Domain knowledge informed unsupervised fault detection for rolling element bearings

Douw Marx[1,2] and Konstantinos Gryllias[1,2]

[1] *Division Mecha(tro)nic System Dynamics, Department of Mechanical Engineering, KU Leuven*
douw.marx@kuleuven.be
konstantinos.gryllias@kuleuven.be

[2] *Dynamics of Mechanical and Mechatronic Systems, Flanders Make*

ABSTRACT

Early and accurate detection of rolling element bearing faults in rotating machinery is important for minimizing production downtime and reducing unnecessary preventative maintenance. Several fault detection methods based on signal processing and machine learning methods have been proposed. Particularly, supervised, data-driven approaches have proved to be very effective for fault detection and diagnostics of rolling element bearings. However, supervised methods rely heavily on the availability of failure data with volume, variety and veracity, which is mostly unavailable in industry. As an alternative data-driven strategy, unsupervised methods are trained on healthy data only and do not require any failure data.

In contrast to supervised and un-supervised data-driven models, physics-based and phenomenological models are based on domain knowledge and not on historical data. Although these models are useful for studying the way in which damage is expected to manifest in a measured signal, they are difficult to calibrate and often lack the fidelity required to model reality. In this paper, an unsupervised data-driven anomaly detection method that exploits informative domain knowledge is proposed. Hereby, the versatility of unsupervised data-driven methods are combined with domain knowledge.

In this approach, supplementary training data is generated by augmenting healthy data towards its possible future faulty state based on the characteristic bearing fault frequencies. Both healthy and augmented squared envelope spectrum data is used to train an autoencoder model that includes regularisation designed to constrain the latent features at the autoencoder bottleneck. Regularisation in the autoencoder loss enforces that the expected deviation of the healthy latent representation towards the augmented latent representation at damaged conditions, is constrained to be maximally different for different fault modes. Consequently, the likelihood of a new test sample being healthy can be evaluated based on the projection of the sample onto an expected failure direction in the latent representation.

A phenomenological and experimental dataset is used to demonstrate that the addition of augmented training data and a specialized autoencoder loss function can create a separable latent representation that can be used to generate interpretable health indicators.

1. INTRODUCTION

1.1. Background on condition monitoring approaches

Condition-based maintenance procedures can help machines operate reliably and continuously by reducing unnecessary maintenance procedures (Lei et al., 2018), and minimizing machine downtime (Lee et al., 2014). Rolling element bearings act as a main source of faults in rotating machinery (Cerrada et al., 2018), and have consequently drawn significant research attention in the condition-based maintenance community. Although impressive fault detection and classification results have been attained using data-driven methods for fault detection in bearings (See Hoang & Kang (2019)), a large majority of these approaches are based on sophisticated supervised learning techniques that require failure data during training.

In contrast, many physics-based, signal processing and unsupervised methods attempt to solve the bearing fault detection problem without the requirement of any fault data. However, these methods bring other challenges. Signal processing methods (Randall & Antoni, 2011) are robust, simple and effective, but often require an expert to interpret the results. Physics-based methods (Cao et al., 2018) and phenomenological (McFadden & Smith, 1984) methods are difficult to design, calibrate and lack the flexibility required to model re-

ality. Finally, unsupervised methods lack interpretability and can identify non-fault-related anomalies as machine faults.

To address the respective limitations of these approaches, hybrid methods have been proposed for diagnostics (Leturiondo et al., 2017) and prognostics (Liao & Kottig, 2014) in order to combine the benefits of physics-based, data-driven, domain knowledge and/or signal processing methods. For example, researchers have designed physics-inspired filters for convolutional neural networks (CNNs) (Sadoughi & Hu, 2019). In this work, the parameters of the convolutional kernels are chosen such that respective kernels are sensitive to different bearing faults manifesting in the envelope signal. Others (Liu et al., 2020) have used transfer learning approaches to learn domain invariant features between measured data, and simulated data, allowing improved remaining useful life prediction and a reduced reliance on real world data. Physics-based knowledge was also included in a CNN (Shen et al., 2021) by adding a penalisation to the network loss if predictions are made that are not compatible with expected bearing fault behaviour.

In the context of bearing fault detection, it could even be argued that the use of convolutional layers in neural networks Jiao et al. (2020), that extract translation-invariant fault patterns in a signal, can be viewed as the addition of domain knowledge into data-driven methods.

In this paper, we propose that unsupervised latent variable models have the potential to facilitate the incorporation of domain knowledge into a fault detection problem. To demonstrate this idea, the loss function of an autoencoder (AE) model is regularized such that, when applied to unseen faulty data, the AE healthy latent representation deviates in a known latent failure direction corresponding to a given fault mode. Ultimately this designed latent space behaviour is then useful for constructing sensitive and interpretable machine health indicators.

In previous work, specialized loss functions have been used to manipulate latent representations for improving supervised classification tasks Li et al. (2018), latent representations have been visualized for different fault modes Booyse et al. (2020), and have been successfully used for constructing informative machine health indicators Balshaw et al. (2022). In other works, augmented training data has been used for data-efficient bearing diagnostics Yu et al. (2021). However, the opportunity of using augmented data, derived from domain knowledge, to shape the latent feature space of an AE with the goal of creating informative and interpretable health indicators has not been widely studied. Therefore, this work intends on making the following contributions.

1.2. Contributions

- Augmented data, as derived from healthy data through a modification of the healthy data towards its expected faulty behaviour, is used as supplementary data for training an AE.
- An AE model is regularized to incorporate domain knowledge into health indicators based on changes in the model's representation with increasing fault severity.
- An interpretable latent representation with diagnostics information is created by including domain knowledge conveyed through augmented data.
- A framework is provided for dealing with the discrepancy between real failure data and a model of the expected failure behaviour.
- The method is applied to a simulated, and experimental dataset.

The remainder of the paper is structured as follows. Section 2 introduces the proposed method, explaining the data preparation, training and evaluation procedures respectively. Results are then presented for a simulated dataset in Section 3 and for the NASA IMS bearing dataset in Section 4. Finally, conclusions and future work are presented in Section 5.

2. METHOD

In this section, a method for incorporating domain knowledge into an unsupervised latent variable model is introduced. Specifically, domain knowledge about bearing fault frequencies is incorporated into an AE model. The intention is to use domain knowledge informed augmented data, as derived from healthy data, to shape the latent space of the AE. As a result, the latent representation of the model should have desirable properties for extracting informative health indicators.

