 38(1): 113-124.

[3] Z. Feng, M.J. Zuo, J. Qu, et al., *Joint amplitude and frequency demodulation analysis based on local mean decomposition for fault diagnosis of planetary gearboxes*, Mechanical Systems and Signal Processing, 2013, 40(1): 56-75.

[4] Z. Fan, H. Li, *A hybrid approach for fault diagnosis of planetary bearings using an internal vibration sensor*, Measurement, 2015, 64(384):71-80.

[5] W. Smith, L. Deshpande, R. Randall, et al., *Gear diagnostics in a planetary gearbox: a study using internal and external vibration signals*, International Journal of Condition Monitoring, 2013, 3(2):36-41.

[6] Y. Guo, L. Zhao, X. Wu, et al., *Tooth Root Crack Detection of Planet and Sun Gears Based on Resonance Demodulation and Vibration Separation*, IEEE Transactions on Instrumentation and Measurement, 2019. doi: 10.1109/TIM.2019.2893011.

[7] Y. Lei, J. Lin, M.J. Zuo, et al., *Condition monitoring and fault diagnosis of planetary gearboxes: A review*, Measurement, 2014, 48:292-305.

[8] P.D. McFadden, *A technique for calculating the time domain averages of the vibration of the individual planet gears and the sun gear in an epicyclic gearbox*, Journal of Sound and Vibration, 1991,144(l):163-172.

[9] P.D. Samuel, J.K. Conroy and D.J. Pines, *Planetary Transmission Diagnostics*, NASA/CR 2004-213068.

[10] J. Antoni, *Fast computation of the kurtogram for the detection of transient faults*, Mechanical Systems and Signal Processing, 2007, 21(1): 108-124.

Towards 3D AFM Using Multiple Vibration Modes

Rubin Eyal[1], Bucher Izhak[1], Davis Solomon[1]

[1]Dynamics Laboratory, Mechanical Engineering, Technion
Haifa 3200003, Israel
bucher@technion.ac.il

Abstract

Atomic force microscopy (AFM) is used for measuring nano-scale topographic features. By exciting a micro-cantilever with a sharp stylus at its tip, at or near resonance, a Frequency Modulated AFM (FM-AFM) device can sense the change of resonance frequency due to the change in tip-surface Van der Waals (VdW) potential. The topography is then retrieved from the relationship between the potential and the distance between stylus and the specimen. To improve the measurement speed and address complex geometries emerging in industrial microchip constructions, several enhancements are introduced. While most FM-AFM devices operate in a single vibrating mode, this article enhances existing sensing methods by extending to multidimensional sensing the resonance frequencies that are modulated by the topology, in several orthogonal vibration modes simultaneously. The latter opens new possibilities, e.g. to measure steep walls and trenches or other complex geometries. An Autoresonance (AR) control scheme for faster excitation, and fast frequency estimation algorithm were used for sensing several modes simultaneously, without the need to wait for steady state settling of the cantilever. The concept was tested on a large-scale experimental system, where VdW forces between tip to surface were replaced by magnetic forces, using a magnetic tip and ferromagnetic samples. Experimental results employ 3D relevant topographies such as inclined surfaces, steep walls and trenches that were reconstructed experimentally with 4 (μm) resolution or better. Downscaling to typical AFM dimensions would theoretically yield sub-nanometer resolution. Numerical and experimental data are shown to demonstrate the advantageous of the new approach.

1 Introduction

Atomic force microscopy (AFM) first presented by Binning et al. [1] is used for measuring nano-scale topography features in the semiconductors industry, and for atomic resolution measurments in physics and biology research. The non-contact configuration of the sensor (NC-AFM) is based on a micro-cantilever with a sharp stylus at its tip, vibrating at or near resonance. The tip interacts with the surface, while the distance dependent Van der Waals (VDW) force between the tip and the surface, acts on the cantilever, altering its resonance frequency. A common technique used today is the frequency modulated AFM (FM-AFM), presented by Albrecht et al. [2]. The latter showed that the measured frequency shift correlate to the surface topology.

The present work enhances the one-dimensional measurement capability of current AFM devices to multidimensional measurement capability by allowing the measuring tip to vibrate at two or more directions simultaneously. Here, each direction of excitation is associates to a distinct mode of vibration, hence has a distinct resonance frequency. By measuring the frequency shift in each mode separately, one can evaluate the tip to surface distance in each measured direction [3]. An Auto-resonance (AR) control scheme combined with modal filtering [4] is used for fast frequency tracking of the different modes, and a fast frequency estimation algorithm [5] is used for evaluating the frequency shift. Using this method for two (or more) spatial modes simultaneously results with a multi-dimensional measurement, that has the ability to measure and reconstruct complex geometries. In addition, Since the AR locks into the natural frequency instantaneously, the measuring process can be expedited. This concept was demonstrated in the lab, by a large-scale sensor, where the VdW forces have been replaced by magnetic ones, and the sensing tip is a passive magnet. The sensing tip interacts with a ferromagnetic specimen having a complex topology.

The paper introduces the multidimensional vibrating sensor, demonstrated on a large-scale experimental system in section 2. Section 3 shows some experimental results that are analysed to assess the sensitivity and resolution of the current system. Finally, Section 4 draws some conclusions.

2 Experimental system layout, actuation and sensing

This chapter explains the multimode actuation and sensing concept on the laboratory large-scale experimental system, and the way to interpret the measured data to reconstruct the topology of the measured specimen.

2.1 Multimode actuation and sensing

The measuring magnetic tip is attached to a rectangular cantilever beam which has two orthogonal modes of vibration with distinct natural frequencies at 447 (Hz) and 549 (Hz). Finite element model of the cantilever showing the two direction of motion is illustrated in Figure 1.

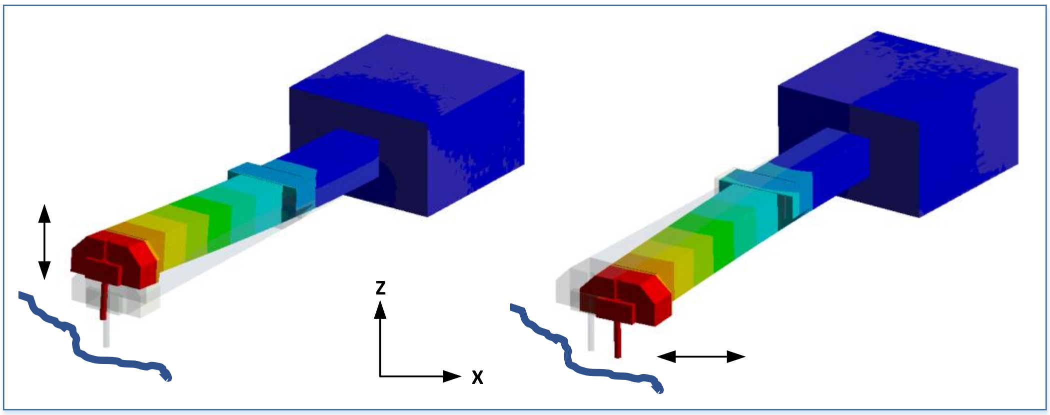


Figure 1: Multidirectional vibration- Left: vetical motion of the tip (z), Right: horizontal motion of the tip (x)

The interaction between the magnetic tip and the surface, due to the magnetic field, contributes to the total potential energy of the cantilever undergoing bending. Hence, one can write:

$$V_{Total} = V_{cantilevel} + V_{ts} = \frac{1}{2}\left(k_x x^2(t) + k_z z^2(t)\right) + V_{ts}(x(t), z(t)) \quad)1($$

where V_{ts} is the tip to surface interaction potential, k_x, k_z are the modal stiffness and $x(t), z(t)$ are the time dependent modal coordinates of the tip associated with the direction of vibration in x, z respectively.

The magnetic potential depends on the distance between the magnetic tip and the ferromagnetic surface. Hence, the interaction contribution can be added to the constant stiffness matrix of the cantilever:

$$\mathbf{K} = \begin{bmatrix} k_x & 0 \\ 0 & k_z \end{bmatrix} + \begin{bmatrix} \frac{\partial^2 V_{ts}}{\partial x^2} & \frac{\partial^2 V_{ts}}{\partial x \partial z} \\ \frac{\partial^2 V_{ts}}{\partial x \partial z} & \frac{\partial^2 V_{ts}}{\partial z^2} \end{bmatrix} \triangleq \mathbf{K}_0 + \Delta\mathbf{K} \quad (2)$$

Non-zero of-diagonal terms in $\Delta\mathbf{K}$ indicates that the specimen is inclined. In the present research, the information in x and z appears at 2 distinct frequencies and we obtain only the diagonal terms of $\Delta\mathbf{K}$, one at a time. The distance of the tip to the specimen is deduced from these terms only.

In order to excite each direction at a different frequency simultaneously we employ modal control according to the configuration shown in Figure 2. The cantilever is excited with two voice-coil actuators placed symmetrically at $45°$ on both sides of the cantilever. Two laser displacement sensors (Keyence LK-H008) are also positioned at $45°$ on both sides of the cantilever measuring the displacement of the cantilever close to the tip. Modal filtering [6] is performed digitally in order to project the measured data and excitation forces on the exact modes directions.

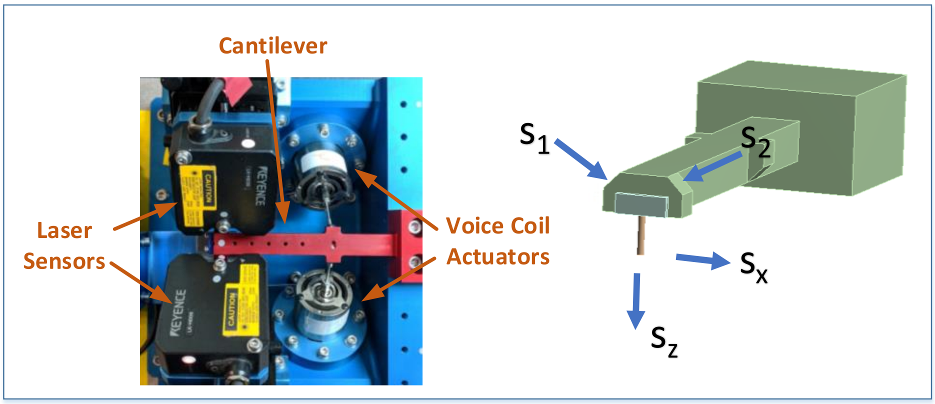


Figure 2: Left: laboratory realization using 2 voice-coil actuators acting at 45 degrees, and two optical (Keyence™) sensors measure in the same directions. Right: Illustration of the measurement of two sensors from inclined surfaces on the cantilever close to the magnetic tip.

The modal filtering (MF) uses the bi-orthogonality between the vibration modes, by taking a linear combination of the two sensors signals so that all modes are canceled but one. An illustration for the modal filtering of the displacement signals, measured by the laser sensors, is presented in

Figure 3 , where S_1, S_2 are real measured signals and S_z, S_x are the modal filtered displacements from the following equation:

$$\begin{aligned} S_x(t) &= S_1(t) - S_2(t) \\ S_z(t) &= S_1(t) + S_2(t) \end{aligned} \tag{3}$$

The separation of modes can be clearly seen on the frequency domain after applying a fast Fourier transform (FFT) to the measured and modal-filtered signals. The latter is demonstrated on the measured signals, and a similar approach is performed on the excitation forces operating on the cantilever by two voice-coil actuators.

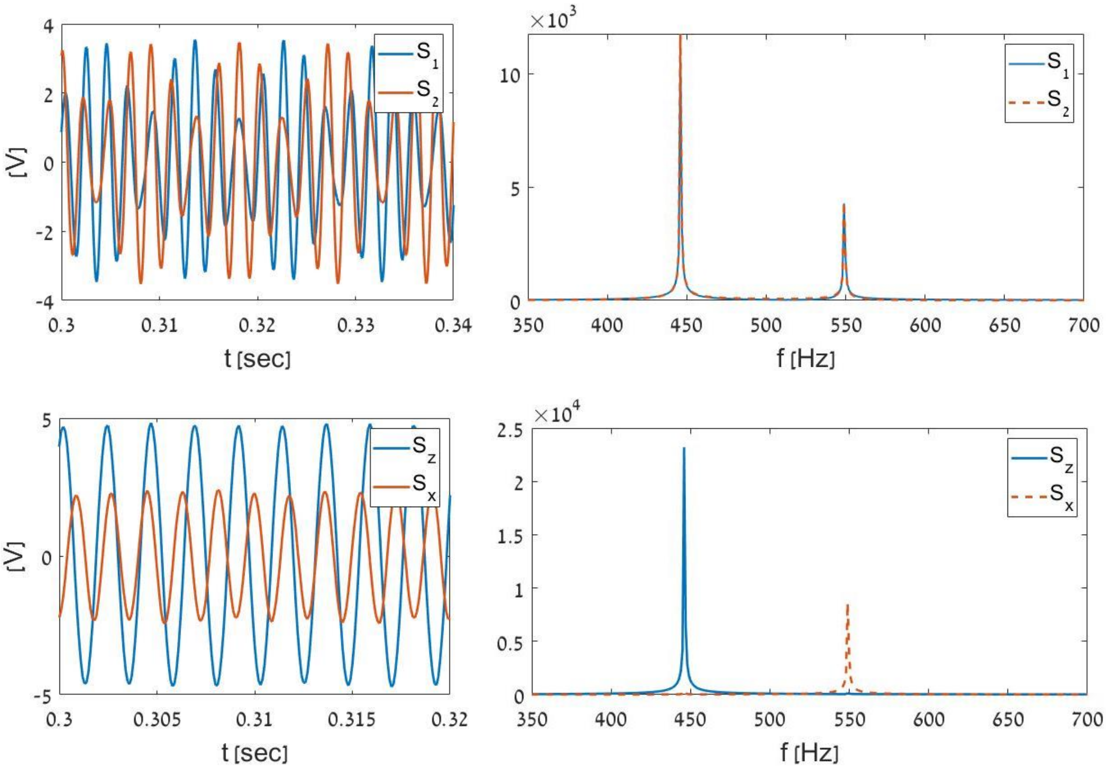


Figure 3: Top-left: measured signals S_1, S_2 from the experimental system. Top-right: S_1, S_2 frequency content showing both resonance frequencies are present in both measured signals. Bottom-left: S_x, S_z after applying modal filtering (3). Bottom-right: S_x, S_z frequency content showing the separation of frequencies in the modal-filtered signals.

Modulation of the natural frequency is detected by auto-resonance (AR) resonance tracking control scheme, combined with a new frequency estimation algorithm [5] , that uses Linear Least Squares (LLS) to fit an instantaneous phase to the noisy signal for several periods, and then fits a time dependent line whose slope is an estimate of the signal's frequency. Autoresonance, also known as self-excitation, is a well-known nonlinear feedback method used for automatically exciting a system at its natural frequency [7], [8]. It locks into resonance from the first cycle, and has the potential to increase the imaging speed compared to the common scheme employed in AFM - Phase Locked Loop (PLL), whose frequency locking speed is greatly reliant on the settling time of the cantilever. Both resonance tracking and frequency estimation are performed simultaneously for the two modes of vibration using modal filtering.

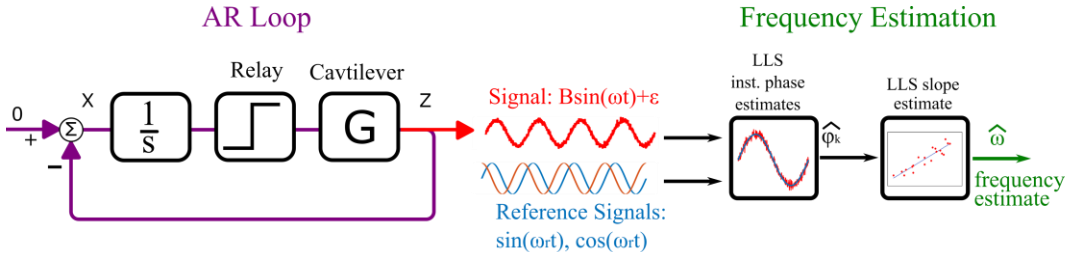


Figure 4: Schematic digram of the resonance tracking and frequency estimation. The AR feedback loop described here consist of a phase shifting element P which shifts the phase of the input signal by 90 degrees, and a digital 'relay' or sign function that forces the amplitude of the input signal to constant values [4].

Assuming white noise, the variance of the frequency estimated using the frequency estimation algorithm [5] depends significantly on the total estimation time t_N, and also on the sampling rate and on the SNR - signal to noise ratio between the variance of the single tone signal without the noise, to the variance of the noise. This variance bounds the minimal frequency shift (Δf_{min}) that can be measured by the sensor. Measurement accuracy and speed should be balanced for a desired working point.

$$\sigma_f^2 \approx \frac{12}{F_s t_N^3 SNR} \tag{4}$$

2.2 From frequency modulation to surface topology identification

Two main scanning methods can be used to scan a specimen [9]: Constant height or constant gap (frequency). In the former the probe is located in a constant height above the surface, and the change is the resonant frequency is measured and then converted into the change in surface topography. In the latter, the gap between the probe and the sample stays constant using a standard PI control scheme that keeps a constant natural frequency for single and dual mode excitation. In the large-scale demonstrator presented in this paper, an XYZ stage instantaneous location (with a $0.2\,\mu m$ resolution) was used for the surface topology estimation.

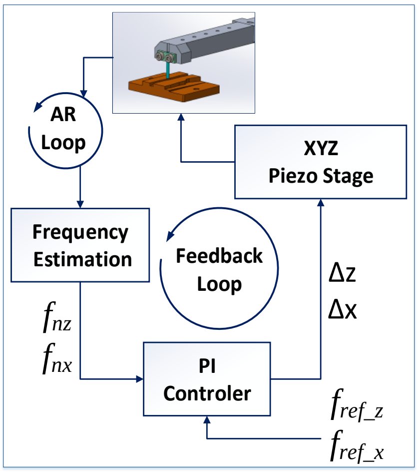


Figure 5: Control scheme for position control. The feedback loop keeps a constant resonance frequency by changing the position of the sample relative to the vibrating tip, keeping a constant gap between them. The XYZ stage displacements are then use to reconstruct the surface topology.

2.3 Laboratory experimental system implementation

The abovementioned measuring methodology is demonstrated using a large-scale experimental system. A schematic layout of the system is presented in Figure 6. The main subsystems consist actuation of the cantilever beam using two voice-coil actuators, measurement of the beam tip displacement using two optical sensors (Keyence™), and specimen displacement using XYZ accurate piezo stage (Nanomotion™). All the vibrating parts are mounted on an optical table through a rigid aluminum base in order to decrease the peripheral vibrations of the system, as detailed in Figure 7. The measured and actuation signals are generated in real time by a Xilinx FPGA connected to a dSpace™ signal processor, and are integrated to Matlab, together with the XYZ stage control.

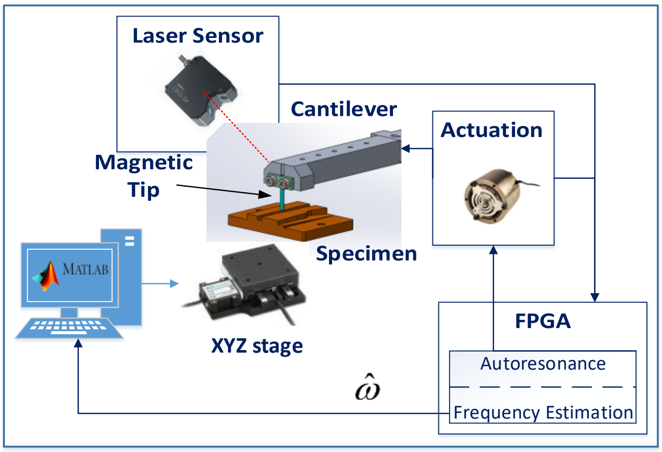


Figure 6: Large-scale experimental system layout

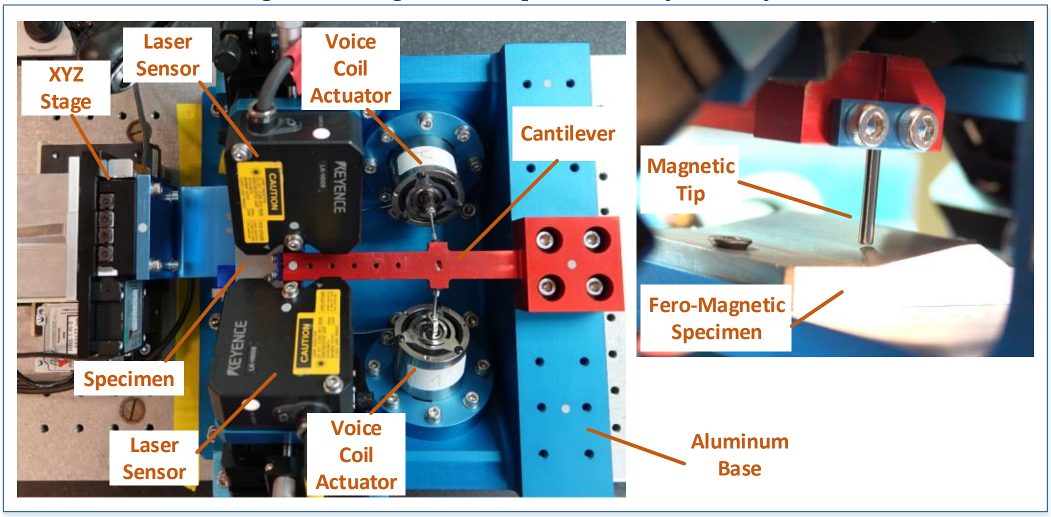


Figure 7: Detailed view of the large scale experimental system. Cantilever beam (in red) length- 130 mm and cross-section- 12mm X 15mm. Magnetic tip diameter- 2 mm and length- 20 mm

3 Experimental results

This chapter will present some of the main measured results of the large-scale experimental system, together with resolution and performance analysis and a relevant down-scale approximation to a real AFM.

3.1 Main results

First result in fig8 show the nonlinear dependency between the tip-surface distance (gap) and resonance frequency that effects the sensitivity of measurement for different gaps. Though, for small gaps and small oscillations (i.e. around 0.1 mm), linear relationship is obtained. Furthermore, once the control system (presented in chapter 2.2) moves the specimen to obtain the same reference frequency, the gap also remains the same, and the nonlinear dependency is effectively eliminated. Hence, the measurement is carried out with the same sensitivity.

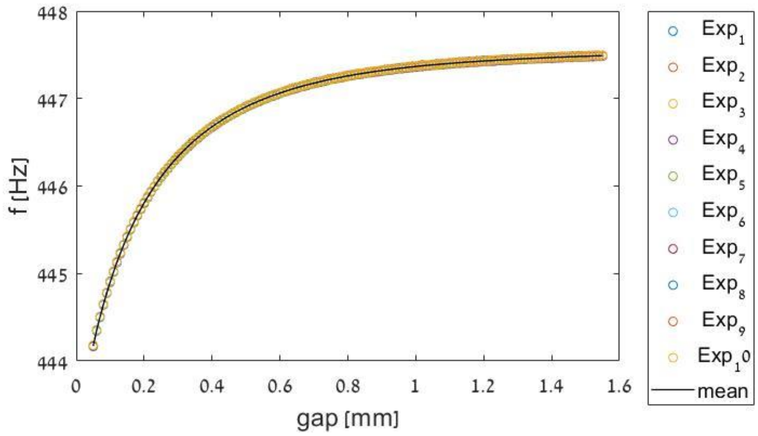


Figure 8: Dependency of resonance frequency on the distance between the measuring tip and the specimen. The experiment was repeated 10 times showing repeatability.

Steep and complex geometries, such as walls, grooves and trenches cannot be fully reconstructed by measuring a single spatial direction. Using the multidimensional measurement method presented in this paper, complex geometries were reconstructed experimentally as demonstrated in Figure 9: Left: measuring tip next to the steep walled and narrow trench specimen. Right: reconstructed geometry showing both the geometry of the wall and of the inclined bottom surface (3D).Figure 9 and Figure 10.

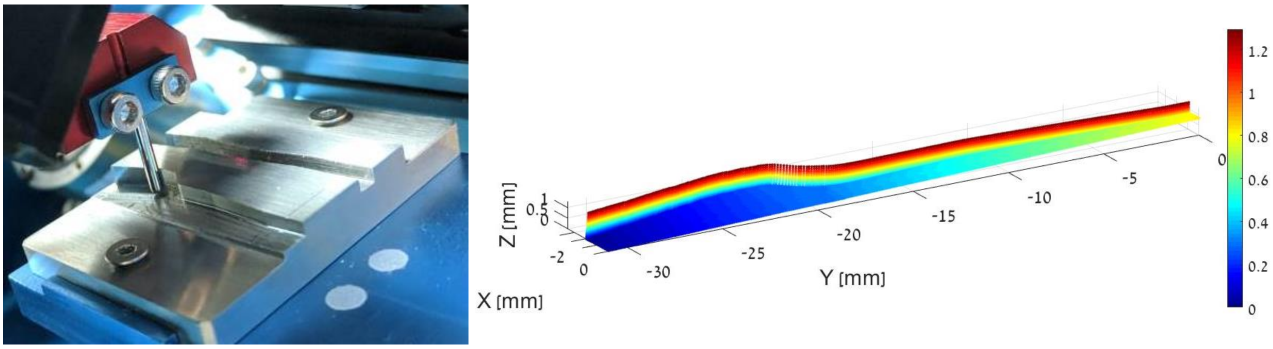


Figure 9: Left: measuring tip next to the steep walled and narrow trench specimen. Right: reconstructed geometry showing both the geometry of the wall and of the inclined bottom surface (3D).

The narrow trench geometry shown on the left in Figure 10 has been reconstructed, including the steep walls curvature and slight machined angle of the bottom surface. Still, since the magnetic field is not as localized as VdW forces in AFM, parasitic cross coupling effects distort the reconstructed geometry. These effects are not expected to occur on the nanoscale where tip-sample interaction, dominated by the local VdW forces, is more localized then the magnetic forces employed here.

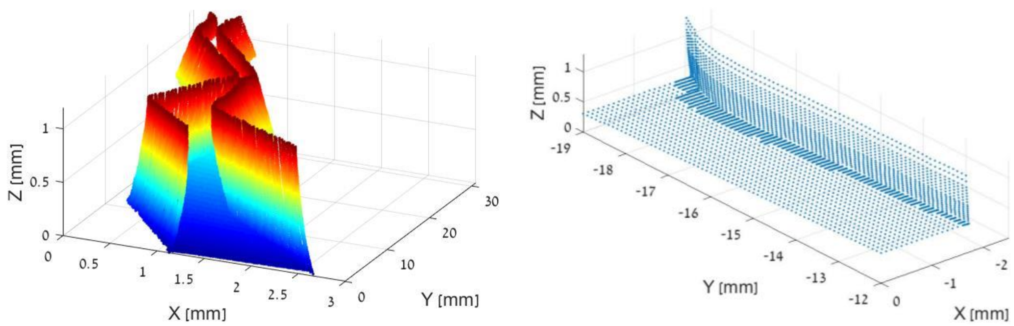


Figure 10: Left: reconstructed geometry of a narrow trench. Right: scanning step size manipulation using multiple mode information. Near the right angled corner, small steps were taken, which was made possible by sensing both vertical and horizontal gaps

The right side of **Error! Reference source not found.** illustrates the use of the additional information in each measuring point for improving the scanning algorithm, e.g. decreasing the scanning step size in advance, while approaching a vertical wall by using the horizontal sensing mode to sense that the geometrical gradient become large.

3.2 Performance analysis

The experimental system is a large scaled AFM, hence the performance should be compared to a commercial FM-AFM [10] using relevant scaling factor. Magnet size, shape and orientation, spatial resolution of XYZ stage, frequency estimation resolution, sensors resolution, external disturbances and noise are some of the many parameters affect the measuring resolution of the large-scale experimental system. The resolution of the commercial AFM was measured for a specific image size, hence a similar method was exploit in the large-scale system. As consequence, a suitable non-dimensional comparison is the ratio between the resolution and the image size. The resolution of the large-scale experimental system was measured by scanning a flat rectangular surface with 10,201 data points, and calculating the standard deviation of the measurements. The RMS measured resolution is 1.3 (μm) for the vertical mode (z), and 4 (μm) for the horizontal mode (x). The reason for the difference between the modes is the spatial effects of the magnetic field created from the cylindrical shape of the magnet. The magnetic forces that affect the frequency shift are larger in the vertical direction, parallel to the axis of the cylindrical magnet, hence the resolution in the vertical direction is better.

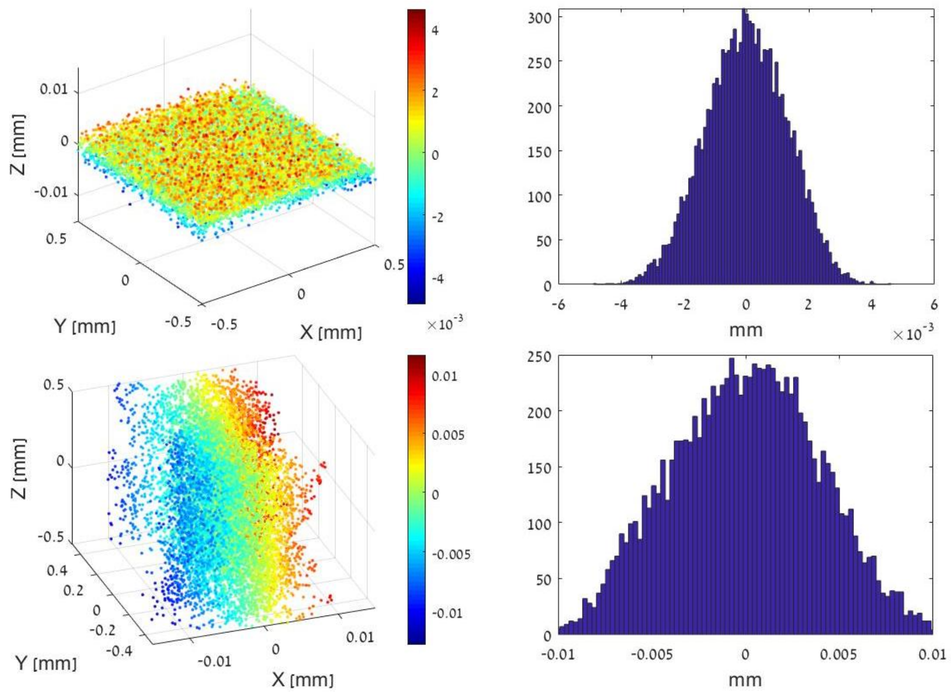


Figure 11: Total resolution measurement of the large-scale experimental system. Top left: vertical mode (z) measured data points. Top right: statistical analysis for the vertical mode showing a histogram representing the measured gap probability distribution. Bottom left: horizontal mode (x) measured data points. Bottom right: statistical analysis for the horizontal mode.

Table 1 compares the large-scale experimental system to a commercial FM-AFM system [10] showing similarity in non-dimensional parameters $\Delta f_{min} / f_0$ and resolution/image size, where f_0 is the basic resonance frequency of the cantilever, and Δf_{min} is the minimal measured frequency shift. The similarity approves the validity of the up-scaling done in the demonstration of the multidimensional measuring method.

Parameter	Commercial AFM [10]	Large Scale System Vertical Mode (z)	Large Scale System Horizontal Mode (x)
f_0 (Hz)	$330\cdot10^3$	447	549
Δf_{min} (Hz)	20	0.012	0.012
$\Delta f_{min} / f_0$	$6.06\cdot10^{-5}$	$2.68\cdot10^{-5}$	$2.18\cdot10^{-5}$
image size	0.2 (μm)	1 (mm)	1 (mm)

resolution	0.38 (nm)	1.3 (μm)	4 (μm)
resolution/image Size	0.0019	0.0013	0.004

Table 1: comparing parameters between a commercial AFM to the large-scale experimental system

4 Conclusions

A method able to measure complex geometries with multidirectional probe has been presented. The method combines modal filtering with Autoresonance, for fast resonance tracking, simultaneously in two or more spatial directions. The method was demonstrated on a large-scale experimental system that was able to reconstruct 3D relevant topographies, such as inclined surfaces, steep walls and trenches, with 4 (μm) resolution or better. A nano-scale system based on the same principles is currently being constructed, and is expected to improve measurement speed and ability to measure complex geometries with Angstroms resolution.

Acknowledgments

This research has been funded by The Israel Innovation Authority, within Multi-Dimensional Metrology (MDM) Consortium, 2018-2020.

References

[1] G. Binnig, C. Gerber, and C. F. Quate, "Atomic Force Microscopy," *Phys. Rev. Lett.*, vol. 56, p. 930, 1986.

[2] T. R. Albrecht, P. Grütter, D. Horne, and D. Rugar, "Frequency modulation detection using high-Q cantilevers for enhanced force microscope sensitivity," *J. Appl. Phys.*, vol. 69, no. 2, pp. 668–673, 1991.

[3] F. J. Giessibl, "A direct method to calculate tip-sample forces from frequency shifts in frequency-modulation atomic force microscopy," *Appl. Phys. Lett.*, vol. 78, no. 1, pp. 123–125, 2001.

[4] S. Davis and I. Bucher, "Automatic vibration mode selection and excitation; combining modal filtering with autoresonance," *Mech. Syst. Signal Process.*, vol. 101, pp. 140–155, 2018.

[5] S. Davis and I. Bucher, "A new approach to single-tone frequency estimation via linear least squares curve fitting," Apr. 2019.

[6] L. Meirovitch and H. Baruh, "Control of Self-Adj oint Distributed-Parameter Systems," *J. Guid. Control. Dyn.*, vol. 5, no. 1, pp. 60–66, 1982.

[7] W. E. Vander Velde, "Multiple-input describing functions and nonlinear system design." McGraw-Hill, New York, 1968.

[8] K. J. ÅStröM, "Oscillations in Systems with Relay Feedback," pp. 1–25, 2011.

[9] N. C. Santos and F. A. Carvalho, *Atomic Force Microscopy Methods and Protocols*. 2019.

[10] H. Seo, D. Goo, G. Jung, R. Stomp, and J. Wei, "Park NX-Hivac : Phase-lock Loop for Frequency Modulation Non-Contact AFM," *Park AFM SYSTEMS*. [Online]. Available: https://www.parksystems.com/index.php/applications/manufacturing/semiconductor/197-high-throughput-and-non-destructive-sidewall-roughness-measurement-using-3-dimensional-afm. [Accessed: 04-Apr-2019].

Fault diagnosis and prognosis for rolling bearings

Early bearing defect detection in a noisy environment based on a method combining singular value decomposition and empirical mode decomposition

Mourad Kedadouche, Zhaoheng Liu and Marc Thomas
Department of Mechanical Engineering, École de technologie supérieure,
Montréal, QC H3C 1K3, Canada
mourad.kedadouche@hotmail.fr, zhaoheng.liu@etsmtl.ca, marc.thomas@etsmtl.ca

Abstract

This paper proposes a new method combining Empirical Mode Decomposition (EMD) and Singular Value Decomposition (SVD) for bearing fault diagnosis. The method includes three steps. First, the signal is decomposed using EMD. Secondly, the instantaneous amplitudes are computed for each component using the Hilbert Transform (HT). Lastly, the Singular Value Vector is applied to the matrix of Cross-Power Spectral Density (CPSD) of the instantaneous amplitude matrix and the SVD versus frequency is analysed. The proposed method is first validated by using various noisy simulated signals. The results show that the proposed method is robust versus the noise to detect the bearing frequencies that are representative of the defect even in a very noisy environment and that the amplitude of the first SVD at each bearing frequency is very sensitive to the defect severity. The proposed method is also applied to two different experimental cases on bearings with very low degradation. The results show that the proposed method is able to detect bearing defects at an early stage of degradation for both experimental cases.

Keywords: Bearing fault, Empirical Mode Decomposition (EMD), Hilbert transform (HT), Cross-Power Spectral Density (CPSD), Singular Value Decomposition (SVD).

1. Introduction

Bearing wear can be considered as a major cause of defects in rotating machinery. Unexpected failures in bearings may cause significant economic losses. Empirical Mode Decomposition (EMD) is an interesting technique for fault diagnosis of rotating machinery. EMD can decompose the signal into several components called Intrinsic Mode Functions (IMFs) [1]. With EMD, the principal “modes” representing the signal can be identified. This method has attracted much attention for signal processing and engineering applications over the past decade [2]. The fundamental idea when using the EMD method is to decompose the vibratory signal into multiple components and the suitable IMF allows for computing the envelope spectrum and analyse their statistical features. Hybrid methods based on EMD and other techniques, like the Wavelet Packet Transform (WPT), the Support Vector Machine (SVM), Spectral Kurtosis (SK) and the Teager-Kaiser Energy Operator (TKEO), have also been applied to bearing fault diagnosis [3–7].

After performing EMD on a signal, some IMFs are associated to bearing faults, others with information unusable for diagnosing such faults. The useful IMF (if it exists) can be selected to perform the Hilbert spectrum. A few studies focus on developing an indicator to select automatically this useful IMF. Wenliao *et al*.[8] used the Wigner-Ville distribution to select the optimum IMFs and the filter bandwidth . Ricci *et al* [9] proposed a new indicator, named Merit Index, to select the appropriate IMF. The Merit Index is a linear combination between the periodicity degree of the IMF and its absolute skewness value. Yi *et al.*[10]

proposed a new indicator called Confidence Index based on combination of correlation coefficient, skewness and kurtosis. Jacek *et al*.[11], Peng et al. [12], Wang *et al*. [13], and Guo and Tse [14] utilized the correlation coefficient as an indicator analysis to select the candidate IMFs.

However, it is well known that the repetitive impacts due to the defect excite all bearing natural frequencies. If only the best IMF is selected, information included in other IMFs excited by the fault is then lost. Selecting all IMFs excited by the fault appears thus more suitable. In [15-16], the authors propose to select all the IMFs excited by the fault. A hybrid method based on EMD and run-up excitation is proposed to select the useful IMFs. By using a swept excitation when running up a rotating machine, the resonance frequency bands of the mechanical system is obtained from the spectrogram of the signal. In [17-18], the authors proposed to select all IMFs selected by the fault for early detection of the defect. The selection is made through an indicator-based kurtosis.

In this study, a new approach exploiting all IMFs of the signal to improve fault diagnosis is proposed. To compress all information extracted from each IMF, Singular Value Decomposition (SVD) is used in this paper. The SVD method has been widely used in fault feature extraction and identification for mechanical systems [19–21]. Before performing SVD, a matrix obtained from the original one-dimensional signal must be constructed. Various matrices exist, for example: the Toeplitz matrix, cycle matrix and Hankel matrix. To improve fault feature extraction, a new approach exploiting the instantaneous amplitude of each IMF obtained by EMD is proposed. The matrix is defined as the Cross-Power Spectral Density (CPSD) of all instantaneous amplitudes of the obtained IMF. Cross-spectral analysis is a powerful tool for investigating the relationship between signals in the frequency domain. Inspired by the frequency domain technique [17], the power spectral density matrix for each frequency is decomposed by applying SVD to the matrix. The singular value plot of the spectral density matrix concentrates information from all spectral density functions. The first singular value should approximately equal the sum of the terms on the diagonal of the PSD matrix. This means that the power of the signals at this frequency can be attributed to the vibratory signature [22]. The following sections give details of the proposed method used for rolling bearings fault diagnosis. The paper first presents the theoretical background of EMD and the proposed approach in Section 2. To validate the approach, the method is applied to a simulated signal and real data from damaged bearing in sections 3 and 4. Section 5 concludes the paper.

2. The proposed approach

2.1 A brief description of EMD

The EMD method can decompose signal in a multiple intrinsic mode functions (IMFs). The decomposed signal may be written as [1-2]:

$$x(t) = \sum_{i=1}^{N} C_i(t) + r_N(t) \tag{1}$$

where $C_i(t)$ is the $ith\ IMF$ and $r_N(t)$ is the residual signal.

This method could suffer of a mixing mode problem and derived methods such as EEMD, CEEMD, CEEMDAN could also be used in this case, but it is not always necessary.

