Importance of temporal preserving latent analysis for latent variable models in fault diagnostics of rotating machinery

Ryan Balshaw[a,*], P. Stephan Heyns[a], Daniel N. Wilke[a], Stephan Schmidt[a]

[a]*Centre for Asset Integrity Management, Department of Mechanical and Aeronautical Engineering, University of Pretoria, Pretoria, South Africa*

Abstract

Latent variable models are important for condition monitoring as they learn, without any supervision, the healthy state of a physical asset as part of its latent manifold. This negates the need for labelled fault data and the application of supervised learning techniques. Latent variable models offer information from which health indicators can be derived for condition monitoring. Namely, information from the latent space and the data space can be used for condition inference. These health indicators are used to explain changes in a physical asset's condition. Conventional black-box approaches only offer information from the data space in the form of reconstruction errors. In contrast, latent variable models offer a latent space and reconstruction space for inference. However, the current application of latent variable models either disregards latent space information or fails to realise its full potential. The full potential can be realised by preserving the time information in the data. Therefore, we propose a model evaluation procedure that specifically preserves time in the latent health indicators. The procedure is generic and can be applied to any latent variable model as demonstrated for Principal Component Analyis (PCA), Variational Auto-Encoders (VAEs) and Generative Adversarial Networks (GANs) in this study. In general, as time information can be discarded or preserved for derived latent health indicators, this study advocates that health indicators that preserve time are more useful for condition monitoring than health indicators that discard time. In addition, it enables the interpretation of the learnt latent manifold dynamics and allows for alternative latent indicators to be developed and deployed for fault detection. The proposed *temporal preservation* model evaluation procedure is applied to three classes of latent variable models using two datasets. Three model-independent latent health indicators that preserve time are proposed and shown to be informative on all three classes of latent variable models for both datasets. The temporal preserving latent analysis procedure is demonstrated to be essential to derive more informative latent metrics from latent variable models.

Keywords: Unsupervised learning, Latent variable models, Temporal preservation, Latent analysis, Fault diagnostics, Time-varying operating conditions

1. Introduction

Vibration-based condition monitoring is an important field that addresses the need to optimise and improve the utilisation of expensive machine assets [1, 2]. The goal of condition monitoring is to uncover the presence, if any, of vibration signatures that are indicative of faults. This goal requires that the performance of condition monitoring techniques be

*Corresponding author
Email address: ryanbalshaw1@gmail.com (Ryan Balshaw)

robust to variations in operating conditions as this may complicate the condition inference procedure [3]. These applications of condition monitoring techniques are commonly referred to as asset degradation assessment or the diagnostics problem and the prognostics problem. In the literature, advanced signal processing and learning-based techniques have been widely investigated. Both are working towards the unified goal of improving the performance capabilities of condition monitoring techniques [4, 5].

For the diagnostics problem, the objective is to use a health indicator (HI) to infer the asset's condition. The development of HIs is an active research endeavour centred around asset degradation assessment techniques [6]. These techniques aim to quantify the asset condition by detecting and enabling fault trending [7]. Examples of traditional HIs include the root-mean-square value, peak value, crest factor and kurtosis [8]. The spectral kurtosis is a statistical quantity used to detect signal impulsiveness, or a deviation from Gaussianity, and is commonly used to determine frequency bands where impulsive signal components dominate [9]. In Antoni and Borghesani [7], statistical HIs were designed to detect stationarity and/or Gaussianity deviations. In Wang and Tsui [6], a generalised dimensionless HI is proposed using the probability density function of a noncentral chi-square distribution defined over the variance-normalised square envelope. In Booyse et al. [10], a Generative Adversarial Network (GAN) was used to develop a HI from the GAN discriminator function, which was used to measure the deviance from a healthy asset state.

Signal processing-based methods and learning-based methods can both provide indicators, but the two fields differ in fault covariate extraction practices. Signal processing techniques are designed using extensive domain knowledge. In contrast, learning-based techniques such as statistical learning and machine learning initially utilised advanced feature engineering to improve model performance in classification or regression tasks [11]. Deep learning introduced powerful nonlinear parametric functions commonly referred to as deep neural networks, which extract features from raw time-series signals for condition inference [5].

However, many learning-based approaches are black-box models for semi-supervised learning optimisation using classification losses, reconstruction losses, or combining the two [12–14]. The term black-box is used in this work to refer to a model in which there is little understanding of the facets of the model other than the output it provides. This induces a natural blind spot for learning-based approaches as certain model facets are never fully addressed as the focus is to maximise method performance. Another issue inherent to supervised learning techniques is the data label problem. The costly and dangerous nature of failure in critical assets introduces limited access to extensive fault data. Usually, a vast amount of healthy asset data is available, as the percentage of time an asset spends in a failure state is low [5]. A large portion of the recent learning-based literature addresses this problem using semi-supervised learning techniques and domain-transfer learning. These techniques require access to some labelled fault classes from the asset of interest or an identical or different machine [5].

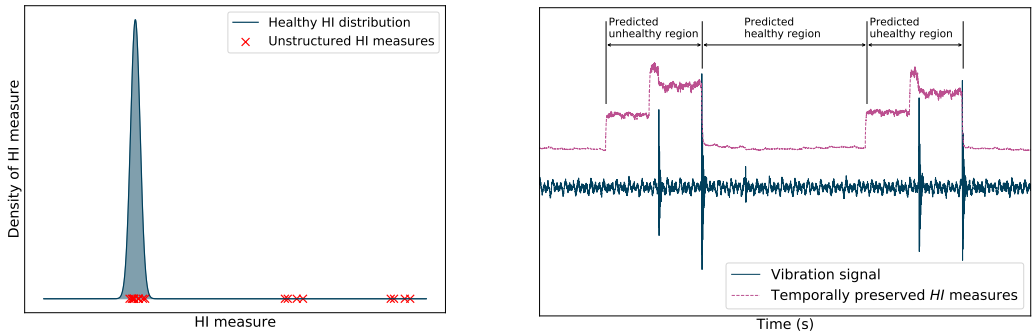
An alternative solution to the data label problem is to use healthy time-series asset data to learn a probabilistic representation of the asset in a healthy state [10]. By learning this representation, the data label problem is bypassed as the primary interest is to measure if unseen data is from this representation. This probabilistic representation learning process manifests through density estimation learning techniques, often classified in the literature under the banner of unsupervised learning [11]. For density estimation, latent variable models (LVMs) are considered a cornerstone [11, 15, 16]. LVMs are techniques that assume that unobserved latent variables are used as conditioning variables for the observed data, and this manifold describes the intrinsic structure of this data. LVMs are used to learn the representation of a

healthy asset state. This approach is powerful, but its potential is not fully realised in current methods due to inadequate data analysis and latent analysis procedures. Specifically, the latent manifold is often left as a black-box facet that is a simple by-product of the model formulation. However, under the manifold hypothesis [17], it would be expected that if the manifold represents healthy data, unhealthy data should not lie in the same area of the manifold.

A common process in the learning-based literature is the use of a set overlap percentage when performing time-series signal segmentation for models trained on raw time-series data [10, 18]. This time-series data evaluation approach is commonly used for both training and evaluation. This may be key to reduce the input feature dimensionality but breaks the temporal structure in the data. The temporal structure referred to here is not the time between recorded time-series data but the time structure within this data. Our work emphasises that this local data timeframe must be preserved. This preservation is atypical but offers much potential in vibration-based CM. Using large overlapping segments is sufficient for model training. However, this can hinder the condition inference procedure during the model evaluation step due to anomalies in time-series data that are localised as the damaged portions move through the loading zone. This is the Achilles' heel of current LVM techniques and disrupts the temporal information in time-series data. Furthermore, present methods cannot readily detect anomalous instances in the latent manifold unless fault labels are supplied. These labels are used to cluster the data, or an adversarial training scheme is required to regularise the encoding transition function to provide an additional health indicator [12, 19].