The methodology can be divided into three main parts (See Figure 1). During data preparation (Section 2.1), raw accelerometer signals are processed and healthy data is augmented towards its expected faulty state. Thereafter, the healthy and the augmented data are used to train an AE (Section 2.2). Regularisation of the AE, enabled by including augmented data in the training procedure, enforces that the healthy latent distribution of the AE should deviate in a specific direction as the severity of the fault increases. Finally, during the evaluation phase (Section 2.3), a health indicator is computed for each of the anticipated fault modes to assess the likelihood of a new sample being healthy or faulty, given a particular fault mode.

2.1. Data preparation

In this work, the Squared Envelope Spectrum (SES) is used as input feature to an auto-encoder since it is simple to modify the healthy envelope spectrum towards an expected damaged

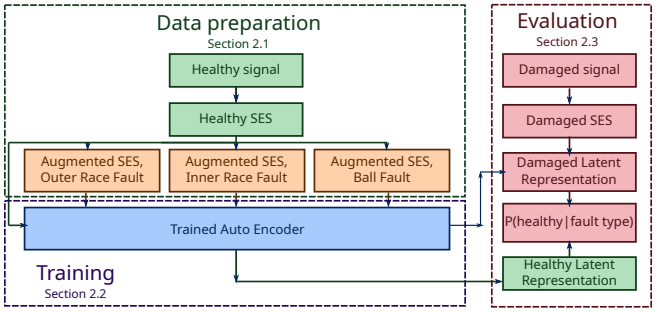


Figure 1. Overview of methodology.

condition. The process of modifying a healthy datapoint towards its expected faulty condition, for a given failure mode, is viewed as data augmentation in this investigation.

Healthy data, $\mathbf{x}_{\text{healthy}}$, is modified by adding a modification signal $\mathbf{x}_{\text{modify}}$ to acquire the augmented data, $\overline{\mathbf{x}}^{(i)}$ for a given fault mode (i).

$$\overline{\mathbf{x}}^{(i)} = \mathbf{x}_{\text{healthy}} + \mathbf{x}^{(i)}_{\text{modify}} \tag{1}$$

Particularly, the amplitude of the healthy squared envelope spectrum at the fault frequency and its first two harmonics corresponding to a given fault mode is increased by $\mathbf{x}_{\text{modify}}$. It should be noted that there are many different ways of achieving this data augmentation, including using the faulty response of a phenomenological model or lumped parameter model or even asking an expert to draw the expected fault behaviour on a graph. In this investigation, simple triangular peaks are added to the expected fault frequency and its harmonics.

A triangular peak x_{peak} as a function of SES frequency f, is defined as:

$$\mathrm{x}_{\text{peak}}\left(f, f_c\right) = \begin{cases} -\left|\frac{-2a}{w}\left(f - f_c\right)\right| + a & \text{if } -\frac{w}{2} \leq f - f_c \leq \frac{w}{2} \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

with a the amplitude of the peak, w the base length of the triangle and f_c the centre frequency for which a high amplitude is associated with a fault.

The total modification signal is obtained by adding the peaks at the fault frequency, f_{fault}, and its harmonics nf_{fault}. The amplitude of the harmonics of the fault frequency decay with frequency and is controlled using the decay parameter α.

$$x_{modify}(f) = \sum_{n=1}^{N} e^{-\alpha f_{\text{fault}}(n-1)} x_{\text{peak}}\left(f, nf_{\text{fault}}\right) \tag{3}$$

For this investigation the number of peaks, N is selected as $N = 3$.

Figure 2 shows an example of a healthy envelope spectrum that was modified by adding triangular peaks at frequencies that are expected to correspond to an outer race fault. The healthy signal and the augmented signal are identical, apart from the sections where peaks were added at the fault frequencies.

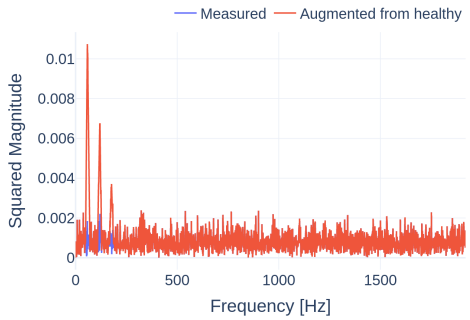


Figure 2. Example of healthy envelope spectrum augmented by adding peaks at expected fault frequencies for an outer race fault.

All healthy training data samples are augmented to additionally obtain an augmented sample for each anticipated fault mode. Both the healthy data and the augmented data are min-max normalized before training.

2.2. Training the auto encoder with specialised loss function

The training phase is applicable during the lifetime of a machine where the bearings are new, have been run in and are assumed to be in a healthy condition. During training, an auto-encoder is used to learn informative latent features from input data. Additionally, the latent space of the auto-encoder is regularised during training such that the augmented data is distributed in the latent space in a way that is beneficial for fault detection.

Specifically, two regularisation terms are used in the loss function in addition to the typical AE reconstruction loss. A first regularisation term enforces that the direction of deviation for the latent healthy data to the augmented data should be maximally different for different fault modes. This ensures that even if the augmentation of the data towards a failed state is not completely representative of reality, the movement of the latent features away from the healthy latent distribution will not be confused with that of another fault mode. This also leads to benefits when computing the projection of the latent representation of a new sample onto an expected failure direction as discussed in Section 2.3. A second regularisation term enforces that the distance from the healthy data cluster in the latent space to the respective augmented clusters should

be similar for different fault modes. This regularisation ensures that the latent space is not disproportionately scaled for a specific fault mode, and simultaneously encodes rudimentary fault severity information into the latent features of the AE since the latent representation of a faulty sample can be compared to that of the augmented samples.

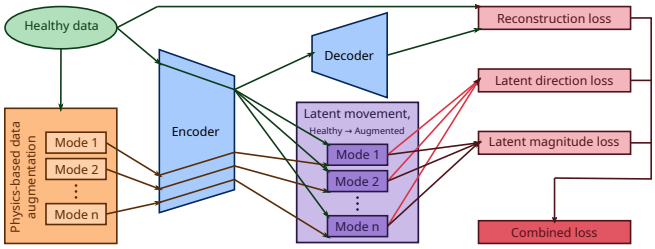


Figure 3. Training Methodology: During forward propagation of the autoencoder, both healthy data and augmented data are separately fed through the encoder. A latent direction loss and latent magnitude loss regularize the latent features whist a reconstruction loss acts on the decoder output.

Figure 3 shows a diagram explaining the training procedure. During the forward propagation step of training the AE, both the healthy data, and the augmented data for each respective fault mode, are fed through the network separately. The healthy data is fed through both the encoder and the decoder, whilst the augmented data for each respective mode are fed through the encoder only. This is since the latent representation of the augmented data is used only to constrain the latent representation, whilst the reconstructed healthy data is additionally used for computing the conventional AE reconstruction error. After forward propagation, the loss is computed for subsequent backpropagation and the update of the model weights.