2.2 The proposed approach based on SVD and EMD

The approach proposed for diagnosing faults on rolling bearings is illustrated in Figure 1. In this paper, the signal matrix obtained from EMD is constructed as follows:

$$[C_1; C_2; C_3; ...C_n] \tag{2}$$

The first step is to compute the instantaneous amplitude of each intrinsic mode function $C_i(t)$. The instantaneous amplitude (*IA*) is computed by means of the Hilbert Transform. The analytical signal is given by the following expression:

$$z(t) = C(t) + j\tilde{\mathrm{C}}(\mathrm{t}) = a(t)e^{i\varphi(t)} \tag{3}$$

where $a(t)$ is the amplitude function given by the expression:

$$IA_i(t) = a(t) = \sqrt{(C_i)^2 + (\tilde{C}_i)^2} \tag{4}$$

The signal matrix of instantaneous amplitudes is constructed as follows:

$$M = [IA_1; IA_2; IA_3; ...IA_n] \tag{5}$$

The second step is to compute the cross-power spectral density of matrix M. The cross-power spectral density is defined by [23]:

$$P_{xy}(w) = \sum_{m=-\infty}^{\infty} R_{xy}(\tau)\, e^{-jw\tau} \tag{6}$$

where $R_{xy}(\tau) = E\{x(t)y(t+\tau)\}$ with $E\,[\cdot]$ denoting the expectation operator on t. (7)

The diagonal elements of the matrix represent the auto-power spectral density (the same *IA*). The off-diagonal elements are the complex cross-spectral densities between two different IA.

$$P(w_i) = \begin{bmatrix} P_{IA_1IA_1}(w_i) & P_{IA_1IA_2}(w_i) & ... & P_{IA_1IA_n}(w_i) \\ P_{IA_2IA_1}(w_i) & P_{IA_2IA_2}(w_i) & ... & P_{IA_2IA_n}(w_i) \\ ... & ... & ... & ... \\ P_{IA_nIA_1}(w_i) & P_{IA_nIA_2}(w_i) & ... & P_{IA_nIA_n}(w_i) \end{bmatrix} \tag{8}$$

The flow chart of the proposed method is shown in Fig. 1.

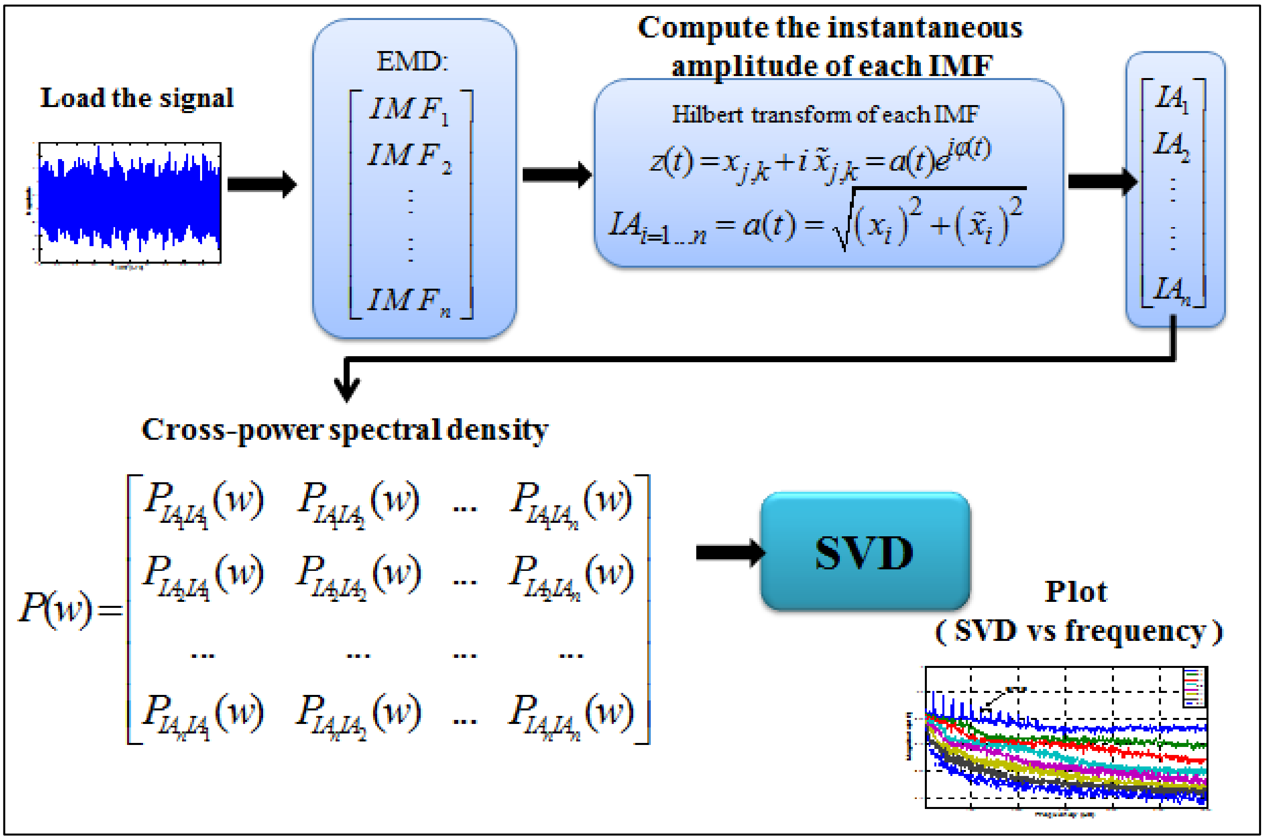


Figure 1: Flow chart of the proposed method

Assuming that $P(w_i)$ is an $n \times n$ matrix, the power spectral density matrix for each frequency (w_i) is decomposed by applying SVD to matrix $P(w_i)$.

$$P(w_i) = U \Sigma V^T \tag{9}$$

where U and V are orthogonal and Σ is a diagonal matrix of the singular values ($\sigma_{1,1} \geq \sigma_{2,2} \geq \sigma_{3,3} \ldots \geq 0$ and $\sigma_{k,j} = 0$ if $k \neq j$).

$$\Sigma(w_i) = \begin{bmatrix} \sigma(w_i)_{1,1} & \ldots & \ldots & 0 \\ 0 & \sigma(w_i)_{2,2} & \ldots & 0 \\ 0 & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & \sigma(w_i)_{n,n} \end{bmatrix} \tag{10}$$

As mentioned in the introduction, the first singular value should approximately equal the sum of the terms on the diagonal of the PSD matrix [22]. The plot of the first singular value versus frequency is thus used to identify the features extracted from the signal.

3. Validation with simulated data

3.1 Fault detection

In order to validate the proposed method and evaluate its effectiveness, a simulated numerical bearing signal is used. The simulated signal is similar to the signal used in [17-18]. The mathematical expression of the signal is given as:

$$x(t) = Ae^{-\alpha t'} \sin(2\pi f_n t) + n(t) \tag{11}$$

where

$$t' = mod\left(t, \frac{1}{F_m}\right) \tag{12}$$

Resonant frequency f_n is set to 1,800 Hz. The BPFO is set to 100 Hz. Amplitude A is set to 1. Sampling frequency F_s is set to 12,000 Hz. A random signal $n(t)$ with variance $\sigma^2 = 0.01$ is added to x(t) .

Figure 2 illustrates the simulated roller bearing signal. Figure 3 shows the 12-IMF obtained by the EMD method. It can be seen from this figure that the shocks related to defect are distributed across the first eight IMFs. As discussed in Section 2, all IMFs are considered in the proposed approach. The result is presented in Figure 4. Figure 4 exhibits the plot of the first singular value versus frequency. The fundamental of the BPFO (100 Hz) and its harmonics up to 1,200 Hz are clearly identified. The initial conclusion is thus that the proposed method can effectively detect the defect. No need to select the useful IMFs to accomplish the diagnosis.

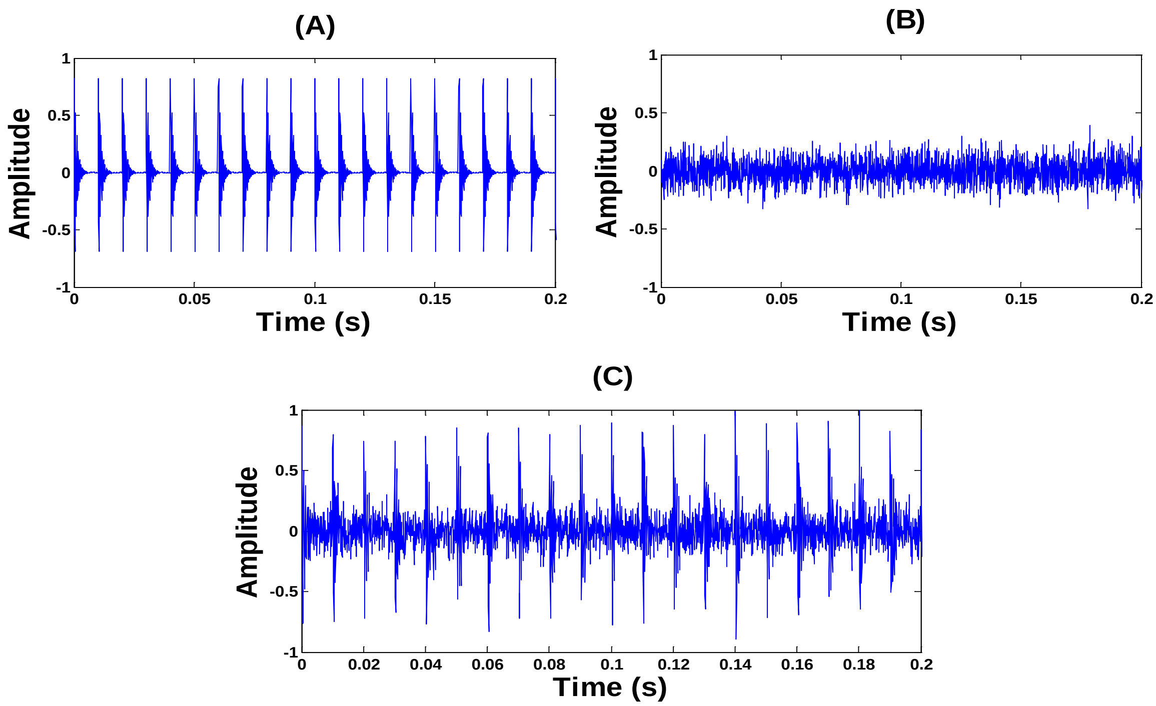


Figure 2: Simulated signal

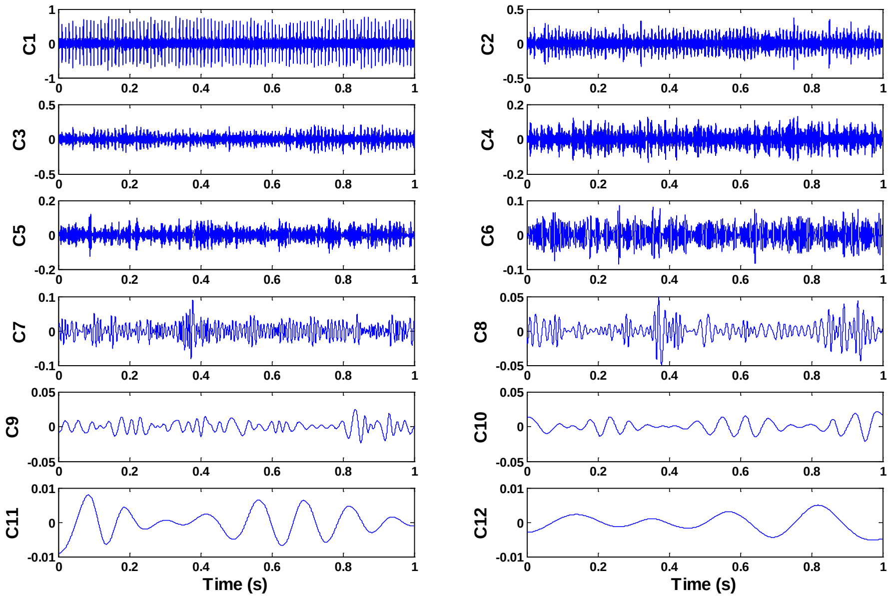


Figure 3: EMD of the simulated signal

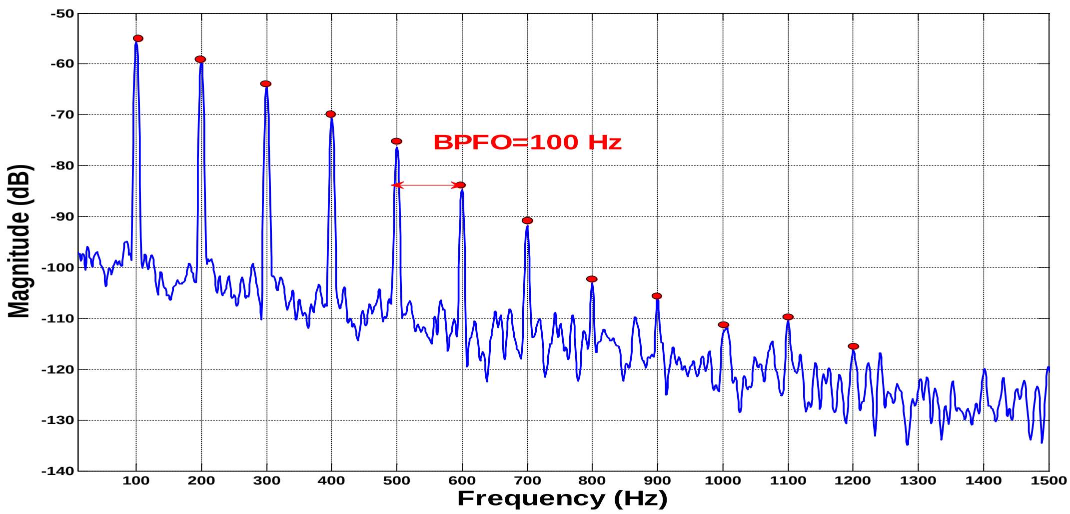


Figure 4: Result obtained for the simulated signal

It is well know that the amplitude of vibration due to bearing defects increases as the fault worsens and high peak levels may be observed. To confirm the efficiency of the method and its sensitivity to the severity of defect-induced vibrations, the simulated signal given by Equation (11) is simulated with A set to 1, 1.3, 1.6 and 1.9. The resulting waveforms are given in Figure 5.

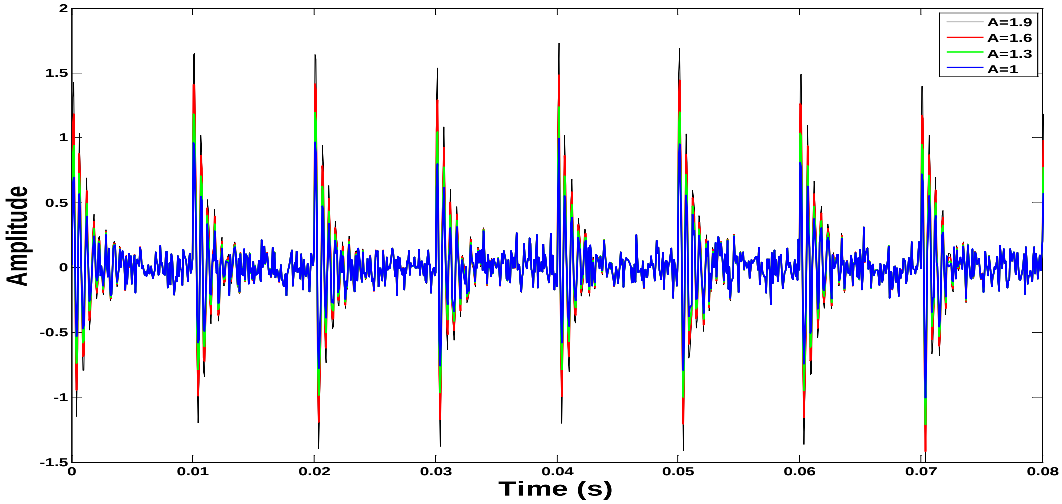


Figure 5: Simulated signal for different levels of A

Figure 6 shows the results obtained by the new approach for different levels of A. The magnitude of the BPFO and number of bearing frequency harmonics increase as the value of A increases. As seen in Figure 7, when $A = 1.3$, the magnitude of the BPFO increases by 6 dB. The same conclusions may be drawn from the cases shown in Figure 7A. Therefore, the method is sensitive to the severity of the shocks. BPFO magnitude and the mean of all harmonics can be used as an indicator to track the severity of the defect.

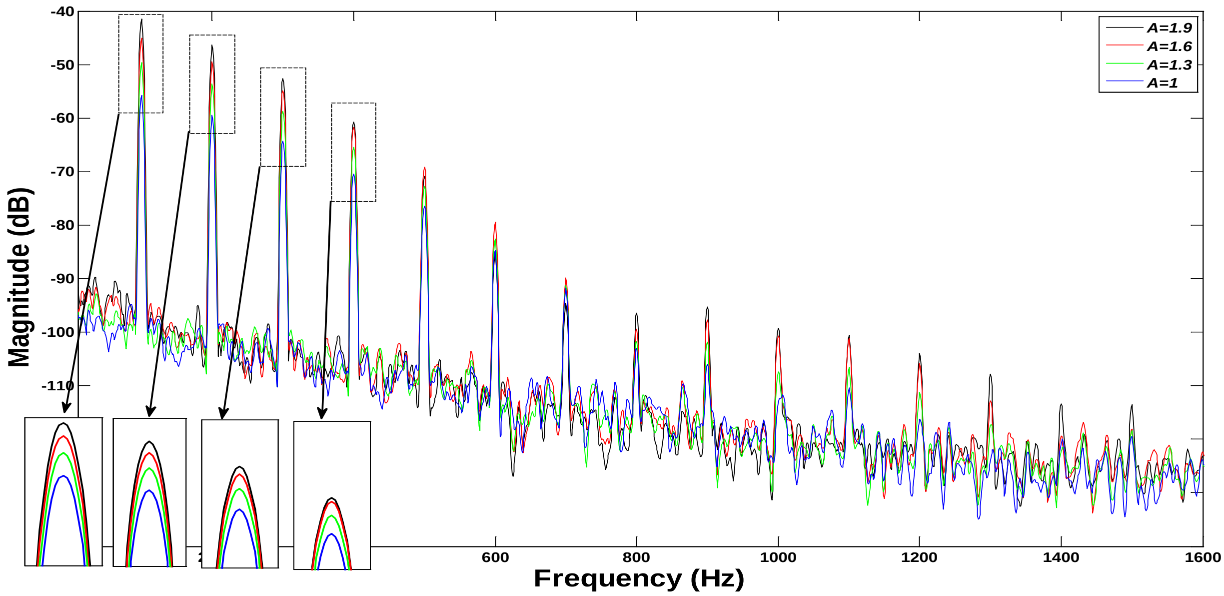


Figure 6: Results obtained for different levels of A

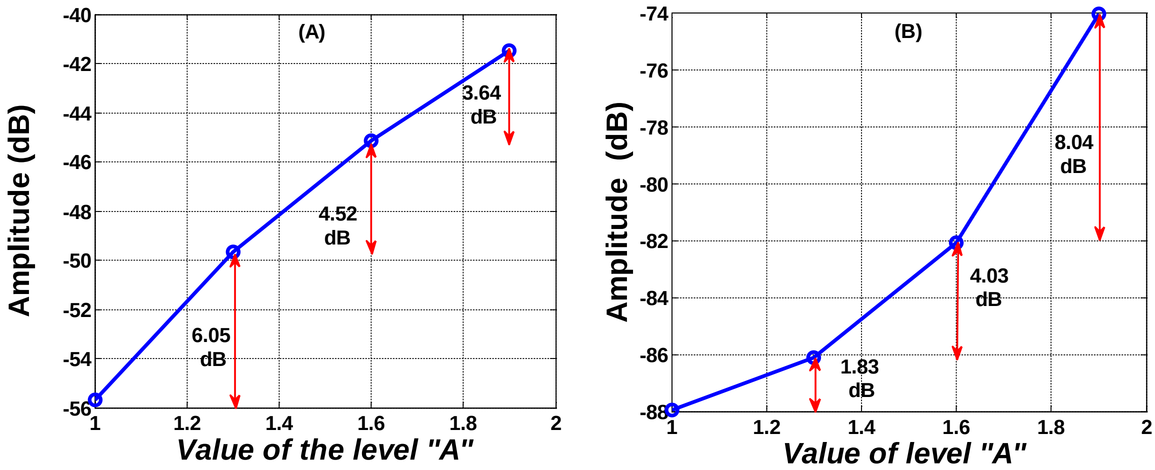


Figure 7: (A) Change in BPFO (B) Change in mean amplitude of all harmonics in Figure 6

The noise level in the simulated signal above is fairly low, though any industrial application would probably involve additive noises, potentially masking the signature of the defect, especially in the case of early bearing degradation. Another test was thus conducted to prove that the proposed method is able to detect the defect even if the noise level is higher.

3.2 Sensitivity to noise

The effectiveness of this method is evaluated based on its ability to extract bearing fault-related information. When a bearing is at an early stage of degradation, the signature of the defect may be masked by noise and difficult to extract. Gaussian white noise signals with variance σ^2 of 0.05, 0.1, 0.2 and 0.4 were thus added to the original simulated signal, $x(t)$, in order to evaluate the ability of the method to extract the defect-related information or signature when it is completely hidden by noise.

Figure 8 shows the simulated signals with the four values of variance σ^2 of added noise. It can be seen that defect-related shocks are masked by noise for $\sigma^2 = 0.1,\ 0.2 \text{ and } 0.4$. The four noisy signals were processed using the proposed method and the results obtained from all signals are exhibited in the Figure 9. Note that even when the noise is very high, the method is able to identify the BPFO, while other tested state of the art methods were not available at this high noise level.

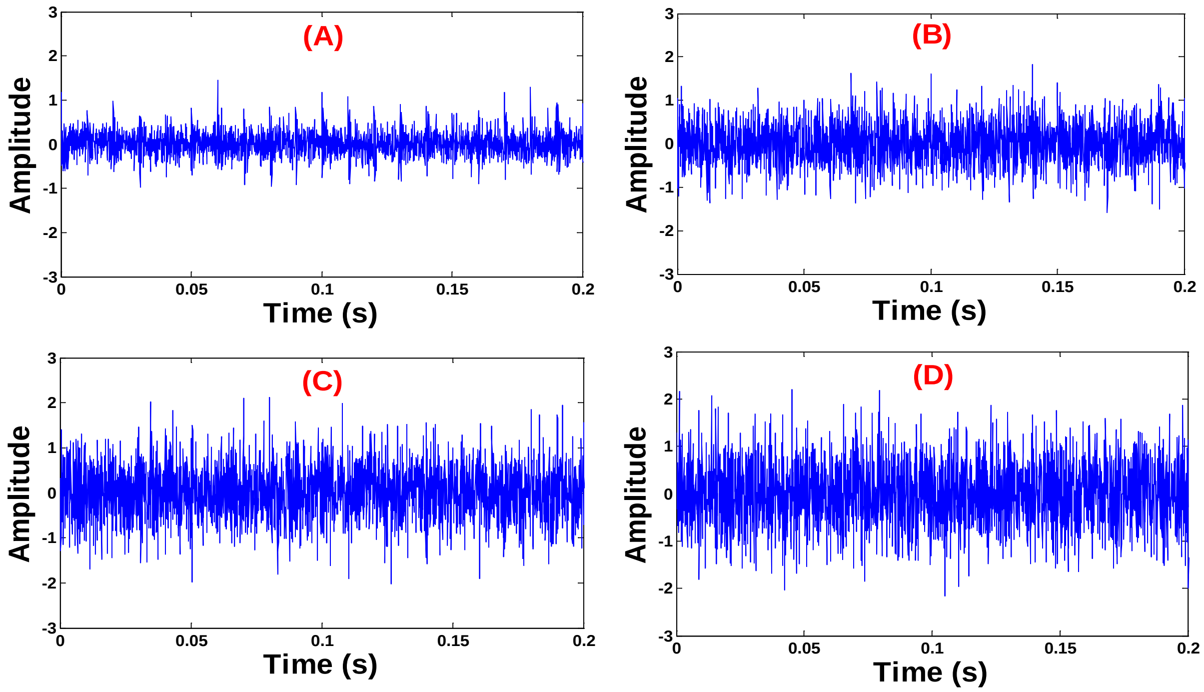


Figure 8: Simulated noisy signals: (A) $\sigma^2 = 0.05$; (B) $\sigma^2 = 0.1$; (C) $\sigma^2 = 0.2$; (D) $\sigma^2 = 0.4$

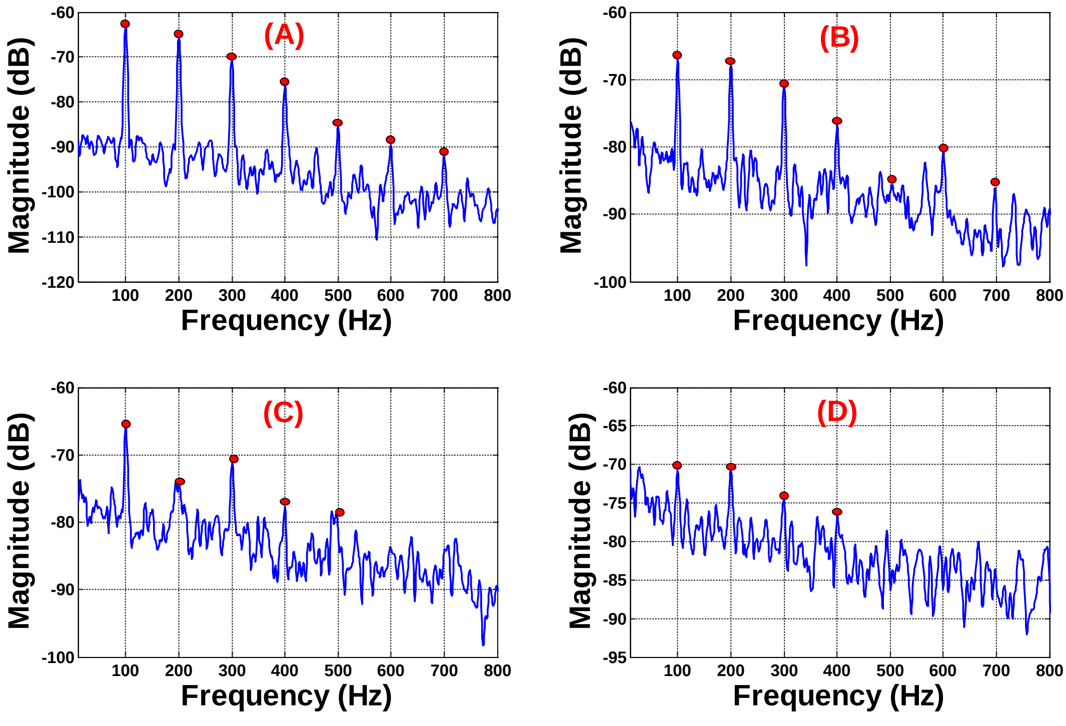


Figure 9: First singular value in frequency domain: (A) $\sigma^2 = 0.05$; (B) $\sigma^2 = 0.1$; (C) $\sigma^2 = 0.2$; (D) $\sigma^2 = 0.4$

4. Experimental data

The proposed method was investigated on two different history cases with different low levels of severity.

4.1 First case study

In the first case study, two defected bearings (SKF 1210 EKTN9) with very low severity levels are investigated. The test bench is shown in (Figure 10A). The first defected bearing noted D1, has an artificial defects about 200 μm deep with a groove width (W) of 50 μm and the second (D2) with $W = 100$ μm. The data were recorded when rotating at 600 rpm, using an accelerometer with a sensitivity of 100 mV/g. The frequency of the BPFO is about 76.46 Hz. The sample frequency is set to 12500 and 64,000 samples are used (acquisition time of 5 seconds).

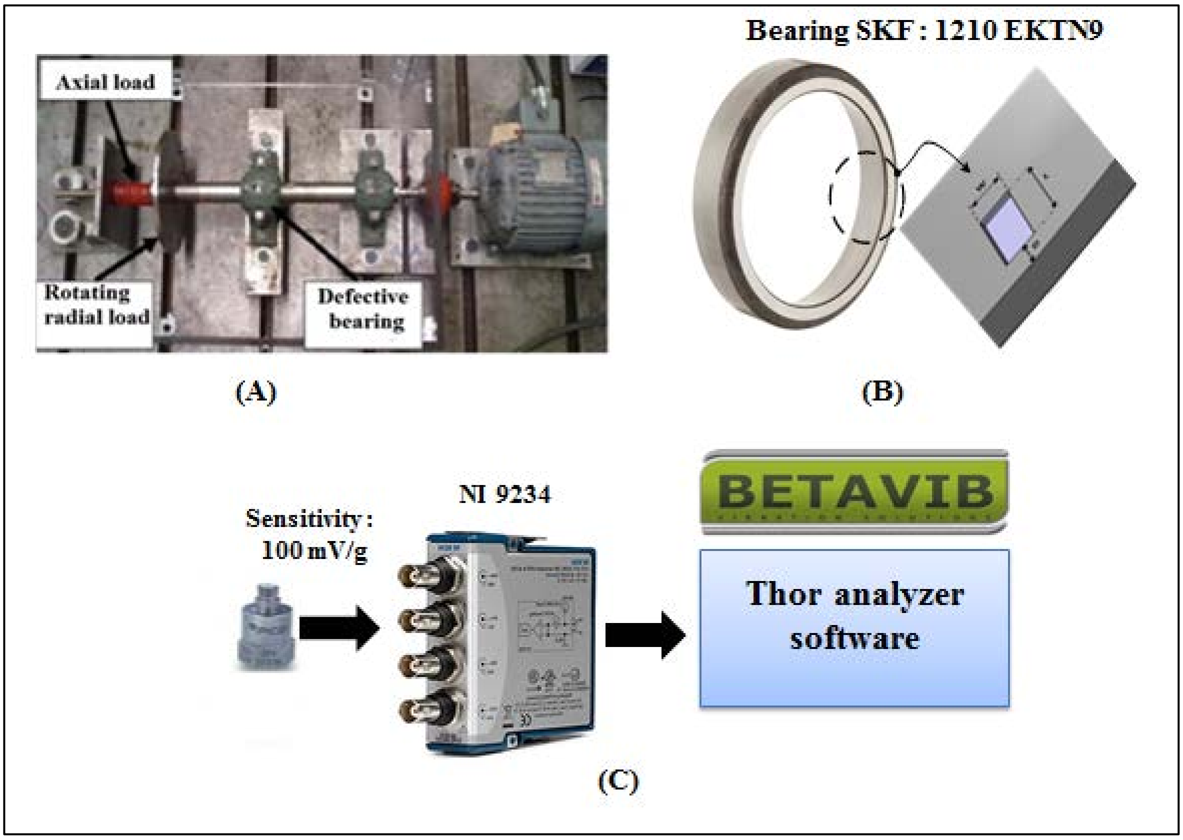


Figure 10: Test bench

Figure 11 shows the time signal of the acquired data. For the healthy bearing (Figure 11-A), quasi-random shock signals are observed. Defect D1 (50 μm) is at early stage of degradation. The time signal of defect D2 (Figure 11C) shows a series of impulse responses at BPFO and the amplitude is modulated periodically at the shaft speed. The shocks caused by defect D2 are more perceptible in the D2 signal than those due to defect D1, which are masked by noise.

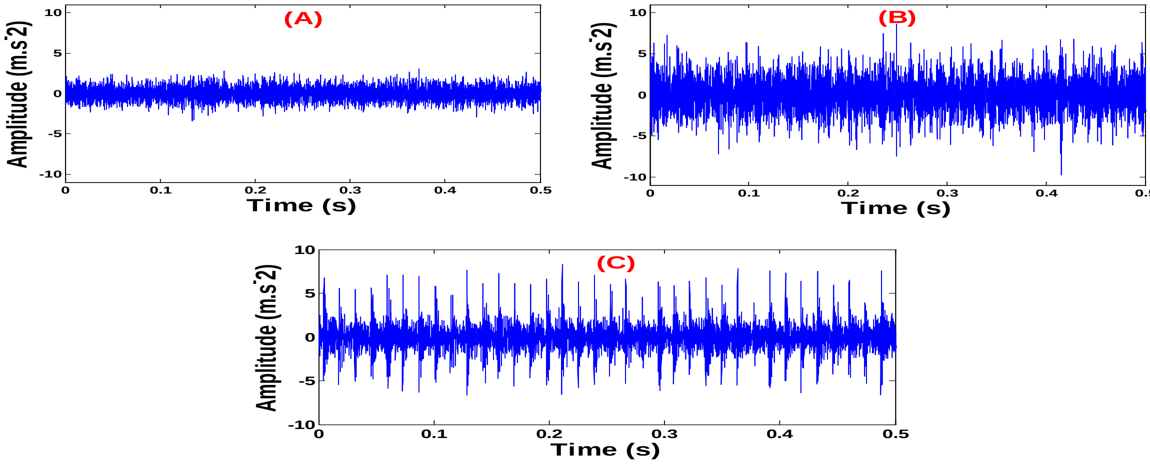


Figure 11: (A) Healthy bearing, (B) Defective bearing D1, (C) Defective bearing D2

The signals were processed using the proposed method. The defective bearing (50 µm) very clearly shows the BPFO (72.46 Hz) and its harmonics (Fig. 12). For defect D2 (100 µm), the BPFO harmonics are clearly identified and an increase in amplitude is observed (Fig. 13). The amplitude of the BPFO increases by 12.43 dB. This reveals that defects are more clearly identified and well-defined using the proposed approach.

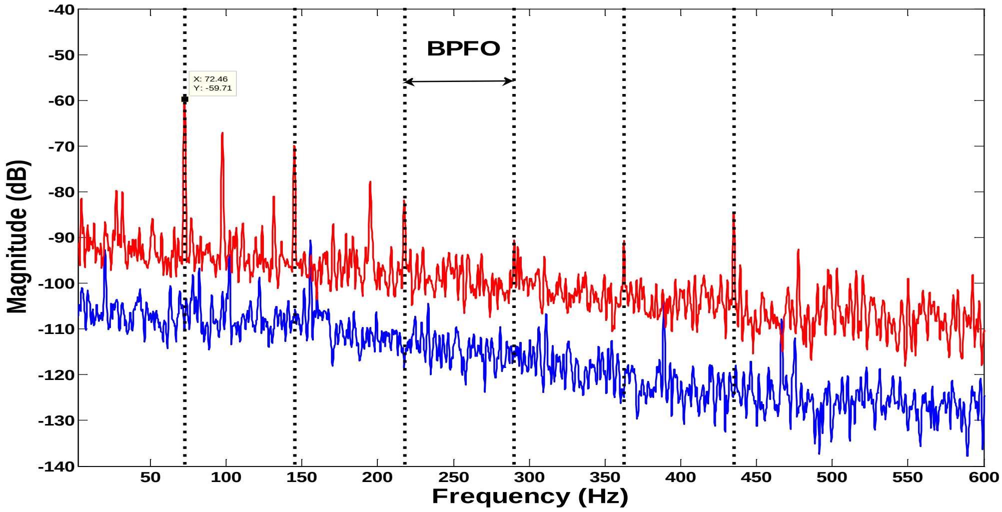


Figure 12: First singular value in frequency domain: Defect D1 (red), healthy bearing (blue)

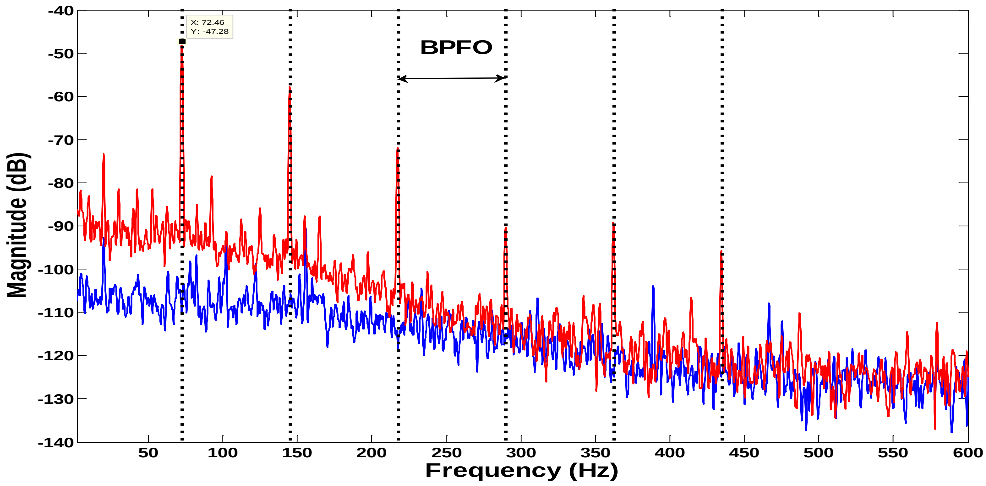


Figure 13: First singular value in frequency domain: Defect D2 (red), healthy bearing (blue)

4.2 Second case study

In this case, bearing (SKF 6205-2RS JEM) with defects in different locations are investigated. The vibration data are available in [24]. The test bench is shown in Figure 14. The first defect has a simulated single-point fault on inner race and the second defect has a simulated single-point fault on outer race. The fault size is 0.007" in diameter and 0.011" in depth.. The vibration data was collected when rotating at 1796 rpm. The frequency of the BPFO is about 107.6 Hz and the BPFI is 161.4 Hz. The sampling frequency (*Fs*) is 12,000 Hz and 60,000 data samples are used.

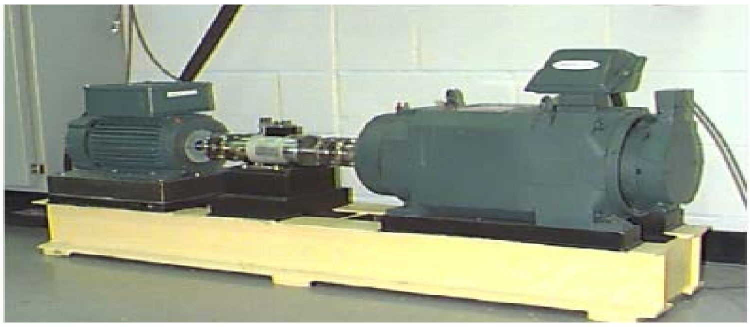


Figure 14: Test bench

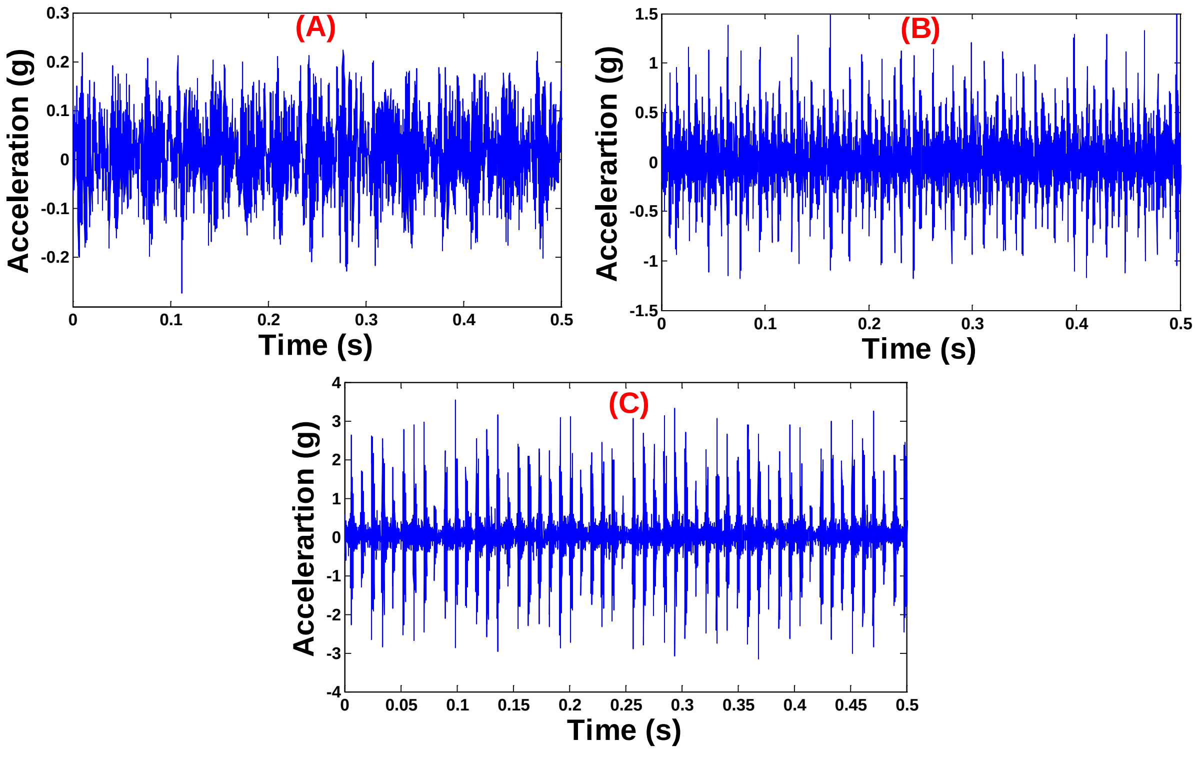


Figure 15: (A) Healthy bearing, (B) Inner ring fault (C) Outer ring fault

The time signals of the healthy bearing, inner ring fault and outer ring fault are shown in Figure 15. The time signal of the inner ring fault in Figure 15B shows that the amplitude is modulated periodically at the shaft speed. This is due to the rate at which the fault passes through the load zone. The signal is processed with the proposed method. Figure 16 illustrates the result. The result shows a series of harmonics of BPFI at 161.4 Hz, with sidebands spaced at the shaft speed to either side of each harmonic, as well as a number of harmonics of the shaft speed. This phenomena is usually related to a bearing looseness [25]. The time signal of outer ring fault is shown in Figure 15C. Unlike that of the inner ring defect, the time signal of the outer ring fault should reveal a series of uniform impulse, but the signal is modulated at the shaft speed. The result obtained by the proposed method is shown in Figure 17. A series of harmonics of the BPFO are detected. As explained in [25], this modulation suggests a rotating load caused probably by mechanical looseness.