A rational guide to overcoming this flaw exists in discrepancy analysis techniques. Localised discrepancy measures are obtained using a healthy probabilistic representation, and these measures quantify the deviation from the healthy data distribution over time [20, 21]. These discrepancy measures are also obtained with crafted signal segmentation approaches that exploit external information such as tachometer signals to carefully design a statistical analysis methodology for specific fault cases [22]. LVMs cover one-half of discrepancy analysis techniques, whereby a model is used to capture the healthy data density. Once this density has been learnt, HIs are used as representative metrics to evaluate whether new data is from this distribution. However, exploiting additional information during the pre-processing stage is counter-productive to deep learning, aiming to only use raw time-series data to train models.

This work proposes that by preserving the temporal structure, LVMs' full potential can be harnessed, and this preservation introduces informative latent health indicators (LHIs). The benefit of the proposed LHIs is that they are generic and can be used with any LVM that makes a continuous and temporally preserved latent space available. The contribution of this work is the augmentation of the LVM analysis approach by preserving the time dimension. This augmentation allows for increased analysis flexibility as time structure is introduced into the HI measures. Figure 1 details this augmentation clearly for a signal with a localised fault. In the conventional latent analysis setting shown in Figure 1(a), the HI measures from the signal contain healthy and anomalous characteristics, while the cause of the anomalies is unclear. In Figure 1(b), the distinct healthy and anomalous (or unhealthy) regions are seen in the proposed analysis setting. Since time is preserved, it is also possible to extract more information from this HI signal using conventional signal analysis approaches (e.g. time synchronous averaging, spectral analysis), making it possible to identify the source of the unhealthy behaviour. By ensuring that the temporal structure in the time-series data is preserved, incipient fault detection, fault isolation, and fault trending become viable avenues of investigation, which are crucial facets of vibration-based fault diagnostics. Temporal preservation is also model agnostic, which introduces a generic analysis setting that exploits the temporal structure in the time-series data for improved diagnostic performance.



(a) Conventional analysis approach.

(b) Proposed analysis approach.

Figure 1: The differences between the conventional analysis approach and the proposed analysis approach for a damaged time-series signal. In the conventional setting, notice how there appears to be HI measures in both the expected healthy density region and some outliers, but this cannot be easily explained. In the proposed setting, it is clear that the unexplained healthy measures are simply those from a healthy local region in the time-series signal. Figure (b) shows that the expanded unhealthy region in the time preserved HI measures over the fault is attributed to the model window length used.

A reformulation of the model evaluation approach used in deep learning practices is essential, and therefore a *temporal preservation* approach is proposed. This reformulation adds improved model interpretability, HI interpretation and the condition inference process. The latent manifold of LVMs is explored, and its responsiveness to anomalous instances is investigated. The main contributions of this research are summarised as follows:

1. A reformulation that exploits the natural temporal structure in the data used in the model evaluation procedure results in HI signals that are more informative. The preservation of temporal structure allows multiple facets of the discrepancy signals to be exploited for condition inference.
2. The latent manifold is shown to be interpretable and responsive to anomalous data samples. A set of LHIs is proposed to be applied to any continuous latent manifold with temporal structure.
3. Temporal preservation enables LHIs to be derived that consider the latent manifold's path information and trajectory dynamics instead of only state information of the latent manifold.

As this work aims to utilise the full potential of LVMs, three different models are considered. These models are Principal Component Analysis (PCA) [11], Variational Auto-Encoders (VAEs) [15] and a Disentangled Latent Space Generative Adversarial Network (DLS-GAN) [23]. Many models are available, but the selected methods are sufficient to investigate the LVM class, model complexity to highlight the temporal preservation approach's potential combined with the proposed LHIs. The performance of common HIs and the proposed LHIs is critically compared on datasets with constant and time-varying operating conditions. The first dataset investigates a gear tooth fault in a gearbox with time-varying operating conditions to highlight the improved condition inference procedure obtained when using the *temporal preservation* approach. This is achieved by using this approach to improve the quality of the HI signals. We show that the proposed LHIs obtained from the proposed framework can detect asset damage in the latent manifold under time-varying operating conditions. The second dataset is a bearing run-to-failure dataset under constant operating conditions. This dataset is used to compare the performance of the various HIs and LHIs developed under the *temporal preservation* approach against state-of-the-art signal processing and deep learning literature. The important techniques relating to LVMs and the *temporal preservation* approach are introduced, and the application to vibration-based condition

monitoring is addressed in Section 2. The HI and LHIs obtained from different models are then evaluated on experimental data and compared in Section 3. Finally, the research is concluded, and recommendations for future work are provided in Section 4.

2. Latent Variable Analysis

2.1. Introduction

This work aims to use LVMs as a method to transform time-series data to a set of HIs from which the condition of an asset may be inferred. LVMs have been used for anomaly detection tasks. However, this work will investigate how anomalous instances are represented in the latent space and propose a set of model agnostic scalar indicators that can be used to detect anomalous instances in the latent manifold. To do this, the basic components of an LVM must be understood, and the components of the model that apply to condition monitoring must be detailed. LVMs are well documented in the literature and are often used for unsupervised learning tasks [11]. Unsupervised learning differs from supervised learning and some semi-supervised learning techniques in the distribution of interest. Unsupervised learning is concerned with density estimation tasks, such as capturing the distribution $p(\mathbf{x})$ over the data space $\mathbf{x}$. In supervised learning, and in most semi-supervised learning cases, the objective is to infer a variable $\mathbf{y}$ given a sample $\mathbf{x}$. These learning methodologies are used in learning-based literature and can be classified into statistical learning, machine learning or deep learning based on model flexibility and pre-processing applicability.

In the density estimation process, LVMs assume that an unobserved latent variable $\mathbf{z}$ describes the data. The assumption is often made that this latent variable space is of a lower dimension to the data space $\mathbb{R}^{latent} \leq \mathbb{R}^{data}$. A generative view of these models is developed by assuming a prior $p(\mathbf{z})$ over $\mathbf{z}$ and a conditional generative distribution $p(\mathbf{x}|\mathbf{z})$ from which data samples are drawn. The data distribution $p(\mathbf{x})$ is then given through the marginalisation of the joint distribution $p(\mathbf{x}, \mathbf{z})$ with respect to the latent variables

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \tag{1}$$

It is then often important to evaluate the posterior distribution using Bayes' theorem for model inference to obtain latent samples from the data samples, which can be obtained through

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}. \tag{2}$$

These two processes are key to LVMs and are succinctly demonstrated in Figure 2, where two parametric functions f_ϕ and g_θ are used to map samples from the data space to the latent space and vice versa. The parametric functions f_ϕ and g_θ are commonly referred to as the encoder and decoder, and these functions are used to parametrise the inference and generative distributions $p(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{z})$. The placement of the latent prior with respect to the latent manifold is demonstrated to enforce the requirement that f_ϕ transforms input feature samples in a region similar to $p(\mathbf{z})$ to ensure that data samples can be generated through sampling $p(\mathbf{x}|\mathbf{z})$ from a sample $\mathbf{z} \sim p(\mathbf{z})$.