The loss function used during training is now described for a randomly selected healthy training example $\mathbf{x}$ and a randomly selected augmented sample $\overline{\mathbf{x}}^{(i)}$, associated with an expected fault mode (i). The combined loss, $\mathcal{L}$ is written as the sum of a reconstruction loss $\mathcal{L}_{\text{reconstruct}}$, acting on the output of the decoder, and the latent feature loss $\mathcal{L}_{\text{latent}}$ acting on the latent representation of the healthy and augmented data.

$$\mathcal{L}\left(\mathbf{x}\right) = \mathcal{L}_{\text{reconstruct}}\left(\mathbf{x}\right) + \mathcal{L}_{\text{latent}}\left(\mathbf{x}, \overline{\mathbf{x}}^{(i)}\right) \tag{4}$$

The mean squared error loss is used as the reconstruction loss as is common in conventional AEs. The purpose of the reconstruction loss is to enforce that low dimensional, informative features are learnt in the latent representation such that the original input can be reconstructed from the latent representation. The reconstruction error is written as:

$$\mathcal{L}_{reconstruct}\left(\mathbf{x}\right) = \left(\mathbf{x} - g(h(\mathbf{x}))\right)^2, \tag{5}$$

where h and g represent the encoder and decoder of the AE respectively.

The combined loss in Equation 4 further includes a regularisation loss $\mathcal{L}_{\text{latent}}$, that acts on the latent features. This latent loss function consists of two parts, namely $\mathcal{L}_{\text{direction}}$ and $\mathcal{L}_{magnitude}$.

$$\mathcal{L}_{\text{latent}}\left(\mathbf{x}\right) = \lambda_1 \mathcal{L}_{\text{direction}}\left(\mathbf{x}\right) + \lambda_2 \mathcal{L}_{magnitude}\left(\mathbf{x}\right) \tag{6}$$

The direction loss enforces that the direction in which a healthy latent cluster is expected to move towards the augmented latent clusters should be maximally different for different fault modes. The magnitude loss ensures that the latent representation is not skewed with respect to a specific latent fault mode. The regularisation hyperparameters, λ_1 and λ_2 scale the importance of the respective loss terms and can be selected based on the loss terms calculated from a validation set.

The magnitude and direction losses are now defined. To simplify the definition of these terms, we introduce $\delta^{(i)}\left(\mathbf{x}\right)$; the difference between the latent encoding of a healthy sample and the latent encoding of an augmented sample for a given mode (i). Furthermore, let $\mathbf{z} = h\left(\mathbf{x}\right)$ represent the latent representation of data fed through the encoder of the AE.

$$\begin{aligned}\delta^{(i)}\left(\mathbf{x}\right) &= h\left(\overline{\mathbf{x}}^{(i)}\right) - h\left(\mathbf{x}\right) \\ &= \overline{\mathbf{z}}^{(i)} - \mathbf{z}\end{aligned} \tag{7}$$

The latent movement direction loss is driven by the dot product between the unit vectors of $\delta^{(i)}\left(\mathbf{x}\right)$ for each fault mode (i). By minimizing the dot project between two vectors, we enforce that the unit vectors of $\delta^{(i)}\left(\mathbf{x}\right)$ and $\delta^{(j)}\left(\mathbf{x}\right)$ for fault modes (i) and (j) are pointing in opposite directions. In this work, the unit vectors of $\delta^{(i)}\left(\mathbf{x}\right)$ are referred to as expected fault directions. The direction loss acting on the latent representation is written as:

$$\mathcal{L}_{direction}\left(\mathbf{x}\right) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\delta^{(i)}\left(\mathbf{x}\right)}{\|\delta^{(i)}\left(\mathbf{x}\right)\|} \cdot \frac{\delta^{(j)}\left(\mathbf{x}\right)}{\|\delta^{(j)}\left(\mathbf{x}\right)\|}, \tag{8}$$

adding the unit vector dot products between fault modes for all n expected fault modes.

Finally, the latent movement magnitude loss enforces that the latent augmentation clusters for each mode are equally far from the healthy latent distribution. This ensures that the latent representation is not more sensitive to movement in one

failure direction as compared to the others. Additionally, this loss term ensures that the latent representation does not collapse to a single point with healthy and augmented data in the same location. The latent movement magnitude loss is given as

$$\mathcal{L}_{magnitude}\left(\mathbf{x}\right) = \sum_{i=1}^{n} \left(\| \delta^{(i)}\left(\mathbf{x}\right) \| - 1 \right)^2 . \tag{9}$$

For each batch in the training dataset, the combined loss can be computed and the weights of the autoencoder can be updated by an optimisation algorithm relying on gradients from backpropagation.

2.3. Evaluation of new samples

After the network has been trained, health indicators can be computed for new samples to assess if a fault is present in the bearing and to determine the fault mode by which the bearing is likely failing. A diagram of the evaluation method is shown in Figure 4.

The goal of the evaluation procedure is to evaluate the likelihood that a new sample is still healthy for a given fault mode. To do this, the latent representation of a new sample is projected onto one of the expected failure directions in the latent space. The likelihood of the projected sample can then be estimated based on the distribution of the healthy latent representation as projected onto the same failure direction. Although the direction loss term (Equation 8) enforced that fault directions should be maximally different during training, these fault directions are not explicitly known after optimisation and should be computed.

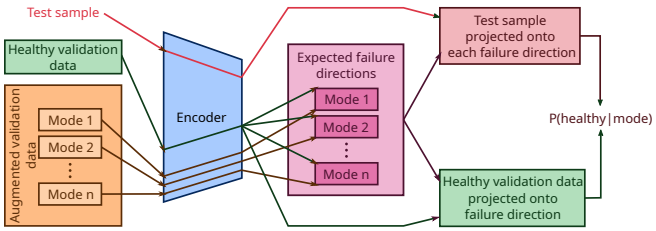


Figure 4. Evaluation method: After computing the expected failure directions in the latent feature space, a new sample is projected onto each of the failure directions so that its likelihood of having failed by a given failure mode can be evaluated.

This is done by computing the expected fault directions from the latent representation of a validation set. The fault directions $\mathbf{v}^{(i)}$ are computed as the normalized unit vector that passes through both the median of the healthy validation data latent representation and the median of the augmented data latent representation for a given failure mode. The latent representation of a new sample from a bearing failing by a given failure mode is expected to deviate from the healthy distribution in a similar direction than the expected failure direction $\mathbf{v}^{(i)}$. This is since the directions in the latent space were designed to correspond to a given failure through the inclusion of augmented data during the training procedure.