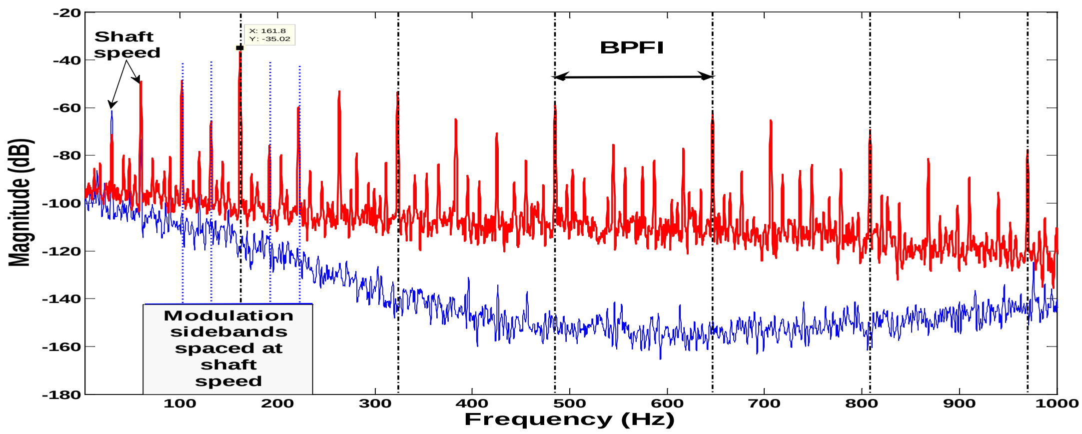


Figure 16: First singular value in frequency domain: Inner race fault (red), healthy bearing (blue)

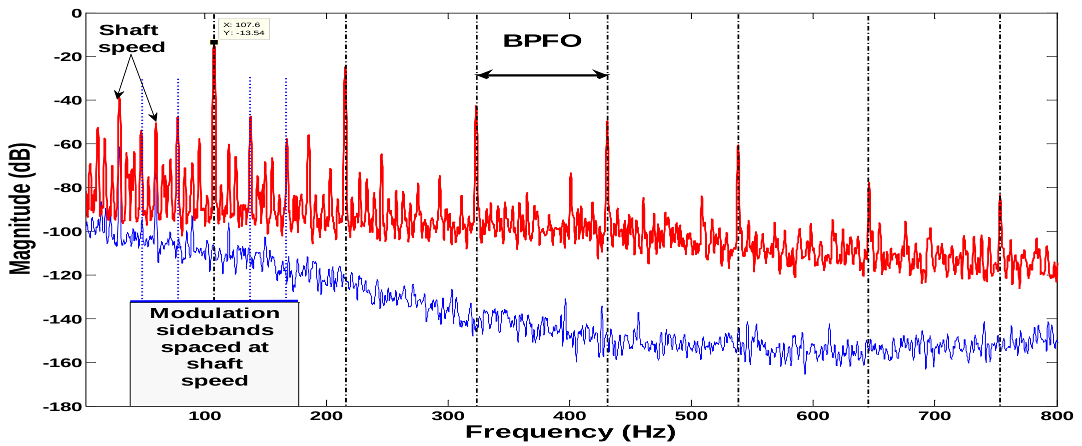


Figure 17: First singular value in frequency domain: Outer race fault (red), healthy bearing (blue)

5 Conclusion

A new feature extraction method for bearing fault diagnosis is presented in this paper. In the proposed method, there is no need to select the useful IMF to accomplish the diagnosis. The matrix of cross-power spectral density of all IMFs is decomposed in the frequency domain using SVD to extract defect-related information. The method was first validated by means of a simulated signal. The results have shown that this method may be used even in a very noisy environment. The proposed method can effectively detect defects even if the induced shocks are completely masked by noise and that the features extracted are sensitive to defect shock amplitude, making them useful indicators to track defect severity. Two test cases are presented to verify the efficiency of the method. Bearings in an early stage of degradation with two levels of defect severity (50 µm and 100 µm) and defects in different locations are investigated. The results show that the method can effectively extract all information related to the defect. In this study, we validated the proposed method using rolling bearings. Future work will extend and generalize the method for fault diagnosis of other rotating machinery, such as gears.

Acknowledgments

The financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC) is gratefully acknowledged. The assistance of the Bearing Data Center of Case Western Reserve University, which provided some experimental results, is also gratefully acknowledged.

References

[1] Huang, N. E., Shen, Z. and Long, S.R. *The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis*. Proceedings of the Royal Society of London, vol. 454, no. 1971 (1998): 903-995.

[2] Lei, Y. G., Lin, J., He, Z. J. and Zuo, M. J. *A review on empirical mode decomposition in fault diagnosis of rotating machinery*. Mechanical Systems and Signal Processing, 35 (1-2), 2012: 108-126.

[3] Peng, Z.K., Tse, P.W. and Chu, F.L. *A Comparison study of improved Hilbert–Hung transform and wavelet transform: application to fault diagnosis for rolling bearing*. Mechanical Systems & Signal Processing, vol. 19 no. 2, (2005): 974-988.

[4] Zhang, Y., Zuo, H. and Bai, F. *Classification of fault location and performance degradation of a roller bearing*. Measurement, vol. 46, no. 3 (2013): 1178-1189.

[5] Guo, W., Tse, P. W. and Djordjevich, A. *Faulty bearing signal recovery from large noise using a hybrid method based on spectral kurtosis and ensemble empirical mode decomposition*. Measurement. vol. 45, no. 5 (2012): 1308-1322.

[6] Yang Y., Yu, D.J. and Cheng, J.S. *A fault diagnosis approach for roller bearing based on IMF envelope spectrum and SVM,* Measurement, vol. 9, no. 40 (2007): 943–950.

[7] Kedadouche, M., Thomas, M. and Tahan, A., *Monitoring machines by using a hybrid method combining MED, EMD, and TKEO*, Advances in Acoustics and Vibration, vol. 2014, 10 pages, 2014.

[8] Wenliao Du, Zhiyang Wang, Xiaoyun Gong, Liangwen Wang, and Guofu Luo, *Optimum IMFs Selection Based Envelope Analysis of Bearing Fault Diagnosis in Plunger Pump*, Shock and Vibration, vol. 2016, Article ID 1248626, 8 pages, 2016. doi:10.1155/2016/1248626

[9] Ricci R., and Pennacchi P. *Diagnostics of gear faults based on EMD and automatic selection of intrinsic mode functions*. Mechanical Systems and Signal Processing 25, no. 3 (2011): 821-838

[10] Yi C, Lin J, Zhang W, Ding J. Faults Diagnostics of Railway Axle Bearings Based on IMF's Confidence Index Algorithm for Ensemble EMD. *Passaro VMN, ed. Sensors (Basel, Switzerland)*. 2015;15(5):10991-11011. doi:10.3390/s150510991.

[11] Jacek D., and Zimroz R. *Rolling bearing diagnosing method based on empirical mode decomposition of machine vibration signal*. Applied Acoustics 77 (2014): 195-203.

[12] Z. K. Peng, P. W. Tse, and F. L. Chu, *An improved Hilbert–Huang transform and its application in vibration signal analysis*, Journal of Sound and Vibration, vol. 286, no. 1-2, pp. 187–205, 2005.

[13] C. Wang, M. Gan, and C. Zhu, *Non-negative EMD manifold for feature extraction in machinery fault diagnosis*, Measurement, vol. 70, pp. 188–202, 2015.

[14] W. Guo and P. W. Tse, *A novel signal compression method based on optimal ensemble empirical mode decomposition for bearing vibration signals*, Journal of Sound and Vibration, 332 (2), 423–441, 2013.

[15] Tsao, W.-C., Li, Y.-F., Le, D.D. and Pan, M.-C.. "An insight concept to select appropriate IMFs for envelope analysis of bearing fault diagnosis." *Measurement*, vol. 45, no. 6 (2012): 1489-1498.

[16] Pan, M.-C. and Tsao, W.-C. *Using appropriate IMFs for envelope analysis in multiple fault diagnosis of ball bearings*. International Journal of Mechanical Sciences, vol. 69, (2013): 114-124.

[17] Kedadouche, M., Thomas, M. and Tahan, A. *A comparative study between empirical wavelet transforms and empirical mode decomposition methods: application to bearing defect diagnosis*. Mechanical Systems and Signal Processing, vol. 81 (2016): 88-107.

[18] Kedadouche, M., Liu, Z. and Vu, V. H. *A new approach based on OMA-empirical wavelet transforms for bearing fault diagnosis*. Measurement, vol. 90, (2016): 292-308.

[19] Cong, F., Chen, J., Dong, G. and Zhao, F. *Short-time matrix series based singular value decomposition for rolling bearing fault diagnosis*. Mechanical Systems and Signal Processing, 34 (1), 2013: 218-230.

[20] Cong, F., Zhong, W., Tong, S., Tang, N. and Chen, J. *Research of singular value decomposition based on slip matrix for rolling bearing fault diagnosis*. Journal of Sound and Vibration, 344 (2015): 447-463.

[21] Kedadouche, M. and Liu, Z. *Fault feature extraction and classification based on WPT and SVD: Application to element bearings with artificially created faults under variable conditions*. Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science (2016): 0954406216663782.

[22] Brincker, R., Andersen, P., and Zhang, L. *Modal identification from ambient responses using frequency domain decomposition*. 18th International Modal Analysis Conference (IMAC), San Antonio, Texas, (2000): 625-630.

[23] Bendat, J.S. and Piersol, A.G. Random Data, *Analysis and Measurement Procedures*, 2nd ed., New York: Wiley-Interscience (1986): 120–121.

[24] http://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website.

[25] Smith, W. A. and Randall, R. B. *Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study*". Mechanical Systems and Signal Processing, 64 (2015): 100-131.

Prognostics of Rolling Element Bearings based on Entropy indicators and Particle Filtering

Junyu QI[1, 2], Alexandre MAURICIO[1, 2], Konstantinos GRYLLIAS[1, 2]

[1]Department of Mechanical Engineering, KU Leuven,
[2]Dynamics of Mechanical and Mechatronic Systems, Flanders Make
Celestijnenlaan 300, BOX 2420, 3001 Leuven, Belgium
Junyu.qi@kuleuven.be, alex.ricardomauricio@kuleuven.be, konstantinos.gryllias@kuleuven.be

Abstract

Rolling element bearings' damage is the main cause for unexpected breakdown in rotating machinery. Therefore there is a continuous industrial interest on condition monitoring of bearings, targeting towards the development and proposal of robust diagnostic techniques which can detect accurately, robustly and early the generation of the fault. On the other hand, industry is not interested only in the proper damage detection and identification of faults but is mainly targeting towards the robust estimation of the Remaining Useful Life (RUL) of machine elements. The proper estimation of the RUL could be linked directly with the maintenance planning and warehouse organization, providing immediately profits in terms of employees health and safety, environmental protection and continuous production. A plethora of diagnostics and prognostics indicators have been proposed during the last decade focusing towards the accurate representation and tracking of the health state of bearings and other machine elements. However, in certain cases (e.g. nonstationary operating conditions), the classic techniques for bearings prognostics (e.g. statistical analysis, frequency analysis and time-frequency analysis) underperform due to the high noise influence or the high machines' complexity. Therefore the classical diagnostics indicators may identify the fault quite late and fail to identify properly the RUL.

In this paper, prognostics indicators based on the measurement of disorder (e.g. entropy) are used in order to track the degradation process of the machinery. An advanced health indicator, termed Info-entropy, is built based on the Spectral Entropy, the Envelope Spectral Entropy and the Spectral Negentropy. It is compared with the Spectral Kurtosis, the Kurtosis and the RMS based on the criterion of trendability. The indicators are estimated on the well-known bearing dataset Prognostia and Particle filtering is used in order to estimate the RUL of bearings.

1 Introduction

In the era of Industry 4.0, Condition Monitoring (CM) is currently attracting unprecedented attention. Rotating machinery systems and components, such as Rolling Element Bearings (REBs), are utilized frequently in almost each mechanical equipment, including electric motors, wind turbines and compressors, etc. The breakdown of large amount of machineries happens due to the failure of bearings, which are mainly caused by misalignment, corrosion, insufficient lubrication and oil contamination, resulting further in accelerated wear in the surface. As the damage accumulates in the continuous operation, the functionality of REBs reduces through the stages of mild (stage 1), moderate (stage 2) and profound fault (stage 3), as presented in Figure 1.

Usually, stage 1 endures for long time, and as the fault reaches to a certain level, the degradation rate starts accelerating and as a result the stage 2 lasts for shorter time compared to stage 1. In stage 3, the functionality reduces to the End of Life (EoL) incredibly fast. Concerning the change of functionality, a number of maintenance concepts have been proposed in the last decades, i.e., the proactive maintenance (to stage 1), preventative maintenance (to stage 2) and reactive maintenance (to stage 3). The reactive maintenance (fault detection and diagnosis) have been explored widely and a number of systematic, automatic and accurate methods are successfully established. However, the period of the reactive maintenance usually sustains shortly till the EoL, as stage 3 shown in Figure 1, and it requires expensive cost. On one hand, the hazardous performance of the bearings influences the normal operating condition of the equipment, which is related with the continuous production quality. On the other hand, it may lead to accidental breakdown of machinery. Preventative maintenance is localized in stage 2, while the defect is still in the impending or the incipient stage and the functionality can be corrected by proper intervention, e.g. cleanness, lubrication and alignment. The proactive maintenance in stage 1 costs less expensive, where the defect incubates slowly and the REB exhibits an acceptable performance. Therefore, if the life of in-service components is able to be predicted in the early time (stage 1), the maintenance strategy and actions can be optimally planned. Compared with traditional CM techniques, the predictive maintenance is recently catching particular attention. The state of the art of methodologies [1] includes mainly three directions, i.e. the Physics model based Prognostics (PbP), the Data-driven based Prognostics (DbP) and the Hybrid-approach based Prognostics (HbP).

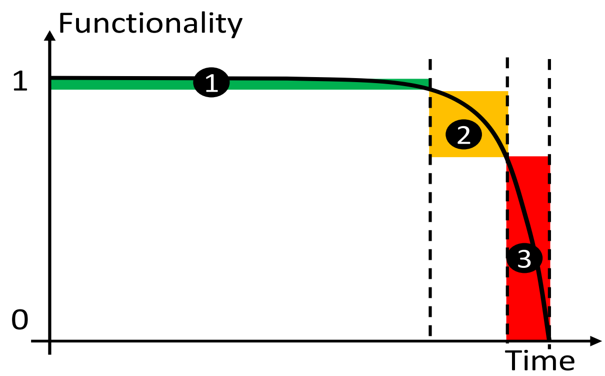


Figure 1: Degradation process and maintenance strategy

The PbP usually describes the degradation phenomenon in an explicit mathematical formulation. The establishment of a physical Finite Element Model (FEM) depends on the available sufficient information, e.g. the geometry, the load and the material, etc. In the past, some physics models have been developed, e.g. spall progression models, crack-growth model, gas path model, mixed lubrication model, etc. [2]. Theoretically, the PbP is built systematically and well adaptive to the degradation evolution process. However, the modeling in the system level requires complete domain knowledge and is expensive to be developed, therefore, the PbP is usually applied to a specific component.

With the advantage of flexibility, applicability and economy, the DbP is preferred by the industrial domain. Unlike the difficulty of building the physics model, an analytic model is directly derived from massive data. In general, the DbP is classified as Machine Learning (ML) and Statistical Model Based (SMB) techniques [3]. ML learns the hidden information from big data. Instead of an explicit model, ML works like a black box and till now often the output is hardly interpretable. In order to reduce the computation cost and enhance the knowledge understanding, the direction of SMB is considered in this paper, e.g. Kalman filter (KF), Particle Filter (PF), linear regression, Hidden Markov Model (HMM), etc. An analytic model, e.g. the exponential model [4], is learned from training data, and further is updated with testing data.

Medjaher et al. [5] presented a method of bearing prognostics based on the regression method. Yet, it is not an optimal solution for the stochastic degradation progress. Alternatively, the Bayesian Recursive Estimation (BRE) is able to represent the model, considering the uncertainty. For example, HMM is used for the RUL prediction, considering different failure modes [6]. However, the probability transition matrix between the model's states need to be presupposed. Qi et al. [7] employed the Unscented Kalman filter (UKF) for the bearing RUL prediction. Additionally, in order to cope with the limitation of the single model, the Switching Kalman Filter was investigated for the prognostics of a helicopter bearing. [8]. PF achieves advantageous performance than KF versions without the restrictions of Gaussian noise, system linearity or existence of low nonlinearities. Li et al. [9] predicted bearing RUL combining the PF with an improved exponential model. If Health Indicator (HI) exhibits a degradation pattern with multiple degradation rates, the single statistical model is probably not adequate anymore. Therefore, a multi-model PF is introduced to track the linear and nonlinear stage, and the corresponding mode is recognized automatically by the calculation of the likelihood [10].

The result of SMB is usually dependent on several factors, e.g., the statistical model, the HI, the failure threshold, etc. To extract high quality HI, different signal processing techniques have been developed in the domain of time, frequency and time-frequency. However, many classic indicators are frequently deficient in the real applications. Ahmad et al. [11] proposed an improved HI removing the fluctuation based on the Linear Rectification Technique (LRT). In order to track the degradation in the best way, some researchers focused on the change in the specific frequency band. Singleton et al. [12] extracted a HI over the specific frequency band of the Choi-Williams Transform. Soualhi et al. [13] built a HI based on the Hilbert Huang Transform, decomposing the signal into several Intrinsic Mode Functions (IMF) in different frequency bands. However, an optimal HI with a remarkable tendency is usually challengeable to attain in reality, especially the tendency in the full lifetime. As a result, classic HIs, such as RMS and Kurtosis, may fail to track the trend in case of high noise, low impulsiveness and varying operating condition. The problem of low trendability triggers the proposal of start prediction point [9], [11] and the RUL prediction by the start prediction point is usually starting in the very late stage, e.g. in stage 3 of Figure 1. Moreover, it also influences the statistical model selection in PF.

In order to enhance the prediction performance, a new HI Info-entropy is proposed for the bearing prognostics in this paper. Following the logic of [12]-[13], Info-entropy assesses the chaos level for a specific frequency band and is composed of different entropy indicators in the frequency domain, i.e. the Spectral Entropy (SpecE), the Spectral Flatness (SpecF) and the Spectral Negentropy (SpecNegE). Based on the criterion of trendability, the best entropy indicators from a certain frequency band is selected. Then, the Info-entropy is collaborated with a classic exponential model and PF. The main contributions of the proposed methodology for SMB prognostics are summarized as follows.

1) Through the chaos measurement in multiple frequency bands, Info-entropy recognizes the hidden signature of the vibration signal of the REB. It provides a new perspective for the HI extraction compared to the traditional impulsiveness based features. Even in cases of ignoring the characteristic bearing frequencies, the low frequency resolution and the repetitive transients, the Info-entropy could detect the inherent change.

2) The in-service machinery degrades with time and its corresponding functionality reduces also, as can be seen in Figure 1. The Info-entropy is sensitive to frequency distributions, demonstrating the characteristic of high trendabillity. With the full lifetime degradation tracking, the aforementioned concept of start prediction point is no longer the practical bottleneck. Consequently, the use of Info-entropy allows for the start of prediction from the initial moment and potentially is useful for long-term RUL prediction.

3) Likewise, as Info-entropy present a high quality of trendability, the corresponding statistical model selection could be simplified. Compared with the complex model, e.g. for the non-monotonic trend, the actual calculation rested on the simple exponential model, becomes faster and efficient.

The rest of paper is organized as follows. The mathematical theory of PF and the construction of Info-entropy is explained in Section 2. The proposed methodology is detailed step by step in Section 3. In Section 4 and 5, the experimental dataset and the RUL prediction results are described. Finally, the paper is concluded in Section 6.

2 Mathematical theory

The theory of Particle filter and the proposed HI based on the Info-entropy are explained in this Section.

2.1 Particle filter

As a kind of BRE method, PF uses a number of particles to estimate the posterior distribution of a stochastic process. It is particularly useful for the estimation of a nonlinear dynamic system with non-Gaussian noise. The principles of PF is based on the Bayesian theory and sampling strategy. Bayesian theory includes steps of the prediction and update. In the prediction step, the priori Probability Density Function (PDF) of the k_{th} predicted state x_k is gained with the posteriori knowledge $p(x_k|z_{1:k-1})$. Then, the posteriori PDF $p(x_k|z_{1:k})$ is updated when the k_{th} measurement $z_k$is available. According to Monte Carlo Sampling (MCS), the state x_k is approximated by a set of particles $x_k^i, i = 1 \dots N$, which are sampled from a certain probability distribution. The posteriori probability of Bayesian estimation is able to be approximated with MCS, $p(x_k|z_{1:k}) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(x_k - x_k^i)$, where δ denotes the Dirac delta function. Each particle has the same weight $\frac{1}{N}$. However, particles degeneracy arises in the sequential importance sampling. The amount of effective particles decreases after some certain recursive steps. Hence, the resampling duplicates the high weight particles and replaces the lower ones. The importance distribution $q(x_k|x_{k-1}, z_k)$ is generally chosen as $p\left(x_k^i \middle| x_{k-1}^i\right)$, then $w_k^i \propto w_{k-1}^i p\left(z_k \middle| x_k^i\right)$.

PF mainly consists of three steps, the particles generation, the weight calculation and the resampling.

1) Particles generation

As the initial step, the amount (N) of particles are firstly generated from a priori distribution and then are transmitted through the propagation model, $x_k^i \sim p\left(x_k \middle| x_{k-1}^i\right), i = 1 \dots N$.

2) Weight calculation

The likelihood of each particle x_k^i is calculated with the measurement z_k, $w_k^i = p\left(z_k \middle| x_k^i\right)$, and is normalized as $\breve{w}_k^i = \frac{w_k^i}{\sum_{i=1}^{N} w_k^i}$.

3) Resampling

Being the most frequently used resampling method, the Multinomial Resampling (MR) compares the weight of particles $\breve{w}_k^i$ with a random threshold between 0 and 1, and only the particles, which are above the threshold, are kept. It is worthy to mention, that the Systematic Resampling (SR) is utilized to replace the classic MR and to enhance the PF performance. Unlike the MR method, SR divides the whole interval into M subspaces $\mathrm{U}^{\mathrm{i}} = ((\mathrm{i}-1)/\mathrm{M}, \mathrm{i}/\mathrm{M}), \mathrm{i} = 1, \dots, \mathrm{M}$, and the particles are taken with a corresponding random threshold in the U^i, higher particle diversity is thus guaranteed.

2.2 Info-entropy extraction

Many classic HIs are extracted based on the impulsiveness of the signal, e.g., the statistical indicators, the Spectral Kurtosis, frequency indicators extracted by the envelope spectrum, after the application of Hilbert transform, etc. Nevertheless, the fault information is easily affected by the noise, non-stationarities and repetitive transients. Moreover, the behavior of rotating machinery performs physically nonlinear, due to the instantaneous variation of the friction, damping and loading conditions. Thus, the chaos assessment is taken into account. In this aspect, entropy is the versatile measurement of disorder, unevenness of distribution and complexity. The concept of entropy, different entropy indicators and the info-entropy will be explained in this Section.

2.2.1 Entropy

As property of thermodynamics, the entropy of an isolated system never decreases. It always evolves toward the thermodynamic equilibrium, which means each of the microstates S_i inside the system has an equal probability p_i.

2.2.2 Entropy indicators

A number of variants of entropy have been proposed in the time series, e.g. the Approximate Entropy, the Sample Entropy, the Kolmogorov Entropy and the Permutation Entropy. Yet, the assessment in the time domain is often insufficient. The frequency information is explored here, as the compensation to capture the latent variation. Thus, the Spectral Entropy, the Spectral Flatness and the Spectral NegEntropy, i.e., the NegEntropy in the Squared Envelope (NegE-SE) and the Squared Envelope Spectrum (NegE-SES) are introduced.

SpecE has been proposed to measure the amplitude distribution of the Fourier spectrum of a signal $x(t)$. It is defined as the same form of entropy:

$$SpecE = -\sum_{i=1}^{L} \mathrm{p_i} log_2\,(\mathrm{p_i}) \tag{1}$$

with,

$$p_i = \frac{X(i)}{\sum_{i=1}^{L} X(i)} \tag{2}$$

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt \tag{3}$$

When the spectral amplitudes distribute evenly, SpecE yields a large value. On the contrary, it outputs a small value if the amplitudes are concentrated around a specific region. To avoid the influence of the data length, the SpecE is usually normalized as in Equation 4, and ranges between 0 and 1.

$$SpecE = \frac{-\sum_{i=1}^{L} p_i log_2 (p_i)}{log_2(L)} \tag{4}$$

SpecF, also called Wiener entropy, has been initially used for the whiteness characterization of speech analysis [14]. It is calculated by the ratio between the geometric mean and the arithmetic mean of the power spectrum, as presented in equation 5.

$$SpecF = \frac{\exp\left(\frac{1}{L}\sum_{i=1}^{L} \ln(|X(i)|)^2\right)}{\frac{1}{L}\sum_{i=1}^{L} \ln(|X(i)|^2)} \tag{5}$$

Equivalent to the SpecE, SpecF approaches to 1 if the spectrum uniformly scatters and to 0 if the spectrum is spiky.

As the opposite of entropy, NegEntropy is used as a measure of distance from normality. The Spectral NegEntropy in the frequency band $(f - 0.5 \cdot \Delta f, f + 0.5 \cdot \Delta f)$ is defined in [15] as:

$$\Delta I_\varepsilon(f, \Delta f) = \langle \frac{\varepsilon_x(t,f,\Delta f)^2}{\langle \varepsilon_x(t,f,\Delta f)^2 \rangle} \ln \left(\frac{\varepsilon_x(t,f,\Delta f)^2}{\langle \varepsilon_x(t,f,\Delta f)^2 \rangle}\right) \rangle \tag{6}$$

where, ε_x is the SE of signal $x(t)$, $\langle * \rangle$ indicates the average operator and ΔI_ε represents the NegEntropy in the SE (NegE-SE). Similarly, NegE-SES is written as:

$$\Delta I_E(f, \Delta f) = \langle \frac{E_x(t,f,\Delta f)^2}{\langle E_x(t,f,\Delta f)^2 \rangle} \ln \left(\frac{E_x(t,f,\Delta f)^2}{\langle E_x(t,f,\Delta f)^2 \rangle}\right) \rangle \tag{7}$$

where, ΔI_E is the NegEntropy in the SES, and E_x denotes SES of $x(t)$. In addition, to overcome the shortcomings of NegE-SE and NegE-SES, the weighted average Spectral NegEntropy (NegE-Ave) has been considered:

$$\Delta I_{1/2}(f, \Delta f) = (\Delta I_\varepsilon(f, \Delta f) + \Delta I_E(f, \Delta f))/2 \tag{8}$$

SpecNegE has reverse behavior with Spectral Entropy. If vibration signal shows more impulsiveness or repetitive transient, it returns a higher value.

2.2.3 Info-entropy

In order to capture the informative event of machinery, spectral NegEntropy has been adopted in Infogram [15]. The signal is firstly decomposed into different frequency bands using a 1/3-binary tree. Then, the demodulation frequency band is localized by the maximal spectral NegEntropy, which is caused from the faulty component.

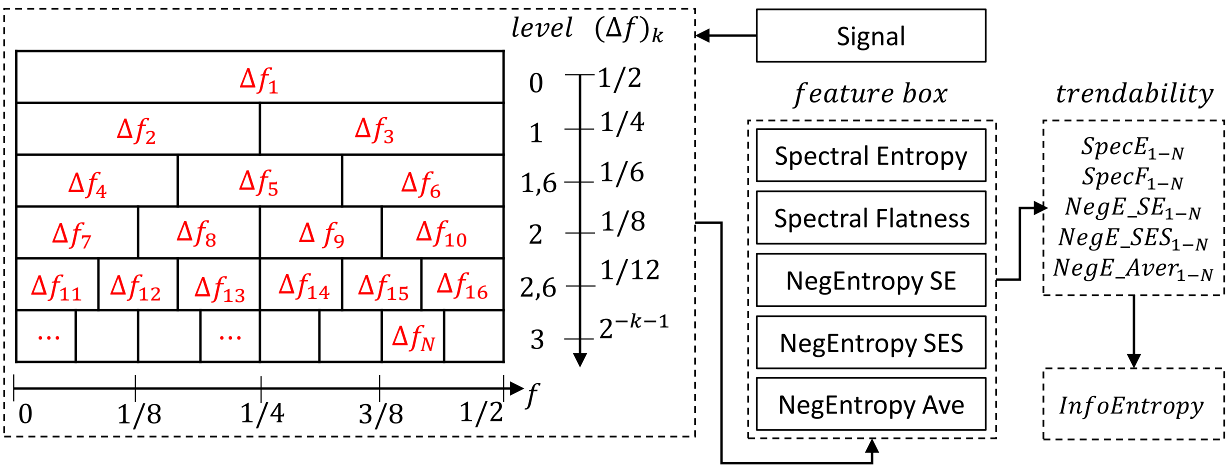


Figure 2: The illustration of Info-entropy proposition

The proposed Info-entropy is structured as presented in Figure 2. At first, the decomposed signals in different frequency bands Δf_1, Δf_2, ⋯, Δf_N are calculated by five kinds of Spectral Entropy and NegEntropy. Then, each entropy indicator has an amount of N values and is collected in a feature box, i.e., $SpecE_{1-N}$,

$SpecF_{1-N}$, $NegE\_SE_{1-N}$, $NegE\_SES_{1-N}$ and $NegE\_Ave_{1-N}$. In order to catch the degradation evolution for the specific operating condition, the anticipant feature is selected based on the trendability, as shown in Equation 9. In this way, the most sensitive and informative frequency band is localized. Finally, the selected Info-entropy can be any of the feature box with an optimal score, based on Equation 10, depicting well the whole degradation process.

$$R = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \tag{9}$$

where x and y represent, respectively, the time index and the feature. n indicates the length of the feature. The trendability measures the correlation between the time and the feature. The value R varies between -1 and 1. The final score of the HI is expressed in the form of equation (10). The perfect score is equal to 0 and signifies high trendability for the full lifetime.

$$Score = 1 - abs(R) \tag{10}$$

3 Proposed methodology

The proposed methodology is detailed in Figure 3. Firstly, the data is recorded in a certain operating condition. However, the vibration signals are easily polluted during the experiment, by e.g. the deterministic composition of gear transmission, anomalies, noise etc. and different preprocessing techniques are widely derived to enhance the bearing signal. In order to solve the problem of outliers, thresholding using the Standard Deviation (SD) is often used. But SD is problematic to the extreme values, e.g.to infinite valueσ. Compared to the criterion of SD, the application of three Median Absolute Deviations (MAD) is more robust to detect and remove anomalies. Therefore, the Moving MAD (MMAD) is proposed as α pre-processing technique. The MAD is calculated w.r.t the D sequence points. The samples beyond the threshold Thr_1 (e.g. $Thr_1 = 3 \cdot MAD$) are recognized as outliers and αρε replaced with Thr_1. Moreover, α prewhitening technique is suggested before calculating the Spectral NegEntropy [15].

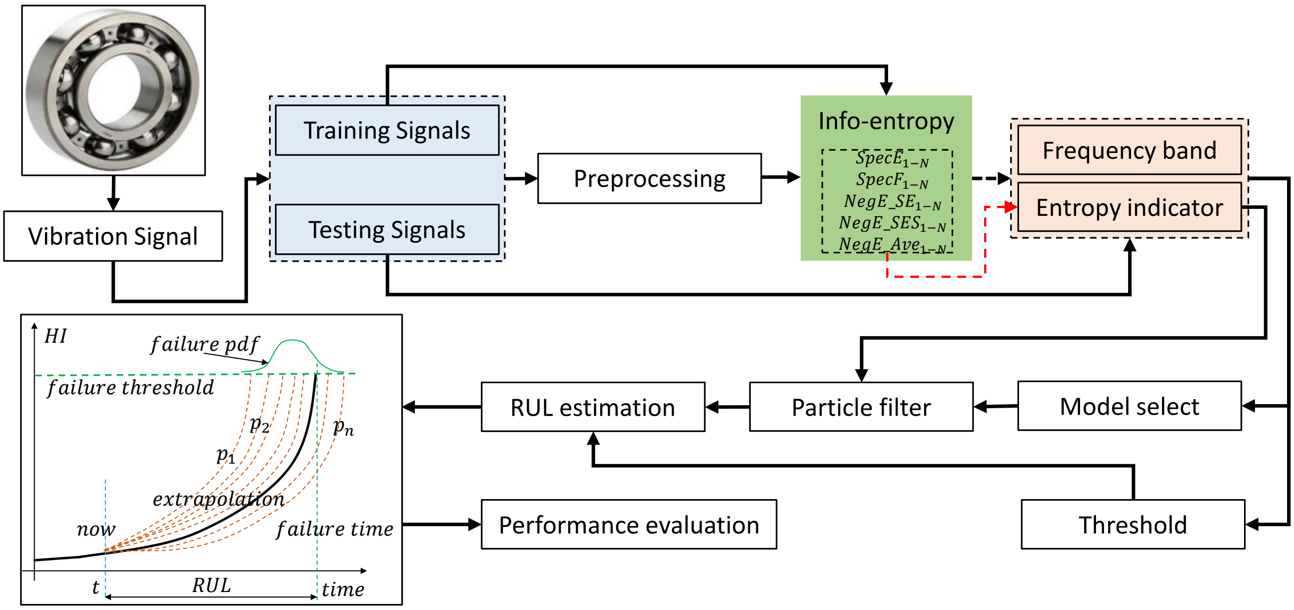


Figure 3: The flowchart of proposed methodology

Then, following the explanation of Info-entropy in Figure 2, the training dataset is utilized to localize the most informative frequency band and the corresponding entropy indicator in the feature box with the best score is selected. Based on the extracted HI, the statistical model is easily determined to represent the degradation process of the bearing. If the model parameter b_i indicates the degradation path of the i_{th} bearing B_i, the similarity level between HIs is measured by the ratio between the slowest degradation b_{min} w.r.t the closest path b_k.

$$\mathrm{ratio} = \frac{\mathrm{b_{min}}}{\mathrm{b_k}} \tag{9}$$

In case, the Info-entropy of all the members of training sets $U(B_i)$ present high similarity (e.g. $ratio > \epsilon$, ϵ is the threshold of similarity level), the failure threshold thr is defined as the average value of the maximum or minimum of the His and the initial model parameters pr_0 are calculated by the average of the fitted values from $U(B_i)$. Otherwise, when the HIs show low similarity in the degradation path (e.g. $ratio < \epsilon$), the thr and pr_0 are set based on the closest HIs. They depend on the two bearings B_{n1} and B_{n2} of $U(B_i)$ if the starting point HI_0 of testing data is localized in the middle of their HI_0. Yet, when HI_0 falls outside the ranges of HI_0 of $U(B_i)$, thr and pr_0 are chosen based on the HI with the closest HI_0.

Finally, after the selection of the statistical model, the initial parameters and the failure threshold, PF starts to update the model parameters with the availability of new measurements. The RUL is calculated recursively by the extrapolation to the thr. As a set of particles is used, the RUL is then expressed as the probability of distribution. The RUL result is evaluated with three classic metrics, i.e., the Cumulative Relative Accuracy (CRA), the Convergence (C_m) and the Error (Err) [16].

The Relative Accuracy (RA) evaluates the error of the RUL prediction in relation to the actual RUL at the specific time. In order to measure the error at multiple time instances, the CRA aggregates the RA of a given time span.

$$\mathrm{CRA}_\lambda = \frac{1}{|l_\lambda|}\sum_{l_\lambda} w_k \mathrm{RA}_k \tag{11}$$

where, l_λ is the set of all time indexes, w_k and RA_k are, respectively, the weight factor and the RA at the time step k. The CRA approaches to 1 in the case of the ideal RUL prediction.

By the measurement of the Euclidean distance, convergence expresses how fast the prediction goes close to the actual RUL.

$$\mathrm{C_m} = \sqrt{(x_c - t_p)^2 + {y_c}^2} \tag{12}$$

where x_c and y_c are the centroid coordinates of the area under the prediction error, such as the accuracy and t_p is the first inspection time. A smaller distance shows that the predicted RUL converges faster to the actual RUL.

In addition, the percentage error is also frequently used, and is defined as follows:

$$\mathrm{Err} = \frac{\mathrm{RUL_{re}} - \mathrm{RUL_{pr}}}{\mathrm{RUL_{re}}} \times 100\% \tag{13}$$

where, RUL_{re} and RUL_{pr} are the actual and the predicted RUL, respectively. The percentage is equal to 0 when the predicted RUL coincides with the actual RUL.

4 Experimental degradation data

The proposed methodology is evaluated on the experimental dataset PRONOSTIA [17]. The setup consists of a motor, a custom gearbox, two support bearings and the test bearing, which bears load from a pneumatic actuator. The experiment is implemented under three constant operating conditions (radial load and rotating speed). The data is measured with a sampling frequency of 25.6 kHz for a duration of 0.1 seconds every 10 seconds. Acceleration signals are recorded simultaneously in the horizontal and vertical direction. The EoL in the experiment is defined based on the vibration amplitude. The measurement stops when the vibration at one of the two measurement directions exceeds 20g in a sequence of samples. In the end, seven measurements are implemented for the first two conditions and three measurements for the condition 3.

5 Results and discussion

Based on the experimental dataset and the established procedure of the proposed methodology, the results of the RUL prediction are discussed. In this section, the experimental dataset is firstly preprocessed before calculating the Info-entropy. In the view of trendability, the extracted HI is compared with some existing papers. Then, the statistical model, the initial parameters and the failure threshold are decided. Finally, the RUL prediction results are presented.

5.1 Data preprocessing

Following the explanation given in section 3, the preprocessing is important to remove the environmental influences. After visual inspection, the time wave of the vibration signals contain outliers and unsymmetric points. The window length D in MMAD is empirically chosen equal to 50. After cleanness, the samples against the degradation behavior are efficiently separated from the normality. In addition, a prewhitening technique based on the linear prediction is utilized before the Spectral NegEntropy. The order is set equal to 50.

5.2 Info-entropy extraction

Based on the description of Section 4, seven bearings have been measured in the first operating condition. In order to demonstrate the priority of the proposed HI in comparison with other state of the art papers [9], [11], only the same 4 bearings (bearing 1-1, bearing 1-3, bearing 1-4 and bearing 1-7) in the first operating conditions are predicted in this paper.