Three common operations can be described when using LVMs. The first is the data generation process which occurs through sampling a point in the latent space $\mathbf{z} \sim p(\mathbf{z})$ and then drawing a data sample from the conditional distribution $p(\mathbf{x}|\mathbf{z})$, which is obtained through $\mathbf{x} = g_\theta(\mathbf{z})$. The second is the input reconstruction process which is obtained from

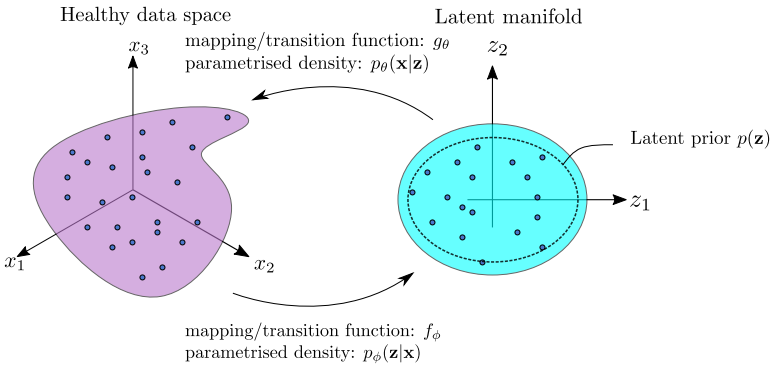


Figure 2: An illustration of the process of a typical application of LVMs for condition monitoring. Notice the location of the latent prior $p(\mathbf{z})$ and the learnt latent manifold through the inference network f_ϕ.

sampling $\mathbf{x} \sim p(\mathbf{x})$ and using both transition functions to produce a reconstructed sample $\tilde{\mathbf{x}} = g_\theta(f_\phi(\mathbf{x}))$. The final process is the model inference step that obtains a latent code for a sample $\mathbf{x} \sim p(\mathbf{x})$ using $\mathbf{z} = f_\phi(\mathbf{x})$. It is important to note that not all LVM methods focus on both parametric functions, and the choice of which parametric function is learnt affects the framework's versatility, as this restricts which model operations are available.

2.2. Proposed latent variable analysis method

In this work, LVMs are used to capture the representation of a healthy asset. This process allows for health indicators to be obtained that track the deviation from this healthy representation. To pre-process the data, typically, a sliding window with an overlap percentage is used to reduce the dimensionality of the data features and increase the training dataset's size [24]. The overlap percentage removes the temporal structure of the time-series data in the signal segments seen by the model, which inhibits the condition inference procedure by limiting the operations that can be performed on the HI or LHI signals. A *temporal preservation* approach is used during model evaluation to preserve time in the obtained condition indicators for a given signal. Figure 3 shows the difference between the conventional approach and the approach proposed in this work. It is emphasised that the model training procedure for LVMs from conventional applications is sufficient. Hence this element of the LVM application remains unchanged. The *temporal preservation* approach proposed in this work requires that time-series data be fed through the model incrementally to obtain condition indicator signals that have temporal structure during model evaluation. This decision is beneficial as it produces informative health indicators, and these signals are now considered a transformation that may enhance the presence of fault covariates in the data. Furthermore, to explore the latent manifold, it is beneficial to preserve time as this allows for the path traversed in the latent space to be investigated.

To develop a set of scalar metrics that can be used to interpret the latent manifold, the deviation paths from the latent space must be conceptually quantified. In Figure 4, the potential paths of latent traversal are highlighted to the reader, and these paths illustrate plausible directions of traversal for data from an asset with a localised fault and data from an asset with a distributed fault, such as distributed gear wear. As highlighted in Figure 4(a), it is difficult to distinguish between the latent space's localised and distributed gear faults. However, as demonstrated in Figure 4(b), we can clearly distinguish between the two fault cases if we use the temporal information.

There are two primary means of anomaly detection in the manifold, one through off-manifold deviance and another

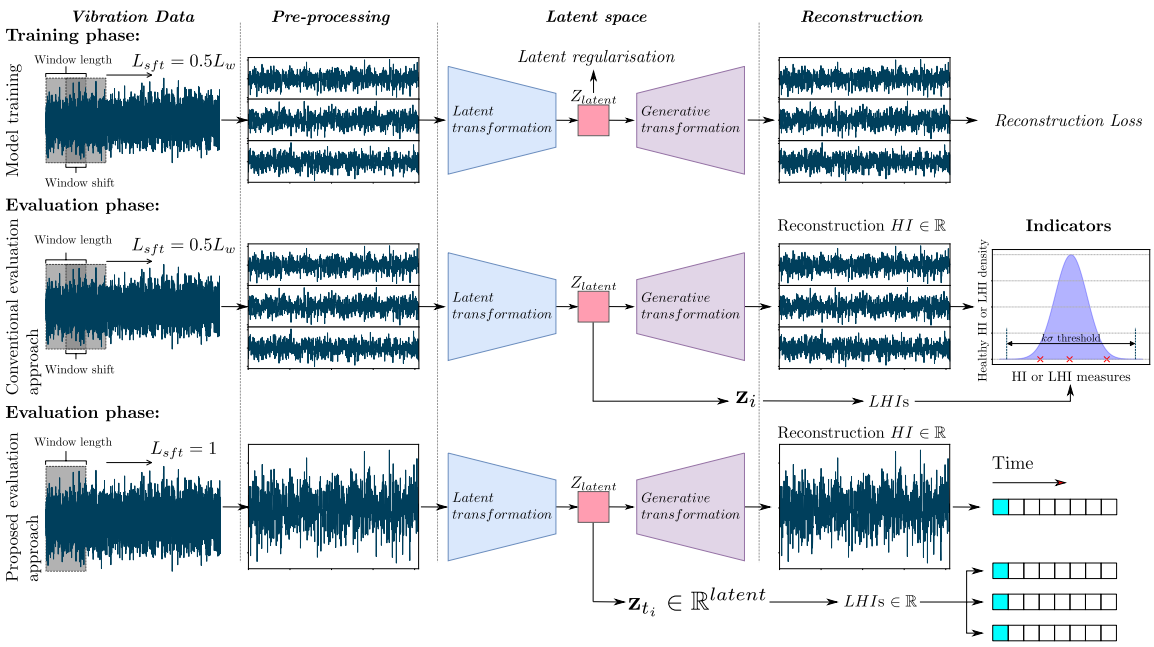


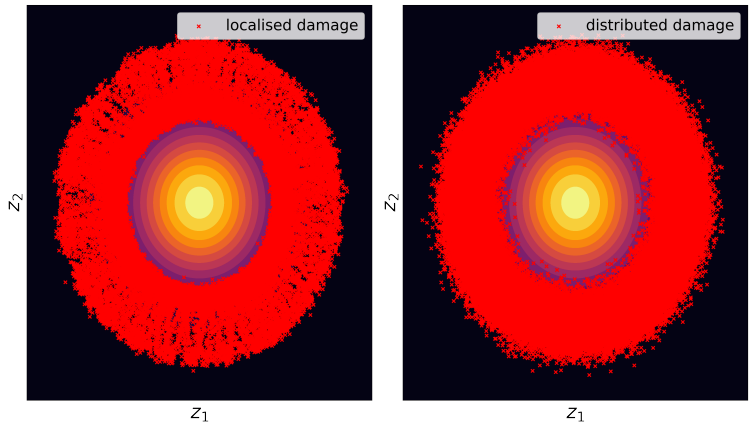
Figure 3: The difference in analysis methodology between current literature approaches and the proposed *temporal preservation* approach. It is key to note that the difference only manifests during the model evaluation stage, where the temporal structure of the vibration data is preserved. This difference allows for the development of more informative HIs and LHIs. The $k\sigma$ rule, where $k \in \mathbb{N}$, is formulated under a Gaussian distribution assumption that is validated under a Gaussian null hypothesis.

through the manifold's path traversal properties. As the latent manifold is representative of healthy data, anomalous data may manifest in the non-healthy manifold regions. As LVMs control the structure of the latent manifold to bear resemblance to a prior distribution $p(\mathbf{z})$, this deviation may be quantified through the log-likelihood of the latent samples under the prior $p(\mathbf{z})$. Furthermore, as the manifold has temporal structure in the *temporal preservation* framework, dynamic properties such as the traversal rate may change without deviating to regions of low likelihood under the prior $p(\mathbf{z})$, which is static with respect to time. The traversal rate may serve as a suitable indicator to detect unhealthy data.