The fault directions $\mathbf{v}^{(i)}$ is given as

$$\mathbf{v}^{(i)} = \text{med} \left\{ \frac{\delta^{(i)}\left(\mathbf{x}\right)}{\| \delta^{(i)}\left(\mathbf{x}\right) \|} \text{for all } \mathbf{x} \text{ in validation set} \right\}. \tag{10}$$

With the expected fault directions $\mathbf{v}^{(i)}$ calculated, the scalar projection of the latent representation in an expected fault direction $\mathbf{v}^{(i)}$ can be computed. This is done by taking the dot product between the latent representation of a sample $\mathbf{z}$ and an expected failure direction $\mathbf{v}^{(i)}$.

$$z_{\text{proj}}^{(i)} = \mathbf{z} \cdot \mathbf{v}^{(i)} \tag{11}$$

This projection is demonstrated on the left hand side of Figure 5.

Finally, the likelihood of a sample being healthy is calculated for each failure direction (i). To do this, the distribution of the projection of the healthy data onto a certain failure direction (i), is assumed to be Gaussian with mean $\mu^{(i)}$ and standard deviation $\sigma^{(i)}$ for positive values of z_{proj}. This ensures that the likelihood of a new sample can be computed based on the healthy distribution parameters $\mu^{(i)}$ and $\sigma^{(i)}$ in the fault direction (i). Furthermore, the likelihood of the sample is assumed to follow a uniform distribution for negative values of z_{proj}, since a deviation of the latent features in the opposite direction of a failure direction is not expected to correspond to a fault. The expression for evaluating the likelihood of a new sample in failure direction (i) is shown in equation Eq. 12.

$$\text{p}(\text{z}_{proj}^{(i)} | \mu^{(i)}, \sigma^{(i)}) = \begin{cases} \frac{1}{\sigma^{(i)}\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{z_{proj} - \mu^{(i)}}{\sigma^{(i)}} \right)^2 \right) & z_{proj} > 0 \\ \frac{1}{\sigma^{(i)}\sqrt{2\pi}} & z_{proj} \leq 0 \end{cases} \tag{12}$$

The evaluation of the likelihood of new samples in the projected dimension is demonstrated on the right hand side of Figure 5. A sample that is considered likely for a given failure direction, can be highly unlikely for a different failure direction.

By evaluating the likelihood of a sample for a given fault mode (projection onto a fault direction) a fault-mode-specific health indicator can be obtained from the latent representation movement with increasing fault severity. The condition that fault directions should be maximally different, as enforced during training, ensures that faulty data from different failure modes are not confused. The projection of new samples

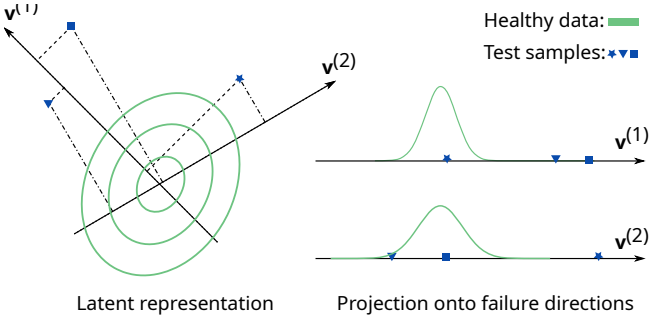


Figure 5. Evaluation of a new test sample: The likelihood of a new sample for a given failure direction $\mathbf{v}^{(i)}$ is calculated based on the projection of the healthy data onto the failure directions in the latent representation.

onto the failure directions further ensures that the discrepancy between the actual fault behaviour and the expected fault behaviour, as communicated to the model by the use of augmented data, is less relevant.

The methodology is now demonstrated on two datasets.

3. Phenomenological Bearing Dataset

In this section, a dataset generated from a phenomenological model based on that of McFadden & Smith (1984) is used to demonstrate the proposed fault detection method. The model used is further extended to include random variations of the amplitude of the transient excitations and random ball slip. The benefit of using a phenomenological model to demonstrate the proposed method is that all potential fault modes can be simulated. This means that the same model, trained on healthy data, can be evaluated on each of the expected degradation paths. The specifications of the phenomenological dataset are listed in Table 1.

The SES for the healthy data is computed from the time domain signal, whereafter the augmented data is computed from the healthy SES as described in the methodology. Figure 6 shows an example of the squared envelope spectrum at the highest fault severity of the data as compared to the augmented data. The figure demonstrates that the augmentation of the data can be imperfect and does not need to be completely representative of reality to improve the separability and interpretability of the latent representation.

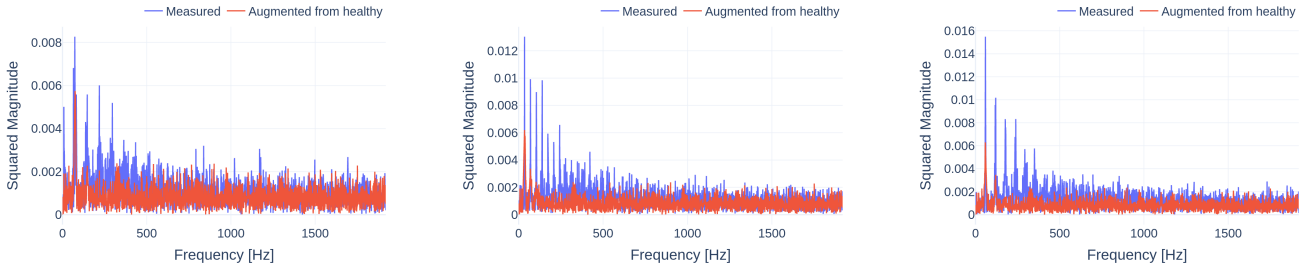
With the healthy and augmented data available, the AE model can be trained. The specifications of the AE models and data augmentation used for each dataset in this investigation are shown in Table 2. In this example, the latent feature representation dimensionality (AE bottleneck size) is chosen as two, so that the latent features can be easily visualized in two-dimensional space. It should be mentioned that the latent representation dimensionality can be viewed as a hyperparameter that needs to be chosen similar to any other AE hyperpa-

Table 1. Specifications for phenomenological dataset.

Model properties	
Transient peak range for different severities	0-1 m/s^2
Modulation amplitude for inner race fault	1
Variance of of slip	0.001 rad
Measurement noise standard deviation	0.2 m/s^2
Transient amplitude standard deviation	0.05 m/s^2
Ball diameter	8.4 mm
Pitch circle diameter	71.5 mm
Number of balls	16
Contact angle	15.7 deg
SDOF stiffness	2×10^{13} N/m
SDOF damping ratio	0.05
SDOF natural frequency	4230 Hz
Constant rotation speed	500 RPM
Sampling frequency	38400 Hz
Signal duration	1 s
Dataset Properties	
Healthy training samples	450
Augmented training samples per fault mode	450
Failure modes considered	3
Healthy validation samples	50
Augmented validation samples per fault mode	50
Damaged test samples	500

Table 2. Specifications AE model and data augmentation used for each dataset.