After conditioning of the vibration signals, the Info-entropy is firstly calculated on the six training sets. The attained informative frequency band and the entropy indicator are then employed into the rest testing data. In the process of Info-entropy calculation, five entropy indicators, i.e. $SpecE_{1-N}$, $SpecF_{1-N}$, $NegE\_SE_{1-N}$, $NegE\_SES_{1-N}$ and $NegE\_Ave_{1-N}$ are simultaneously achieved. In this paper, only the example of NegE-SE of the bearing 1-1 is presented in Figure 4. The NegE-SE in each frequency band performs an increasing tendency and it coincides with the degradation evolution and the accumulation of damage. The Info-entropy is able to find the highest trendability for all training datasets in certain frequency band, which represents the characteristic information of the same operating condition.

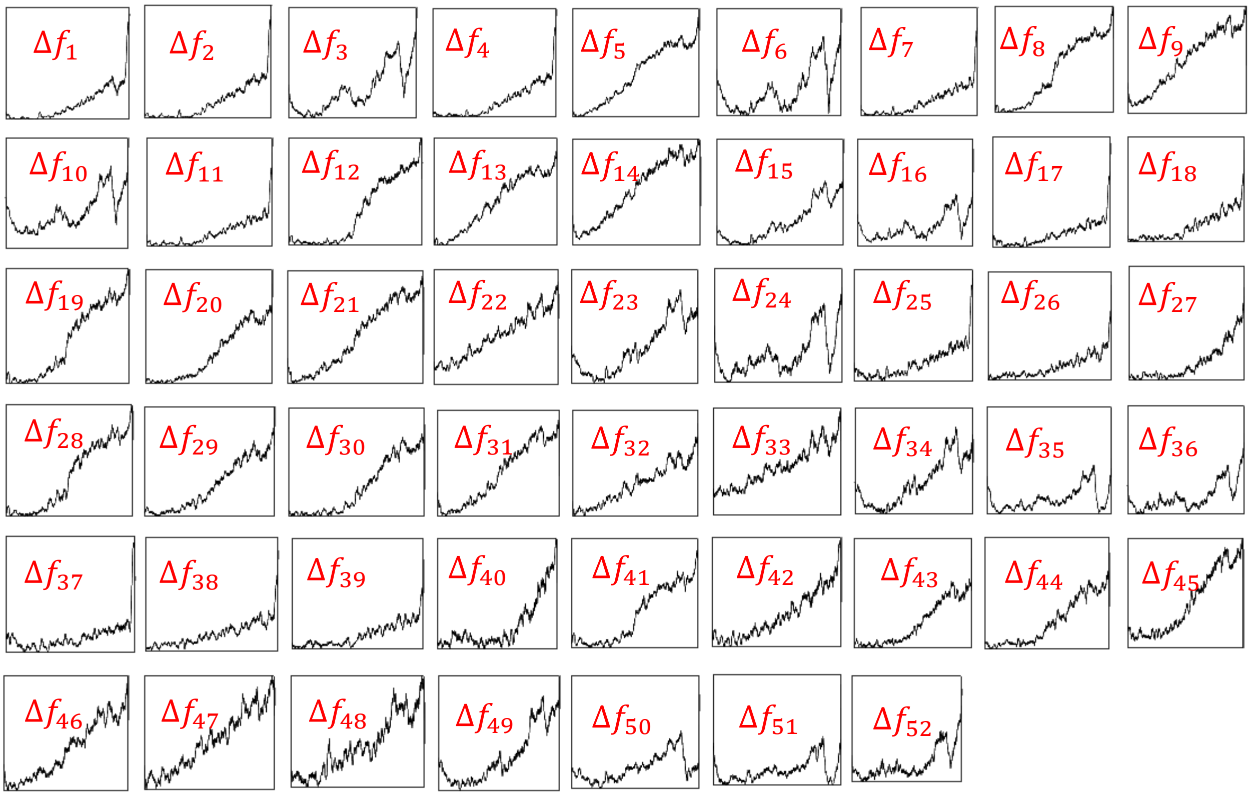


Figure 4: Spectral NegE-SE of decomposed exemplary signal (bearing 1-1)

	Bearing 1-1	Bearing 1-3	Bearing 1-4	Bearing 1-7
RMS	0.3387	0.4676	0.2707	0.6127
Kurtosis	0.3576	0.1863	0.3633	0.2786

SK	0.1840	0.1692	0.5237	0.5131
Info-entropy	0.0215	0.0467	0.1778	0.0428

Table 1: Trendability comparison of different His

The extracted feature in the frequency bands $\Delta f_1 - \Delta f_{52}$ in Figure 4 is generated by a 1/3 binary-tree, where the decomposition level L_v is empirically set equal to 4. Theoretically, the larger level L_v is able to capture more information with more decomposed signals. However, it requires also more computation cost. Based on the score of Equation 15, NegE-SE is selected as the Info-entropy among five different entropy indicators over the frequency band Δf_5, i.e., 4267 Hz-8533 Hz. The NegE-SE of four to be tested bearings are shown in Figure 5d. To highlight the superiority of the proposed HI, two statistical features Kurtosis and RMS for the same bearings in [9] and the Spectral Kurtosis (SK) are introduced as a comparison, and their trends are shown together in the Figure 5.

Analysing Figure 5a, RMS indicates the stable trend in the early stage. Especially for bearing 1-4 and 1-7, it suddenly reaches to failure after the long term ‘healthy’ running. Compared with RMS, Kurtosis in Figure 5b is sensitive to the impulsiveness and the slope of degradation is more distinct, e.g. in bearing 1-7. However, its trend drops in the late stage, where the fault size may accumulate and the vibration signal becomes less impulsive than the early stage fault. SK in Figure 5c has similar results as Kurtosis, but with slight higher slop. It is obvious that Info-entropy tracks well the full lifetime degradation for the four bearings. The slope of each bearing is high and cluster together. In addition, their corresponding scores based on the Equation 10 are listed in Table 1. The values of Kurtosis of bearings 1-3 and 1-7 approach closer to 0 than RMS. However, the SK of the bearing 1-1 and 1-3 triumphs over Kurtosis. It should be mentioned that the Info-entropy reveals the best score compared to the other indicators for each bearing. Therefore, with the advantage of trendability, the Info-entropy will be applied in the next section.

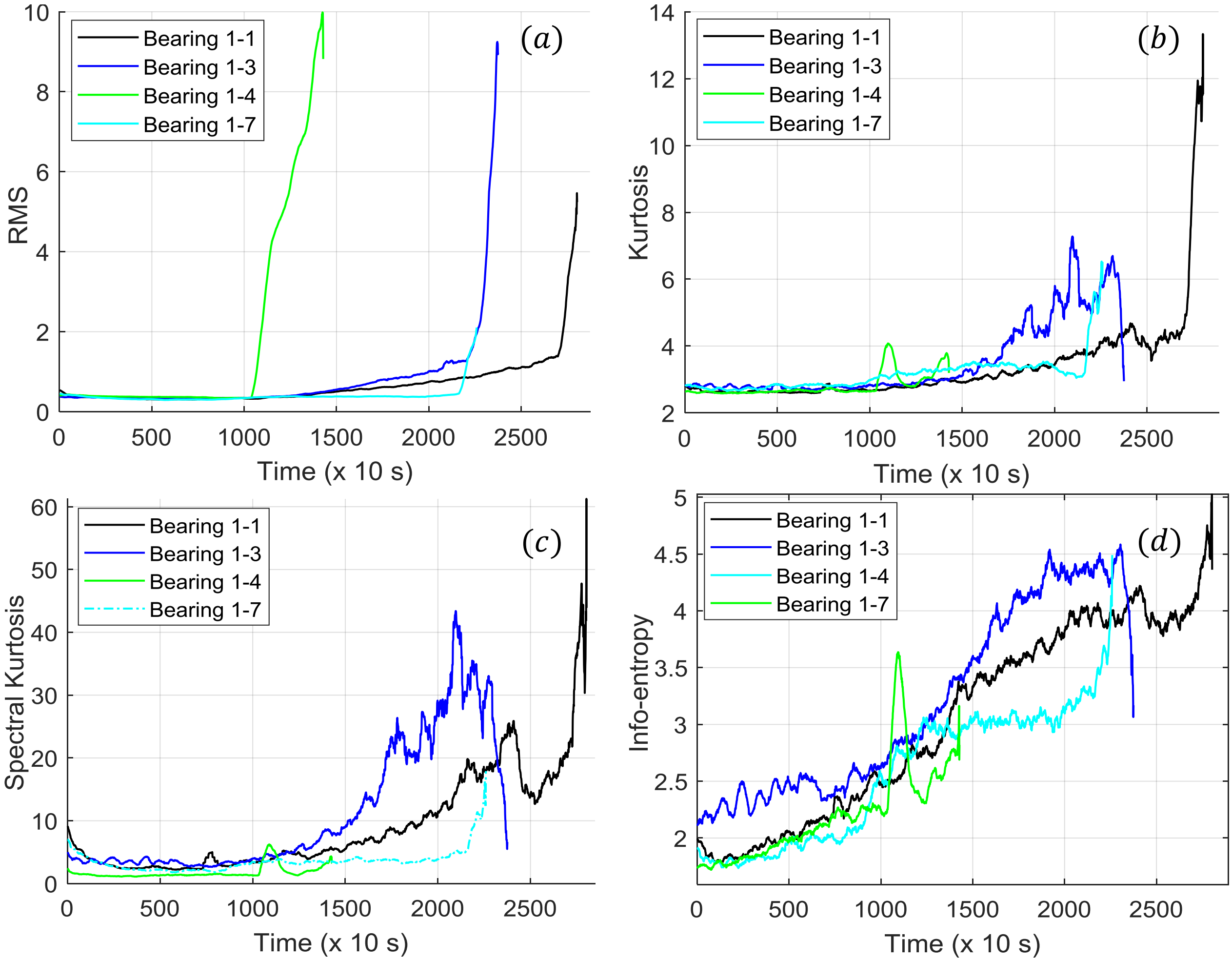


Figure 5: Comparison of different HIs

5.3 RUL results

5.3.1 Parameters definition

From the extracted Info-entropy, the degradation trend of each bearing generally follows an exponential path. Hereby, a classic exponential model is thus chosen for PF:

$$f(t) = ae^{b \cdot t} \tag{14}$$

where, a and b are the model parameters.

The failure threshold and the initial parameters are realized based on the strategy in Section 3. The threshold of similarity level ϵ is set equal to 0.6. The ratio is lower than 0.6 for the bearings in condition 1. In addition, as known, Spectral Entropy and Spectral Flatness range between 0 and 1. To facilitate the implementation, Spectral NegEntropy is normalized by the maximal value of HIs from the training dataset.

5.3.2 Results

After the selection of the HI, the statistical model and the parameters definition, the RUL of the aforementioned four bearings (length $Len = 1400, 1802, 1139, 1502$, duration $d = Len \times 10\ s$) is predicted based on the proposed methodology. In order to evaluate the effectiveness of different algorithms, the PF based methods, e.g., the Classic Particle Filter (CPF) with MR strategy and the Enhanced Particle Filter (EPF) with SR, are compared with other SMB methods, i.e. the CKF, the EKF, the UKF and the simple regression (REG). Finally, their performance are evaluated using the performance metrics, e.g. the CRA, the C_m and the Err.

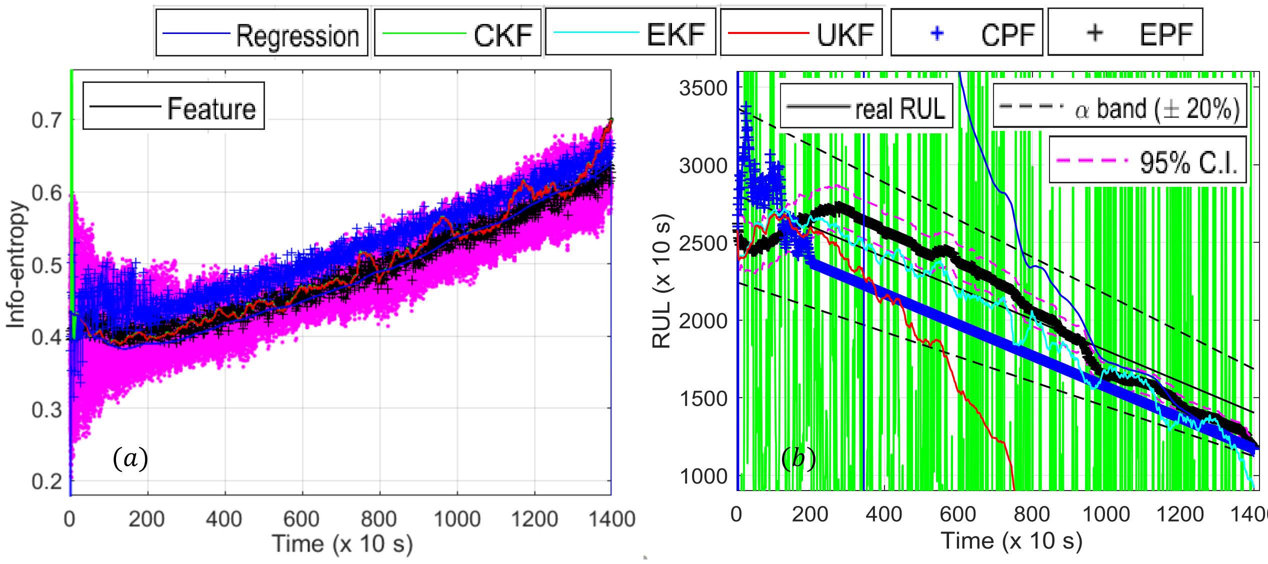


Figure 6: Feature estimation (links) and RUL estimation (right) by different methods for bearing 1-1

The results of bearing 1-1 are presented in Figure 6. In the Figure 6a, the feature is estimated perfectly using the KF versions (i.e. CKF, EKF and UKF). The particle filter generates a number of particles to approximate the posterior distribution of the state in each step. The mean value of the CPF and the EPF is plotted in the Plus (+) sign. From the Figure 6a, the PF tracks the general trend of the feature till the end, and the EPF approximates closer to the feature than CPF. By updating the model parameters in each step, the RUL result is able to be calculated by the extrapolation to the predefined threshold. The corresponding RUL estimation is presented in Figure 6b. The CKF estimates the RUL with severe fluctuation without convergence, while EKF and UKF predict in a similar manner before 200 (x 10 s). Afterwards, UKF diverges from the actual RUL, probably because of the random generated sigma points. However, the EKF follows almost the actual RUL before 900 (x 10 s) and then deviates to the outside of the (+/-) 20% error band. As comparison, the EPF performs the comparable prediction as EKF before 900 (x 10 s) and much closer to the actual RUL than the simple regression. Between 900 (x 10 s) and the end, the EPF goes slightly far from the actual RUL and has a better performance than EKF. It should be noticed that the full lifetime prediction by EPF falls within the error band. Moreover, the uncertainty of the RUL prediction is evaluated with the 95% confidence band, starting in wide uncertainty and decreasing with the time passing by. In contrast, the performance of CPF seems reliable with the stable distance to the actual RUL line, yet, it has the severe problem of particle degeneracy after 200 (x 10s).

The result of the second example bearing 1-4 is presented in Figure 7. Similarly with the results in Figure 6, the feature estimation of KFs demonstrates better than PFs in Figure 7a, especially for the part after 1000 (x 10 s). In Figure 7b, the performance of different algorithms is shown. The worst results are achieved by the CKF. UKF demonstrates good results before 400 (x 10 s) but then diverges to the outside of the error band. The RUL prediction by EKF seems more reliable than REG, CKF and UKF, but with large biased error before 1000 (x 10 s). Then, it converges fast to the actual RUL and diverges again in the very late stage. Analogously to the Figure 6b, CPF predicts in the form of a straight line after 180 (x 10 s). Although the observed prediction error is small, yet, it has a serious problem of particle degeneracy. As comparison, the EPF performs better results than EKF and REG before 1000 (x 10s), and converges with a biased error w.r.t the actual RUL. Looking back to the Info-entropy in Figure 5d, the failure threshold is set based on the bearing 1-7, which is much higher than the real life of bearing 1-4. This is the reason that the prediction of EPF and EKF has a biased error in this case. Therefore, it can be concluded that obviously the failure threshold set has a critical influence to the final RUL estimation.

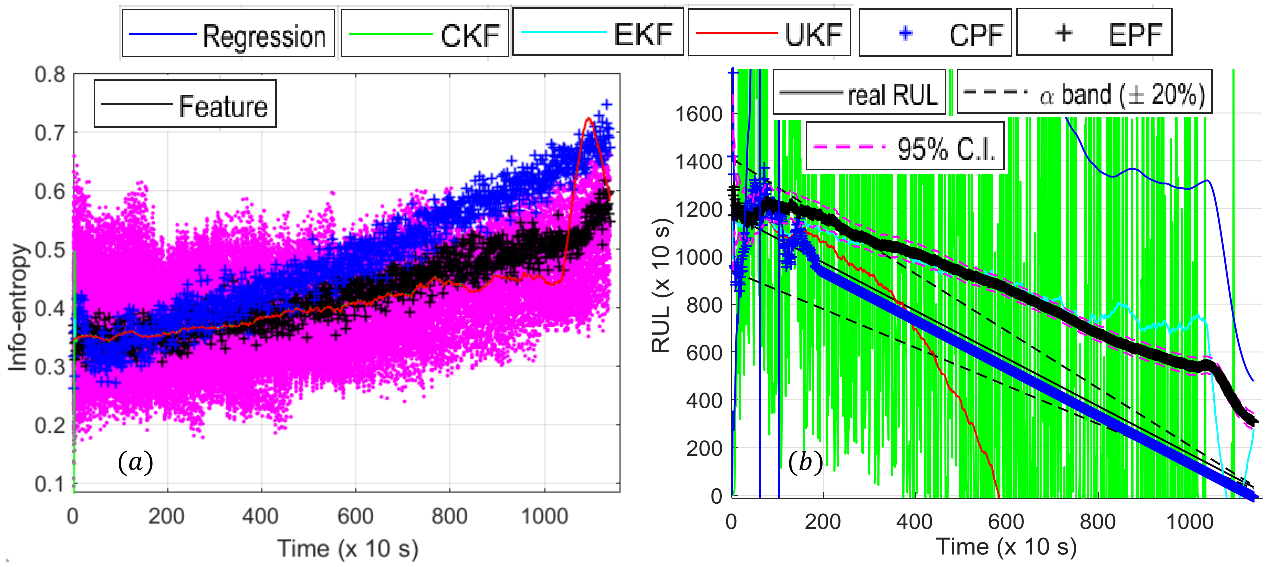


Figure 7: Feature estimation (links) and RUL estimation (right) by different methods for bearing 1-4

In addition to the visual depiction, the performance of these four bearings are presented in Table 2. For all bearings, the CRA of the EPF is the best ,closest to the 1, while the EKF performs better than the rest. With the huge fluctuation, CKF gains a negative CRA value for all bearings here. In the view of the Err, EPF performs the best score almost in total, with the exception of bearing 1-4. REG shows advantage than CKF, EKF and UKF in some bearings as it is calculated however based on the all history data in each step. Moreover, the convergence of EPF is not always best at CRA and Err and it may be surpassed by the CPF and CKF. However, the CPF is easily caught with the problem of degeneracy and the CKF usually has the larger prediction error.

Unlike the RUL prediction using the concept of start prediction point in [9], [11]-[13], with its help the good RUL results are possible to be achieved at the very late stage, i.e. the stage 3 of Figure 1. Using the proposed HI, the RUL in this paper is predicted from the beginning. The prediction of some bearings performs even better than [9] and [11], for example, the CRA of bearing 1-1 has better score than an exponential model in [9] and the bearing 1-3 and 1-7 has smaller error than the LRT technique in [11]. Therefore the proposed Info-entropy is able to track the degradation very early and can achieve good results in the long terms. The prognostics in the proposed methodology becomes more meaningful than state of the art papers, because the maintenance operators have plenty of time to schedule the repair plan, further avoiding a hazardous machinery breakdown.

6 Conclusion

In this paper, a statistical model based prognostics methodology for the RUL estimation of rolling element bearings is proposed. The methods of HI construction, statistical model selection and RUL estimation are explained in detail. Traditionally, a HI with a high trendability is expected, being capable to reflect the full lifetime degradation. However, the degradation status is commonly not easily tracked, especially in case of less impulsiveness and low frequency resolution. Many classic signal processing techniques may fail to extract high quality HI, which is critical to the selection of the failure threshold, of the model and even of the concept of start prediction point. In order to overcome these problems, entropy, as a chaos measurement, is considered. An HI Info-entropy is constructed based on the entropy indicators in the frequency domain, i.e., the Spectral Entropy, the Spectral Flatness and the Spectral NegEntropy. After the signal decomposition, the most informative frequency band is localized and the entropy indicator that follows the optimal degradation trend is selected.

Through the validation on the experimental dataset, Info-entropy is effective to capture the bearing degradation trend in the full lifetime. It provides a new perspective to construct the high trendability HI in a complex operation and further simplify the prognostics process, e.g. the model selection, the elimination of the concept of the start prediction point. In the end, the Info-entropy is combined with a classic exponential model, and is incorporated with six different methods, i.e. mathematical regression, classic Kalman filter, Extended Kalman filter, Unscented Kalman filter and Particle filter. Through the analysis, systematic resampling based Particle filter is efficient to better enhance the particle diversity than the classic resampling method and it achieves more accurate prediction results than others.

		Bearing 1-1	Bearing 1-3	Bearing 1-4	Bearing 1-7
REG	CRA	-0,07	0,02	-3,69	-5,41
	Err	20,20	-10,19	-1308,02	5,84
	C_m	912,92	855,53	356,43	1110,79
CKF	CRA	-1,59	-7,26	-8,12	-1,77
	Err	617,46	727,73	411,52	649,89
	C_m	167,54	357,92	3125,91	176,38
EKF	CRA	0,91	0,30	-0,30	0,35
	Err	34,38	-36,49	-717,90	-52,45
	C_m	681,80	892,43	410,18	804,69
UKF	CRA	-0,08	-3,46	-10,70	-0,86
	Err	496,72	546,77	283,19	546,74
	C_m	224,40	767,07	7740,37	391,44
CPF	CRA	0,87	0,64	0,81	0,34
	Err	16,64	64,57	119,43	101,15
	C_m	687,63	848,54	541,94	671,61
EPF	CRA	0,93	0,73	-0,50	0,44
	Err	15,52	-0,34	-808,25	-4,37
	C_m	696,84	949,01	407,38	792,10
PF [9]	CRA	0,87	0,76	0,87	0,93
	Err	/	/	/	/
	C_m	8,99	4,28	275,40	290,20
LRT [11]	CRA	0,94	/	/	/
	Err	/	7,00	0	19,00
	C_m	/	/	/	/

Table 2: Comparison of RUL estimation by different methods

References

[1] K. Javed, R. Gouriveau and N. Zerhouni, "State of the art and taxonomy of prognostics approaches, trends of prognostics applications and open issues towards maturity at different technology readiness levels," *Mech. Syst. Signal Process.*, vol. 94, pp. 214–236, 2017.

[2] A. Cubillo, S. Perinpanayagam and M. Esperon-Miguez, "A review of physics-based models in prognostics: Application to gears and bearings of rotating machinery," *ADV. MECH. ENG.*, vol. 8, no. 8, pp. 1-21, 2016.

[3] B. Sun, S. Zeng, R. Kang and M. G. Pecht, “Benefits and Challenges of System Prognostics,” *IEEE Trans. Rel.*, vol. 61, no. 2, pp. 323-335, Jun. 2012.
[4] N. Gebraeel, M. Lawley, R. Li and J. Ryan, “Residual Life Distribution from Component Degradation Signals: A Bayesian Approach,” *IIE Trans.*, vol. 37, no. 6, pp. 543-557, 2005.
[5] K. Medjaher, N. Zerhouni, and J. Baklouti, “Data-driven prognostics based on health indicator construction: Application to PRONOSTIA’s data,” in *Proc. Eur. Control Conf.*, 2013, pp. 1451–1456.
[6] D. Tobon-Mejia, K. Medjaher, N. Zerhouni, G. Tripot, “Hidden Markov models for failure diagnostic and prognostic.,” in *Proc. Prognostics Syst. Health Manage. Conf.*, Shenzhen, China, 2011 May, pp. 1–8.
[7] J. Qi, A. Mauricio, M. Sarrazin, K. Janssens and K. Gryllias, “Remaining Useful Life Prediction of Rolling Element Bearings Based on Unscented Kalman Filter,” in *Proc. Int. Conf. on Condition Monitoring of Machinery in Non-stationary operations*, Santander, Spain, 2018 June 20-22, pp. 123–138.
[8] C. K. R. Lim and D. Mba, “Switching Kalman filter for failure prognostic,” *Mech. Syst. Signal Process.*, vol. 52-53, pp. 426-435, Feb. 2015.
[9] N. Li, Y. Lei, J. Lin, and S. X. Ding, “An improved exponential model for predicting remaining useful life of rolling element bearings,” *IEEE Trans. Ind. Electron.*, vol. 62, no. 12, pp. 7762–7773, Dec. 2015.
[10] P. Wang, R. Qiang and R. X. Gao, “Multi-Mode Particle Filter for Bearing Remaining Life Prediction.,” in *Proc. Int. Manufacturing Science and Engineering Conf.*, Texas, USA, 2018 June.
[11] W. Ahmad, S. Khan and J. Kim, “A Hybrid Prognostics Technique for Rolling Element Bearings Using Adaptive Predictive Models.,” *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, Feb. 2018.
[12] R. K. Singleton, E. G. Strangas, and S. Aviyente, “Extended Kalman filtering for remaining-useful-life estimation of bearings,” *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1781–1790, Mar. 2015.
[13] A. Soualhi, K. Medjaher and N. Zerhouni, “Bearing Health monitoring based on Hilbert-Huang Transform, Support Vector Machine and Regression.,” *IEEE Trans. Instrum. Meas.*, vol. 64, no. 1, pp. 52–62, Jan. 2015.
[14] A. Gray and J. Markel, “A Spectral-Flatness Measure for Studying the Autocorrelation Method of Linear Prediction of Speech Analysis.,” *IEEE Trans Acoust.*, vol. 22, no. 3, pp. 207-217, Jun. 1974.
[15] J. Antoni, “The infogram: Entropic evidence of the signature of repetitive transients.,” *Mech. Syst. Signal Process.*, vol. 74, Nov. 2018.
[16] A. Saxena, J. Celaya, B. Saha, S. Saha and K. Goebel, “On Applying the Prognostic Performance Metrics.,” in *Proc. Prognostics Health Manage. Conf.*, San Diego, USA, 2009 September 27- October 1, pp. 1-16.
[17] P. Nectoux, R. Couriveau, K. Medjaher and E. Ramasso, “PRONOSTIA: An experimental platform for bearings accelerated degradation tests.,” in *Proc. IEEE Int. Conf. on Prognostics Health Manage.*, Denver, USA, 2012 June, pp. 1-8.

Spall Evolution in a Rolling Element Bearing

Dmitri Gazizulin[1], Roni Schneck[2], Renata Klein[3], Jacob Bortman[1]

[1]PHM Lab, Department of Mechanical Engineering, Ben-Gurion University of the Negev, P.O. Box 653, Beer Sheva 8410501, Israel
dima.gzizo@gmail.com
jacbort@bgu.ac.il

[2]Department of Materials Engineering, Ben-Gurion University of the Negev, P.O. Box 653, Beer Sheva 8410501, Israel
roni.shneck@gmail.com

[3] R.K. Diagnostics, Gilon, P.O.B. 101, D.N. Misgav 20103, Israel
renata.klein@rkdiagnostics.co.il

Abstract

Rolling element bearing (REB) is one of the basic mechanical components in rotating machinery. It is common to divide the REB wear into two stages, damage initiation and damage propagation. There has been a growing awareness of the need to understand the damage mechanism during the propagation phase. The current work includes a discussion on the ongoing research and the methodology for the development of the prognostic method for damage propagation. The methodology integrates experiments, diagnostic methods, and physics-based models. Endurance tests were conducted in order to learn about the damage propagation process and model validation. Furthermore, Finite Element model of the spalled bearing was developed and validated. The FE model aims to investigate and simulate the damage propagation process. The simulation results are in good agreement with the experimental observation.

1 Introduction

Failure prognosis of rolling element bearings (REBs) is crucial in rotating machinery PHM. The damage evolution in REBs consists of two main phases: damage initiation and propagation. The conventional REB life models address the lifetime of the bearing to the damage initiation, i.e. first spall formation [1]. However, after the first spall formation, the bearing might be fully operational for millions of cycles. After the first spall formation, it propagates in the circumferential direction of the raceway, until the bearing becomes non-operational [2].

Many diagnostic tools have been developed in order to monitor the spall propagation. However, there is lack of reliable prognostic of the remaining useful life (RUL). Physic-based prediction of the damage propagation in the REB, after the first spall generation. The difficulties in prognosis of the propagation stage necessitate deep understanding of the damage mechanisms, the stochastic nature of the spall propagation process, and its modeling [3, 4]. The main goal of the current work is to develop a physics-based prognostic method for the spall propagation in REBs. Herein, we present a methodology for the development of the prognostic method, which combines experiments, diagnostic methods, and physics-based models.

2 Methodology

The proposed methodology for prognostic of the RUL is based on a combination of physics-based models, diagnostic methods and experiments [7]. The concept is presented in Figure 1. This section includes a description of the objectives that need to be reached in order to achieve the main goal, RUL prognostics.

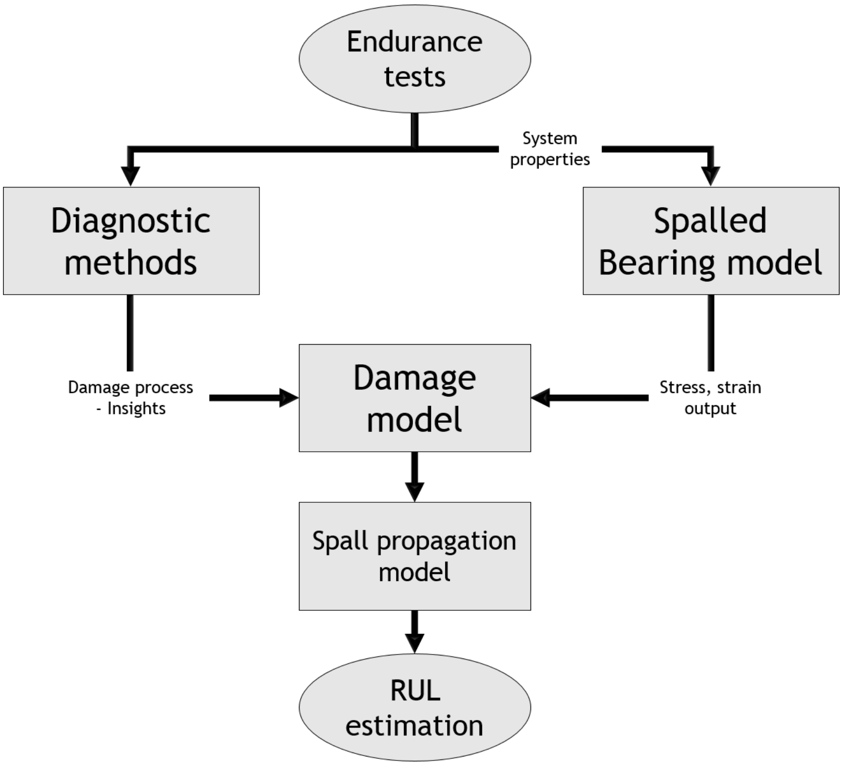


Figure 1: Research flow chart describing the steps toward the development of the prognostic method.

The first objective is the quantitative and qualitative understanding of the damage-driven mechanisms, e.g. plastic strains, residual stresses, etc., of the spall propagation process. In addition, it is important to learn, based on the existing literature and experiments, about the effects of the bearing's features (e.g. hardness, ball mass), and operational conditions (e.g. speed, load) on the propagation process and the trend of the spall growth [8, 9]. This objective can be attained by conducting endurance tests. Figure 2 shows an example of a diagnostic indicator evolution, BPFO Z-score distance vs. time. The tests can add insight regarding the spall propagation process and can be used for the physical model validation.

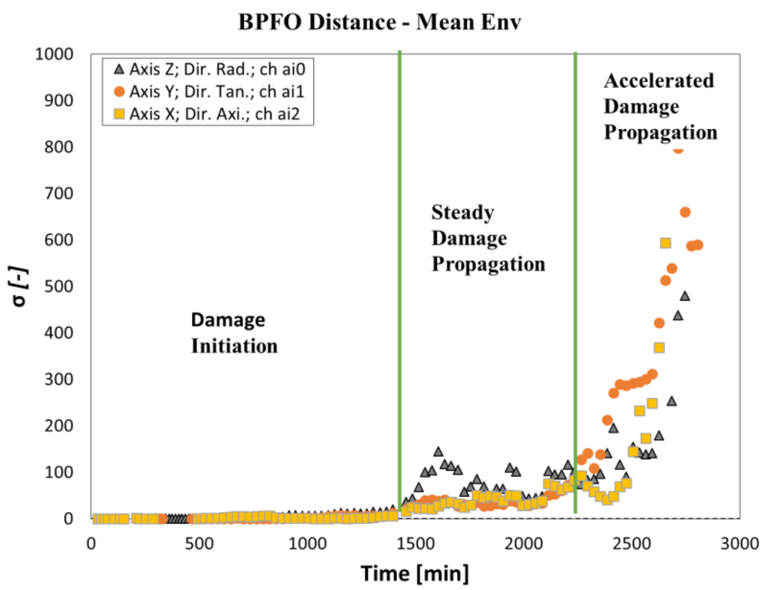


Figure 2: Damage evolution trajectory during the endurance test. Each data point correspond to the vibration measurement during the test.

The second objective is the development of a model for the damage propagation process. First, the material response in the presence of a defect must be analyzed. This analysis, coupled with the endurance test results, will shed light on the mechanism governing the damage propagation process. For this purpose a physics-based model of a spalled bearing has been developed. The model is used for analysis of the material response in the presence of a defect. Moreover, the model was used for damage evolution simulation. Figure 3 shows examples of a simulation result and the metallurgical analysis of a bearing from an endurance test. The accomplishment of this objective will provide a mean estimated damage trajectory. However, the damage propagation is a stochastic process. Hence, dispersed results are expected.

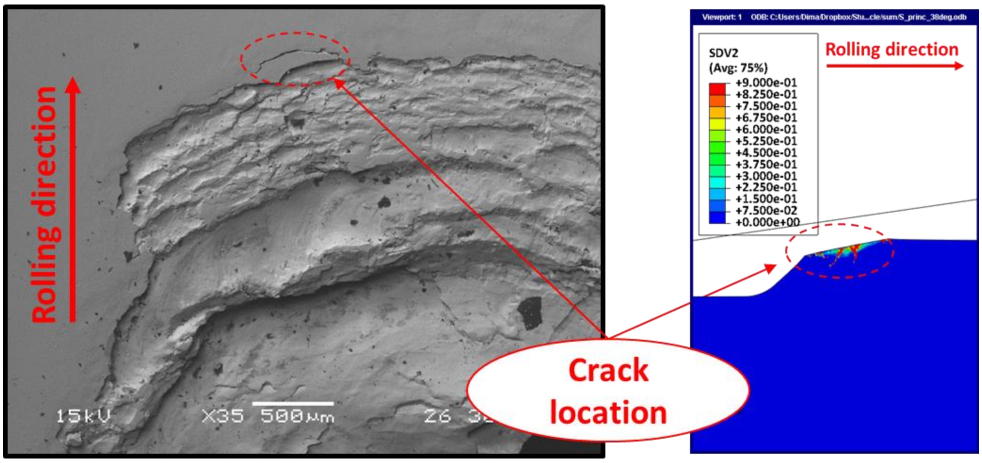
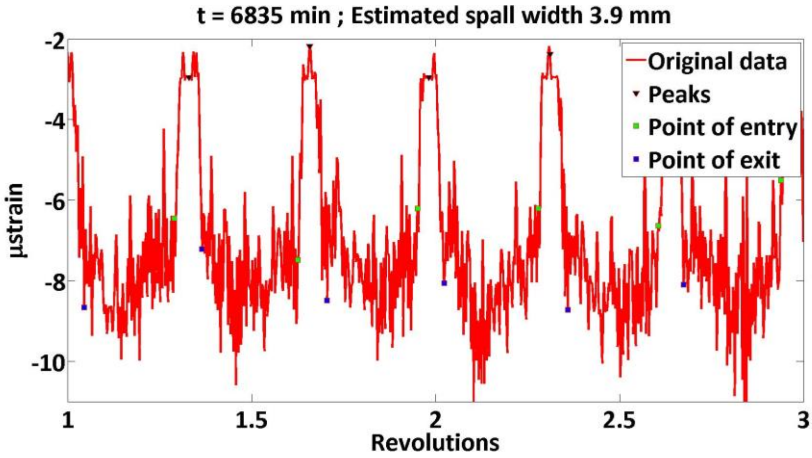
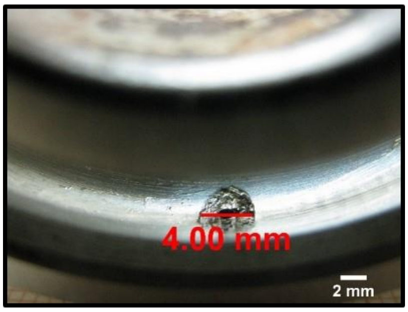


Figure 3: SEM Image of the spall trailing edge and simulated cracks.

The third objective of the research, and probably the most challenging, is to model the stochastic nature of the damage propagation. The propagation of the spall in the REB is a stochastic process. Even under well controlled experimental conditions, using allegedly identical bearings, the results of the endurance tests vary [2]. The prognostic method must consider the uncertainties and the progress of the probability distribution. One of the common methods is to use diagnostic condition indicators in the early stages of the damage in order to monitor its propagation, e.g. oil debris, vibration level, etc. For this purpose, methods for the spall width estimation via time domain analysis were developed [10]. The spall width is estimated based on the acceleration and strain signatures, Figure 4. The spall width and vibration based CIs will be used for the estimation of the damage model parameters by the trend identification of the spall propagation process. The accomplishment of this goal will complete the development of the physics-based prognostic method.



(a)



(b)

Figure 4: Strain data vs. cycle, based on which the spall size was estimated. The (a) estimated size is 3.9 *mm*, and the (d) measured size is 4.0 *mm*.

3 Summary and Conclusions

The remaining useful life (RUL) estimation using physics-based prognostic method is schematically illustrated in Figure 5. The calibration of the damage model can be implemented by comparing the simulations results with the data extracted from endurance tests. For example, a diagnostic method for the defect severity estimation (fault size, vibration level, etc.) can be implemented during the first stages of the tests. The results obtained by the diagnostic method can be used for the estimation of the damage model parameters and their uncertainties. The integration of the prognostic and the diagnostic methods, has the potential for reliable online estimation of the RUL including probability distribution of the result.

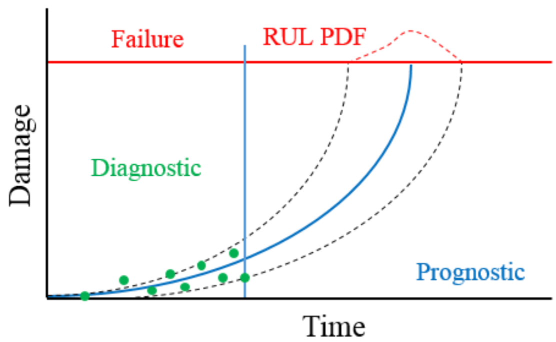


Figure 5: Bearing prognosis - first the damage propagation process is monitored; next, the model parameters are estimated; and the RUL is calculated [7].

Acknowledgments

The authors would like to express their deepest appreciations to Pearlstone Center for their support and funds of this work.

References

[1] F. Sadeghi, B. Jalalahmadi, T. S. Slack, N. Raje, and N. K. Arakere, "A review of rolling contact fatigue," *Journal of tribology*, vol. 131, no. 4, p. 041403, 2009.