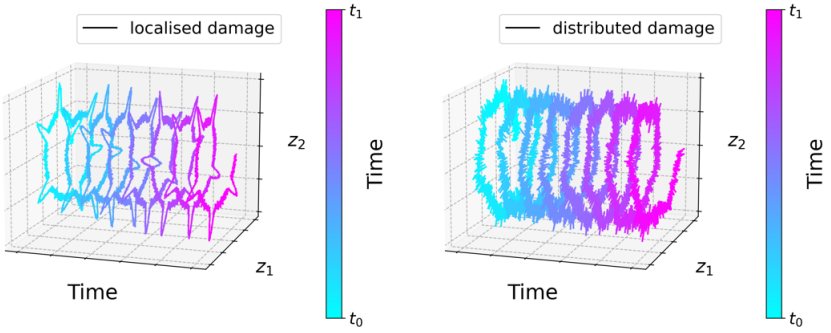
Temporal preservation is key to interpret changes in the latent manifold as the temporal structure of the data can be used to track the path traversed through this manifold. The temporal preservation approach introduces structure to our data analysis. As a result, the data structure is preserved in a low-dimensional representation in the latent manifold. This configuration is used to interpret the latent manifold, thereby allowing time to be untangled. Deviations in the static and dynamic properties of the latent manifold may now be quantified.

2.3. Data pre-processing

In this work, the raw time-series data must be reduced in dimensionality to satisfy computational limitations and to produce a dataset that can be used to optimise the model parameters. Firstly, the available healthy vibration data is processed according to a sliding window scheme [10, 18]. The data is developed into a Hankel matrix $\mathbf{H}$ [24] by storing segments of a size designated by the model window length L_w and overlap controlled by the shifting parameter L_{sft}. L_w and L_{sft} control the amount of signal information seen by the model and the amount of overlap between signal segments, respectively. L_w is the model window length or segment length, and L_{sft} is the segment overlap length. Signal segments



(a) Manifold representation for faulty time-series data in $\mathbb{R}^2$.



(b) Manifold representation for faulty time-series data with time in an arbitrary time interval $t \in [t_0, t_1]$.

Figure 4: The potential manifold traversal paths that can be obtained from LVMs for different data cases, shown in Figure (a) the latent space and Figure (b) the latent space with time. Note that in Figure (a), the contours refer to the prior $p(\mathbf{z})$. It is clear from Figure (a) that the different damage cases manifest similarly, and it is non-trivial to discern between the two. In Figure (b), where the temporal structure is preserved, the manifold responds differently to different fault types. The preservation of time allows for manifold interpretation.

are developed by sliding the window and storing segments for measurement $\mathbf{x}_i$ through

$$\mathbf{H}(\mathbf{x}_i) = \begin{bmatrix} \mathbf{x}_i(0) & \mathbf{x}_i(1) & \mathbf{x}_i(2) & \cdots & \mathbf{x}_i(L_w) \\ \mathbf{x}_i(L_{sft}) & \mathbf{x}_i(L_{sft}+1) & \mathbf{x}_i(L_{sft}+2) & \cdots & \mathbf{x}_i(L_w+L_{sft}) \\ \mathbf{x}_i(2L_{sft}) & & \cdots & & \mathbf{x}_i(L_w+2L_{sft}) \\ \vdots & & \ddots & & \vdots \end{bmatrix}, \tag{3}$$

where $\mathbf{H} \in \mathbb{R}^{N_{obs} \times L_w}$ and N_{obs} is the number of windows extracted from a single signal. This process is then repeated for the n available healthy vibration data signals and each matrix is concatenated to develop the data matrix $\mathbf{X}_{data}$ through

$$\mathbf{X}_{data} = \begin{bmatrix} \mathbf{H}(\mathbf{x}_1) \\ \mathbf{H}(\mathbf{x}_2) \\ \vdots \\ \mathbf{H}(\mathbf{x}_n) \end{bmatrix} \tag{4}$$

where $\mathbf{X}_{data} \in \mathbb{R}^{n*N_{obs} \times L_w}$. Common to any learning-based application, the $\mathbf{X}_{data}$ is then partitioned into a training and validation set that is used to train the model. It is a requirement for this work that the training data consist only of

signal segments from the asset in a healthy condition. For model training in this study, the shift parameter is selected as $L_{sft} = 0.5L_w$, and the data is standardised using

$$\mathbf{X}_{data} = \frac{\mathbf{X}_{data} - \boldsymbol{\mu}_{train}}{\boldsymbol{\sigma}_{train}}, \tag{5}$$

where $\boldsymbol{\mu}_{train} \in \mathbb{R}^{L_w}$ and $\boldsymbol{\sigma}_{train} \in \mathbb{R}^{L_w}$ are the healthy $\mathbf{X}_{data}$ matrix feature mean and standard deviation. The set of model parameters are then optimised using $\mathbf{X}_{data}$. Finally, it is conventional to use Equation (3) with $L_{sft} = 0.5L_w$, $\boldsymbol{\mu}_{train}$ and $\boldsymbol{\sigma}_{train}$ to process and standardise all available vibration data to evaluate HI metrics for each segment. The statistical properties of the HI measures are analysed to infer the condition of the asset.

In this work, the proposed *temporal preservation* approach preserves the temporal structure in the segments of $\mathbf{X}_{data}$ during the model evaluation process by re-evaluating Equation (3) with $L_{sft} = 1$. This step ensures that when HI and LHI measures are obtained, the temporal structure in the resulting HI or LHI signal is preserved. This preservation ensures the structure in the time-series data is now retained in the evaluation step. Furthermore, the HI and LHI signal is a transformed representation of the time-series signal. This allows for each segment to be consistently quantified as there is now a preserved relationship to time. To highlight additional considerations into the choice of model window length, consider Figure 5.

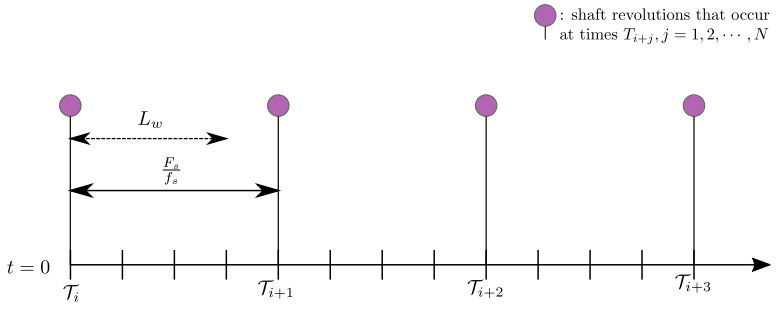


Figure 5: The interaction between model window length L_w, signal sampling frequency F_s and the shaft speed f_s visualised. Note here that the shaft speed is used as a proxy for the presence of faults, as fault frequencies are often proportional to the shaft speed.