Specifications	Phenomenological Dataset	IMS Dataset
Model specifications		
Input size	1920	1024
Encoder layer 1	754	402
Bottleneck layer	2	2
Decoder layer 1	754	402
Output size	1920	1024
Activation central layers	ReLU	ReLU
Activation output layer	Tanh	Tanh
Direction regularisation, λ_1	1×10^{-2}	1×10^{-1}
Magnitude regularisation, λ_2	1×10^{-2}	1×10^{-1}
Augmentation specifications		
Peak amplitude a	5×10^{-3}	5×10^{-3}
Decay parameter, α	2×10^{-2}	2×10^{-2}
Base width w	20Hz	20Hz



(a) Inner Race Fault. (b) Ball Fault. (c) Outer Race Fault.

Figure 6. Phenomenological dataset: Comparison of augmented data squared envelope spectrum with failure data at maximum severity. Although there are significant discrepancies between the expected and true fault behaviour, the augmented data is still useful for constraining the latent representation of the autoencoder model

rameter. There is no requirement for a relationship between the dimensionality of the latent space and the number of fault modes that the model accounts for.

Figure 7 shows the latent representation of data after the model has been trained. The latent representation of the faulty test data is shown in Figure 7a together with augmented data from the validation set. The expected failure directions that pass through the median of the healthy and augmented data for each mode are also shown as straight lines. The failure directions are separated in the latent space and the augmented clusters are equally far away from the healthy data cluster, demonstrating that the direction and magnitude loss in Equation 6 was successful in constraining the latent representation. For a given fault mode, the initially healthy latent distribution moves in a direction in the latent representation with increasing fault severity that is consistent with the expected fault direction. For instance, if an outer race fault is present, the latent distribution will move in the general direction indicated by the expected fault direction of the outer race fault.

Therefore, including domain knowledge through augmented data lead to a separated and interpretable latent representation for the AE. As a result, this ensures that informative health indicators can be computed from the latent representation. Additionally, this makes the latent representation interpretable, since anomalous samples not related to bearing faults will likely be distributed in the latent representation in a way that is not consistent with what is expected from the encoded bearing fault behaviour.

In the next step of the methodology, the samples are projected onto the failure directions as visualised in Figure 7 and the likelihood of a test sample is evaluated from Equation 12.

Figure 8 shows the negative log-likelihood for each of the expected failure directions. Each sub-figure shows the result for a certain ground truth fault mode. For a given fault mode, the negative log-likelihood health indicator rises sharply for faulty data projected onto the failure direction that corresponds to the ground truth fault mode. For example, in Figure 8a the ground truth fault mode is a ball fault. Consequently, the negative log-likelihood of the data projected onto the ball fault direction increases with increasing fault severity. In contrast, the negative log-likelihood of data projected onto fault directions that are not associated with a ball fault remains comparatively low. Thereby, an informative healthy indicator is obtained that can indicate faulty behaviour and simultaneously provide diagnostics information about the fault mode.

In the next section, a similar analysis is conducted on the IMS dataset.

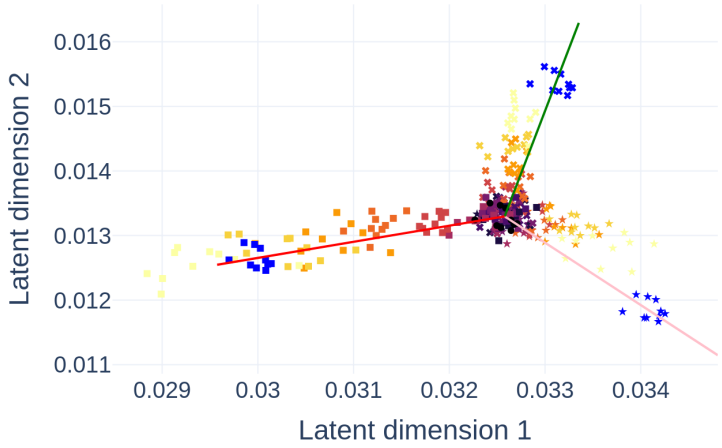
4. Application on NASA IMS experimental dataset

The NASA IMS dataset (Qiu et al., 2006) is a popular dataset used in bearing condition monitoring. It consists of three separate run-to-failure tests, each including data for four bearings. Ground truth labels of the failure modes in which a bearing had failed (Inner Race, Outer Race or Ball Fault) are available for four of the 12 bearings. This investigation will focus on the four datasets with labelled ground truth labels in order to check if the proposed method is successful in detecting a fault associated with a particular fault mode. Information about the datasets that are used in this investigation is shown in Table 3.

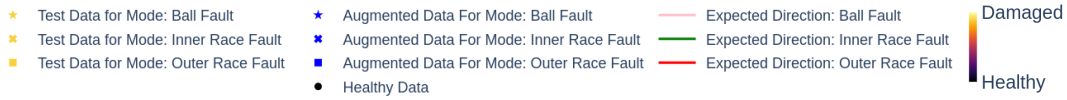
The healthy records for training are chosen in accordance with that of Liu & Gryllias (2020) with some run-in records not used for training. The remaining records are used as the test set during the evaluation phase. Table 3 shows the record numbers used for training as well as the ground truth label for each of the datasets considered.

To illustrate the data augmentation process, Figure 9 shows an example of test samples from the IMS dataset at a high severity compared to the augmented data for the same fault mode.

The SES seems to be an effective way of extracting fault information from the two outer race fault datasets, but is less



(a) Test samples (shaded), augmented validation samples (blue), healthy validation samples (black) and expected failure directions (colored lines).



(b) Legend.

Figure 7. Latent representation after training. Shaded data points represent test data at a certain fault severity. Shapes represent different fault modes

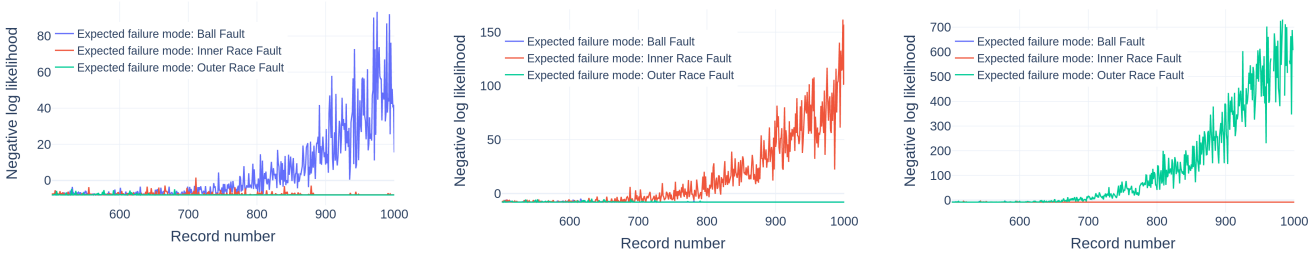
Table 3. IMS dataset information.