[2] L. Rosado, N. H. Forster, K. L. Thompson, and J. W. Cooke, "Rolling contact fatigue life and spall propagation of aisi m50, m50nil, and aisi 52100, part i: experimental results," *Tribology Transactions*, vol. 53, no. 1, pp. 29–41, 2009.

[3] Y. Li, S. Billington, C. Zhang, T. Kurfess, S. Danyluk, and S. Liang, "Adaptive prognostics for rolling element bearing condition," *Mechanical systems and signal processing*, vol. 13, no. 1, pp. 103–113, 1999.

[4] J. Qiu, B. B. Seth, S. Y. Liang, and C. Zhang, "Damage mechanics approach for bearing lifetime prognostics," *Mechanical systems and signal processing*, vol. 16, no. 5, pp. 817–829, 2002.

[5] M. Hoeprich, "Rolling element bearing fatigue damage propagation," *ASME, Transactions, Journal of Tribology*, vol. 114, no. 2, pp. 328–333, 1992.

[6] M. L. French and W. M. Hannon, "Angular contact ball bearing experimental spall propagation observations," *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, vol. 229, no. 8, pp. 902–916, 2015.

[7] D. Gazizulin, R. Klein, and J. Bortman, "Physics based methodology for the estimation of bearings' remaining useful life: Physics-based models, diagnostic methods and experiments," in *Proceedings of the European Conference of the PHM Society*. PHM Society, 2018.

[8] N. K. Arakere, N. Branch, G. Levesque, V. Svendsen, and N. H. Forster, "Rolling contact fatigue life and spall propagation of aisi m50, m50nil, and aisi 52100, part ii: stress modeling," *Tribology Transactions*, vol. 53, no. 1, pp. 42–51, 2009.

[9] N. A. Branch, N. K. Arakere, V. Svendsen, and N. H. Forster, "Stress field evolution in a ball bearing raceway fatigue spall," *Journal of ASTM international*, vol. 7, no. 2, pp. 1–18, 2009.

[10] G. Kogan, J. Bortman, and R. Klein, "A new model for spall-rolling-element interaction," *Nonlinear Dynamics*, vol. 87, no. 1, pp. 219–236, 2017.

[11] J.-L. Chaboche, "Continuum damage mechanics: Part i - general concepts," *Journal of Applied Mechanics*, vol. 55, no. 1, pp. 59–64, 1988. [Online]. Available: http://dx.doi.org/10.1115/1.3173661

[12] S. Marble and B. P. Morton, "Predicting the remaining life of propulsion system bearings," in *Aerospace Conference, 2006 IEEE*, 2006, pp. 8 pp.–.

Data Mining Classification & Machine Learning methods

Multi-label fault diagnosis based on Convolutional Neural Network and Cyclic Spectral Coherence

Zhuyun Chen[1], Alexandre Mauricio[2,3], Weihua Li[1], Konstantinos Gryllias[2,3]

[1]School of Mechanical and Automotive Engineering, South China University of Technology, 510640, GuangZhou, China
(mezychen@gmail.com; whlee@scut.edu.cn)
[2]Department of Mechanical Engineering, KU Leuven, Celestijnenlaan 300, Leuven, 3001, Belgium
[3]Dynamics of Mechanical and Mechatronic Systems, Flanders Make, Leuven, Belgium
(alex.ricardomauricio@kuleuven.be; konstantinos.gryllias@kuleuven.be)

Abstract
Rotating machines are widely used in manufacturing industry, where sudden failures of key components such as bearings may lead to unexpected breakdown of machines and cause economic loss and human casualties. In addition, machines usually are operating under different working conditions leading to the dynamic changes of fault characteristic, thus presenting big challenges of reliable and accurate fault diagnosis. Data-driven based Deep Learning (DL) fault diagnosis methods are powerful tools to capture hierarchical features from raw input to classify fault patterns by stacking multiple non-linear transformation layers. It constructs and trains deep models relying on huge historical data and requiring less expert knowledge to obtain decision-making. These techniques present effectiveness and advantages in many intelligent fault diagnosis tasks. However, many DL methods are developed for the diagnosis of single fault type without considering the correlations of fault modes. In this paper, we develop a novel fault diagnosis method based on cyclostationary tool and Convolutional Neural Networks (CNN) to tackle these problems. The proposed method presents three characteristics: 1) Cyclic Spectral Coherence (CSCoh) is adopted to provide bearing discriminative patterns for specific type of faults. 2) A fault occurred on the same component (fault pattern), but having different fault severity levels can be regarded a multi-label classification problem, where the fault pattern and the fault severity level are considered to be two specific faults. 3) A novel CNN is constructed by introducing a sigmoid activation output and binary-cross entropy loss function to conduct the multi-label classification task. Specifically, CSCoh is adopted to capture correlation features of periodic phenomenon in the frequency domain. CSCoh is a bi-variable map of two frequency values, which could be used to enhance signatures masked by strong noise, characterizing the fault vibration signals obtained from the rotating machinery under different operating conditions. Then a CNN is developed for multi-label fault classification, which includes fault patterns and fault severity levels identification. The proposed method is evaluated in the experimental study of rolling element bearing fault diagnosis, in which data are collected under different working conditions. The experiment results demonstrate that the proposed method could presents good classification performance and superiority compared with other approaches.

1 Introduction

Rotating machines play important role in manufacturing industry. Rolling element bearings, as the key components of machines easily suffer from the sudden failures due to the long operating under the harsh conditions. This unexpected breakdown of machines may lead to economic loss and even human casualties. Thus, it is essential to develop the condition monitoring techniques for the early and accurate defect detection of such components.

Recently, data driven based-DL intelligent fault diagnosis methods have achieved increasing attentions, due to the powerful feature leaning capability from raw input. DL algorithms refer to deep neural networks, where multiply non-linear transformation layers are stacked to construct the hierarchical architectures. Each layer can be regarded as a data pre-processing unit, where the input is converted into abstract features. With the increase of the layers, the high-level layer can learn more discriminative representations which are helpful for the diagnosis tasks [1]-[2]. The typical DL algorithms, such as Auto Encoder (SAE), Deep Belief Network (DBN), Convolutional Neural Network (CNN) and Long Short-Term Memory network (LSTM), have been applied for the fault diagnosis and detection [3]-[5].

Wang [6] took advantages of the CNN to learn features automatically from the raw vibration signals. Then, Hidden Markov Models (HMM) were employed as strong stability tools to classify rolling element bearing faults. Chen [7] proposed a SAE-DBN method for fault diagnosis by utilizing the multiply sensor information. In the first step, eighteen statistic indexes were extracted from the raw vibration signals, which were fed into sparse autoencoders for feature fusion. Finally, the fused features were input into DBN for fault diagnosis. Janssens [8] extracted the frequency spectra from two vibration accelerometers, and then a 2D CNN with one convolutional layer was designed to learn useful features for bearing fault detection. The proposed method demonstrated its advantages compared to that with hand-crafted features. Sun [9] presented a sparse Deep Stacking Network (DSN) to improve motor diagnosis performance, where the output label of DSN was coded as binary 0 and 1, which leads to more accurate and robust classification results. Chopra [10] adopted SAE for unsupervised features extraction from the engine data, and the majority voting based criteria was used to determine the engine fault type. Althobiani [11] utilized both the Teager Kaiser energy operator and statistical measures to reveal the fault patterns contained in collected signals, and then further adopted DBN for diagnosis of reciprocating compressor. The proposed method provided highly reliable and applicable. Tamilselvan [12] applied DBN for aircraft engine health diagnosis and electric power transformer health diagnosis, which obtained high classification accuracy and presented good generalization performance. In addition, Ince [13] developed a 1D CNN to conduct end-to-end motor fault diagnosis from raw signal input. Jia [14] proposed a normalized CNN for improving the bearing diagnosis performance under imbalanced data by embedding normalized layers and weighted Softmax loss.

From those works mentioned above, different measurements, such as raw time-series signals, frequency spectra, time domain and frequency domain statistical indexes, were adopted as the input of the DL algorithms, which obtained high diagnosis results. However, most of the studies are focusing on distinguishing different fault patterns while ignoring the diagnosis of fault severity levels. The fault severity identification is meaningful to detect the early fault occurrences and guide the decision-making.

Gan [15] adopted the Wavelet Packet Transform (WPT) to extract representative features and then designed a two-layer Hierarchical Diagnosis Network (HDN) for rolling bearing faults. In this study, different DBNs are stacked together, respectively, for classifying fault patterns and fault severity levels. However, multiply DBNs should be designed and trained for meeting the classification tasks. In addition, Wen [16] proposed a Hierarchical Convolutional Neural Network (HCNN), which can be used to recognize the fault patterns and fault severity levels at the same time. In this work, two fully-connected branches are designed at the end of pooling layer to conduct diagnosis. The first branch is used for the diagnosis of fault patterns, and the second branch is used for the recognition of fault severity levels. However, the drawback is that each branch should be constructed with an independently loss function, and a parameter was introduced to balance the trade-off of two losses of HCNN, which requires much expertise knowledge and computational cost for training.

Inspired by previous works, a novel approach integrating CNN and CSCoh is proposed for the multi-label fault classification of rolling element bearings. Firstly, CSCoh is adopted to capture correlation features of periodic phenomenon in the frequency domain, which provides a good discriminative input for CNN. Then a novel CNN is constructed for implementing multi-label fault classification by introducing a novel activation function and loss function. Compared to the other methods, the proposed method only requires a CNN to obtain the fault patterns and fault severity levels at the same time. In addition, compared to other methods, the proposed only replaces the activation function and loss function, without introducing extra parameters and computational cost, which is more suitable for the real industrial applications.

The remaining part of the paper is organized as follows. In Section 2, the theory of the cyclic spectral analysis and the CNN are provided. The diagnosis procedure using the proposed method is introduced in

Section 3. In Section 4, a comprehensive experimental description and the analytical are introduced. Section 5 describes the conclusion of this paper.

2 Introduction to cyclic spectral analysis and Convolutional Neural Networks

2.1 cyclic spectral analysis

In rotating machines, the bearing defects usually generate modulated signals by the characteristic frequencies of the bearings. Such signal, though not periodic, usually can be described as cyclostationary, whose statistical properties vary periodically with time [17]-[18]. The common spectral analysis technique is. Fast Fourier Transform (FFT). It is based on the assumption that the analysed signals are stationary, which can not accurately describe the cyclostationary. To deal with the problems, the cyclic spectral analysis techniques are developed to detecting and identify the hidden periodic behaviour of signals [19]-[20].

For a cyclostationary signal $x(t)$, the second-order moment of cyclostationarity can be defined as an instantaneous AutoCorrelation Function (ACF) with a cyclic T, which is defined as:

$$R_{xx}(t,\tau)=R_{xx}(t+T,\tau)=E\{x(t+\tau/2)x(t\text{-}\tau/2)^{*}\} \tag{1}$$

Then, the second-order statistical descriptor of cyclostationarity, called the Cyclic Spectral Correlation (CSC), can be estimated by implementing the double Fourier transform on the ACF, which is given by:

$$\mathrm{CSC}(\alpha, f)=\int\int R(t,\tau)e^{-j2\pi(\alpha t+f\tau)}dtd\tau \tag{2}$$

The CSC is a bi-variable map of two frequency values. The parameters f and α are called the spectral frequency and cycle frequency, respectively. Contrary to the classic spectral analysis, it provides an additional frequency dimension, revealing both the carriers and their modulations. Spectral frequency f is linked to the carrier component, and the cyclic frequency α is linked to its modulation. It can be observed that for α is equal to zero, it is the classical power spectrum. Furthermore, for α is not equal to zero, it is the power spectrum for that specific cyclic component. Then the Cyclic Spectral Coherence (CSCoh) can be used to measure the degree of correlation between two spectral components given estimated by:

$$\mathrm{CSCoh}(\alpha, f)=\frac{\mathrm{CSC}(\alpha, f)}{\mathrm{CSC}(0, f)\mathrm{CSC}(0, f-\alpha)} \tag{3}$$

The CSCoh can be interpreted as the CSC of a whitened signal, which tends to equalize regions with very different energy levels, magnifying weak cyclostationary signals [20].

2.2 Convolutional Neural Network

CNN as a category of multi-layer neural network has achieved great success in areas such as image recognition, image classification, object detections, recognition faces [21]. A typical CNN usually is constructed by the four main operations, convolutional layer, activation layer, pooling or sub sampling layer, and fully-connected layers. Different kinds of layers play different roles. By stacking multiply convolutional, pooling and fully-connected layers, CNN can learn from low-level features to high-order or more abstract features. The layer types considered in this work are introduced.

2.2.1 Convolutional layer

Convolution is the first layer of CNN. The primary purpose of convolution is to extract the features by implementing the convolution operation on raw input data with learned convolution kernel/weights [22]. For each input $\mathbf{x}_i$ and convolution kernel $\mathbf{k}_j$, the output feature map can be calculated as follows,

$$\begin{gathered}\mathbf{y}_{i,j}=f(\mathbf{b}_j+\textstyle\sum_i \mathbf{k}_j * \mathbf{x}_i)\\ f(\mathbf{x})=\max(0,\mathbf{x}),\ \mathbf{x}>0\end{gathered} \tag{4}$$

where, * denotes the convolution operation, **k** and **b** are the value of the kernel and the bias. $f(\ )$ is the activation function, which is usually selected as the Rectified Linear Unit (ReLU) to enable better training of CNN.

2.2.2 Pooling layer

In the second step, the pooling layer is followed, which is used to reduce the spatial dimension and gain computation performance and some translation invariance. This is achieved by summarizing the feature responses in a region of neurons in the previous layer. For an input feature map $\mathbf{x}_i$, the output feature map is obtained,

$$\mathbf{y}_i = \max_{r\times r}(\mathbf{x}_i) \tag{5}$$

where r is the pooling size, and the common pooling operation adopted is known as max-pooling, which slides a window, and gets the maximum on the window as the output.

2.2.3 Fully-connected layer

In the fully-connected layer, the neurons are fully connection to all activations in the previous layer, and a Softmax classifier is usually attached to compute the class score. For the input vector $\mathbf{z}_i$ (i=1, 2, …, N), where N is the number of samples. the Softmax computes the exponential of the given input vector, and the sum of exponential values of all the values in the inputs. Then the ratio of the exponential of the input value and the sum of exponential values is the output of the Softmax function, which can be defined as,

$$Softmax(\mathbf{z}_i) = \frac{\exp(\mathbf{z}_i)}{\sum_j \exp(\mathbf{z}_i)} \tag{6}$$

The output corresponds to the probabilities of each class, and the target class will have the high probability. Softmax will enforce that the total sum of all the probabilities equals to one. That means, in order to increase the probability of a particular class, the module will correspondingly decrease the probability of at least one of the other classes. Thus, the final output will only have one true label. In order to effectively update the neural network, the Cross-Entropy (CE) loss can be adopted by minimizing the loss function between the probability output and the true target class, which is defined as,

$$Loss_{CE} = \sum_{i=1}^{C} \mathbf{y}_i (\log(\hat{\mathbf{y}}_i)) \tag{7}$$

where $\mathbf{y}$ is the true label of the data set, $\hat{\mathbf{y}}$ is the Softmax output, and C is the number of class.

3 The proposed CSCoh-CNN fault diagnosis framework

3.1 The architecture of the proposed CNN

In this section, the architecture of the proposed CNN is designed. Compared to the traditional CNN architecture, the proposed architecture introduces a new activation function in the output layer and a new loss function of CNN.

3.1.1 Sigmoid activation function

In the traditional CNN, the Softmax is usually regarded as the final fully-connected layer to predict the classes. While in the proposed CNN, it is replaced with Sigmoid activation function, which can be defined as,

$$Sigmoid(\mathbf{z}_i) = \frac{1}{1+\exp(-\mathbf{z}_i)} \tag{8}$$

For each value of the Sigmoid input, the Sigmoid function returns an independently real-valued output, which can be used to estimate the true output. For the Softmax output, the high value will have the higher probability than other values. That means for a classification problem, there is only one right class output, the outputs are mutually exclusive. While for the Sigmoid output, since the output are independently, it allows to have high probability for all of the classes, and the high value will have the high probability but not the higher probability. That means, for a multi-label classification, Sigmoid can output multiply correct classes, once a probability of one of the output nodes is above the threshold which is usually set to 0.5.

In order to better explain the differences of Softmax and Sigmoid, a fault diagnosis case is taken for example, presented in figure 1. When a Ball Fault with defect diameters of 14 mil (BF14) of rolling element bearing occurs, it can be observed that the traditional Softmax can correctly predict the BF fault with a probability of 85%. But it can only provide a true class output, which fails to diagnosis the severity level at the same time. On the contrary, for the Sigmoid activation output, it not only can obtain the fault pattern: BF with a probability of 92%, but also the fault severity level which is estimated with a probability value of 84%.

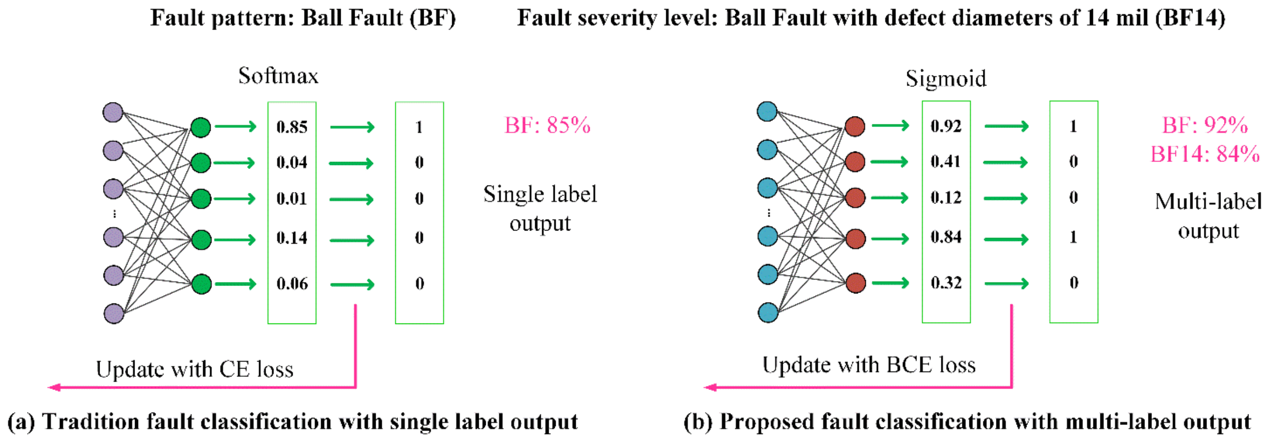


Figure 1: The differences between Softmax and Sigmoid activation function

3.1.2 Binary Cross-Entropy Loss

In order to optimize the CNN with multi-label classification tasks, the Binary Cross-Entropy (BCE) loss is adopted by splitting a multi-label classification problem in C binary classification problems. Unlike the CE loss, BCE is independently for each class, meaning that the loss computed for every output component is not affected by other output class. The loss function can be defined as,

$$Loss_{BCE} = \sum_{i=1}^{C} \mathbf{y}_i \log(\hat{\mathbf{y}}_i) + (1-\mathbf{y}_i)\log(1-\hat{\mathbf{y}}_i) \tag{9}$$

where the update of the weight can be easily implemented by Back-Propagation (BP) algorithm, which is the same as that of the traditional CNN.

3.2 Fault diagnosis based on the proposed CSCoh-CNN

In this section, a multi-label fault diagnosis framework combining CSCoh and CNN is constructed as presented in figure 2. Inspired by the typical LeNet-5 [22], the proposed CNN architecture is designed by stacking two convolutional layers, two pooling layers, one fully-connected layer, and one Sigmoid classification layer.

In the convolutional layer configurations, a small kernel size (3 x 3) is applied in each convolution layer to capture the detail information and reduce the number of parameters. The convolution stride is fixed to 1 pixel. The number of filters is set equal to 6 in the first convolutional layer and the second one is doubled (12) to increase the feature learning capability. In the pooling layers, the max pooling is carried on the input over a (2 x 2) pixel window with stride 2. Therefore, the feature map size is halved to reduce the time complexity. Then the fully-connected architectures is set to 100 neurons. The number of Sigmoid output nodes correspond to the number of predicted classes.

The specifically training procedure can be summarized as follows,

Step 1: The raw vibration data are collected from the test rig, which are pre-processed by cyclic spectral analysis to obtain 2D CSCoh maps. The maps are further downsized to size 112×112 by balancing the computational cost and accuracy.

Step 2: The CNN is constructed by stacking multiply convolutional and pooling layers. Especially, Sigmoid activation function is adopted to predict the independently probability of each class. Accordingly, the BCE is adopted for measuring the distribution between the multi-label output and the target output.

Step 3: CNN network is updated by minimizing the BCE loss to improve the performance of the model in each epoch. The training procedure is the same as that of the traditional CNN.

Step 4: At the testing phase, the testing samples are fed into the trained CNN model to obtain the final diagnosis result.

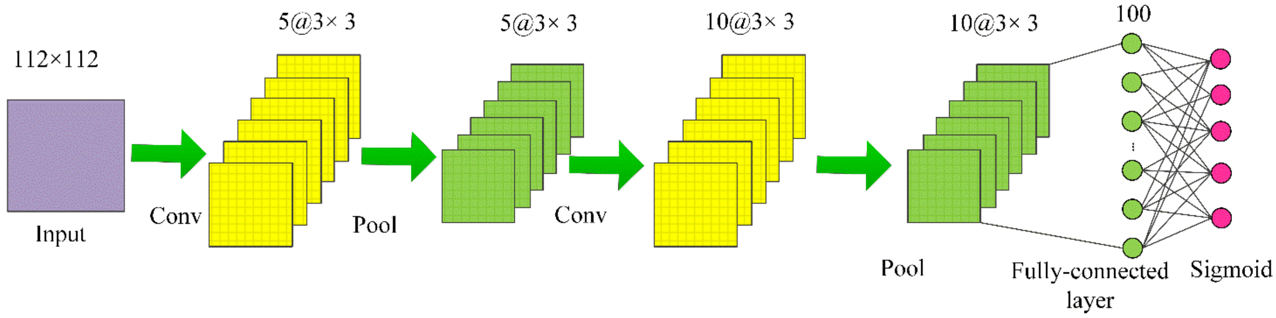


Figure 2: The fault diagnosis framework of the proposed method

4 Experiment verification

4.1 Experiment Setup

The experimental data of rolling element bearings have been acquired from the public bearing data center provided by Case Western Reserve University (CWRU), which is regarded as a benchmark dataset. The test rig is mainly composed of an induction motor, a transducer and a dynamometer. The vibration data are collected near the driving end of motor with a sampling frequency of 48 kHz. The motor bearings were seeded with faults using electro-discharge machining (EDM). In addition to the Normal Condition (NC), bearings with defect diameters of 7 mil, 14 mil and 21 mil have been introduced separately at the inner raceway, ball and out raceway. Each bearing is tested under four different loads (0, 1, 2 and 3 hp). Thus, the faults with two different diagnosis levels ('Level 1' and 'Level 2') can be detected. 'Level 1' means to discriminate the bearing with different fault patterns. While 'Level 2' denotes to further diagnosis the bearing with different severity levels, such as BF7 and BF17 cases, which is more challenge. The detailed description of data is listed in Table 1.

Level 1: Fault pattern	Level 2: Fault severity levels (mil)	Class encoding
Normal Condition (NC)	0	[1,0,0,0,0,0,0,0,0,0,0,0,0]
Ball Fault (BF)	7	[0,1,0,0,1,0,0,0,0,0,0,0,0]
	14	[0,1,0,0,0,1,0,0,0,0,0,0,0]
	21	[0,1,0,0,0,0,1,0,0,0,0,0,0]
Inner race Fault (IF)	7	[0,0,1,0,0,0,0,1,0,0,0,0,0]
	14	[0,0,1,0,0,0,0,0,1,0,0,0,0]
	21	[0,0,1,0,0,0,0,0,0,1,0,0,0]
Outer race Fault (OF)	7	[0,0,0,1,0,0,0,0,0,0,1,0,0]
	14	[0,0,0,1,0,0,0,0,0,0,0,1,0]
	21	[0,0,0,1,0,0,0,0,0,0,0,0,1]

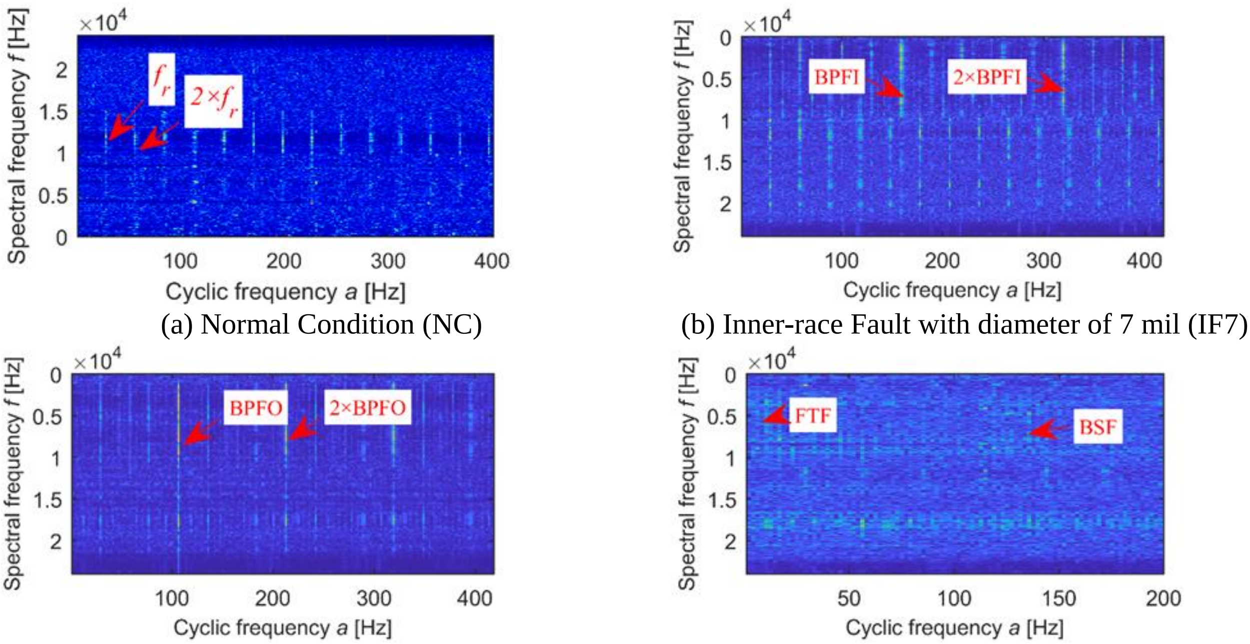
Table 1: Description of the bearing health conditions

4.2 Analysis of the CSCoh 2D maps

In order to obtain the 2D CSCoh maps from the vibration measurements, 24000 data points (time duration of 0.5 seconds) are considered from the time-series signals to form one sample. There are 20 samples obtained from each health condition under every working load. In addition, it should be noted that, due to the limited sampling time, there are only 14 samples obtained for IF14 under the load 1.

When defects of bearings occur, the bearing fundamental fault frequencies can be detected to analyse their dynamic behaviors. In order to validate the effectiveness of CSCoh in revealing the discriminative information of different fault types, four health conditions including the NC, IF with defect diameter of 7 (IF7), OF with defect diameter of 7 (OF7) and BF with defect diameter of 14 (BF14) are presented in figure 3.

It can be seen that the CSCoh maps provide unique representations for given fault types. In figure 3 (a), the fundamental shaft frequency (f_r) and its harmonic presents in the lower frequency are clearly observed, which is consistent with the dynamic behavior of the normal condition. In figure 3 (b) and figure 3 (c), Ball Pass Frequency of Inner-race (BPFI) and the Ball Pass Frequency of Outer-race (BPFO) and its harmonic can be clearly captured respectively, corresponding to the occurrence of the specific faults. It should be noted that in the case of BFs, presented in figure 4 (d), the weak amplitude of the Fundamental Train Frequency (FTF) and the Ball Spin Frequency (BSF) can be detected only in a few of samples of BF14, which reveal the existence of the ball fault. This demonstrates that the proposed is able to provide a good discriminative features when defects of bearings occurred.



(a) Normal Condition (NC)

(b) Inner-race Fault with diameter of 7 mil (IF7)

(c) Outer-race Fault with diameter of 7 mil (OF7)

(d) Ball Fault with diameter of 7 mil (BF14)

Figure 3: The 2D CSCoh maps of different health conditions

4.3 A Fault diagnosis using the proposed method

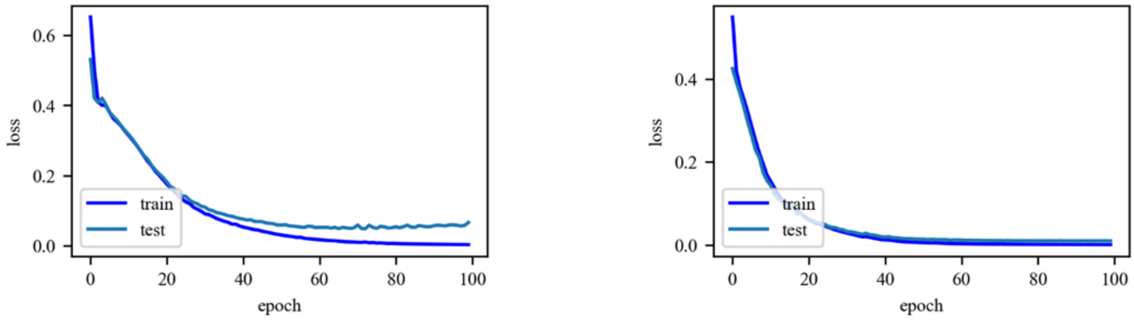
4.3.1 Effect of training sample ratio on classification performance

In order to study the influence of the training sample size on the classification rate, two datasets (dataset A and dataset B) are constructed to evaluate the performance. In dataset A, 20% of the samples are taken as the training data, while the rest for testing. In dataset B, 50% of samples are taken as training, and the rest for testing. Dataset A is constructed to simulate the insufficient training for the network. While dataset B is designed to sufficient training of network.

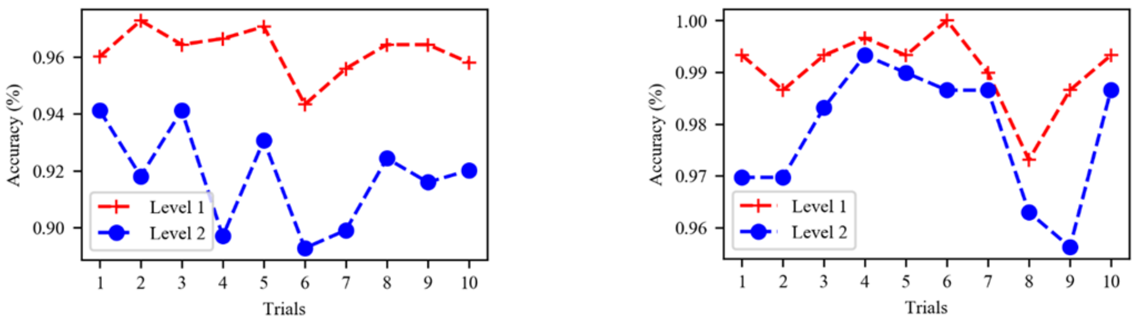
For CNN training, Adam algorithm is utilized to adjust the network weights with a batch size of 50. The epoch is set to 100, Ten trials are implemented to reduce the random. And the loss curves are shown in figure 4. From the figure 4, it can be seen that the training losses in both datasets are smooth, and remain stable, when they reach a certain number of iterations. It reveals that the models are well trained under the two training data. In the test stage, the loss curve in dataset A decreases slowly than that of the train stage, and it is close to a fixed value, and keep stable. It is possible that CNN with a large number of parameters trained on the insufficient training data suffers from the overfitting problem. By adding more training samples, the loss curve as shown in figure 4(b), are obviously decreasing.

In addition, the diagnosis results of ten trials are presented in figure 5. “Level 1” denotes the classification accuracy of the fault patterns, where NC, BF, IF and BF are correctly recognized. “Level 2” reflects the total recognition accuracy, where the fault patterns and fault severity levels are all correctly classified.

From the results, it can be observed that, the results of all ten trials present relative high accuracy in both datasets. In addition, ‘Level 1’ is much higher than ‘Level 2’, since the former only needs to diagnosis the specific fault patterns, while the later requires to discriminative the fault severity levels of each fault patterns. Moreover, CNN with dataset A is much lower than that of dataset B, especially in ‘Level 2’. This is because that ‘Level 2’ contains more discriminative classes, which is more difficult for diagnosis. Therefore, the results of dataset B is able to obtain better classification performance compared to that of dataset A, since more of the training samples are contained.



(a) Loss curve of CNN using dataset A (b) Loss curve of CNN using dataset B
Figure 4: The loss curve of CNN with different datasets



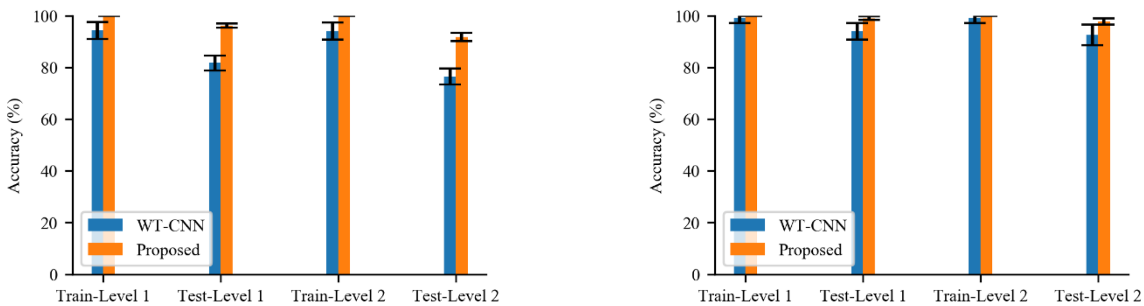
(a) Results of 10 trials using dataset A (b) Results of 10 trials using dataset B
Figure 5: The results of ten trials of different datasets

4.3.2 Comparison with other methods

Wavelet Transform (WT), which decomposes the signals into wavelets, is usually considered as an effective tool for pre-processing nonstationary and transient signals [23]. The two-dimensional time frequency representation of WT provides a high resolution in both the time-domain and frequency-domain, which provides good information about the health conditions of rotating machinery.

In this section, a comparison of the WT scalograms and the CSCoh maps is carried out. Morelet wavelet basis is adopted to implement the time-frequency transformation. In order to provide a relative fair comparison, all the pre-processing data are fed into CNN for training, and the results are conducted with ten trials. Final results are averaged. The training and testing accuracy of different methods are shown in figure 6.

It can be seen that, the training accuracies of the proposed method in both datasets are 100.0%. While the training accuracies of WT are relative lower, and present larger standard deviations. In addition, the testing accuracy, especially, in ‘Level 2’, the proposed method also obviously performs better than that of WT-CNN.



(a) Result comparison of different methods using dataset A (b) Result comparison of different methods using dataset B
Figure 6: Result comparison of different methods

Furthermore, the testing accuracy of each types under different severity levels are further listed in Table 2 and Table 3, respectively. It can be observed that WT-CNN achieves average testing accuracy of 76.7% and 92.8%, in dataset A and dataset B, respectively. On the contrast, the proposed method achieves average testing accuracy of 92.2% and 97.9%, respectively, which is better that that of the WT-CNN. In addition, the diagnosis accuracies of the BF7 and the BF14 are obviously lower that of the other fault types. This is because that the characteristic frequencies in those cases are not obvious in the 2D CSCoh maps, which makes CNN difficulty to obtain good classification performance. From the analysis, it can be concluded that the proposed method is effective in extracting discriminative features and conducting the multi-label classification tasks.

Methods	Accuracy (%) of each fault severity level using dataset A									
	NC	BF7	BF14	BF21	IF7	IF14	IF21	IF7	IF14	IF21	AVG
WT-CNN	100	52.7	71.7	64.2	100.0	59.8	99.5	93.4	39.5	86.5	76.7
Proposed	100	100.0	79.0	58.4	100	96.1	95.9	100.0	96.4	96.5	92.2

Table 2: The average testing accuracy of each fault severity level using dataset A

Methods	Accuracy (%) of each fault severity level using dataset B										
	NC	BF7	BF14	BF21	IF7	IF14	IF21	IF7	IF14	IF21	AVG
WT-CNN	100.0	95.7	85.0	82.3	100.0	92.0	100	95.4	79.5	98.0	92.8
Proposed	100.0	100.0	95.7	89.6	100.0	96.5	98.3	100.0	99.0	99.4	97.9

Table 3: The average testing accuracy of each fault severity level using dataset B

Conclusion

In this work, a new DL-based fault diagnosis framework, combining CSCoh and CNN is proposed for multi-label fault classification. Firstly, CSCoh is considered, as a pre-processing step, to reveal the fault nature of each fault types. Then, a novel CNN is constructed for conducting fault classification with multiply labels by introducing the Sigmoid activation function and BCE loss function. The proposed method is verified on the data collected from the CWRU motor bearing test rig. Two different datasets including the insufficient training and sufficient training data are designed to evaluate the effectiveness of the methodology. It has been demonstrated that the proposed method not only achieves high classification performance, but also presents better generalization performance compared to WT-CNN fault diagnosis method.

References

[1] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, R.X. Gao, *Deep learning and its applications to machine health monitoring*, Mech. Syst. Signal Process. 115 (2019) 213-237.

[2] S. Khan, T. Yairi, *A review on the application of deep learning in system health management*, Mech. Syst. Signal Process. 107 (2018) 241-265.

[3] F. Jia, Y. Lei, J. Lin, X. Zhou, N. Lu, *Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data*, Mech. Syst. Signal Process. 72-73 (2016) 303-315.

[4] D. Verstraete, A. Ferrada, E. L. Droguett, V. Meruane, M. Modarres, *Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings*. Shock and Vibration, vol. 2017, Oct. 2017.

[5] B. Zhang, S. Zhang, W. Li, *Bearing performance degradation assessment using long short-term memory recurrent network*, Computers in Industry. 106 (2019) 14-29.

[6] S. Wang, J. Xiang, Y. Zhong, Y. Zhou, *Convolutional neural network-based hidden Markov models for rolling element bearing fault identification*, Knowledge-Based Systems, 144 (2018) 65-76.

[7] Z. Chen, W. Li, *Multisensor Feature Fusion for Bearing Fault Diagnosis Using Sparse Autoencoder and Deep Belief Network*, Ieee T Instrum Meas. 66 (2017) 1693-1702

[8] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, S. Van Hoecke, *Convolutional Neural Network Based Fault Detection for Rotating Machinery*, J. Sound Vibrat. 377 (2016) 331-345.

[9] C. Sun, M. Ma, Z. Zhao, X. Chen, *Sparse Deep Stacking Network for Fault Diagnosis of Motor*, IEEE Transactions on Industrial Informatics. 14 (2018) 3261-3270.

[10] P. Chopra and S. K. Yadav, *Fault detection and classification by unsupervised feature extraction and dimensionality reduction*, Complex Intell. Syst., vol. 1, no. 1, pp. 25–33, Jan. 2015.

[11] F. Althobiani and A. Ball, *An approach to fault diagnosis of reciprocating compressor valves using Teager–Kaiser energy operator and deep belief networks*, Expert Syst. Appl., vol. 41, no. 9, pp. 4113–4122, Dec. 2013.

[12] [P. Tamilselvan and P.Wang, *Failure diagnosis using deep belief learning based health state classification*, Rel. Eng. Syst. Safety, vol. 115, pp. 124–135, Jul. 2013.

[13] Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, *Real-time 751 motor fault detection by 1-d convolutional neural networks*, IEEE Trans. 752 Ind. Electron., vol. 63, no. 11, pp. 7067–7075, Nov. 2016

[14] F. Jia, Y. Lei, N. Lu, S. Xing, *Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization*, Mech. Syst. Signal Process. 110 (2018) 349-367.

[15] M. Gan, C. Wang, C.a. Zhu, *Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings*, Mech. Syst. Signal Process. 72-73 (2016) 92-104.

[16] L. Wen, X. Li, L. Gao, Y. Zhang, *A New Convolutional Neural Network-Based Data-Driven Fault Diagnosis Method*, IEEE Trans. Industr. Electron. 65 (2018) 5990-5998.