Figure 5 describes the relationship between L_w, the signal sampling frequency F_s and the shaft speed f_s. The shaft speed and sampling frequency give an important lower bound to L_w, which develops due to the periodicity of localised faults in relation to the shaft speed. This bound assumes that the faults occur at least once per shaft revolution. If the window length L_w is less than the ratio $\frac{F_s}{f_s}$, one may obtain signal segments from a damaged signal that contains no fault covariates. Naturally, a combination of F_s, f_s and fault frequency proportion that ensures the presence of fault covariates within signal segments. However, we are not aware of a universal rule that overcomes this issue.

The *temporal preservation* approach offers an improved methodology that can be used to better analyse time-series signals and can improve the model diagnostic performance. By being aware of this phenomenon, a variety of signal analysis techniques can be performed, allowing for further exploitation and investigative flexibility for condition inference. Techniques that may be applied to the HI or LHI signals for condition inference interpretation are, for example *i)* signal statistics, *ii)* the time-synchronous average, and *iii)* a spectral analysis. This combination of deep learning and signal processing techniques enhances the condition inference procedure. It introduces a natural unification where deep learning highlights the fault covariates, and signal processing can be used to extract the fault information.

2.4. Latent variable models

In this study, we classify LVMs into three classes: i) explicit density linear models, ii) explicit density nonlinear models, and iii) implicit density nonlinear models that can be described within the framework depicted in Figure 6. The difference between linear and nonlinear models is a function of the encoding and decoding parametric functions. Explicit models commonly assume that distributions are Gaussian which introduces reconstruction losses. In contrast, implicit models use an adversarial scheme with a discriminative function to quantify the difference between the real and generated data. The simplest LVM[1] that is available is Principal Component Analysis (PCA) that assumes that $p(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{z})$ are explicitly parametrised by Gaussian distributions with linear transition functions [11]. In the PCA framework, Tipping and Bishop [27] show that the transition functions reduce to $f_\phi = \mathbf{U}$ and $g_\theta = \mathbf{U}^T$ where $\mathbf{U}$ is the hierarchically organised data covariance eigenvector matrix, often referred to as the Principal Component (PC) matrix. The Variational Auto-Encoder (VAE) framework generalises the linear PC transition function to learn a nonlinear parametric encoder and decoder network that represent the explicitly assumed Gaussian distributions $p(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{z})$. This data reconstruction loss guides this process $\mathcal{L}(\theta, \phi) = \frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2$, often referred to as the L_2 or mean-squared error (MSE) loss, and the latent Kullback-Liebler (KL) divergence loss $D_{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$, which regularises the encoder to enforce that data is encoded into a space similar to the assumed latent prior. GANs aim to improve the generative capacity of g_θ by replacing the explicit generative log-likelihood with the likelihood ratio $r = \frac{p_{data}(\mathbf{x}|\mathbf{z})}{q_\theta(\mathbf{x}|\mathbf{z})}$ [16]. This decision allows for the generative distribution to implicitly capture the data distribution, providing natural flexibility to g_θ extending past Gaussian distributions. However, the original GAN framework cannot perform model inference. Thus, various techniques have been proposed in the literature to produce a framework that benefits from the GAN adversarial framework while still recovering latent codes. In some frameworks, a combination of the adversarial auto-encoder (AAE) [28] and the InfoGAN framework [29] is used to improve the generative capacity of g_θ, and Mutual Information maximisation is used to introduce explicit latent disentanglement losses into the model [4, 23]. In Figure 6, a generic LVM training framework highlights different latent regularisation strategies and highlights how the GAN framework is typically incorporated for nonlinear implicit formulations. It is important to note that some implicit techniques still incorporate the auto-encoder losses, which induces a trade-off between the explicit and implicit frameworks. A suitable latent prior $p(\mathbf{z})$ for latent regularisation is an isotropic Gaussian distribution. Its isotropic nature can provide some implicit independence on the learnt posterior distribution to ensure that each latent dimension is independent [30].

To use LVMs for fault detection, the various facets of the model are exploited. For example, models that make explicit Gaussian distribution assumptions, such as PCA or VAEs, can access the reconstruction log-likelihood to measure the likelihood that a sample is from the generative distribution $p(\mathbf{x}|\mathbf{z})$. For models that use or incorporate implicit distribution assumptions such as GANs, a data discriminator $D_\chi(\mathbf{x})$ is made available to measure how likely an observed sample is in relation to the learnt healthy data distribution, as shown in Booyse et al. [10]. In Baggeröhr [19], an auto-encoder with an adversarial optimisation scheme was used as a latent regularisation methodology for the latent manifold alongside an InfoGAN framework to improve the generative capacity of the generative distribution. This technique uses a Wasserstein metric latent critic function $D_\omega(\mathbf{z})$ to measure how likely a latent sample was from the learnt latent space [31–33]. The

[1] It is emphasised at this stage that the reader is assumed to be familiar with LVM methods. Should this not be the case, we refer the reader to [25] and [26].

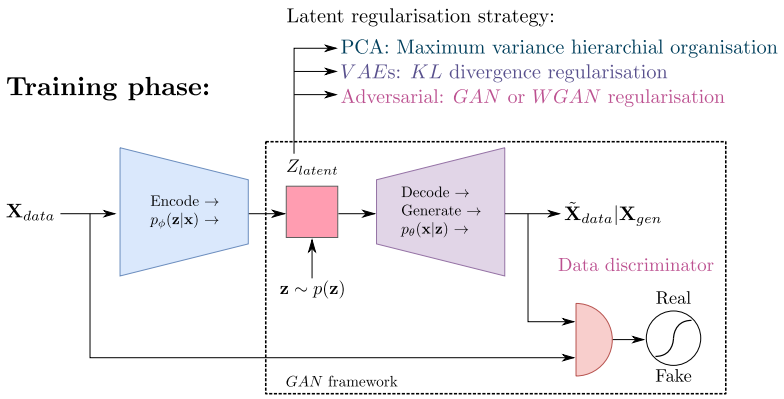


Figure 6: The LVM model training phase detailing the different model operations and potential latent regularisation strategies. For the GAN framework, model inference is typically unavailable as an encoder is never learnt, and there is no data encoding stage. The method is used as a pure data generative model.

current application of LVMs can consider both the data space and the latent manifold. However, the latent manifold has never been analysed to consider the path traversed through the manifold. Furthermore, latent critic regularisation is not LVM agnostic as the KL divergence regularisation term was explicitly replaced.

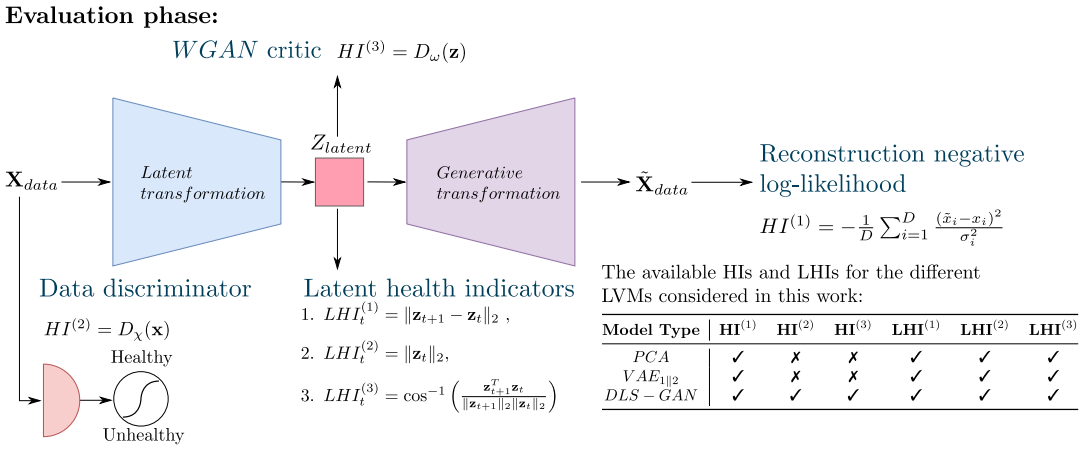
Once a latent variable model has been trained, it offers indicators from its latent and data space by feeding data through the model, as highlighted in Figure 3. Consider the case where a signal with a fault is fed through an LVM, with HI measures obtained using the current approach and the proposed *temporal preservation* approach. In Figure 1, an illustrative example of the differences between the conventional and proposed evaluation procedures is detailed. To assess whether this data is healthy, a healthy reference distribution, is created and the likelihood of the collected HI measures is evaluated, as detailed in Figure 1(a). However, suppose there are healthy signal segments in the processed vibration signal. In that case, one may obtain likelihoods that are not indicative of a fault, and the signal may be incorrectly deemed healthy. In the proposed procedure, the introduction of time can be used to visualise the HI measures through time. This process will correctly clarify the presence of the healthy segments due to the type of fault present in the data and the chosen model window length. Furthermore, as the HI signal is a transformation of the original time-series signal, alternative techniques such as the HI signal Fourier transform can confirm the source of the measure deviations, or the statistics of the HI signal can be trended through time.