Dataset	Channel	Recorded Failure Mode	Healhty record numbers	Training Samples	Validation Samples
1	Bearing 3, Channel 5	Inner Race	200-600	360	40
1	Bearing 4, Channel 7	Ball	200-600	360	40
2	Bearing 1, Channel 1	Outer Race	50-300	225	25
3	Bearing 3, Channel 3	Outer Race	50-300	225	25

effective for inner race and ball fault datasets, leading to large discrepancies between the true and augmented data. As a result, the latent representation for the fault data shown in Figure 10 is not well structured in the latent representation for all of the datasets. However, for the outer race fault of test 2, bearing 1 the latent features clearly move along the outer race failure direction with increasing fault severity.

The negative log-likelihood of a sample projected onto a given failure direction is shown in Figure 11. The method appears to be effective for the outer race and inner race fault datasets, with the negative log-likelihood increasing for the ground truth failure direction. However, the negative log likelihood associated with the ball fault does not seem to be more sensitive to the ball fault as compared to the negative log-likelihood associated with the other fault modes. This is due to the latent space being uninformative after the completion of training, since the envelope spectrum that is used as input is not sensitive to the ball fault.

The effectiveness of the method is reliant on a correspondence between the augmented data and the true failure behaviour so that the latent representation can be sensitive to a given failure mode. Consequently, it is expected that the effectiveness of the method is dependant on how informative the input features are. In this analysis, where a simple input feature such as the envelope spectrum was used without any additional pre-processing such as band pass filtering around informative frequency bands, it is clear that the envelope spectrum is not sensitive to certain fault types, and as a result the latent feature representations were not informative for these failure modes. In future work this limitation could be addressed by training models on the time domain data directly, or by using more advanced features from signal processing.



(a) Ball Fault.

(b) Inner Race Fault.

(c) Outer Race Fault.

Figure 8. Phenomenological dataset: Negative log-likelihood with increasing fault severity. Sub figures show data for different ground truth fault modes. Each trace on a sub-figure shows the log-likelihood of the test data for an expected failure mode.

5. Conclusion and Future Work

This paper presents a new way of encoding domain knowledge into the latent representation of an AE by using augmented data. Results on a phenomenological dataset demonstrate that incorporation of domain knowledge leads to an interpretable latent representation that is useful for constructing informative health indicators for fault detection and diagnostics. Furthermore, the method is applied to the experimental NASA IMS dataset. The method proves to be effective for three of the four IMS datasets considered, with the success of the method being reliant on how sensitive the input features are to damage.

In future work, the proposed method can be extended to act on time-frequency maps or even directly on time series data, where the data augmentation can be facilitated by a phenomenological model. This can ensure that hidden fault information is not withheld from the model whilst still allowing for the incorporation of domain knowledge. Furthermore, the sensitivity of the method to inaccuracies in modelling the expected fault behaviour, the chosen size of the latent representation, training batch sizes and the model architecture needs to be determined.

Acknowledgment

The authors gratefully acknowledge the support of the European Commission under the Marie Sklodowska-Curie program through the ETN MOIRA project (GA 955681) and the support of the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme.

References

Balshaw, R., Heyns, P. S., Wilke, D. N., & Schmidt, S. (2022, April). Importance of temporal preserving latent analysis for latent variable models in fault diagnostics of rotating machinery. *Mechanical Systems and Signal Processing*, *168*, 108663. doi: 10.1016/j.ymssp.2021.108663

Booyse, W., Wilke, D. N., & Heyns, S. (2020, June). Deep digital twins for detection, diagnostics and prognostics. *Mechanical Systems and Signal Processing*, *140*, 106612. doi: 10.1016/j.ymssp.2019.106612

Cao, H., Niu, L., Xi, S., & Chen, X. (2018, March). Mechanical model development of rolling bearing-rotor systems: A review. *Mechanical Systems and Signal Processing*, *102*, 37–58. doi: 10.1016/j.ymssp.2017.09.023

Cerrada, M., Sánchez, R.-V., Li, C., Pacheco, F., Cabrera, D., Valente de Oliveira, J., & Vásquez, R. E. (2018, January). A review on data-driven fault severity assessment in rolling bearings. *Mechanical Systems and Signal Processing*, *99*, 169–196. doi: 10.1016/j.ymssp.2017.06.012

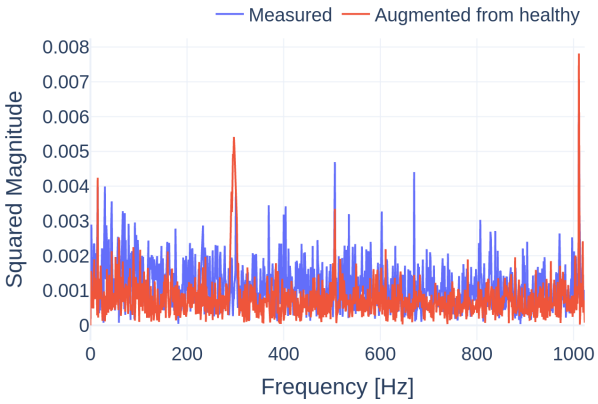
Hoang, D.-T., & Kang, H.-J. (2019, March). A survey on Deep Learning based bearing fault diagnosis. *Neurocomputing*, *335*, 327–335. doi: 10.1016/j.neucom.2018.06.078

Jiao, J., Zhao, M., Lin, J., & Liang, K. (2020, December). A comprehensive review on convolutional neural network in machine fault diagnosis. *Neurocomputing*, *417*, 36–63. doi: 10.1016/j.neucom.2020.07.088

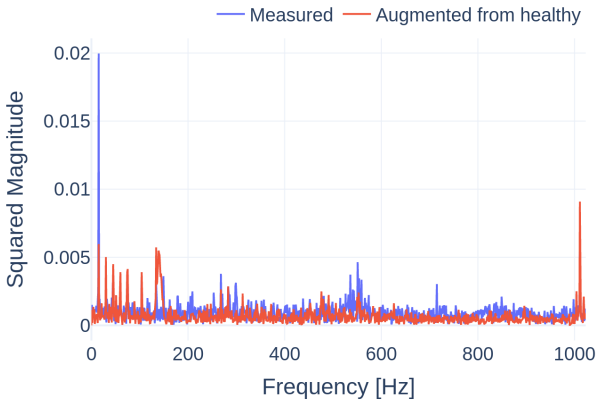
Lee, J., Wu, F., Zhao, W., Ghaffari, M., & Liao, L. (2014). Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. *Mechanical Systems and Signal Processing*, 21. doi: http://dx.doi.org/10.1016/j.ymssp.2013.06.004