[17] P. Borghesani, J. Antoni, *A faster algorithm for the calculation of the fast spectral correlation*, Mech. Syst. Signal Process. 111 (2018) 113-118.

[18] A. Mauricio, J., Qi, K. Gryllias, 2019. *Vibration-Based Condition Monitoring of Wind Turbine Gearboxes Based on Cyclostationary Analysis*. Journal of Engineering for Gas Turbines and Power, 141(3), p.031026.

[19] K. Gryllias, A. Mauricio, J. Qi, 2018, September. *Advanced cyclostationary-based analysis for condition monitoring of complex systems*. In 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 385-389). IEEE.

[20] A. Mauricio, J. Qi, W. Smith, R. Randall, K. Gryllias, *Vibration Based Condition Monitoring of Planetary Gearboxes Operating Under Speed Varying Operating Conditions Based on Cyclo-non-stationary Analysis*, 61 (2019) 265-279.

[21] Y. LeCun, G. Bengio, Hinton, *Deep learning*, Nature, 521 (2015) 436-444.

[22] Y. LeCun, L. Bottou, Y. Bengio, Y, P. Haffner, P. (1998). *Gradient-based learning applied to document recognition*. Proceedings of the IEEE, 86(11), 2278-2324.

[23] Yan, R.X. Gao, X. Chen, *Wavelets for fault diagnosis of rotary machines: A review with applications*, Signal Processing, 96 (2014) 1-15.

A Semi-supervised Support Vector Data Description- based Fault Detection Method for Rolling Element Bearings based on Cyclic Spectral Coherence

Chenyu LIU [1,2], Konstantinos GRYLLIAS [1,2]

[1]Department of Mechanical Engineering,
Faculty of Engineering Science, KU Leuven
[2]Dynamics of Mechanical and Mechatronic Systems, Flanders Make
Celestijnenlaan 300, BOX 2420, 3001 Leuven, Belgium
chenyu.liu@kuleuven.be, konstantinos.gryllias@kuleuven.be

Abstract

Modern Internet-of-Things (IoT)-driven condition monitoring exploits data from various surveillance tools to reflect the health status of machinery. However, many diagnostic tasks are still hardly achievable, considering the complex operating environment of mechanical components. Detecting rolling element bearing faults, for timely repair and maintenance before a total breakdown, is one of the challenges which has a significant impact on rotating machinery. Nowadays advanced signal processing techniques are combined with high level machine learning approaches, focusing towards automatic fault diagnosis. A plethora of diagnostic indicators have been proposed to track the system degradation. Cyclic Spectral Correlation (CSC) and Cyclic Spectral Coherence (CSCoh) have been proved as powerful tools in signal processing community. Due to the periodic mechanism of the bearing faults' impacts, the diagnostic indicators extracted from CSCoh domain can better detect bearing faults. On the other hand, supervised machine learning approaches with labeled training and testing datasets cannot be realistically obtained under industrial conditions. In order to overcome this limitation, a novel fault detection approach based on semi-supervised learning and Support Vector Data Description (SVDD) is proposed in this paper. The proposed SVDD model utilizes CSCoh domain indicators to build the feature space and fits a hyper-sphere to calculate the Euclidean distances in order to isolate the healthy and faulty data. Meanwhile a systematic fault detection decision strategy is proposed to estimate the bearing status simultaneously with the detection of fault initiation. Two run-to-failure bearing datasets are used to implement the proposed fault detection framework and the results show that the SVDD model with CSCoh indicators can precisely detect the bearing faults. Moreover, the CSCoh based indicators show superior characteristics in the detection process comparing to time and Fast Kurtogram (FK) based Squared Envelope Spectrum (SES) domain indicators.

1 Introduction

Rolling element bearings are important but fragile components and are highly sensitive to external operation environment. Premature bearing failure caused by mounting errors, poor lubrication, debris contamination etc. may trigger a breakdown of machinery or even a fatal accident [1]. Therefore, the early bearing fault detection arises as a critical mission in the frame of predictive maintenance, receiving extensive attention in recent years.

Data-driven based fault detection methods have been widely applied to monitor rolling element bearings in rotation machinery. Its framework consists of four major steps: 1) data acquisition, 2) construction of Health Indicators (HI), 3) determination of decision threshold, 4) detection of anomaly [2]. Vibration signals are widely used in the data acquisition step to detect machine faults. HIs, derived from different domains, are used to track the bearing fault progression. Time domain HIs, such as root mean square (RMS), kurtosis, skewness etc., are widely employed [3]. They represent the statistical features of the time sequence but cannot characterize nonlinearities of bearing vibrations. On the other hand, a plethora of signal processing methods have been developed to enhance the embedded defect information, including Envelope Analysis, Spectral Kurtosis, Short-Time Fourier Transform, Cyclic Spectral Analysis etc. [4]. These approaches allow the extraction of sophisticated HIs which correlate with bearing fault characteristic frequencies with more expressiveness and robustness in practical applications.

Among the above mentioned signal processing methods, cyclic spectral analysis recently gained much attention due to its capability of revealing the second-order cyclostationary periodicities of rolling element bearing signals. It has been proved that the vibration signals of faulty bearings, especially with localized fault, will exhibit cyclostationary behaviour [6]. Cyclic Spectral Correlation (CSC) and Cyclic Spectral Coherence (CSCoh) have been developed as efficient tools in cyclostationary analysis [7]. They represent the potential fault modulation information into frequency-frequency domain bivariable maps. The integration over the spectral frequency, leads to the estimation of the Enhanced Envelope Spectrum (EES), which demonstrates the modulation frequencies and their harmonics. Due to the periodic mechanism of the bearing faults' impacts, the EES can provide a clearer representation of bearing faults compared to other methods such as the Squared Envelope Spectrum (SES) [8]. Some attempts have been proposed to construct HIs from EES. In [9], a self-running bearing diagnosis framework with scalar indicators derived from EES has been set up . Moreover the summation of the amplitudes of three harmonics, which appear at the characteristic fault frequencies of EES, has been used initially as diagnostic indicators [10], and further as bearing prognostic indicators.

Another key problem is the definition of a threshold, which can trigger the detection based on the HIs, estimated on signals emitted by faulty bearings. On one hand, traditional statistic decision threshold strongly depends on the expert knowledge of the indicators' statistical distribution as well as on the empirical understanding of bearing faults. On the other hand, the detection threshold should be able to precisely isolate the faults meanwhile tolerating outliers to maintain a robust performance of the false alarm and misdetection rate under different external environments. Nowadays machine learning approaches show great potential to achieve this comprehensive target. In the research of automatic fault diagnosis, machine learning algorithms are considered as black-box models which can exhibit system behaviours from data. Therefore, it can provide continuous information for the rapid deterioration of bearings and detect anomalies. Support vector machine (SVM) is one of the initial machine learning attempts in real-world fault detection application [11]. The major principle of SVM is to seek an optimal hyper-plane as the decision surface that can maximize the margin of separation between two classes. In the case of bearing fault detection, the hyper-plane of SVM has been used as the detection threshold to separate indicators extracted from bearing signals into healthy and faulty class [12, 13]. Artificial neural network (ANN) is also widely applied to fulfill the task of fault detection. A supervised non-linear autoregressive neural network with exogenous input, was trained in [14], to model the healthy condition of bearings meanwhile setting up the detection threshold using the Mahalanobis distance to detect the bearing faults.

In real industrial conditions, the bearings are operating for most of the time in healthy conditions and the faulty datasets are rather limited and usually not labelled. Therefore standard supervised machine learning approaches, which are based on known labels of the classes, cannot be used for fault detection. In order to overcome this limitation, semi-supervised fault detection techniques have been proposed, based on the training of the methodology exclusively on healthy data. Support Vector Data Description (SVDD) has been proposed by Tax and Duin [15] and has been successfully used in one-class classification problems for various applications. SVDD is inspired by SVM but uses a hyper-sphere, instead of a hyper-plane, to solve the non-linear separable problem. SVDD provides a geometric description of the observed data by calculating the Euclidean distances between the data and the center of the hyper-sphere. It is trained by minimizing the volume of the sphere with most training data. When testing with the trained SVDD, any data found to have larger distance than the boundary of the sphere, are considered to be anomalies. In the bearing fault detection, the hyper-sphere of the SVDD can be used as the detection threshold.

In this paper, the CSCoh based diagnostic indicators and the SVDD classifier are combined in order to detect bearing faults. A comprehensive bearing status decision strategy is proposed within this framework. The methodology is tested and evaluated on experimental data from run-to-failure bearing datasets. The results demonstrate the efficacy of the method, presenting high detection rate with low false alarm and misdetection rate. The rest of the paper is organised as following: the background of Cyclic Spectral Coherence is discussed in Section 2 and the theory of SVDD is introduced in Section 3. The proposed methodology and the experimental setup are illustrated in Section 4 and 5, respectively. The results are presented and analysed in Section 6. Finally, the paper closes with some conclusions.

2 Cyclic Spectral Coherence

The concept of cyclostationarity was first proposed in the field of telecommunications and later was introduced to the mechanical engineering community. As a subcategory of non-stationary processes, cyclostationarity is a stochastic process that exhibits hidden periodicities embedded in systems. When a fault is generated in rolling element bearings in different locations (inner race, outer race or rolling element), a series of shocks in repetition form will be generated simultaneously, modulated by other frequencies like the shaft speed. This phenomenon can be described as cyclostationary and can be exploited to detect bearing damage. Different orders of cyclostationarity are defined based on the order of moments. The first-order cyclostationarity (CS1) is described as:

$$R_{1x}(t) = R_{1x}(t+T) = \mathbb{E}\{x(t)\} \tag{1}$$

where the first-order moment, called the statistical mean, $R_{1x}(t)$ is periodic with period T. $x(t)$ represents the signal function at time t. $\mathbb{E}\{x(t)\}$ is the ensemble average, which represents the average of the same stochastic process of repeated experiments. CS1 exhibits as periodic waveforms in vibration signals of rotating machinery and can be generated by imbalances, misalignments or flexible coupling. Moreover, second-order cyclostationary (CS2) depicts the periodicity of second-order moments with the autocorrelation function:

$$R_{2x}(t,\tau) = R_{2x}(t+T,\tau) = \mathbb{E}\{x(t)x(t-\tau)^*\} \tag{2}$$

where τ is the time-lag. CS2 provides a distinction of the stochastic process with amplitude or frequency modulation. CS2 has been proved extremely effective to achieve diagnosis on rotating components that are not completely phase-locked with shaft speeds, such as rolling element bearings.

The Cyclic Spectral Correlation (CSC) is designed as a tool to describe CS1 and CS2 signals in frequency-frequency domain. It is defined as the autocorrelation function of two frequency variables using the two-dimensional Fourier transform, as shown in Eq. 3:

$$CSC_x(\alpha, f) = \lim_{W\to\infty} \frac{1}{W}\mathbb{E}\{\mathscr{F}_W[x(t)]\mathscr{F}_W[x(t+\tau)]^*\} \tag{3}$$

where α is the cyclic frequency related with the modulation and f is the spectral frequency representing the carrier. $\mathscr{F}$ stands for the Fourier transform and W is the time duration. Furthermore, a normalization procedure can be added to the CSC, to minimize uneven distributions, which is known as Cyclic Spectral Coherence (CSCoh). The CSCoh is estimated as:

$$CSCoh_x(\alpha, f) = \frac{CSC_x(\alpha, f)}{\sqrt{CSC_x(0,f)CSC_x(0,f-\alpha)}} \tag{4}$$

The integration of the CSCoh over the spectral frequency f from zero to the Nyquist frequency can lead to the Enhanced Envelope Spectrum (EES) which is an improved version of the Squared Envelope Spectrum, as described in Eq. 5:

$$IES(\alpha) = \frac{1}{F_2 - F_1}\int_{F_1}^{F_2} |CSCoh_x(\alpha, f)|df \tag{5}$$

3 Semi-supervised SVDD

3.1 Theory

Semi-supervised learning approaches focus on partially labelled datasets. The objective of a semi-supervised model is to classify some of the unlabelled data, leveraging information from the labelled part. SVDD has been developed as one-class classification method [15] in the way of semi-supervised learning. It can be seen as an extension of SVM, which uses a hyper-sphere instead of a hyper-plane as the classification decision surface. The hyper-sphere is characterized by the radius R and the center a, which are the decision variables. The labelled data "targets" are regarded as the training set of SVDD. The data, being exactly on the boundary, are the support vectors and the data outside are "outliers". The primary goal of SVDD is to construct a hyper-sphere with a minimum radius, which simultaneously contains the maximum number of targets. The objective function can be described by Eq. 6:

$$F(R,a) = \min R^2 + C\sum_{i=1}^{n} \xi_i \tag{6}$$

This function subjects to two constraint conditions: $\|x_i - a\|^2 \leq R^2 + \xi_i, \forall i = 1,\dots,n$ and $\xi_i \geq 0, \forall i = 1,\dots,n$. Here $x_i \in \mathbb{R}$ is the training dataset with n samples. ξ_i is the slack variable which represents the tolerance of targets being outside the boundary. The slack variable is used to avoid an extremely large radius hyper-sphere that reduces the descriptive ability. C is the penalty factor that controls the trade-off between the radius and the rejected number of data points. It can also be described as $C = \frac{1}{nf}$, where f is the outlier fraction. Eq. 6 is a convex quadratic programming problem which cannot be solved for unknown R. However, it can be transformed to an equivalent dual problem with Lagrange duality. By applying Lagrange multipliers, the constraint conditions can be fused into the objective function, which turns it to a dual form as Eq. 7:

$$L = \max \sum_{i=1}^{n} \alpha_i (x_i \cdot x_i) - \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) \tag{7}$$

where $(x_j \cdot x_i)$ stands for the inner product of x_i and x_j, $\alpha_i \in \mathbb{R}$ are the Lagrange multipliers. New conditions are set for the dual form: $\sum_{i=1}^{n} \alpha_i = 1$ and $0 \leq \alpha_i \leq C, \forall i = 1,\dots,n$. The dual form is composed of the data itself which makes the problem solvable. The training data x_i and its corresponding α_i are related with the radius R and the center a of the hyper-sphere according to Eq. 8:

$$\begin{cases} x_i - a^2 < R^2 \iff \alpha_i = 0 \\ x_i - a^2 = R^2 \iff 0 \leq \alpha_i \leq C \\ x_i - a^2 > R^2 \iff \alpha_i = C \end{cases} \tag{8}$$

3.2 Flexible data description

The inner product $(x_i \cdot x_j)$ in Eq. 7 can be replaced with kernel functions $K(x_i, x_j)$ to make a more flexible description for non-linear datasets. The kernel function can map the data to a higher dimension feature space which makes the non-linear data separable. The kernel functions have various choices such as Gaussian, linear, polynomial etc. Among all the functions, Gaussian kernel is the most common choice and is also adopted in this paper as described in Eq. 9:

$$K(x_i, x_j) = \exp\frac{-\|x_i - x_j\|^2}{2s^2}; \tag{9}$$

where s represent the kernel width parameter. By using the kernel function, the distance of any observation z to the center of the sphere can be described as follows:

$$\text{dist}^2(z) = K(z,z) - 2\sum_i \alpha_i K(x_i, z) + \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \tag{10}$$

Considering the previous spherical data boundary conditions in Eq. 8, $\text{dist}^2(z) < R^2$ represents the position inside the sphere and $\text{dist}^2(z) = R^2$ is on the boundary of the sphere. When $\text{dist}^2(z) > R^2$, then z is recognized as an outlier. It should be noticed that the width of the kernel function K will influence on the calculation of the distance and is optimized in the proposed methodology, as explained in the following section.

4 Proposed methodology

4.1 Health Indicator construction

In order to realise the fault detection, HIs need to be constructed to evaluate the current condition as well as the degradation level of the bearing. The goodness of HIs can be characterized by monotonicity, trendability and prognosability. In real applications, the HIs should be able to detect bearing faults in early stage meanwhile minimizing the false alarm and misdetection rate. Time domain statistic features, such as Root Mean Square (RMS), kurtosis, skewness etc. are the most used HIs since they are effective to capture instant faults, such as using kurtosis in the detection of bearing spall initiation. However, the effectiveness of time domain HIs is restricted when the faults are distributed across the bearing surface. During the developing of damage, the vibration signals exhibit more cyclostationarity with masked fault signatures. To overcome this limitation, the HIs, in this paper, are constructed using EES features from the CSCoh domain, as shown in Tab. 1. The CSCoh domain HIs are defined as the amplitude of the harmonics of the bearing characteristic frequencies extracted from the EES, i.e. the Shaft Rotation speed (SR), the Ball Pass Frequency Outer race (BPFO), the Ball Pass Frequency Inner race (BPFI) with sidebands and the Ball Spin Frequency (BSF) with sidebands. Four (4) harmonics and their sum are chosen to keep the robustness of the HIs and to facilitate the early start of detection.

S1R	EES peak at SR	S2I	EES peak at 2nd harmonic BPFI and sidebands
S2R	EES peak at 2nd harmonic of SR	S3I	EES peak at 3rd harmonic BPFI and sidebands
S3R	EES peak at 3rd harmonic of SR	S4I	EES peak at 4th harmonic BPFI and sidebands
S4R	EES peak at 4th harmonic of SR	SIR	Sum of 4 EES peaks of BPFI harmonics and sidebands
SSR	Sum of 4 EES peaks of SR harmonics	SIOR	Sum of 4 EES peaks of BPFO and BPFI harmonics
S1O	EES peak at BPFO	S2B	EES peak at two times BSF and sidebands at FTF
S2O	EES peak at 2nd harmonic of BPFO	S4B	EES peak at four times BSF and sidebands at FTF
S3O	EES peak at 3rd harmonic of BPFO	S6B	EES peak at six times BSF and sidebands at FTF
S4O	EES peak at 4th harmonic of BPFO	S8B	EES peak at eight times BSF and sidebands at FTF
SOR	Sum of 4 EES peaks of BPFO harmonics	SBF	Sum of 4 EES peaks of BSF and sidebands
S1I	EES peak at BPFI and sidebands		

Table 1: Health indicators from CSCoh domain

To compare the classification performance by using different health indicators, 9 statistic features from the time domain are also extracted during the test, as shown in Tab.2. Meanwhile, the same 21 frequency domain features from the Fast Kurtogram (FK) based SES are also extracted as another comparison group. The results from these indicators are illustrated in the following section.

RMS	$RMS = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{n}}$	Variance	$VR = \frac{\sum_{i=1}^{n}(x_i-m)^2}{(n-1)\sigma^2}$	Shape Factor	$SF = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i{}^2}}{\frac{1}{n}\sum_{i=1}^{n}\lvert x_i\rvert}$
Kurtosis	$KU = \frac{\sum_{i=1}^{n}(x_i-m)^4}{(n-1)\sigma^4}$	Mean	$m = \frac{\sum_{i=1}^{n} x_i}{n}$	Impulse Factor	$IF = \frac{\max\lvert x_i\rvert}{\frac{1}{n}\sum_{i=1}^{n}\lvert x_i\rvert}$
Skewness	$SK = \frac{\sum_{i=1}^{n}(x_i-m)^3}{(n-1)\sigma^3}$	Crest Factor	$CF = \frac{\max\lvert x_i\rvert}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i{}^2}}$	Clearance Factor	$CLF = \frac{\max\lvert x_i\rvert}{(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\lvert x_i\rvert})^2}$

Table 2: Health indicators from time domain

4.2 Kernel width optimization using cross validation with grid search

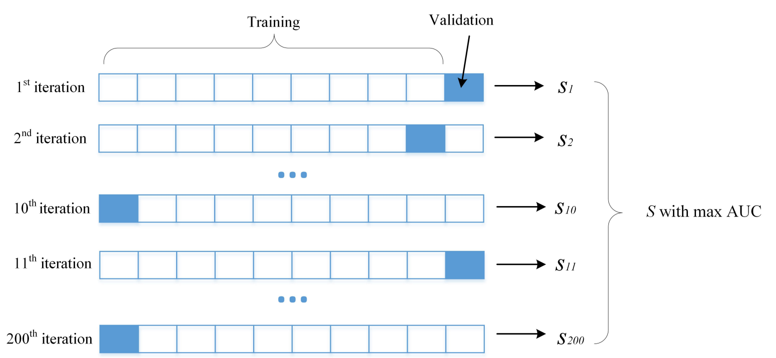


Figure 1: Optimization of kernel width using 10-fold Cross validation with grid search

The boundary of the hyper-sphere can be described with a kernel function for the SVDD classifier. In this paper, the Gaussian kernel is used since it has been proved effective in former research [15]. On the other hand, the kernel width acts as a hyperparameter in the SVDD model and influences the classification performance. The Receiver Operating Characteristic (ROC) curve is commonly used to evaluate the performance of a binary classifier. The corresponding area under the ROC curve (AUC) is the key metric. The classifier with a larger AUC value performs better than the others. In order to tune the kernel width meanwhile proceeding cross validation during the training process, cross validation with grid search is adopted to search for the kernel width, that can maximize the AUC of the classifier. With 10-fold cross validation method, 200 iterations are proceeded to search for the kernel width s, which gives the maximum AUC values as depicted in Fig. 1. In this way, the model could achieve an automatic parameter optimization to avoid a stochastic selection of the hyperparameter. The results from the ROC curve will be discussed in Section 6.

4.3 Bearing fault detection procedure

The practical need for the bearing fault detection approach is a systematic decision strategy as a complement of the SVDD classifier based on CSCoh indicators. The flowchart of the proposed bearing fault detection process is shown in Fig. 2. The entire procedure consists of three major parts: feature extraction, classification and fault detection. The CSCoh based HIs are extracted from the EES to construct the feature space in the feature extraction stage. The features are then normalized by subtracting the mean value and dividing by the standard deviation to avoid the calculated Euclidean distances be governed by particular features in the SVDD classifier. Semi-supervised methods are able to learn from fewer labelled data points with the help of a large number of unlabelled data points, therefore the training data are selected from the health conditions which normally can be gathered at the beginning period of the bearing operation. In the classification stage, the entire dataset is split into training and testing sets with 25% and 75% of the whole dataset, respectively. The training set includes the validation part to realise the parameter optimization and meanwhile is labelled as healthy. A banana dataset is generated as artificial outliers around these real data to train the hyper-sphere as the classifier.

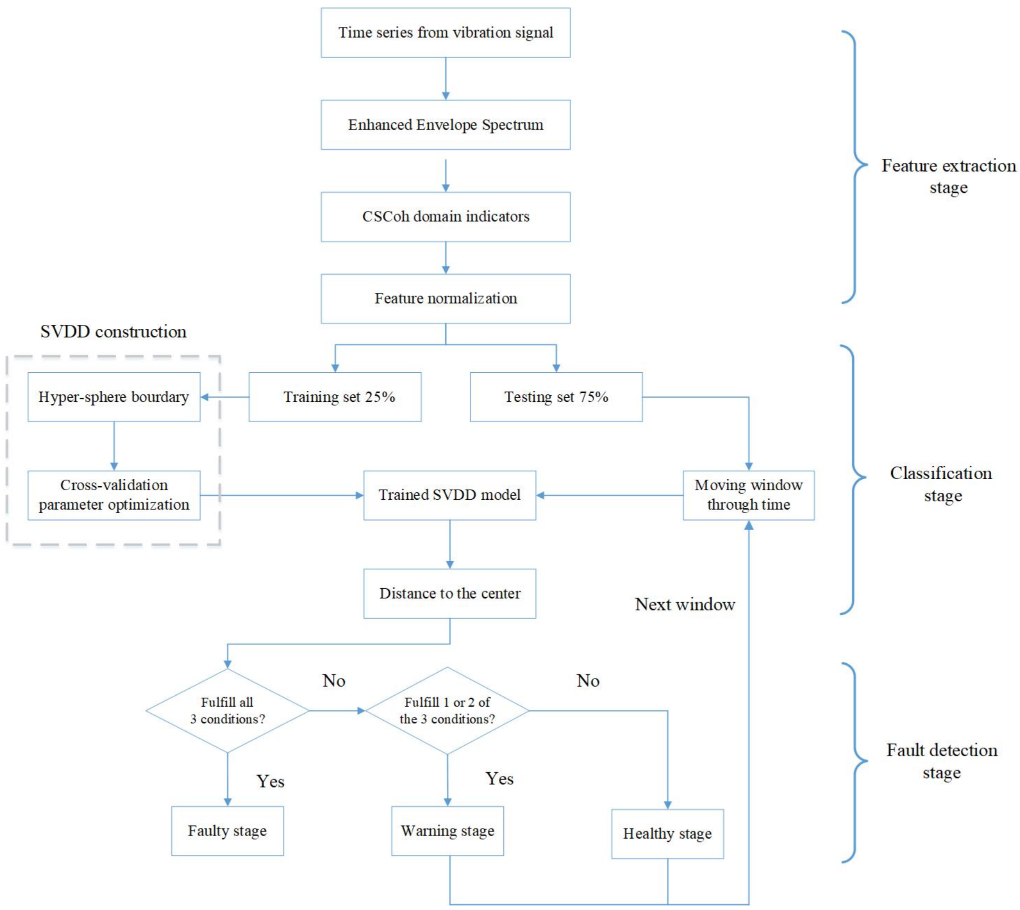
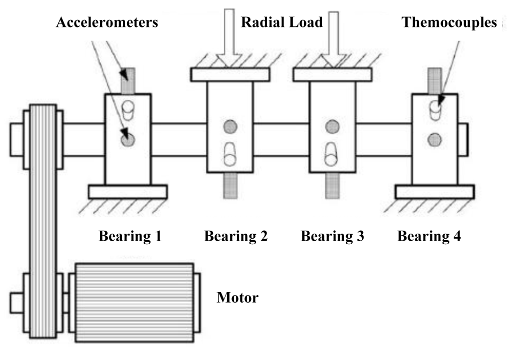


Figure 2: Procedure for the SVDD-based fault detection

Once the SVDD classifier is trained, a moving window is applied on the testing set. The data in the window are sent to the SVDD model to calculate the Euclidean distances to the center of the hyper-sphere. Considering the distances and the threshold, i.e. the boundary of the hyper-sphere, a fault detection decision strategy is built following 3 conditions: 1) 50% or more of the distances in the window are above the threshold. 2) 50% or more of the distances in the window continuously pass the threshold. 3) The average distance of the data in the window is equal or greater than the threshold. These conditions can keep the detection threshold robust from the influence of random dominant outliers meanwhile reducing the misdetection rate. Three bearing statuses are then defined, according to the conditions. When all the three conditions are fulfilled, the status of the bearing is considered as faulty. In contrast, if less than 3 conditions are satisfied, then the status is considered as warning. If the distances from the bearing data obey none of the three conditions, it is considered as in a healthy status. Additionally, the starting point of the warning status represents the existence of a premature anomaly and the starting point of faulty status is the detection point of a mature bearing defect.

5 Experimental setup and dataset

The proposed SVDD based semi-supervised methodology is tested and validated using the NASA Intelligent Maintenance Systems (IMS) dataset [16]. The IMS dataset was collected during the rolling element bearing endurance experiment by using a dedicated test rig in the University of Cincinnati. The layout of the test rig is shown in Fig. 3, which consists of an electric motor coupled with a rub belt with 2000 rpm stationary speed, 4 double row Rexnord ZA-2115 bearings mounted on a common shaft lubricated by a circulation system, a 6000 lbs radial load applied on the bearing 2 and 3, a shaft and PCB 353B33 High Sensitivity Quart ICP accelerometers. The collecting of data was stopped when the accumulated debris exceeded a certain level on a magnetic plug and was considered as the end of bearing life. The physical parameters and the characteristic frequencies can be found in Fig. 3.



Physical Parameters	
Pitch diameter	71.5 mm
Rolling element diameter	8.4 mm
Number of rolling element per row	16
Contact angle	15.7
Static load	26690 N
Characteristic Frequencies	
Shaft rotation speed	33 Hz
Ball pass frequency outer race (BPFO)	236 Hz
Ball pass frequency inner race (BPFI)	297 Hz
Ball spin frequency (BSF)	139 Hz
Fundamental train frequency (FTF)	15 Hz

Figure 3: IMS test rig layout (left) and characteristics of bearings (right)

The run-to-failure experimental results were published in three datasets with numerous of files. Some details should be noticed when processing these datasets. According to the document describing the experiment, the sampling frequency should be 20 kHz but the file for one second includes 20,480 points so the sampling frequency probably was 20.48 kHz and this one is used in this paper. Moreover, the measurement for Dataset 1 is not continuous but with several interruptions during recording. These interruptions influences the behaviour of some HIs and this issue will be discussed in the following section. The signals from the dataset will be marked with time (Day) and number of signal (# signal number).

6 Results and discussion

6.1 Case 1: Dataset 1 Bearing 3

The 9 time domain HIs, the 21 FK based SES domain HIs and the 21 CSCoh domain HIs estimated processing the Dataset 1 Bearing 3, after normalization, are presented in Fig. 4. It is noticeable that the time axis is not in equal intervals for this map and other figures involving Dataset 1 in this section, due to the recording interruptions. The time axis represents the real time points in days from beginning (Day 0) to the end of the experiment (Day 34.5) but not the recording time. Analysing the time and the FK based SES features in the map, it can be visually observed that there is a significant changing step before Day 7.1 (#157), caused by a long period interruption. Therefore the data before Day 7.1 are abandoned to avoid contamination of the

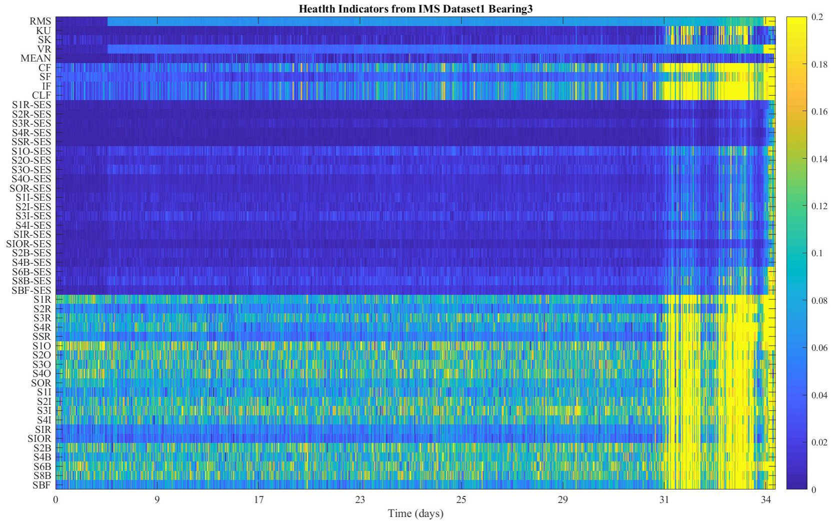


Figure 4: Health indicator map for Dataset 1 Bearing 3

healthy training set, which is selected from Day 7.4 (#200) to Day 17 (#600). The remaining data are used for testing.

The distances of the CSCoh indicators to the SVDD center are shown on the left of Fig. 5. Based on the proposed decision framework, the fault detection result is shown on the right of Fig. 5. The warning stage starts from Day 29.2 (#1570) till Day 30.9 (#1795) and the fault is detected also at Day 30.9 (#1796).

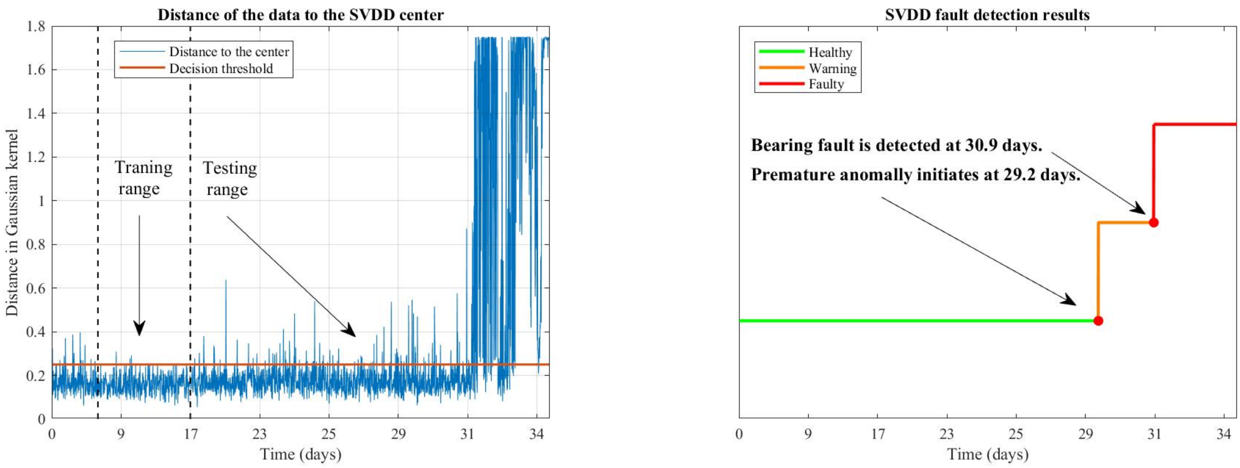


Figure 5: Fault detection restuls (left) and testing ROC curve (right)
of CSCoh indicators of Dataset 1 Bearing 3

The evaluation of the detection performance uses results from the study [17] as reference. The testing ROC curve of the SVDD classifier is plotted in Fig. 6. As discussed in Section 4, the ROC curve is employed to evaluate the classification ability of a binary classifier. The horizontal and vertical axis of a ROC curve represent the False Positive rate (FPr) and the True Positive rate (TPr). The classifier behaves better when the curve is closer to the left up point where FPr=0, TPr=1. Additionally, the area under ROC curve (AUC) can be used as a quantitative index in the evaluation. A larger AUC represents a better classification performance. Furthermore, more indexes are used to evaluate the classifier in the later part.

The HIs from the time and the FK based SES domain are also used to train the SVDD model as shown in Fig. 7 and 8. Compared to the results of CSCoh, it is obvious that the distances of the time and the FK based SES domain indicators exhibit wide fluctuations around the threshold. The outliers above the threshold spread in the whole testing range for both the two HI groups. Hence, the fault detection results show longer warning stages for the two groups with 7.9 and 8 days respectively. The time indicators detected the fault at Day 31 (#1820) and the FK based SES indicators give the detection at the same time as the CSCoh indicators at Day

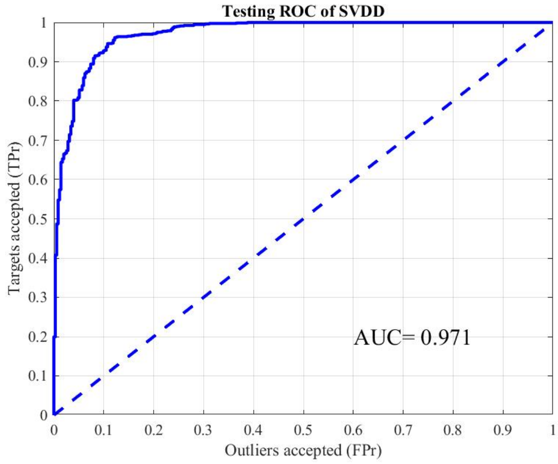


Figure 6: ROC curve for SVDD with CSCoh indicators of Dataset 1 Bearing 3

30.9 (#1796). The detection results show that time and FK based SES indicators are sensitive to the changing of external environment due to the recording interruptions. On the contrary, the distances of SVDD results from CSCoh indicators keep steady during most of the testing period which indicates that the CSCoh HIs are robust in detecting the bearing faults.

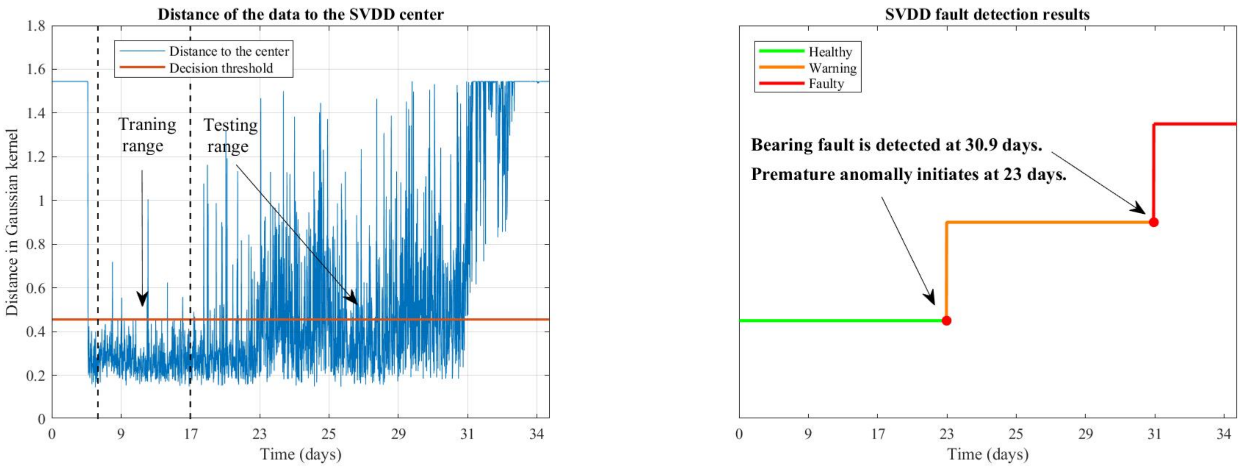


Figure 7: Distances to SVDD center (left) and fault detection results (right) for time domain indicators of Dataset 1 Bearing 3

The metrics derived from the classification confusion matrix are adopted to compare the performance of SVDD with different HIs, including the False Positive Fraction (FPF), the False Negative Fraction (FNF), the False Positive Rate (FPR) and the False Negative Rate (FNR). The calculation of the four values is shown in Fig. 9. The metrics represent the classification performance, measured from the estimated labels and the true labels, based on [17]. More specifically, FPR can be seen as the evaluation of false alarms and FNR represents the misdetection of the classifier. Lower values of these metrics indicate a better classification performance. The metrics together with the AUC value of all three group of HIs are listed in Tab. 3. The SVDD classifier apparently performs better with indicators from CSCoh than the other two groups, presenting superiority in the accurately detection of bearing faults.

6.2 Case 2: Dataset 2 Bearing 1

The HI map of Dataset 2 Bearing 1 is shown in Fig. 10. Since the measurement of this dataset is continuous without any interruption, all the indicators behave more monotonic than in Dataset 1. A clear separation can be seen at Day 4.9 for all the indicators which represent a dramatic increasing of the HIs. On the other hand, the HIs from the CSCoh show a forehead increase at Day 3.8 for the indicators constructed with the harmonics and

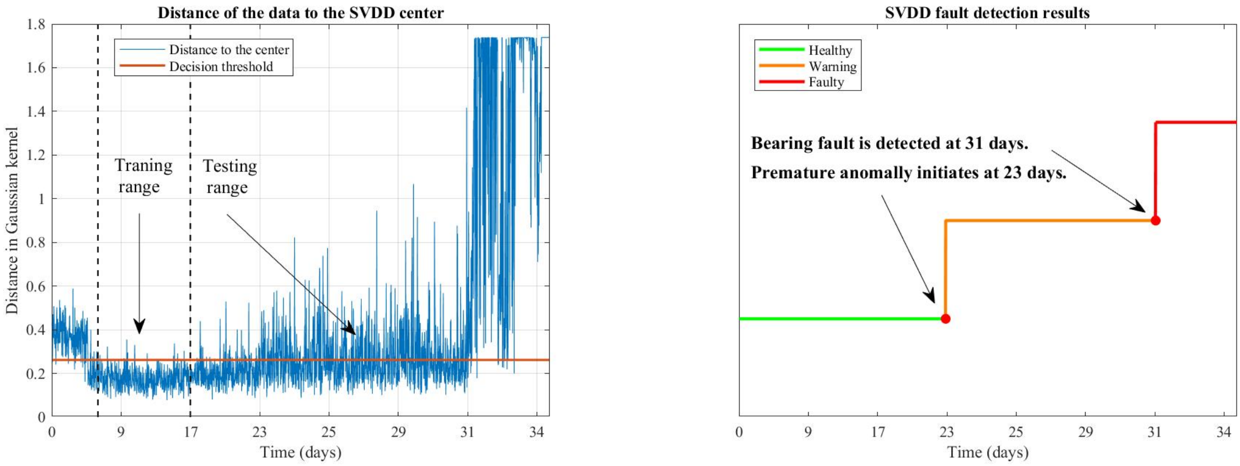


Figure 8: Distances to SVDD center (left) and fault detection results (right) for FK based SES indicators of Dataset 1 Bearing 3

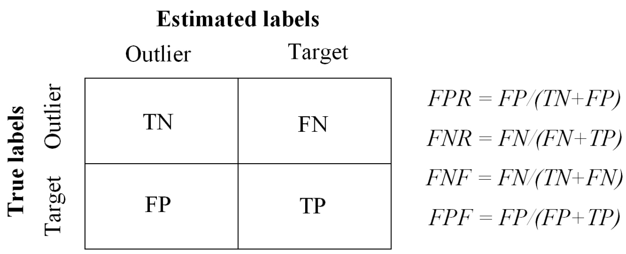


Figure 9: Calculation of the evaluation metrics.