It is emphasised here that the methods used in this paper are merely incidental. Numerous alternative techniques exist in the literature, but investigation of these techniques is beyond the scope of this work. This work primarily focuses on the temporal preservation analysis of latent spaces to derive temporally structured HIs from LVMs applied to time-series data. The aim is not to identify the best LVM or to propose an improved LVM. Unsupervised latent variable models are used to capture an asset's healthy manifold under the condition monitoring objective to monitor changes in the asset condition over time. In this study, the effect of model complexity is explored by considering linear and nonlinear parametric functions, and the impact of latent disentanglement [30] is investigated through implicit and explicit latent disentanglement regularisation. More powerful parametric functions may offer improved mapping flexibility to the latent manifold and may allow the manifold to capture data space nonlinearities. Latent disentanglement is crucial to latent manifold anomaly detection as an entangled manifold may hinder the condition inference procedure. Suppose damage

is present in the time-series data. In that case, the detection of these instances in an entangled latent manifold may be corrupted by changes in other generative factors such as the asset operating condition. Besides, time-preserved latent analysis enables time-preserved health indicators that preserve the temporal structure that allows for more distinct and informative HI to be developed and derived.

2.5. Proposed health indicators

It is possible to collectively identify potential HIs commonly available to LVMs, as detailed in Figure 7. These HIs can be informative, but important information in the data is lost when the temporal information is not preserved. This restriction on the data impedes the condition inference task and stops the full potential of LVMs from being realised. To assess the potential of LVMs, the temporal information in the data must be preserved to improve the quality of the common HIs and to provide access to a set of LHIs to track the representation of time-series data through the latent manifold.



Model Type	$\mathbf{HI}^{(1)}$	$\mathbf{HI}^{(2)}$	$\mathbf{HI}^{(3)}$	$\mathbf{LHI}^{(1)}$	$\mathbf{LHI}^{(2)}$	$\mathbf{LHI}^{(3)}$
PCA	✓	✗	✗	✓	✓	✓
$VAE_{1\|2}$	✓	✗	✗	✓	✓	✓
$DLS-GAN$	✓	✓	✓	✓	✓	✓

Figure 7: A detailed schematic of the HIs and LHIs obtained during LVM evaluation phase for the models considered in this work. To preserve the temporal information in the data to utilise the LHIs, $\mathbf{X}_{data}$ must be obtained using Equation (3) with $L_{sft} = 1$.

Commonly used HIs are functions of the model framework. For models that assume Gaussian generative distributions, a reconstruction metric is developed as

$$HI^{(1)}(\mathbf{x}, \tilde{\mathbf{x}}, \boldsymbol{\sigma}^2) = -\frac{1}{D}\sum_{k=1}^{D}\frac{(\tilde{x}_k - x_k)^2}{\sigma_k^2}, \tag{6}$$

which is the reconstruction log-likelihood for an input $\mathbf{x}$, its reconstructed mean $\tilde{\mathbf{x}}$ and variance $\boldsymbol{\sigma}^2$. This term can also be considered to be the negative squared Mahalanobis distance for a factored Gaussian distribution. The second HI that is available for implicit density estimation techniques is the GAN data discriminator

$$HI^{(2)}(\mathbf{x}) = D_\chi(\mathbf{x}), \tag{7}$$

which is based on the input feature space $\mathbf{x}$. For condition monitoring applications, the data discriminator $D_\chi(\mathbf{x})$ estimates the likelihood that any observed data is from the healthy asset data distribution. The third health indicator is that of the Wasserstein metric estimate

$$HI^{(3)}(\mathbf{z}) = D_\omega(\mathbf{z}), \tag{8}$$

which may be based in the input feature space or, as indicated, based on the latent feature space $\mathbf{z}$ if it is used as an adversarial latent regularisation strategy. This learnt metric measures whether any latent representation $\mathbf{z}$ is from the prior $p(\mathbf{z})$.

The *temporal preservation* approach allows for the development of LHIs that can be used to interpret where data lies in the latent manifold and detects off-manifold traversal. These LHIs form the foundation for latent analysis, where latent analysis refers to the analysis of the latent manifold. To detect changes in the latent manifold, three latent metrics are proposed in this work and are given as

$$LHI_t^{(1)} = \|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2, \tag{9}$$

$$LHI_t^{(2)} = \|\mathbf{z}_t\|_2, \tag{10}$$

$$LHI_t^{(3)} = \cos^{-1}\left(\frac{\mathbf{z}_{t+1}^T \mathbf{z}_t}{\|\mathbf{z}_{t+1}\|_2 \|\mathbf{z}_t\|_2}\right), \tag{11}$$

where $\mathbf{z}_t$ indicates the latent space representation at any point $t \in [0, L_d - 1]$ from the *temporal preservation* analysis procedure where $L_d = L_{signal} - L_w$ and $\|\cdot\|_2$ is the L_2 norm. $LHI^{(1)}$ is the latent distance norm between two time-continuous latent representations and allows one to interpret the inter-time distance characteristics of the latent space. $LHI^{(1)}$ is also a proxy to the latent traversal velocity if the data sampling rate F_s is constant. $LHI^{(2)}$ is the trivial calculation of the latent Euclidean norm and measures the distance of a latent sample from the Cartesian origin. $LHI^{(1)}$ was derived using the Mahalanobis distance, a commonly used distribution distance metric, for the isotropic Gaussian prior distribution $p(\mathbf{z})$. $LHI^{(3)}$ can be interpreted as the angle between two points in the latent space and allows one to interpret the directional characteristics of the latent manifold. Many alternative LHIs can be developed as this interpretation of the latent manifold has not been rigorously optimised and improved. The optimisation of the LHI set is beyond the scope of this work.

We use the proposed LHIs to gain insight into how unhealthy time-series data may be represented in a latent space representative of healthy data. The static properties of the manifold are quantified through the negative log-likelihood of a latent sample $\mathbf{z}$ using the prior $p(\mathbf{z})$, as the prior is time-invariant. The dynamic properties of the manifold can be quantified by considering the positional behaviour of the path followed through the latent space. $LHI^{(1)}$ quantifies the positional behaviour in a Cartesian coordinate system, while $LHI^{(3)}$ quantifies the positional behaviour in a non-Cartesian coordinate system. Static and dynamic behaviour can be distinctly isolated as the *temporal preservation* approach untangles time in the manifold, which allows for time axis exploitation to produce scalar metrics.