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018, May). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, *104*, 799–834. doi: 10.1016/j.ymssp.2017.11.016

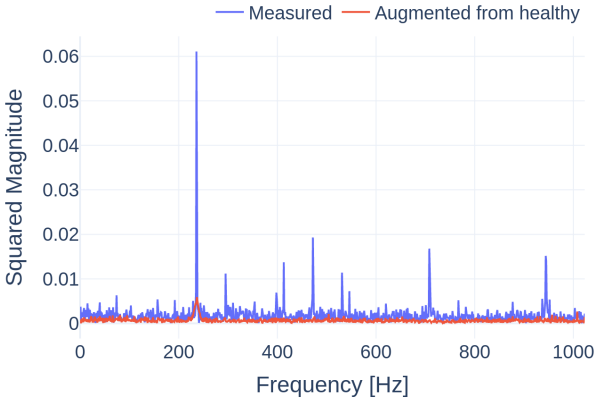
Leturiondo, U., Salgado, O., Ciani, L., Galar, D., & Catelani, M. (2017, October). Architecture for hybrid modelling and its application to diagnosis and prognosis with missing data. *Measurement*, *108*, 152–162. doi: 10.1016/j.measurement.2017.02.003



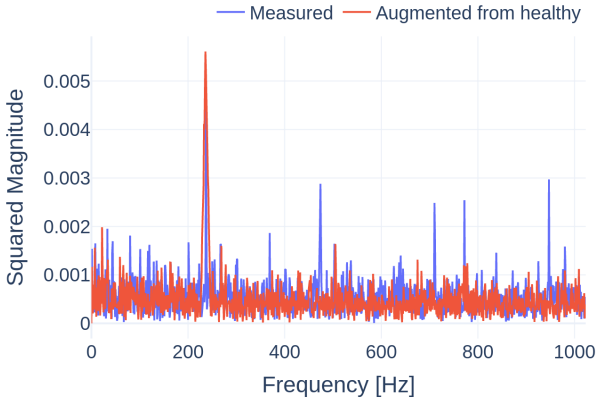
(a) IMS test 1, Bearing 1: Inner Race Fault.



(b) IMS test 1, Bearing 4: Ball Fault.



(c) IMS test 2, Bearing 1: Outer Race Fault.



(d) IMS test 3, Bearing 3: Outer Race Fault.

Figure 9. Squared envelope spectrum for augmented data and data at high severity.

Li, X., Zhang, W., & Ding, Q. (2018, October). A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning. *Neurocomputing*, *310*, 77–95. doi: 10.1016/j.neucom.2018.05.021

Liao, L., & Kottig, F. (2014, March). Review of Hybrid Prognostics Approaches for Remaining Useful Life Prediction of Engineered Systems, and an Application to Battery Life Prediction. *IEEE Transactions on Reliability*, *63*(1), 191–207. doi: 10.1109/TR.2014.2299152

Liu, C., & Gryllias, K. (2020, June). A semi-supervised Support Vector Data Description-based fault detection method for rolling element bearings based on cyclic spectral analysis. *Mechanical Systems and Signal Processing*, *140*, 106682. doi: 10.1016/j.ymssp.2020.106682

Liu, C., Mauricio, A., Qi, J., Peng, D., & Gryllias, K. (2020, November). Domain Adaptation Digital Twin for Rolling Element Bearing Prognostics. *Annual Conference of the PHM Society*, *12*(1), 10. doi: 10.36001/phmconf.2020.v12i1.1294

McFadden, P., & Smith, J. (1984, September). Model for the vibration produced by a single point defect in a rolling element bearing. *Journal of Sound and Vibration*, *96*(1), 69–82. doi: 10.1016/0022-460X(84)90595-9

Qiu, H., Lee, J., Lin, J., & Yu, G. (2006, February). Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. *Journal of Sound and Vibration*, *289*(4-5), 1066–1090. doi: 10.1016/j.jsv.2005.03.007

Randall, R. B., & Antoni, J. (2011, February). Rolling element bearing diagnostics—A tutorial. *Mechanical Systems and Signal Processing*, *25*(2), 485–520. doi: 10.1016/j.ymssp.2010.07.017

Sadoughi, M., & Hu, C. (2019, June). Physics-Based Convolutional Neural Network for Fault Diagnosis of Rolling

Element Bearings. *IEEE Sensors Journal*, *19*(11), 4181–4192. doi: 10.1109/JSEN.2019.2898634

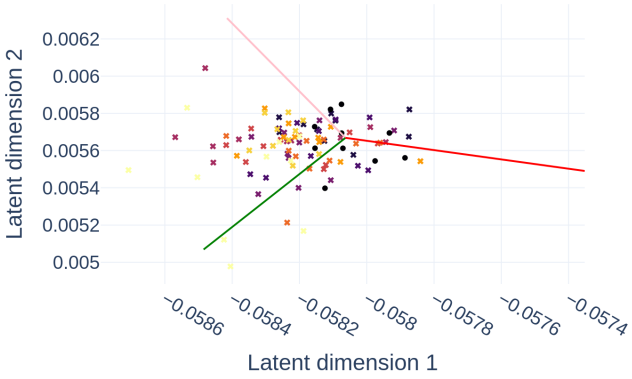
Shen, S., Lu, H., Sadoughi, M., Hu, C., Nemani, V., Thelen, A., ... Kenny, S. (2021, August). A physics-informed deep learning approach for bearing fault detection. *Engineering Applications of Artificial Intelligence*, *103*, 104295. doi: 10.1016/j.engappai.2021.104295

Yu, K., Lin, T. R., Ma, H., Li, X., & Li, X. (2021, January). A multi-stage semi-supervised learning approach for intelligent fault diagnosis of rolling bearing using data augmentation and metric learning. *Mechanical Systems and Signal Processing*, *146*, 107043. doi: 10.1016/j.ymssp.2020.107043

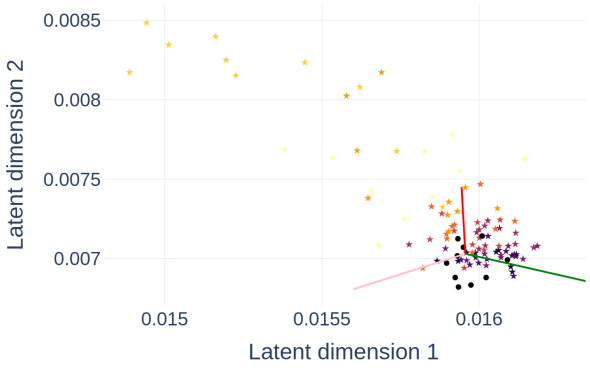
BIOGRAPHIES

Douw Marx received his B.S. and M.Sc degree in mechanical engineering from the Univeristy of Pretoria, South-Africa. He joined the Noise and Vibration Research Group in the Department of Mechanical Engineering, KU Leuven, Belgium as a PhD researcher in 2021. His research interests include hybrid approaches for fault detection, signal processing and deep learning.