	FPF	FNF	FPR	FNR	AUC
Time	0.474	0.195	0.370	0.085	0.905
FK based SES	0.248	0.040	0.525	0.023	0.963
CSCoh	0.076	0.023	0.177	0.019	0.971

Table 3: Performance of SVDD classifiers with different indicators for Dataset 1 Bearing 3

the sum of BPFO, which is the fault type in this case.

The distances from the SVDD and the fault detection results of the CSCoh indicators are shown in Fig. 11. The training set is selected from Day 0.3 (#50) to Day 2.1 (#300) and the remaining data are used for testing. Less distances pass the threshold which represent the monotonicity of the indicators. The warning stage is 0.1 days from Day 3.6 (#521) to Day 3.7 (#533) which is relatively short due to the indicators with more tendency. Then the bearing fault is detected at Day 3.7 (#534).

The indicators from the time and the FK based SES are also sent to the SVDD model to realise fault detection and the results are shown in Fig. 12 and 13. The CSCoh indicators provide the earliest detection compared to the time indicators (Day 3.8, #550) and the FK based SES indicators (Day 3.7, #541). The amplitudes of the characteristic frequencies and their harmonics from EES present more monotonicity and trendability which provide evidence in the construction of feature space during the training of SVDD model. As a result, the detection performs better using the CSCoh indicators.

The metrics of the classifiers' performance among different groups of HIs show less variance, compared to the first case. The CSCoh indicators gain lower FPF, FPR and FNR. The time domain indicators get lower FNF but the AUC is still lower than for the CSCoh.

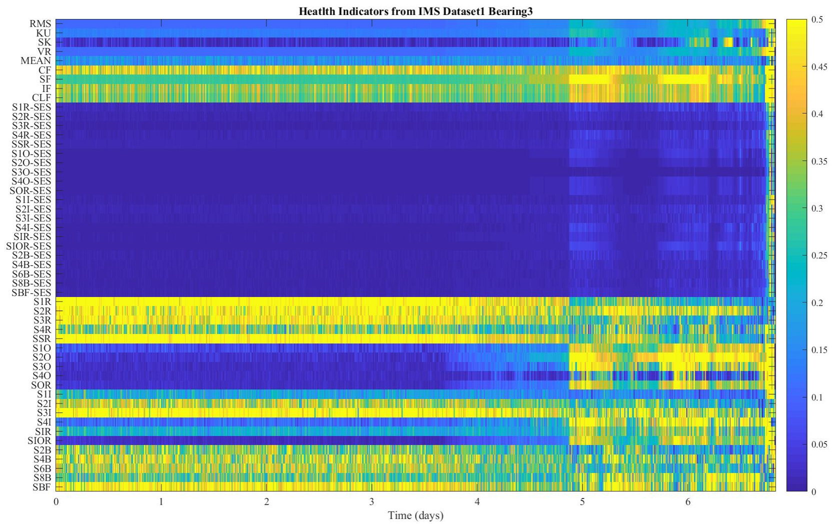
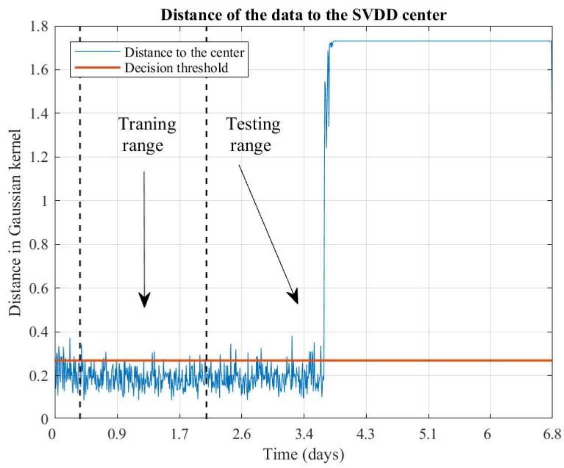


Figure 10: Health indicator map for Dataset 2 Bearing 1



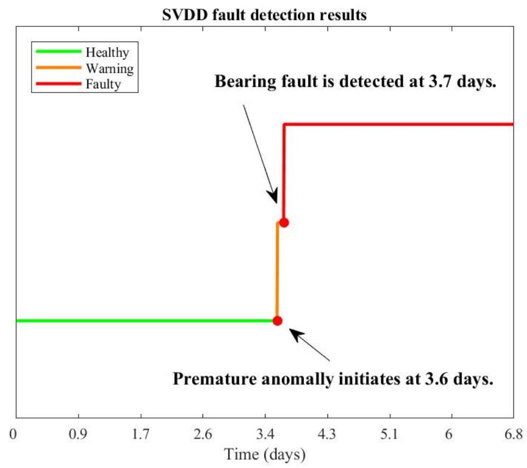
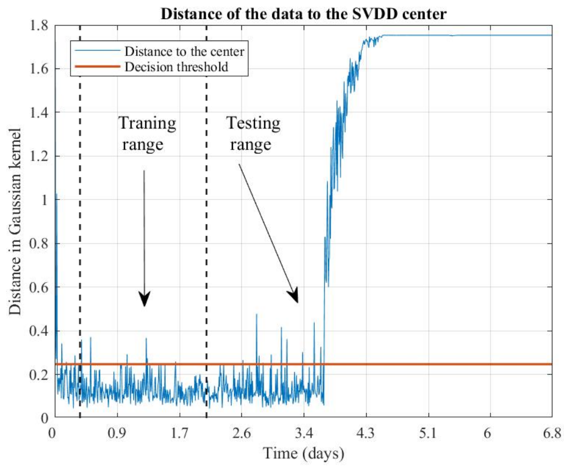


Figure 11: Distances to SVDD center (left) and fault detection results (right) for CSCoh indicators of Dataset 2 Bearing 1



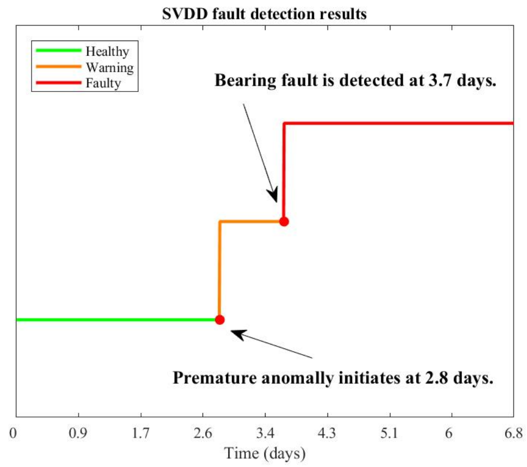


Figure 12: Distances to SVDD center (left) and fault detection results (right) for time domain indicators of Dataset 2 Bearing 1

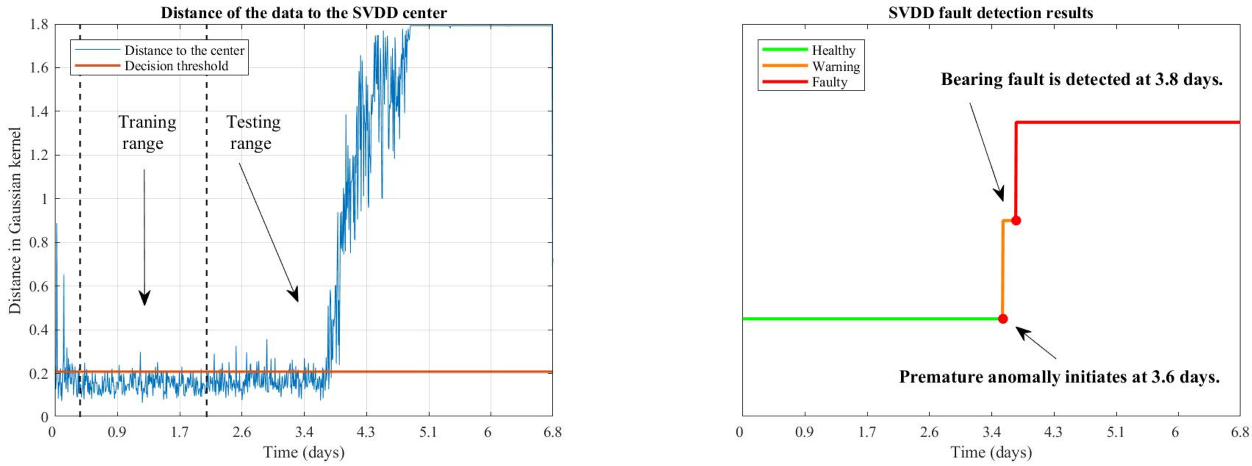


Figure 13: Distances to SVDD center (left) and fault detection results (right) for FK based SES indicators of Dataset 2 Bearing 1

	FPF	FNF	FPR	FNR	AUC
Time	0.078	0.018	0.028	0.041	0.991
FK based SES	0.090	0.020	0.038	0.005	0.992
CSCoh	0.056	0.020	0.028	0.005	0.992

Table 4: Performance of SVDD classifiers with different indicators for Dataset 2 Bearing 1

7 Conclusion

In this paper, a novel rolling element bearing automatic fault detection approach is proposed by combining CSCoh domain indicators and a semi-supervised SVDD technique. Experimental results from run-to-failure bearing datasets prove that the bearing faults can be accurately detected with the proposed methodology. Compared to the indicators from time domain and the FK based SES domain, CSCoh diagnostic indicators are more robust to the changing of external environment and improve the performance of the SVDD classifier. The proposed semi-supervised methodology therefore has a strong practical significance in industrial applications since it can be effective in both bearing failure warning and damage detection.

Acknowledgements

Chenyu Liu would like to acknowledge the support from China Scholarship Council.

References

[1] R. Randall , J. Antoni, *Rolling element bearing diagnostics-A tutorial*, Mechanical Systems and Signal Processing, 2011 February, 1, vol. 25, no. 2, pp.485-520.

[2] B. Zhang, G. Georgoulas, M. Orchard, A. Saxena, D. Brown, G. Vachtsevanos, S. Liang. *Rolling element bearing feature extraction and anomaly detection based on vibration monitoring*, 16th Mediterranean Conference on Control and Automation, 2008 June 25, pp.1792-1797.

[3] K. Gryllias, S. Moschini, J. Antoni, *Application of cyclo-nonstationary indicators for bearing monitoring under varying operating conditions*, Journal of Engineering for Gas Turbines and Power, 2018 Jan 1, vol 140, no. 1.

[4] C. Yi, D. Wang, W. Fan, KL. Tsui, J. Lin, *EEMD-based steady-state indexes and their applications to condition monitoring and fault diagnosis of railway axle bearings*, Sensors, 2018 Feb 27, vol. 18, no. 3, pp.704.

[5] Y. Lei, F. Jia, J. Lin, S. Xing, SX. Ding, *An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data*, IEEE Transactions on Industrial Electronics, 2016 May, vol. 63, no. 5, pp.3137-3147.

[6] J. Antoni, F. Bonnardot, A. Raad, M. El Badaoui, *Cyclostationary modelling of rotating machine vibration signals*, Mechanical Systems and Signal Processing, 2004 Nov 1, vol. 18, no. 6, pp.1285-1314.

[7] A. Mauricio, A. Miguel, W. Smith, R. Randall, J. Antoni, K. Gryllias, *Cyclostationary-based tools for bearing diagnostics*, Online Proceedings of ISMA 2018 including USD 2018, 2013, pp.905-918.

[8] A. Mauricio, J. Qi, K. Gryllias. *Vibration-Based Condition Monitoring of Wind Turbine Gearboxes Based on Cyclostationary Analysis*, Journal of Engineering for Gas Turbines and Power, 2019 Mar 1, vol. 141, no. 3.

[9] S. Kass, A. Raad, J. Antoni, *Self-running bearing diagnosis based on scalar indicator using fast order frequency spectral coherence*, Measurement, 2019 May 1, vol. 138, pp.467-484.

[10] J. Qi, A. Mauricio, K. Gryllias, *Enhanced Particle Filter and Cyclic Spectral Coherence based Prognostics of Rolling Element Bearings*, In PHM Society European Conference, 2018 Jun 28, vol. 4, no. 1.

[11] K. Gryllias, I. Antoniadis, *A Support Vector Machine approach based on physical model training for rolling element bearing fault detection in industrial environments*, Engineering Applications of Artificial Intelligence, 2012 Mar 1, vol. 25, no. 2, pp.326-44.

[12] T. Shon, Y. Kim, C. Lee, J. Moon, *A machine learning framework for network anomaly detection using SVM and GA*, In Proceedings from the sixth annual IEEE SMC information assurance workshop, 2005 Jun 15, pp.176-183.

[13] R. Perdisci, G. Gu, W. Lee, *Using an Ensemble of One-Class SVM Classifiers to Harden Payload-based Anomaly Detection Systems*, In ICDM, 2006 Dec 18, vol. 6, pp.488-498.

[14] P. Bangalore, LB Tjernberg, *An artificial neural network approach for early fault detection of gearbox bearings*, IEEE Transactions on Smart Grid, 2015 Mar, vol. 6, no. 2, pp.980-987.

[15] D. Tax, R. Duin, *Support vector data description*, Machine learning, 2004 Jan 1, vol. 54, no. 1, pp: 45-66.

[16] H. Qiu, J. Lee, J. Lin, G. Yu. *Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. Journal of sound and vibration*, 2006 Feb 7, vol. 289(4-5), pp.1066-1090.

[17] K. Gryllias, I. Antoniadis, C. Yiakopoulos, *A novel semi-supervised mathematical morphology-based fault detection & classification method for rolling element bearings*, In Proceedings of the 22nd International Congress on Sound and Vibration, 2015.

Machine teaching to optimize algorithms performances on restricted dataset.

Alexandre CARBONELLI[1], Maurin WOUSSEN [1]

[1]Vibratec, 28 Chemin du Petit Bois, 69130 Écully, France
alexandre.carbonelli@vibratec.fr

Abstract

Machine Learning techniques open important ways of development of physical models in almost every field. Performances reached rely on two major pillars: the (physical) model and the data. When a huge amount of data is available, the intrinsic characteristics of the chosen algorithm become less critical. On the other hand, with limited amount of data, all the human knowledge on the system to be modelled becomes critical to exploit.

One of the multiple applications of Machine Learning (ML) technics concerns the meta-models. Indeed, in this paper, we show how we can bypass a computation scheme by using clever regression models. The approach is performed on a system from which we want to know instantaneously the first natural frequencies without performing each time finite elements (FE) computation. We study the performance reached by studying the number of training cases required to teach the algorithm how to link inputs to outputs within a satisfying accuracy. Different algorithms are tested with very encouraging results as going into higher dimensional problem. The final aim of this study is to provide global guidelines for the most efficient Machine Teaching.

1 Introduction

Reducing the computation time is an objective is many cases. Whether it is for having instantaneously a result or to integrate the meta-model in a bigger loop, the idea remains the same: to link inputs to outputs without going through a whole computation process. The best way to achieve that is for instance to have an analytical formula based on physics consideration which can represent exactly the phenomenon studied. According to the task, it is unfortunately not always possible to use an analytical formula. The typical representative case of such a task would be for instance a classification problem: to determine if a given picture represents a cat or a dog, it seems obvious that we will not find an analytical formula able to do that. To illustrate our idea we can consider Figure 1.

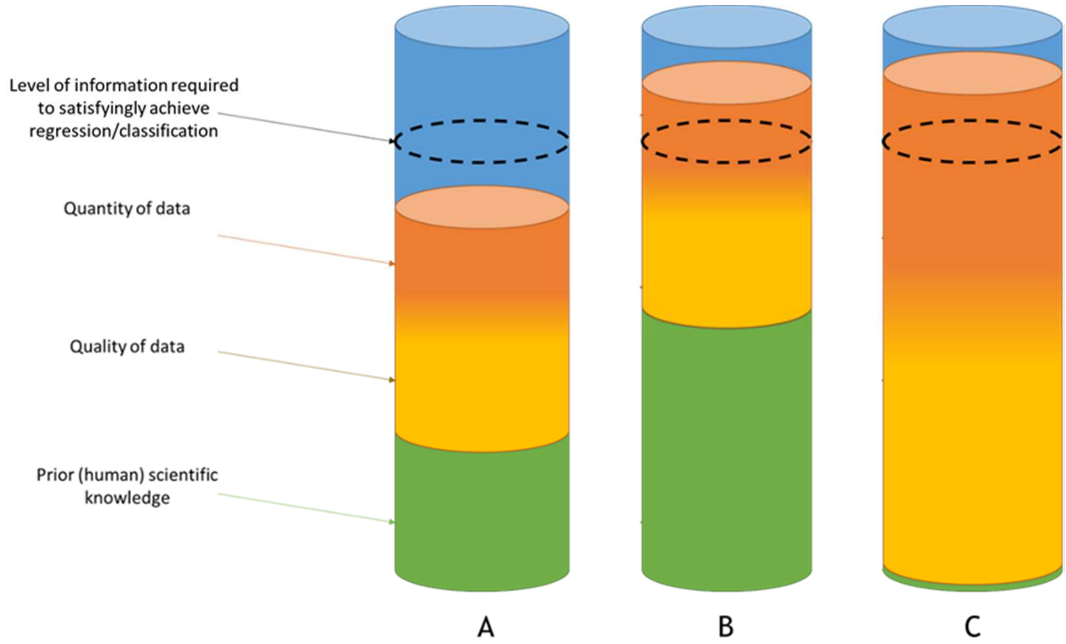


Figure 1: General approach of a regression/classification problem

When we lack information (case A), we can either use scientific knowledge (case B) or generate more data (case C) or of course both in order the fill the gap.
The general parameters suggested to be considered are:

- The level of information required to ***satisfyingly*** achieve the regression/classification implies to set a threshold beyond which we estimate that the task is a success,
- The quantity of data is directly linked to their quality (we consider that in the next part), meaning that many data of poor quality could not be that helpful according to the accuracy aimed for,
- The prior (human) scientific knowledge establishes the base of the whole stack (for instance, an adapted analytical formula fills the gap to the level of information required by its own).

This approach will be later included into a classification task consisting in detecting in a block made of carbon is cracked or not by studying its natural frequencies. We focus in this paper on the determination (through regression) of the natural frequencies of the block without any additional FE computation than those used for the model building. The accuracy of the reference natural frequencies (those obtained for a healthy block) must be high in order to be the more sensitive possible in the incoming classification task.

2 The concept of Machine Teaching

The Machine Teaching concept as we thought it relies on a direct analogy on human learning. Let's consider a student who has a set of examinations.

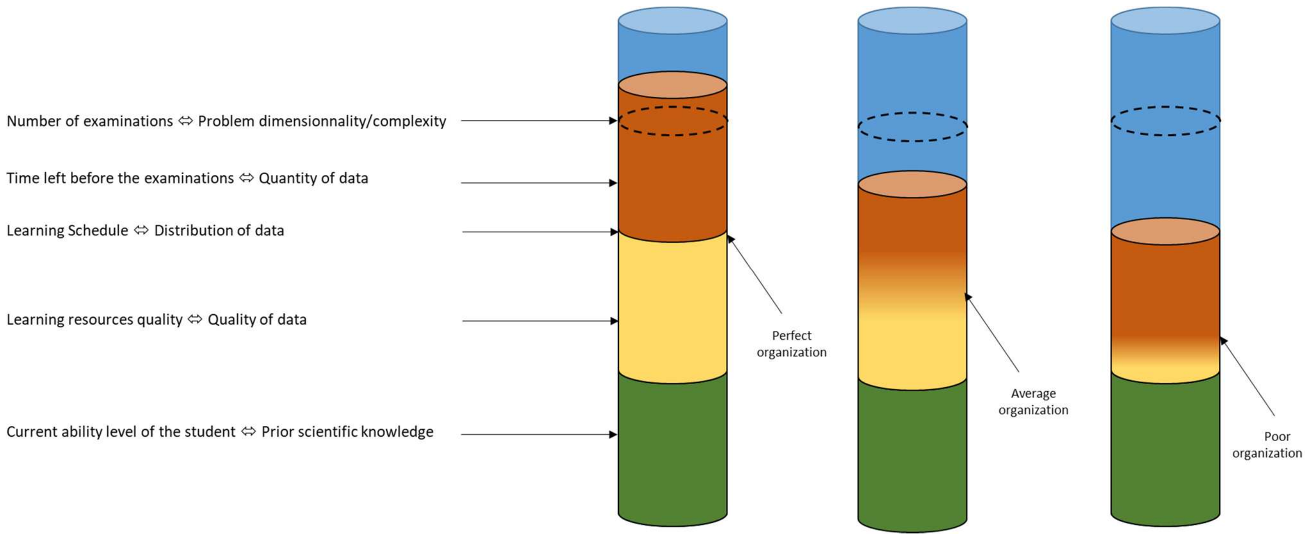


Figure 2: Illustration of the Machine Teaching concept

We suggest on Figure 2 the following symmetries. The number of examinations can be related to the complexity of the task or its dimensionality. The time before the examinations can be related to the quantity of data that one can generate in a given time. The learning resources (writing quality, the teacher implication and so on) can be related to the quality of data (does one need to correct and to complete manually the dataset, taking risk to insert errors…). The current ability of the student can be related on the priori scientific knowledge: what does he knew before any learning? His knowledge basis. Finally, the learning schedule is directly related for us to the concept of Machine Teaching. Indeed, having well distributed data (not necessarily uniform) permits to optimize the supply of each new data, avoid generalization problem, eliminate redundancies and perform better. To continue the analogy, for a student with a limited amount of time, a good schedule learning during which he will be able to the see different cases. The most representative cases will definitely improve his learning compared to spend most of the time on one chapter and ignoring the six left. Therefore a good teaching can be related to one ability to understand the knowledge compulsory to integrate and to provide consistent organization of the data.

3 Presentation of the study case

3.1 The free-free beam

The study of the free vibration of a beam is not that challenging. Nevertheless, classical approach using Euler-Bernoulli's and Timoshenko's beam theories [1-2] provides satisfying as long as some assumptions are satisfied, such as the ratio between the length of the beam and its cross-sectional dimensions (in which case the Euler-Bernouilli's beams present an important error). In our case, we don't want to make any assumptions about the beam behaviour and we use therefore FE model with solid elements to feed the meta-model. The Figure 3 shows two beam samples, which can be rectangular or circular. The Table 1 lists the different parameters considered and their variation ranges and Table 2 sums up the meshing details used.

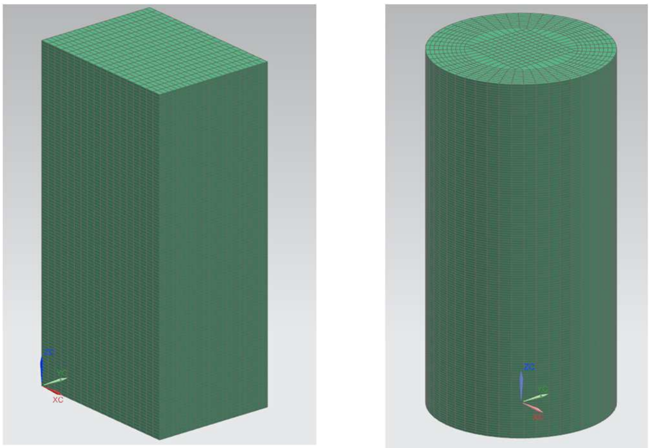


Figure 3: Examples of considered rectangular and circular beams.

Parameters	Minimum	Maximum
Length (m)	2	9
Section[1] (m)	0.3	1
Young's modulus (MPa)	7.92	11.88
Poisson's ratio	0.1	0.4

Table 1: Parameters studied and their variation ranges

	Rectangular beams	Circular beams
Element's type	Cube	Rectangular cuboid
Type of grid	Structured	Structured
Number of elements	50 112	93 600
Number of nodes	55 575	97 999

Table 2: Meshing details

3.2 Automatic generation of samples

The study was realized in two steps, a first naïve approach to the problem, and then a second step where we developed and improved a method. The different cases firstly used to feed the meta-model are generated automatically and are divided in N_1 circular and N_2 rectangular. We separate two different types of set, the training data used for the learning procedure, generated in a grid way, we will explain later the reason of this choice. And the test data, used for the meta-model validation, which are randomly generated, to emphasize the generalization on the entire study domain. We recall the different subsets in Table 3 and Table 4*: Subsets labels of the second step***Erreur ! Source du renvoi introuvable.**.
L, b, h, E, ν respectively stand for the length, the height, the thickness, the Young's Modulus and the Poisson's coefficient of the beam.

[1] Section corresponds to the larger length of the section's beam, i.e. diameter for circular beams or maximum between thickness and height for rectangular beams.

	First Step			
	Rectangular beams		Circular beams	
	Train	Test	Train	Test
Variables involved	b, h, L	b, h, L	b, h, L	b, h, L
Number of samples	512	141	400	75
Generation	Grid	Random	Grid	Random

Table 3: Subsets labels of the first step

	Second Step			
	Rectangular beams		Circular beams	
	Train	Test	Train	Test
Variables involved	b, h, L, E, ν	b, h, L, E, ν	b, h, L, E, ν	b, h, L, E, ν
Number of samples	100	50	100	50
Generation	Grid	Random	Grid	Random

Table 4: Subsets labels of the second step

We begin with 512 samples of rectangular beams for the training process, such as we have eight different values of length, thickness and height to make the grid. The initial distribution is naively uniform.
Its test set counts 141 random samples, with a criteria on the ratio, length over the working dimension of the section, where it can't overstep 3.33 for each configuration. The first training data for the circular beams contains 400 samples, such as we have twenty values of length and twenty values of diameter, to make a grid. We have 75 samples for its test set.

For the second step, we include the material properties in the study and we try to reduce the number of samples needed for the learning. For both of the circular and rectangular cases, we generate a grid such as we have 5 values of length, 5 values of thickness, 2 values of Young's modulus and 2 values of Poisson's ratio. We'll explain later the reasons of that configuration choice. As we add two new features, we also have to create new test sets of 50 samples each, generated randomly.

Finally, a total of 1428 samples have been automatically generated. To give an idea of the length of the process, one computation takes about three minutes. Then, to generate the complete set of input data, about 72 hours have been required.

3.3 Automatic extraction of data

The outputs required are simply the three first traction/compression (Z axis) eigenfrequencies and the three first flexural eigenfrequencies in both directions (X and Y axis). Each computation output file are post-process and for each case the 9 outputs are written related to the parameters defined in part 3.1.

4 Methodology for building the meta-model

Facing a lack of information, we firstly choose the random forest as our starting strategy to predict the eigenfrequencies. Indeed, the ensemble methods can have good results on small subsets, where they start from a high variance and thanks to bagging algorithm, decreases it during the learning process to improve the generalization [4]. However, many parts of the response values are underrepresented, which distorts the training phase, leading to a bad accuracy. We had to change our point of view on the procedure and try to develop a way to guide the algorithm during the training so that we can control how it learns along the process and "help" it. That's why we choose to build a custom learning method, closer to the machine teaching area. This method is based on nonlinear least squares [5] and the use of a specific grid as training data.
We can see in the Table 5: Comparison of relative error between random forest and a custom method a comparison on the worst mode's results, between random forest and our custom method. To do so, we calculate the relative error e such as:

$$e = \sqrt{\frac{(y-\tilde{y})^2}{y^2}} \times 100 \tag{1}$$

with y, the observation value and $\tilde{y}$, the prediction from the modelling function. This error is determined for each point of the test subset.

	Random Forest			Custom Method		
	F1X	F1Y	F1Z	F1X	F1Y	F1Z
Mean (%)	9.57 ; 29.1	13.7 ; 29.3	0.210 ; 5.38 10^{-2}	9.35 10^{-2} ; 0.122	0.100 ; 0.122	0.212 ; 0.265
Max (%)	46.54 ; 154	39.98 ; 156	2.42 ; 0.346	0.264 ; 0.554	0.255 ; 0.554	0.414 ; 0.843

Table 5: Comparison of relative error between random forest and a custom method
Rectangular beams / Circular beams

This table shows the limit of classical machine learning algorithms in this context. We had to modify existing methods to have a better fitting and more complex regression curve.

The main idea behind the method used is to deconstruct the learning procedure, so that we can control the tendency of our response depending on each features separately, during each step. We want to approximate the output with a function such as:

$$f: \begin{array}{ccc} I_1 \times \ldots \times I_N & \to & \mathbb{R} \\ (v_1 \times \ldots \times v_N) & \to & f((v_1 \times \ldots \times v_N), \beta) \end{array} \tag{2}$$

Where N is the number of features, $I_k \subset \mathbb{R}, k = \{1, \ldots, N\}$ the set where the feature is taken from, and β a set of parameters used by the function f. In our situation, we can choose the configuration $(v_1 \times \ldots \times v_N)$ we want to append in our samples set. Then, we choose to build a grid, as a combination of values from chosen subsets. The strategy is given in Appendix.

5 First step

5.1 Strategy

To start the study, several comments have been done to make easier the resolution. We choose to consider each eigenfrequency separately, instead of using multioutput regression algorithms which tends to be less accurate. To approximate the transverse modes along the Y axis, we used the symmetric properties of these modes and take the regression function on the transverse modes along X axis, where we inverse the thickness and height values. We now had to build 6 modelling functions. We also consider that the height does not influence the transverse modes along X axis, so the approximation will involve only two dimensions. For the longitudinal modes, we noticed that the eigenfrequencies are highly correlated with the length, so the regression function will only depend on this feature.
Our custom method were mainly used to build the transverse modes interpolation. So we have two layers, according to the features involved. The thickness layer is approximated with polynomials of second order and the parameters to estimate as related to a rational function such as:

$$f(x) = \frac{\beta_1 x^2 + \beta_2 x + \beta_3}{\beta_4 x^2 + \beta_5 x + 1} \tag{3}$$

Where x is the feature and $\beta_i, i = 1, \ldots, 5$ the parameters. The final function looks like:

$$f(b,L) = \frac{\alpha_1 L^2 + \alpha_2 L + \alpha_3}{\alpha_4 L^2 + \alpha L + 1}\, b^2 + \frac{\beta_1 L^2 + \beta_2 L + \beta_3}{\beta_4 L^2 + \beta_5 L + 1}\, b + \frac{\gamma_1 L^2 + \gamma L + \gamma_3}{\gamma_4 L^2 + \gamma L + 1} \quad (4)$$

With $\alpha_i, \beta_i, \gamma_i, i = 1, \dots, 5$ the final parameters.
For the longitudinal modes, we did a log-log transformation then a linear regression on the result, such as $f(L) = L^{\beta_1} e^{\beta_2}$.

5.2 Results

5.2.1 Longitudinal modes

The relative error mentioned (1) is used to explore the function's accuracy and validate the model. We deduct some statistic information of its tendency, such as the mean above the values, the standard deviation around it and maximum and minimum value of the error data. This procedure is kept for all the results in the study.
The mentioned data are presented in **Erreur ! Source du renvoi introuvable.**.

	Rectangular beams			Circular beams		
	F1Z	F2Z	F3Z	F1Z	F2Z	F3Z
Count	141	141	141	75	75	75
Mean	0.212	0.202	0.198	0.265	0.263	0.296
Standard Deviation	$9.31\ 10^{-2}$	$8.99\ 10^{-2}$	0.138	0.215	0.241	0.295
Min	$7.63\ 10^{-2}$	$4.86\ 10^{-2}$	$9.36\ 10^{-3}$	$1.12\ 10^{-2}$	$2.01\ 10^{-3}$	$1.53\ 10^{-3}$
Max	0.414	0.432	0.608	0.843	0.973	1.42

Table 6: Error (%) for longitudinal modes with circular beams during the study's first step

We can already be satisfied by the error which is less than 1%, except for the third longitudinal modes with the circular beams, where we notice a peak at 1.42%. All the means are around 0.2% of error which is promising for the efficiency of the method used.

We also represent the error depending on the ratio, to enhance the influence of ratio as major error vector. The results are represented in the Figure 4.

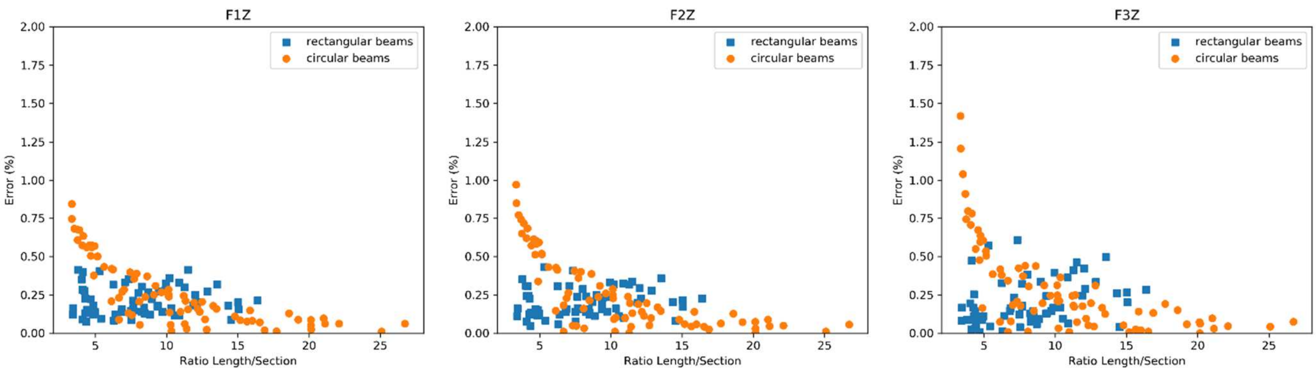


Figure 4: Errors in function the ratio for longitudinal modes during the study's first step

The error has a dependency with the ratio which increases when this ratio decreases. The difference of behavior between circular and rectangular beams is due to the difference of the grid's construction used.

5.2.2 Transverse modes

Table 7: Error (%) for transverse modes with rectangular beams during the study's first step and Table 8*: Error (%) for transverse modes with circular beams during the study's first step* contain the statistic of the error on the transverse modes. We still have an error less than 1% except a peak for the third modes which results of its higher complexity. The means tends to evolve around 0.15% which is still promising.

	F1X	**F2X**	**F3X**	**F1Y**	**F2Y**	**F3Y**
Count	141	141	141	141	141	141
Mean	$9.35\ 10^{-2}$	0.128	0.126	0.100	0.141	0.139
Std. Dev.	$6.76\ 10^{-2}$	$8.58\ 10^{-2}$	$8.98\ 10^{-2}$	$6.55\ 10^{-2}$	$8.53\ 10^{-2}$	$8.31\ 10^{-2}$
Min	$1.04\ \ 10^{-3}$	$2.44\ 10^{-3}$	$1.45\ 10^{-3}$	$5.45\ 10^{-3}$	$6.17\ 10^{-3}$	$5.08\ 10^{-4}$
Max	0.264	0.323	0.305	0.255	0.324	0.344

	F1X	**F2X**	**F3X**	**F1Y**	**F2Y**	**F3Y**
Count	75	75	75	75	75	75
Mean	0.122	0.160	0.185	0.122	0.153	0.187
Std. Dev.	0.130	0.153	0.187	0.130	0.153	0.187
Min	$9.52\ 10^{-8}$	$1.21\ 10^{-3}$	$2.92\ 10^{-3}$	$9.52\ 10^{-8}$	$1.21\ 10^{-3}$	$2.92\ 10^{-3}$
Max	0.554	0.877	1.23	0.554	0.877	1.23

Table 7: Error (%) for transverse modes with rectangular beams during the study's first step

Table 8: Error (%) for transverse modes with circular beams during the study's first step

We have globally the same accuracy as for the longitudinal modes, the distribution is wider which can be explained by the more complex behavior of these quantities. The use of 2 layers here can also add some deviation for some values.

We have a good accuracy, even if we have some peaks for small ratio which confirming the changing behavior of this area as shown Figure 5.

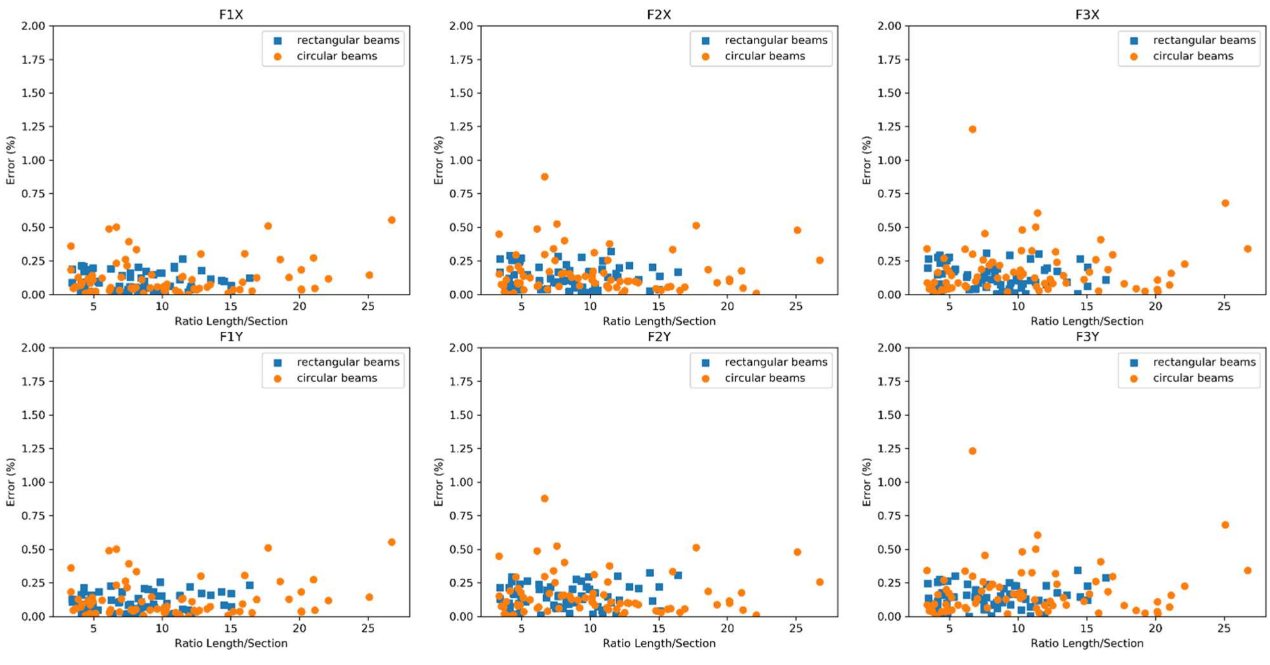


Figure 5: Error in function of the ratio for transverse modes during the study's first step

5.3 Critics and way of improvement

Even if the accuracy could be higher, we have satisfying results for this first meta-model, which gives us 1% error on our generalization data. This results is satisfying for the model. As we are in a problem with maximum two dimensions, having only one feature to fit the longitudinal works well in this specific case.

Moreover, we want to include in the second step the material properties of the beam, so we will have to be more flexible on the complexity of our regression function, even for the longitudinal modes. A more specific way of determining the training samples has to be built, so that we minimize its size. The next step will be to set a guideline to choose an initial subset and optimize the number of samples needed to have a great accuracy.

6 Second step

6.1 Strategy

As we have two new features involved, another grid needs to be made. We want to minimize the number of samples, so the first idea is to keep as much variance as possible in the domain of study, while maintaining the properties of a grid. Nevertheless, during the study, we set a constrain on the ratio, where for a beam too close of a cube, we are not sure to find a third longitudinal mode, due to its geometry. To respect this constrain, instead of taken the length between 2 and 9 meters, we set its boundaries between 3.33 and 9 meters.