The combination of the *temporal preservation* analysis and the proposed LHIs can lead to an interpretable latent space and introduce a level of intuition into how different latent manifolds respond to anomalous data. It is also key to note that $LHI^{(1)}$ and $LHI^{(3)}$ cannot be intuitively developed without preserving the temporal structure of the data. The term interpretation is used in the context of a latent manifold to describe how these metrics quantify the manifold response to anomalous data instances.

3. Experimental investigation

The experimental investigation demonstrates informative generic HIs used with various LVMs when preserving time structure during the models' evaluation. The temporal preservation approach ensures that the model's full potential is used during the condition inference stage for LVM-based CM. This comes through augmenting conventional LVM analysis with signal processing techniques, an endeavour that is only possible through temporal preservation, which naturally allows for more fruitful insights. Not only can fault detection and trending occur, but the type of fault can also be inferred. In this way, the emphasis is not to identify the best HI, LHI and LVM combination or to propose the best combination, as this is problem-dependent. Rather, this study highlights the benefit offered from the temporal preservation approach over conventional approaches. The proposed *temporal preservation* approach and LHIs are evaluated using the considered LVMs on two experimental datasets. The first dataset has time-varying operating conditions, and the second dataset has stationary operating conditions. Health indicator denoising, albeit interesting, is beyond the scope of this work. For those interested, results provided by [34–36] are relevant.

3.1. Latent variable model application

For the experimental investigations in this work, the models considered are PCA, VAEs and the DLS-GAN method [11, 15, 23]. For PCA, a cumulative contribution rate (CCR) is defined as

$$CCR = \frac{\sum_{i=1}^{M} \lambda_i}{\sum_{i=1}^{N} \lambda_i}, \tag{12}$$

where λ_i is the i_{th} largest eigenvalue of the training data covariance matrix. This ratio allows for the data variance capture rate to be enforced to rationalise the number of M eigenvectors used, $M \leq N$. This work will investigate the effect of learning on output variance from the VAE decoder $\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I} = g_\theta(\mathbf{z})$. This variance may improve the discrepancy metric responses from $HI^{(1)}$ by quantifying the expected feature changes. Models with and without this learnt variance are denoted with subscripts VAE_2 and VAE_1, respectively. For the DLS-GAN method, the methodology proposed in Ding and Luo [23] is augmented to introduce a continuous latent code $\mathbf{s}$ with an isotropic latent prior $p(\mathbf{s}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. This decision splits the latent space Z_{latent} into three components, $Z_{latent} = [\mathbf{c}, \mathbf{s}, \mathbf{n}]$ where $\mathbf{c}$ is a discrete categorical code and $\mathbf{n}$ is a continuous random noise code. This decision adds two L_2 losses to the DLS-GAN formulation, one included in the generative InfoGAN element of the model and one added when enforcing latent $\mathbf{n}$ separation, thereby adding two additional β weighting parameters to the final DLS-GAN objective function. Furthermore, the latent component $\mathbf{n}$ is regularised with an adversarial training scheme to match an isotropic Gaussian prior $p(\mathbf{n}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ using a Wasserstein GAN critic with gradient penalty, which provides access to $HI^{(3)}$ for this model [31, 33]. This decision replaces $\mathcal{L}_{MMD}(\phi)$ with $\mathcal{L}_{WGAN}(\phi, \omega)$, which introduces an additional parametric critic function $D_\omega(\mathbf{z_n})$ that is used to regularise the encoder. Furthermore, the GAN generator loss is changed to use the KL divergence loss proposed in Sønderby et al. [37] to produce favourable generator gradients during training. Finally, all network architectures, hyper-parameters and activation functions required for work reproducibility are detailed in Appendix A.

3.2. Gearbox dataset

The experimental gearbox dataset used in this research was obtained from the Centre for Asset Integrity Management laboratory at the University of Pretoria. This dataset has been extensively analysed in the works of Schmidt et al. [34, 38].

This dataset was used as it exhibits time-varying operating conditions. As depicted in Figure 8, the experimental setup consists of an electrical motor, one step-down helical gearbox, two-step up helical gearboxes, and an alternator. The centre gearbox from Figure 8 was used in the experiments, and a tri-axial accelerometer was mounted to the bearing casing such that the axial acceleration could be measured. The input and output operating condition states of the gearbox are shown in Figure 9. Note that the Bayesian Geometry Compensation (BGC) method proposed by Diamond et al. [39] was used for the input shaft speed. The gearbox dataset consists of measurements from two distinct experiments. The first experimental dataset contains one hundred vibration data samples from the test gearbox in a healthy condition. The second experimental dataset contains two hundred measurements from the gearbox in a damaged state. To introduce damage, a single tooth had a slot seeded into its width. The vibration data in the second dataset was obtained until the gear tooth failed. The vibration signals from each dataset were $20s$ in length and were sampled at a rate of $F_s = 25.6kHz$. A proximity probe and an optical probe were mounted to the test gearbox's output and input shafts, respectively. The sampling rate of the optical probe was set to $F_s = 51.2kHz$.

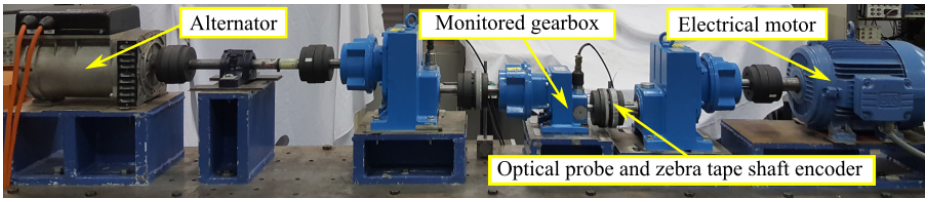
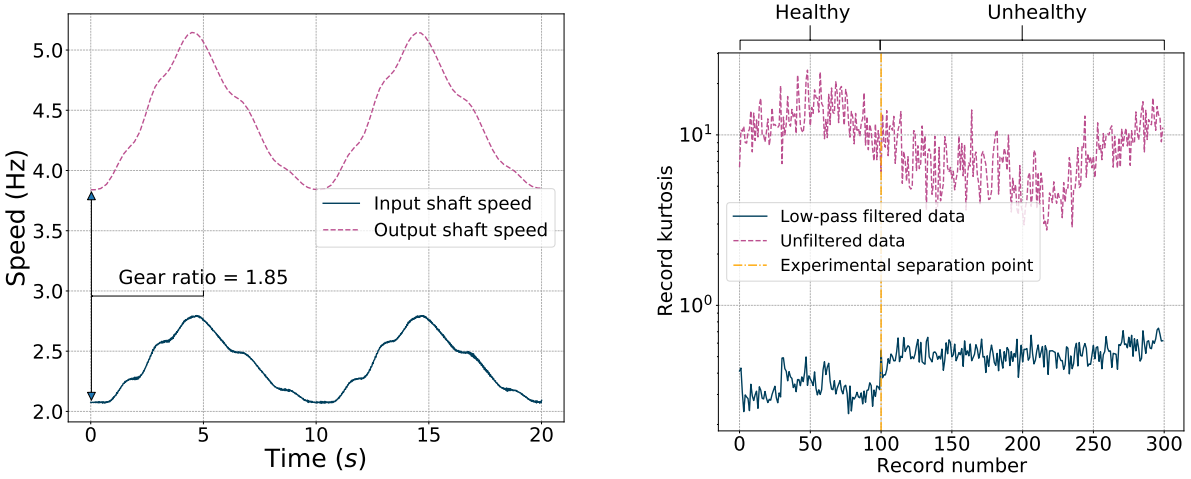


Figure 8: The C-AIM test gearbox experimental setup.