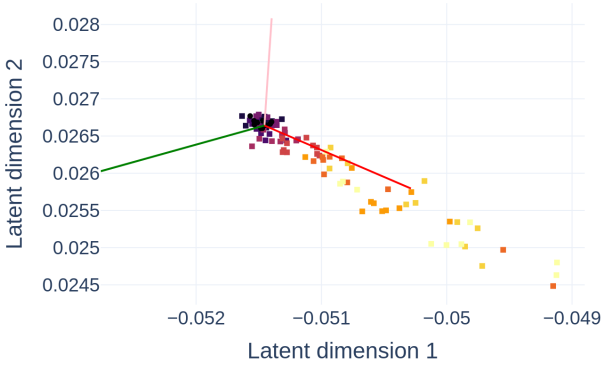
Konstantinos Gryllias holds a 5 years engineering diploma degree and a PhD degree in Mechanical Engineering from National Technical University of Athens, Greece. He holds an associate professor position on vibroacoustics of machines and transportation systems at the Department of Mechanical Engineering of KU Leuven, Belgium. He is also the manager of the University Core Lab Dynamics in Mechanical & Mechatronic Systems DMMS-M of Flanders Make, Belgium. His research interests lie in the fields of condition monitoring, signal processing, prognostics and health management of mech. & mechatronic systems.



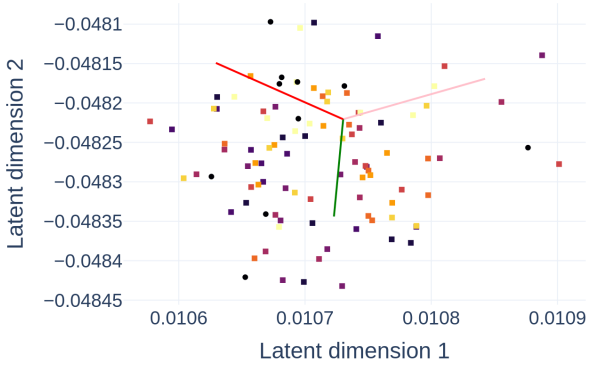
(a) IMS test 1, Bearing 1: Inner Race.



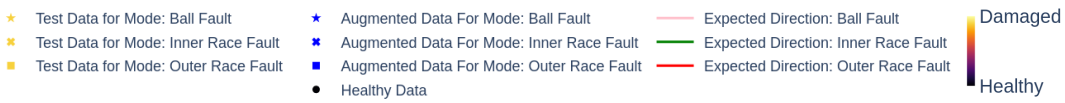
(b) IMS test 1, Bearing 4: Ball.



(c) IMS test 2, Bearing 1: Outer Race.

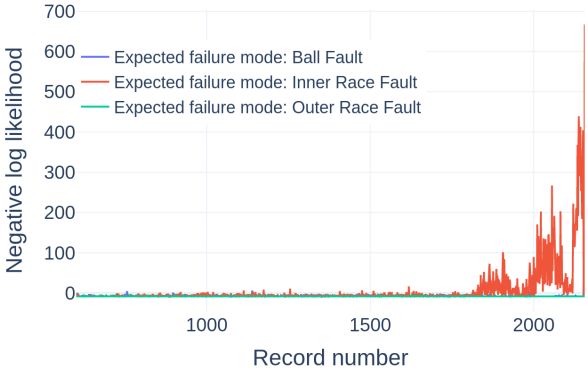


(d) IMS test 3, Bearing 3: Outer Race.

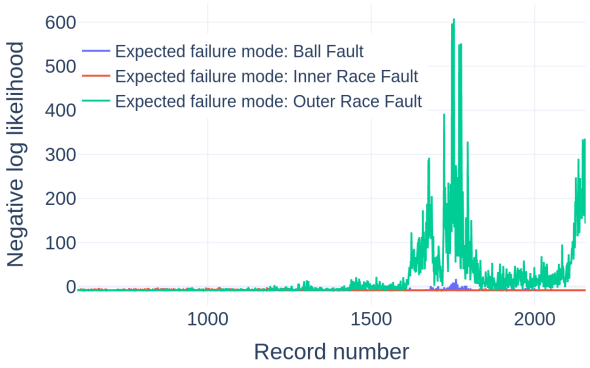


(e) Legend.

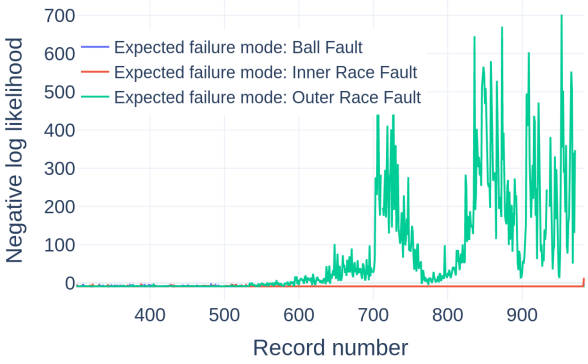
Figure 10. IMS dataset latent representation at different severities.



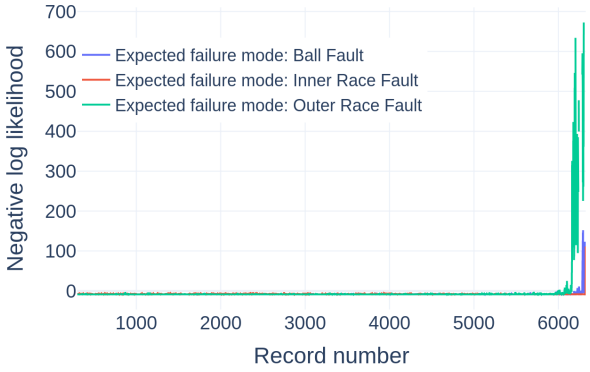
(a) IMS test 1, Bearing 1: Inner Race Fault.



(b) IMS test 1, Bearing 4: Ball Fault.



(c) IMS test 2, Bearing 1: Outer Race Fault.



(d) IMS test 3, Bearing 3: Outer Race Fault.

Figure 11. IMS dataset: Negative log likelihood of measurements onto projected onto respective fault modes