The optimization is mainly done by the number of values we choose for each feature. Thanks to theoretical properties, we know that the eigenfrequencies are linearly dependent with Young's modulus and Poisson ratio [6]. So for the layers associated to material features, we needed two values each to make a linear regression. Then for the other features, we want to approximate data with curves which have a certain complexity, so we decided to choose five values as our beginning training set. The five values will allow us to choose function including five parameters and this gives us a first flexibility on function complexity.

We took the same comments as the previous step, where only six modes are approximated and the height is not included as a relevant feature for both the transverse and longitudinal modes. Moreover, for the rectangular beams, we generate randomly the height always strictly lower than the thickness, so that we avoid redundancy in the training data.

We also improve the custom method, now for each layer and for each parameters, the function to interpolate can be chosen. This increases a lot the complexity of the final function and allows us to fit more precisely our data and have a visualization of the fitting at each step of the process. We chose among the following functions to adjust the data:

- Linear : $f(x) = \beta_1 x + \beta_2$
- Polynomial of degree 2 : $f(x) = \beta_1 x^2 + \beta_2 x + \beta_3$
- Polynomial of degree 3 : $f(x) = \beta_1 x^3 + \beta_2 x^2 + \beta_3 x + \beta_4$
- Rational function of first order : $f(x) = \frac{\beta_1 x + \beta_2}{\beta_3 x + 1}$
- Rational function of second order : $f(x) = \frac{\beta_1 x^2 + \beta_2 x + \beta_3}{\beta_4 x^2 + \beta_5 x + 1}$
- Exponential : $f(x) = \beta_1 e^{\beta_2 x}$

With $\beta_i, i = 1, \dots, 5$ the parameters to adjust.

6.2 Results

6.2.1 Longitudinal modes

In the Table 9*: Error (%) for longitudinal modes with rectangular beams during the study's second*

	Rectangular beams			Circular beams		
	F1Z	F2Z	F3Z	F1Z	F2Z	F3Z
Count	50	50	50	50	50	50
Mean	0.378	0.412	0.582	0.215	0.278	0.351
Standard Deviation	0.138	0.223	0.442	0.221	0.263	0.291
Min	$1.62\ 10^{-2}$	$6.31\ 10^{-3}$	$4.20\ 10^{-4}$	$5.03\ \ 10^{-4}$	$6.00\ 10^{-5}$	$8.71\ 10^{-4}$
Max	0.570	0.954	2.30	0.565	0.763	0.886

step, we present the results of longitudinal modes for the second step of the study.

Table 9: Error (%) for longitudinal modes with rectangular beams during the study's second step

The error is slightly higher than the first step and increases with the order of the mode, but the means are under 0.6% of error, which is a promising accuracy for this initial grid.

The Figure 6: Error in function of the ratio for longitudinal modes during the study's second presents

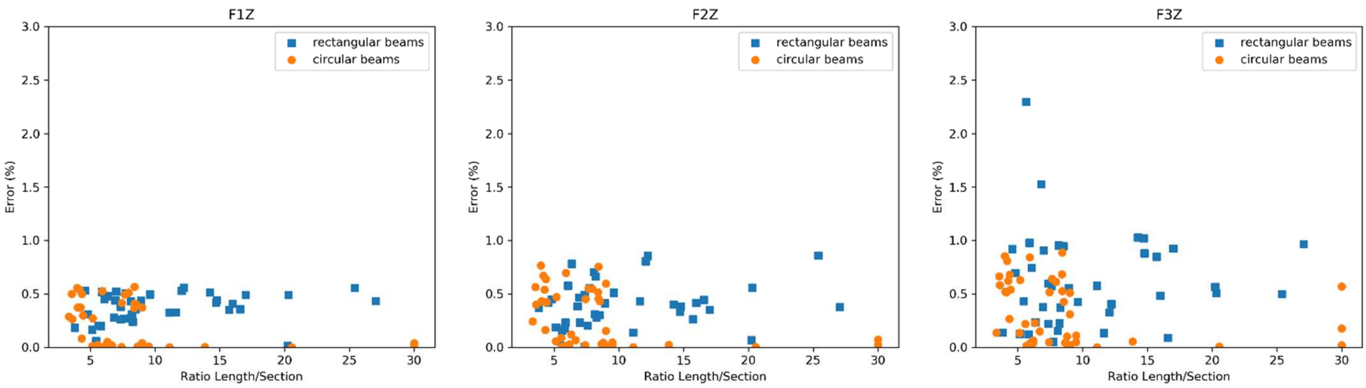


Figure 6: Error in function of the ratio for longitudinal modes during the study's second

the error depending on the ratio of the beams.
The rise of the error depending on the mode's order is more prominent here, with still some perturbations for small ratio.

6.2.2 Transverse modes

In the Table 10*: Error (%) for transverse modes with rectangular beams during the study's second step* and
Table 11: Error (%) for transverse modes with circular beams during the study's second step, we present the results of transverse modes for the second step of the study.

	F1X	F2X	F3X	F1Y	F2Y	F3Y
Count	50	50	50	50	50	50
Mean	0.430	0.487	0.512	0.399	0.678	0.829
Standard Deviation	0.203	0.250	0.379	0.166	0.494	0.623
Min	$1.93\ \ 10^{-2}$	$6.66\ 10^{-2}$	$2.76\ 10^{-2}$	$3.79\ 10^{-2}$	$1.45\ 10^{-2}$	$6.23\ 10^{-1}$
Max	1.05	1.24	1.92	0.837	2.48	3.03

Table 10: Error (%) for transverse modes with rectangular beams during the study's second step

	F1X	F2X	F3X	F1Y	F2Y	F3Y
Count	50	50	50	50	50	50
Mean	0.186	0.309	1.36	0.189	0.309	1.36
Standard Deviation	0.173	0.472	1.00	0.172	0.470	1.00
Min	$4.02\ \ 10^{-3}$	$1.26\ 10^{-4}$	$2.73\ 10^{-2}$	$4.02\ 10^{-3}$	$1.26\ 10^{-4}$	$3.73\ 10^{-2}$
Max	0.475	3.25	3.64	0.475	3.24	3.63

Table 11: Error (%) for transverse modes with circular beams during the study's second step

The same remark as the previous section can be made, the error is higher than the first step but still acceptable, considering the number of training data.

In the Figure 7*:* Error in function of the ratio *for transverse modes during the study's second*, the graphs show the limits of the method and the grid, where we have generalization issues, especially for small ratios and third modes.

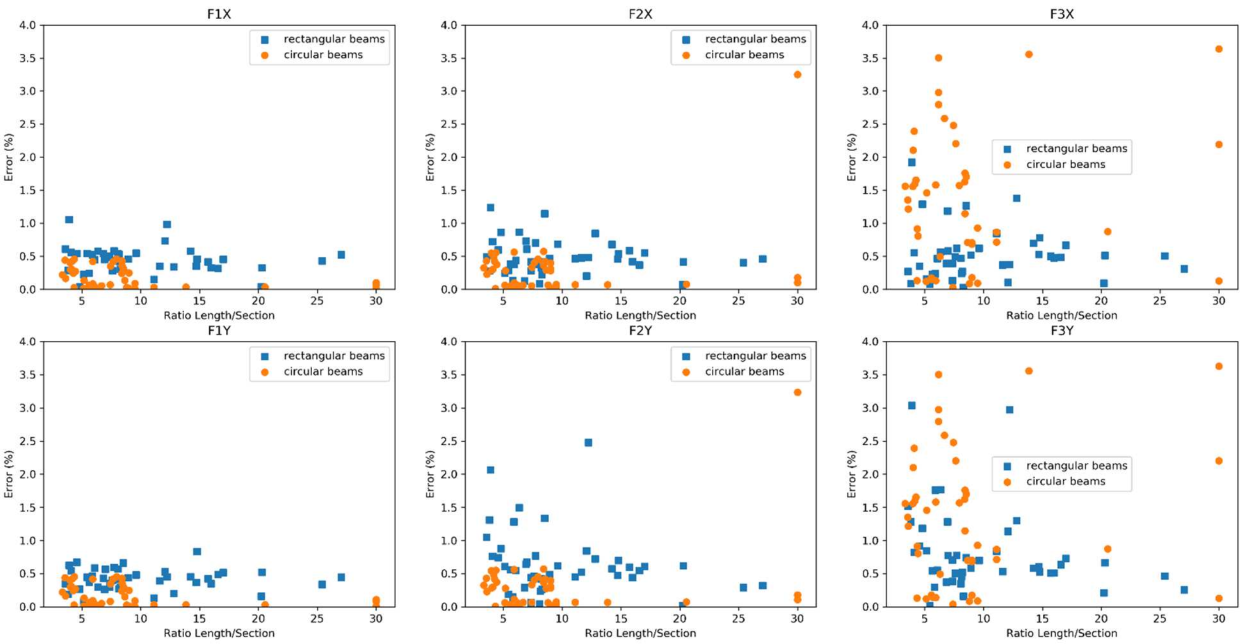


Figure 7: Error in function of the ratio for transverse modes during the study's second

6.3 Critics and ways of improvement

The error increases but is still acceptable. Moreover, we have to consider the context of the learning which is more complex than the previous step. The error is twice higher than the first step, under 3% error of accuracy. It is worth remembering that we added two features and decreases the training data from 512 to 100 samples.

A first improvement would be to stabilize generalization's error. To do so, we will propose a guideline to select leading samples to add in the training data which will increase the accuracy. The criteria to build the initial grid can also be improved. Indeed, we didn't train the model with length under 3.33

meters and it impacts the generalization as we see in the results. A new methodology could spare the lack information in this ratio's area.

6.4 Guideline for sample selection

We propose a first guideline to increase the precision by adding some relevant samples. The initialization starts with the construction of a first grid, with a minimum number of samples. Here it will depends on the nature of the features and the domain of study. If some follows a specific tendency, then we know the minimum number of requisite values to approach them.
In order to get a satisfying learning, we need to have a uniform representation of the response on its boundaries. Then, with the modelling function learnt with the initial grid, the purpose is to find the missing configuration with that function, to restore a uniformly represented response.
We can either use directly the reciprocal function of our final regression function, or generate a set of values with that function and determine the closest ones to our wanted configuration. Once we have all the inputs needed, we have to adjust those values to a grid which will be added to the initial grid and used in the next iteration.

7 Discussion and perspectives

The first aim of this study was to build a model to obtained directly the 9 first eigenfrequencies of interest of a circular or rectangular beam in order to lately integrate this model in a larger loop which purpose is to classify carbon blocks whether there are susceptible to contained cracks or not, and this the more precisely possible. This has been achieved with success.
The second aim was to investigate the concept of Machine Teaching in the sense that we were searching for guidelines to minimize the number of input data required conserving a given accuracy goal. To illustrate the idea, we can think about Figure 1, and more particularly at the link between the quality and the quantity of data. Indeed, by structuring cleaver the distribution of the input data taking into account the produced output data, we strongly believe that we can reach higher levels of regression or classification ability while minimizing the number of initial computations required to feed the model. That is of first importance when the computation time last hours or days.
This study based on a simple model permitted to not be hampered by too long computational times in order to find the optimal strategy. The further work will consist in applying those guidelines on a more complex problem and to implement it for optimization purposes.

References

[1] Méndez-Sánchez, R. A., Morales, A., & Flores, J. (2005). Experimental check on the accuracy of Timoshenko's beam theory. *Journal of Sound Vibration*, *279*, 508-512.
[2] Mokhtari, A., Sarvestan, V., Ghayour, M., & Mirdamadi, H. R. Free vibration of undamped Euler-Bernoulli and Timoshenko beams by using spectral finite element method for different boundary conditions.
[3]Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.
[4] Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble machine learning* (pp. 307-323). Springer, Boston, MA.
[5] Kenneth Levenberg (1944). A Method for the Solution of Certain Non-Linear Problems in Least Squares.
[6] Michel Lalanne, Patrick Berthier, Johan Der Hagopian. (1984). Mechanical vibrations for engineers. *J. Wiley.*

A Appendix

The grid used is built with the following pattern:

$$
\begin{aligned}
&choose\ m_1\ points\ in\ I_1, such\ as\ , \left\{v_1^{(1)}, \dots, v_{m_1}^{(1)}\right\} \\
&\qquad\qquad\qquad\qquad \vdots \\
&choose\ m_N\ points\ in\ I_N, such\ as\ , \left\{v_1^{(N)}, \dots, v_{m_N}^{(N)}\right\} \\
&for\ i_1 = 1, \dots, m_1 : \\
&\quad \ddots \\
&\qquad for\ i_N = 1, \dots, m_N : \\
&\qquad\quad Add\ \ (v_{i_1}^{(1)}, \dots, v_{i_{N_1}}^{(N)})
\end{aligned}
$$

This kind of combination leads us to $\prod_{i=1}^{N} m_i$ samples. In this way of data generation, we can explore a domain study with a chosen discretization, which can be useful in our case. A second benefit of the grid, is the possibility to observe the response depending on only one feature, where all the others are fixed. We will use this ability for the visualization.

We start by unrolling the grid's construction operations, by fixing each feature, then we have a 2D graph with the response associated to the set $\left\{v_1^{(N)}, \dots, v_{m_N}^{(N)}\right\}$. Then, we interpolate the graph by a chosen function, with the nonlinear least squares algorithm, which gives us the best parameters to fit the interested values. After repeating this curve fitting for each configuration on the last loop, we interpolate the parameters of the chosen function with the set $\left\{v_1^{(N-1)}, \dots, v_{m_{N-1}}^{(N-1)}\right\}$. The process is done when all the features were involved in the learning process.

We can illustrate that decomposition with an example. Let's consider a problem with three real input features, *A*, *B* and *C,* and *F* as a real output feature. In order to make a grid, we choose arbitrary two points for each input, and by applying the previous building method, we get the following combinations (Table).

A	B	C	F
$\boldsymbol{A_1}$	B_1	C_1	F_1
$\boldsymbol{A_1}$	B_1	C_2	F_2
$\boldsymbol{A_1}$	B_2	C_1	F_3
$\boldsymbol{A_1}$	B_2	C_2	F_4
$\boldsymbol{A_2}$	B_1	C_1	F_5
$\boldsymbol{A_2}$	B_1	C_2	F_6
$\boldsymbol{A_2}$	B_2	C_1	F_7
$\boldsymbol{A_2}$	B_2	C_2	F_8

Table 12: One grid example

Then, we can apply the decomposition, which can be illustrated with a graph, like the one in Figure 8. The set of nodes corresponding to one feature is a step of the algorithm, which can also be described as a feature layer. Only the last layer will proceed the response, corresponding to the first step of the method. Then all the layers will deal with the parameters of the function fitted in the layer before.

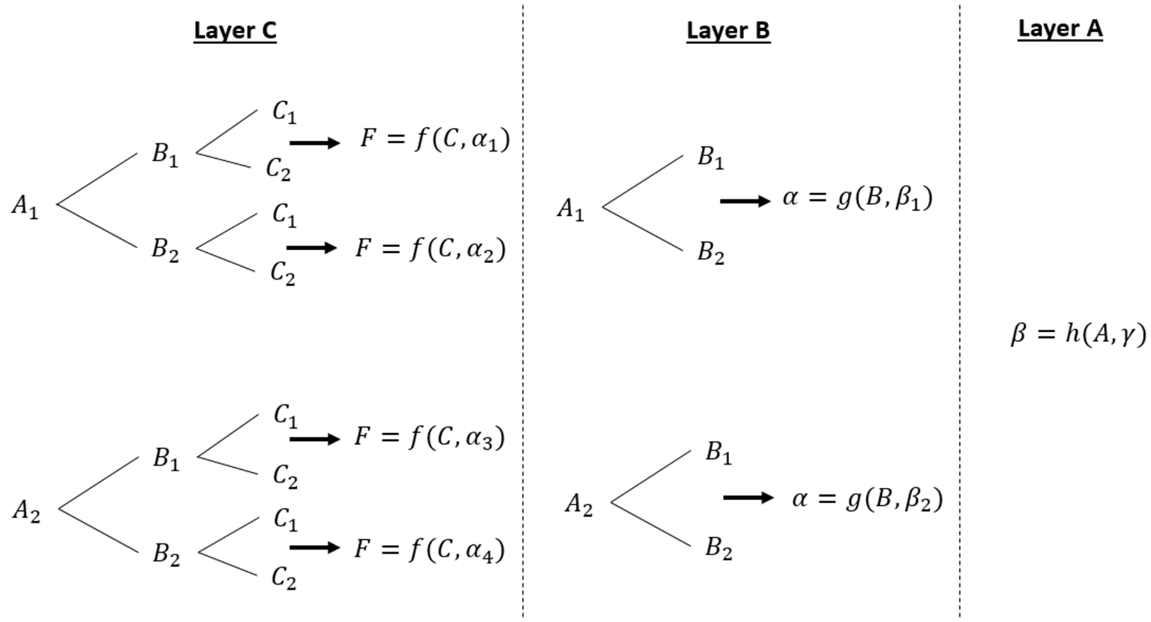


Figure 8: Graph of each layer of the example

The purpose is to manually select the best fitted function among a lot of possibilities, in terms of intrinsic properties such as convexity, asymptotic behavior or the level of complexity. The construction of the final modelling function is done by a composition of all the functions involved during each feature's layer. In our example, the final regression function would be:

$$\tilde{F}\,(A, B, C) = f\big(C, g\big(B, h(A, \gamma)\big)\big) \tag{5}$$

The interest of this method, is firstly the high complexity of the final function, often with several input features, where we adjust several functions on a 2D graph. It also decomposed the training phase in several steps, where a human can control the level of accuracy. The possibility of choosing the modelling function at each layer for each parameter, depending of the tendency followed by the values is a determinant way to obtain maximum precision with a highly restricted data set.

Effects of Particle Swarm Optimization Algorithm Parameters for Structural Dynamic Monitoring of Cantilever Beam

Xiao-Lin LI[1,2], Roger SERRA[1,2] and Julien OLIVIER[1,3]

[1] Institut National des Sciences Appliquées Centre Val de Loire, Blois France,
[2] Laboratoire de Mécanique Gabriel Lamé - EA 7494, Blois, France
[3] Laboratoire d'Informatique Fondamentale et Appliquée de Tours, EA 6300, Blois, France

Abstract

Nowadays, particle swarm optimization (PSO) algorithm has become a widespread optimization method. However, it is well known that its main parameters (inertia weight, two learning factors, velocity constraint and population size) have a critical effect on its performance. Currently the effects of PSO parameters on structural health monitoring have not been comprehensively studied. Therefore, in this paper, the PSO algorithm is used for damage detection assessment of a cantilever beam, and the simulation results are used to analyze the effects of PSO parameters. There are five levels for each parameter in our experiment, mean fitness value and success rate for each level are used as criteria to measure the convergence and stability of the PSO algorithm. Considering the effect of population size on CPU time, a trade-off strategy is presented to further determine the selection of population size.

1 Introduction

Over the last few years, there have been increasing demands to develop structural dynamic monitoring system over different kinds of aerospace, mechanical and civil engineering structures because of the huge economic and life-safety benefits. Vibration testing is the widest used method for structural damage detection[1, 2]. The main idea behind damage detection techniques based on structural dynamic changes is the fact that the modal parameters of a structure are functions of the physical parameters (such as mass, stiffness and damping) thus the existence of damage leads to changes in the modal properties of the structure. The inverse method for damage detection using vibration data and solving by optimization algorithms have received extensive attention in recent years. The usual approach is to minimize an objective function, which is defined in terms of discrepancies between the predicted model parameters and the initial model parameters. Using classic optimization methods to solve it often meet some difficulties. However, the particle swarm optimization (PSO) algorithm can be used on complicated optimization problems that cannot be expressed explicitly. PSO algorithm is one of the newest intelligent method, this parallel evolutionary computation technique was developed by Kennedy and Eberhart in 1995[3].The basic idea comes from the study of group behaviors such as predation of birds. PSO algorithm was first used for function optimization and neural network training[3]. Since the algorithm has many advantages such as comparative simplicity, easy to implement and few parameters to be adjusted, PSO has found its application in many complex engineering optimization problems, including structural damage detection of beam structure[4, 5].

It is well known that in various optimization methods, parameters is one of the key factors which have a great effect on the performance. For different kinds of optimization problems, the matching and cooperation modes between parameters are different. Even for the same type of optimization problem, if problem scales are different, parameter selections are not completely the same. Although PSO algorithm has few parameters to adjust, how to determine them is also an important problem. However, currently the effects of PSO parameters on structural damage detection have not been comprehensively studied. In this paper, the damage detection of a cantilever beam by PSO algorithm for two damage patterns are simulated, and the experimental results are analyzed to study the effects of PSO parameters. There are five levels for each parameter in our experiment, the mean fitness value and success rate for each level are used as criteria to measure the convergence and stability of the PSO algorithm. Considering the effect of population size on CPU time, a *Ratio* compromise strategy is proposed to further determine the selection of population size.

2 Problem formulation

2.1 Structural dynamic finite element formulation

The governing equation for an Euler-Bernoulli beam with negligible damping is given by:

$$\frac{\partial^2}{\partial x^2}\left[EI(x)\frac{\partial^2\omega(x,t)}{\partial x^2}\right]+m(x)\frac{\partial^2\omega(x,t)}{\partial t^2}=f(x,t) \quad (1)$$

where $EI(x)$ denotes the flexural stiffness, $m(x)$ denotes the mass per unit length of the beam, $w(x,t)$ represents the transverse displacement of the beam. The beam is discretized into a number of elements, with displacement and slope as nodal degrees of freedom and cubic interpolation function.For an n-degree of freedom system of uniform beam, the stiffness matrix and the consistence mass matrix are given in [6], respectively.

The characteristics of the beam are given in the table below. The total mass is 3.237 kg.

Young modulus E (Pa)	Poisson ratio ν	Density ρ $(Kg.m^{-3})$	Length L (m)	Width b (m)	Thickness h (m)
2e+11	0.33	7850	1	5.3e-3	2.5e-2

Table 1: Beam properties

For a properly modeled structure, the structural dynamic eigenvalue equation is given by:

$$(K-\lambda M)\Phi=0 \quad (2)$$

where K and M are the global stiffness and mass matrices respectively, and λ and Φ represent the natural frequency and vibration mode shape vectors. It can be assumed that when a structural modification occurs, the local stiffness of the structure changes whereas the change in mass may be neglected. Hence, the equation (2) could be rewritten for a modified system as:

$$(K_d-\lambda_d M)\Phi_d=0 \quad (3)$$

where K_d and M are the global modified stiffness and mass matrices respectively, and λ_d and Φ_d represent the new natural frequency and vibration mode shape vectors for the modified structure respectively. In many studies on structural health monitoring, the structural modification has been simulated by decreasing one of the local element 's stiffness parameters like a inertia moment I, cross sectional area S or elasticity (Young) modulus E. In this work, the structural modification of each element was simulated using Young modulus reduction factor (also called stiffness reduction factor) x as a scalar variable between $[0,1]$ where zero value corresponds to no modification and a value near to one corresponds to failure condition introduced as follows[5]:

$$x_i=\frac{E-E_i}{E} \quad (4)$$

where E is the initial Young modulus and E_i is the reduced Young modulus of the i^{th} element. In this case, the stiffness matrix K will be modified as:

$$K_d=\sum_i(1-x_i)K_i \quad (5)$$

2.2 Optimization problem

Equation (3) forms the basis of the structural modification detection method through an inverse procedure giving the new vibration natural frequencies and the mode shapes. As the structural modification causes change in vibration natural frequencies and which are easier to measure than mode shapes (limited number of accelerometers) and the error associated is comparatively less. Hence, they are used as structural modification indicators in this study. PSO algorithm is used to search a particular stiffness reduction factor x so that the predicted numerical natural frequencies exactly match with the initial natural frequencies. The initial numerical model of the structure is generally considered for the optimization. When the exact match between modified and initial natural frequencies is observed, the value of stiffness reduction factor represents the actual modification location and amount. The usual approach to solve the inverse problem of structural dynamic monitoring involves minimization of the fitness function (or objective function) which is defined in terms of discrepancies between the predicted natural frequencies and the initial natural frequencies. In this study, the fitness function can be presented just like in literature [7]:

$$F = \sum_{s=1}^{n} \left(\frac{(f_s^r)^2 - (f_s^p)^2}{(f_s^r)^2} \right)^2 \tag{6}$$

where f_s^r and f_s^p are the initial and predicted natural frequencies respectively. n is the number of input response parameters chosen (natural frequencies) and for this study is taken as five.

2.3 Particle swarm optimization algorithm

The particle swarm optimization technique is a population based stochastic technique in nature (bio-inspired) so-called evolutionary computational model which is based on swarm intelligence. PSO is developed by Kennedy and Elberhart [3] and primarily used to tackle continuous optimization problems. The system is initialized firstly in a set of randomly generated potential solutions, and then performs the search for the optimum one iteratively by swarms following the best particle. Compared to others evolutionary algorithms, PSO has much more profound intelligence background and could be performed more easily. Based on its advantages, the PSO is suitable for engineering applications, in the fields of evolutionary computing, optimization and many others.

As suggested in literature [8], a fully connected topology is elected as PSO algorithm topology. Set for the D-dimensional search space, mth particles compose a population $\{X_1, X_2, \ldots, X_m\} \subset R^n$, and the ith particle position is $X_i = (x_{i1}, x_{i2}, \ldots, x_{iD})^T$, the velocity of the particle X_i can be represented by another D-dimensional vector $V_i = (v_{i1}, v_{i2}, \ldots, v_{iD})^T$. The best position previously visited of the particle X_i is denoted as $P_i = (p_{i1}, p_{i2}, \ldots, p_{iD})^T$, and the best position among all particles in the population is $P_g = (p_{i1}, p_{i2}, \ldots, p_{iD})^T$. Each particle adjusts its position dynamically according to the principle of following the current optimal particle, the particle X_i updates its speed and position according to (7) and (8).

$$v_{id}^{t+1} = \omega v_{id}^t + c_1 r_1 (p_{id}^t - x_{id}^t) + c_2 r_2 (p_{gd}^t - x_{id}^t) \tag{7}$$

$$x_{id}^{t+1} = x_{id}^t + v_{id}^{t+1} \tag{8}$$

where t is iteration time, d represents the dimension of the particle, $d = 1, 2, \ldots, D$, i represents the number of the particle, $i = 1, 2, \ldots, m$. r_1, r_2 are random between 0 and 1, ω is the inertia weight, c_1 and c_2 are the learning factors to adjust each iteration step length.

many literatures has emphasized the importance of ω, the linear decreasing inertia weight ω^t has been widely used[9], which is defined as follow:

$$\omega^t = \omega_{max} - \frac{\omega_{max} - \omega_{min}}{t_{max}} \times t \tag{9}$$

where t_{max} is the maximum number of iteration. In this strategy, ω^t changes with iteration. At beginning, the value of ω^t is ω^{max} , and ω^t decrease during the execution of the algorithm. At the end the value of ω^t is ω_{min}. ω_{max} and ω_{min} are set to 0.9 and 0.4 respectively.

3 Analysis of parameters on the algorithm performance

In the experiment, a steel cantilever beam is considered for structural damage detection. Figure 1 shows the sketch of the beam with element number using in the finite element simulations, 30 equal Euler-Bernoulli beam elements are chosen for finite element modeling.

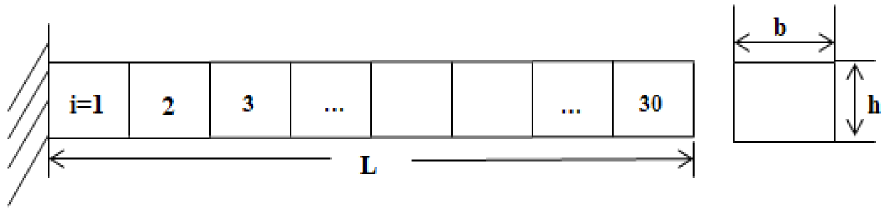


Figure 1: The Euler-Bernoulli cantilever beam model

PSO algorithm has some important parameters, such as population size (m), inertia weight (ω), two learning factors (c_1, c_2) and maximum velocity ($Vmax$). The effects of PSO parameters are analyzed by simulating structural damage detection of a cantilever beam. 200 iterations and 1000 runs are set for two damage patterns in Table 2.

Damage Pattern I		Damage Pattern II	
Element	Damage(%)	Element	Damage(%)
5	10	5	10
		7	10

Table 2: Simulated damage patterns in cantilever beam

Table 3 shows the parameter setting. The value of $Vmax$ is set to $Vmax = \gamma Xmax$ and $\gamma \in (0, 1]$, $Xmax$ denotes the dynamic range of the variable on each dimension. In each test, there are five levels for each parameter, mean fitness value is equal to the average value of fitness function F, success rate is equal to the ratio of the number of successful runs to total number of runs, which are all considered as criteria for parameter performance measurement. When fitness function provides smaller mean fitness value it shows better convergence performance of the algorithm. The higher success rate provides the stronger stability of the algorithm. The convergence and stability for each parameter at five levels are compared.

Table 3: parameter setting for five levels

level	population size m	inertia weight ω	learning factors (c_1, c_2)	maximum velocity $Vmax$
1	20	0.25	$(0,4)$	$0.2Xmax$
2	40	0.5	$(1,3)$	$0.4Xmax$
3	60	0.75	$(2,2)$	$0.6Xmax$
4	80	1.0	$(3,1)$	$0.8Xmax$
5	100	ω^t	$(4,0)$	$Xmax$

3.1 Effect of population size m

The choice of population size is related to the complexity of the problem. As the complexity of the problem increases, the population size also grows. The five levels of population size are given in Table 3, the other parameters are chosen as $\omega = \omega^t, c_1 = c_2 = 2, Vmax = Xmax$ [9]. Mean fitness value and success rate for two damage patterns are shown in Figure 2. It is clearly that along with the increase of population size m, the convergence and stability of the PSO algorithm is becoming stronger and stronger.

However, for a given problem, the parameters that affect CPU time are mainly finite element number, population size, maximum iterations and number of PSO runs, they are given except for population size. Therefore, when the effect of population size are analyzed, CPU time needs to be taken into consideration. The larger the population size represents the longer CPU time. Then, a trade-off strategy (10) is proposed as a criterion to further determine the population size m. Obviously, a larger $Ratio$ means a better performance with selected population size. Thus, it can be seen in Figure 3 that for two damage patterns,the optimal choices are $m = 20$

and $m = 60$,respectively.

$$Ratio = \frac{\text{success rate}}{\text{population size}} \tag{10}$$

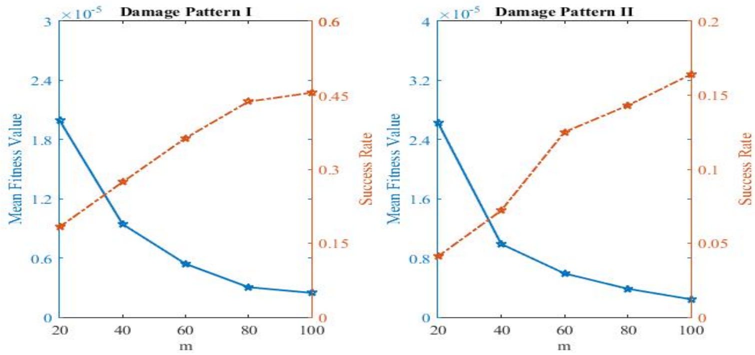


Figure 2: Effect of population size on PSO algorithm for two damage patterns

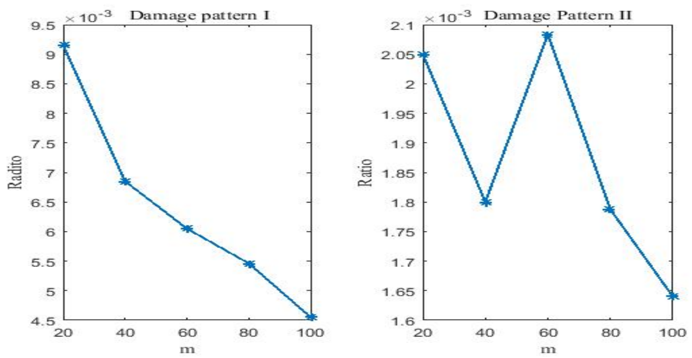


Figure 3: Selection of population size for two damage patterns

3.2 Effect of inertia weight ω

The inertia weight ω affects the particle's global and local search ability. The stochastic process theory in [10] shows that the range of ω is $[0,1]$. From the above,the best setting for population size are: $m = 20$ for damage pattern I, $m = 60$ for damage pattern II. And $c_1 = c_2 = 2, Vmax = Xmax$. In order to examine the balance between global and local exploration, the five levels of inertia weight are compared, and the simulation results are shown in Figure 4.

When ω is small, PSO algorithm hardly converges and the success rate is low. Along with the increase of ω, PSO algorithm has a better convergence and stability. Although a better convergence and stability can also been obtained when ω takes linear-decreasing strategy, the optimal choice is $\omega = 1$ which is different from the general linear-decreasing strategy.

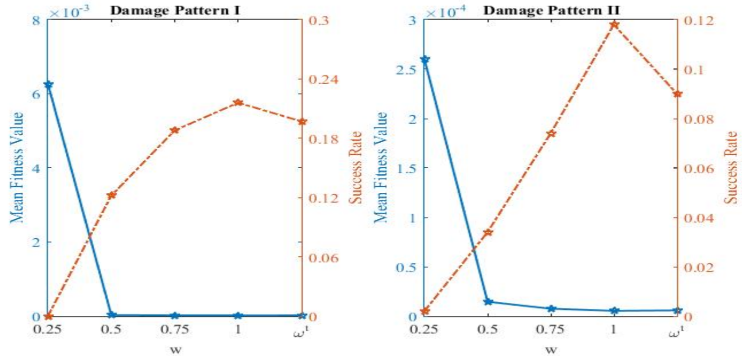


Figure 4: Effect of inertia weight on PSO algorithm for two damage patterns

3.3 Effect of learning factors c_1, c_2

A general rule for setting the two learning factors is $c_1 + c_2 <= 4$, there is a close relationship between c_1 and c_2. Therefore, the values of the two learning factors are considered simultaneously. Considering the simplicity of the experiment, the following relation between c_1 and c_2 will be used:$c_1 + c_2 = 4$, the five levels of the learning factors are compared with $\omega = 1, Vmax = Xmax$, $m = 20$ for damage pattern I, $m = 60$ for Damage Pattern II.

From figure 5, it can be seen that $(c_1, c_2) = (3, 1)$ is the optimal choice for the convergence and stability of PSO algorithm. That means, for structural damage detection of the cantilever beam, the algorithm shows a good performance when the population put much more attracted to the best location found by itself.

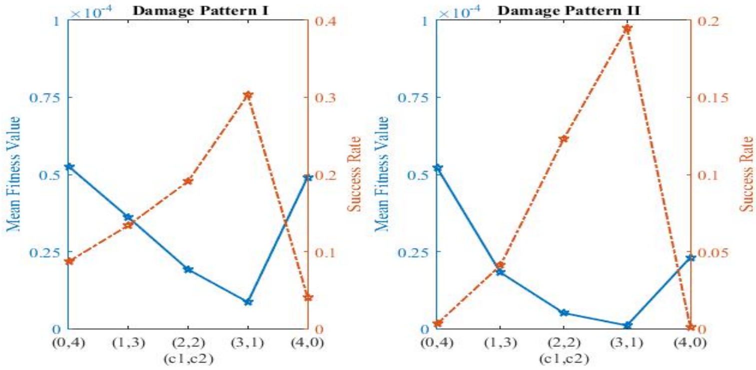


Figure 5: Effect of leaning factors on PSO algorithm for two damage patterns

3.4 Effect of maximum velocity $Vmax$

The velocity of the particles can be limited to $[-Vmax, Vmax]$ by a maximum velocity, which acts as a constraint to control the global exploration capability of the population, It is clearly that $\omega = 1, c_1 = 3, c_2 = 1$,$m = 20$ for damage pattern I,$m = 60$ for damage pattern II are optimal choices for structural damage detection. Then, the convergence and stability of $Vmax$ for the five levels are compared.

The simulation results for mean fitness value and success rate are shown in Figure 6. Along with the increase of the maximum velocity, mean fitness value is decreasing and success rate is increasing, which means the convergence and stability of the algorithm is becoming stronger and stronger. Usually set $Vmax$ as a constant, $Vmax = Xmax$ is the best choice for structural damage detection of the cantilever beam.

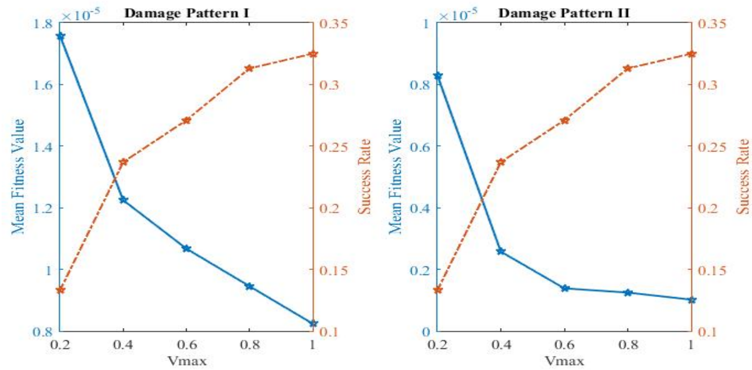


Figure 6: Effect of maximum velocity on PSO algorithm for two damage patterns

4 Validation

From the above paper, the optimal parameter configuration for damage pattern I is $(m, \omega, c_1, c_2, Vmax) = (20, 1, 3, 1, Xmax)$, at this time, mean fitness value $= 8.60e-06$, success rate $= 0.36$. For damage pattern two is $(m, \omega, c_1, c_2, Vmax) = (60, 1, 3, 1, Xmax)$, the corresponding mean fitness value and success rate are $9.78e-07$

and 0.18, respectively. Under the optimal configuration, PSO shows a better convergence and stability than other configurations used in our experiment by using relatively low time costs. Therefore, the effectiveness is verified by simulating structural damage detection of cantilever beam.

5 Conclusion

In this paper, first of all, the five levels for each parameter are designed to perform damage detection of cantilever beam. Then mean fitness value and success rate obtained from simulation results are used as criteria to evaluate the convergence and stability of the algorithm. Considering CPU time, the *Ratio* strategy is proposed to further determine the selection of population size. A parameter guideline are given for structural damage detection of cantilever beam.

References

[1] S. W. Doebling, C. R. Farrar and M. B. Prime, *A Summary Review of Vibration-Based Damage Identification Methods*, The Shock and Vibration Digest, 1998.

[2] Y.J. Yan, L. Cheng, Z.Y. Wu and L.H. Yam, *Development in Vibration-based Structural Damage Detection Technique*, Mechanical Systems and Signal Processing, 2007.

[3] J. Kennedy and R.C. Eberhart, *Particle swarm optimization*, Proceedings of IEEE international conference on neural networks, 1995.

[4] B. Nanda, D. Maity and D. K. Maiti,*Vibration Based Structural Damage Detection Technique using Particle Swarm Optimization with Incremental Swarm Size*, International Journal of Aeronautical and Space Sciences, 2012.

[5] G. Ghodrati Amiri, A. Zare Hosseinzadeh and S.A. Seyed Razzaghi, *Generalized Flexibility-based Model Updating Approach via Democratic Swarm Optimization Algorithm for Structural Damage Prognosis*, International Journal of Optimization in Civil Engineering, 2015.

[6] J. Lee, *Identification of Multiple Cracks in a Beam Using Natural Frequencies*, Journal of Sound and Vibration, 2009.

[7] L. Rubio, J. Fernández-Sáez and A. Morassi. *Identification of Two Cracks with Different Severity in Beams and Rods from Minimal Frequency Data*, Journal of Vibration and Control, 2014.

[8] A. Carlisle and G. Dozier, *An Off-the-shelf PSO Proceedings of the Workshop on Particle Swarm Optimization*,Proceeding of the 2001 Workshop on Particle Swarm Optimization, Indianapolis IN, 2001.

[9] R. Perera, S-E Fang and A Ruiz, *Application of Particle Swarm Optimization and Genetic Algorithms to Multiobjective Damage Identification Inverse Problems with Modelling Errors*, Information Processing Letters, 2007.

[10] M. Jiang, Y. P. Luo and S. Y. Yang, *Stochastic Convergence Analysis and Parameter Selection of the Standard Particle Swarm Optimization Algorithm*, Information Processing Letters, 2007.