(a) shaft speeds f_s

(b) Record kurtosis

Figure 9: The shafts speeds and record kurtosis for the low-pass filtered and unfiltered versions of the gearbox dataset. The low-pass filtering was conducted at a cut-off frequency of $3200Hz$ with a third-order Butterworth filter.

The objective of this dataset is to highlight that by using the *temporal preservation* approach, more informative HIs can be obtained from LVMs, the presence of damage can be uncovered in the HI signals and in the latent manifold and this damage can be visualised using a unification of signal processing and deep learning techniques. To investigate this dataset, a simple statistical investigation was performed into the signals from the two datasets. To quantify the dataset complexity, two versions of the gearbox dataset were considered. The first version is when the time-series data were low-pass filtered

using a third-order Butterworth filter around $3200Hz$, and the second version is the original unfiltered dataset. This was done as the unfiltered dataset's data contained significant impulses in both the healthy and unhealthy states. In Figure 9(b), the record kurtosis is given. A detectable change in the record kurtosis for the filtered dataset occurs around the transition from the healthy dataset to the unhealthy dataset. The effect of low-pass filtering is clear in Figure 9(b), where the unfiltered record kurtosis is significantly larger and clearly shows the presence of the signal impulsiveness that was filtered out of the filtered dataset. Due to this clear impulsivity, the authors chose to investigate the performance on the low-pass filtered dataset to compare the different metrics' performance without complicating the signal information with non-damage impulsive components, as was done in the work of Chen et al. [40]. The investigation of LVM performance under complex non-Gaussian signal information will be considered in future work.

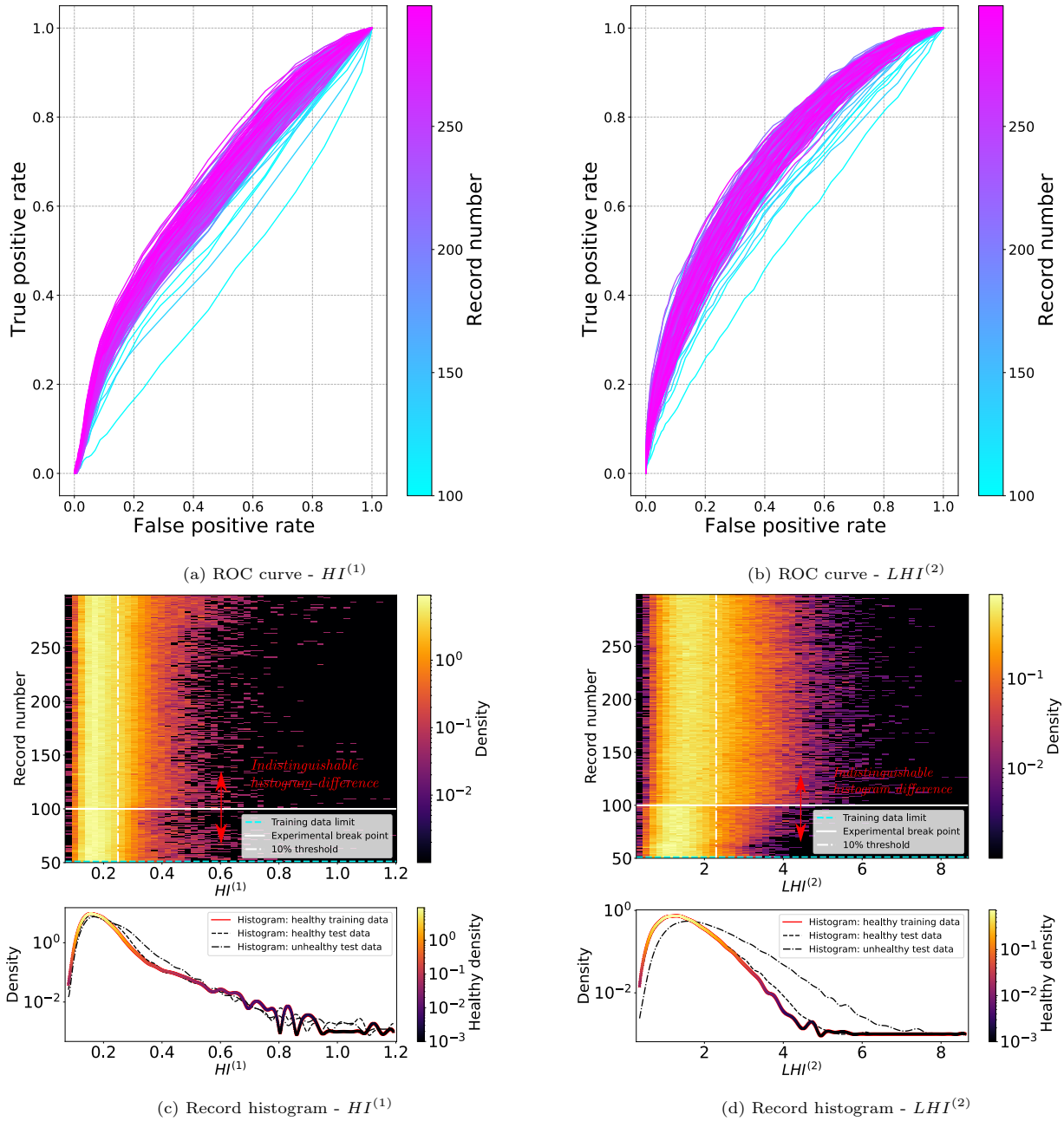
3.2.1. Model Development

To develop the training and validation datasets for model optimisation, the first fifty records from the healthy dataset were low-pass filtered and then processed using Equation (3) with $L_{sft} = 0.5L_w$. The dataset $\mathbf{X}_{data}$ was then partitioned into $\mathbf{X}_{train}$ and $\mathbf{X}_{valid}$ using an $80-20\%$ ratio with signal segments chosen randomly. Once the model parameters were optimised, $\mathbf{X}_{data}$ was re-evaluated using all the available data for the proposed *temporal preservation* approach from Section 2 and standardised using $\boldsymbol{\mu}_{train}$ and $\boldsymbol{\sigma}_{train}$ from the original training dataset. This re-evaluation step is not performed in the conventional setting, and the only LHI measure available is $LHI^{(2)}$. Condition indicator signals were obtained from each model by passing $\mathbf{X}_{data}$ through the models and developing discrepancy signals using the HIs and LHIs as detailed in Figure 7. To quantify the benefit of the proposed approach, the results from the conventional analysis setting will be shown.

3.2.2. Conventional analysis

In the conventional setting, LVM-based CM analysis is a histogram analysis procedure, as shown in Figure 3(a). This procedure uses the histograms of the healthy training data HI and LHI measures to compare and quantify whether any new measures are from the healthy asset state. To this effect, the receiver operating characteristic (ROC) curve, area under the ROC curve (AUC) and the classification accuracy become features of interest. The ROC curve tracks the false positive rate (FPR) against the true positive rate (TPR) as the healthy histogram threshold is varied against a set of test HI measures. In this work, the AUC metric is favoured over the ROC curve as it summarises the ROC curve information in a single scalar. A fault classification analysis is also performed. Following [10], a healthy/unhealthy threshold is chosen to induce a 10% FPR. For CM applications, a 10% FPR on the healthy data implies that 10% of the healthy HI or LHI measures used for model training are classified as damaged. The classification accuracy under the 10% threshold reflects the likelihood of health state deviation in the signals of interest, which serves as a proxy to fault detection.

To showcase the ROC curve and histogram results for this dataset, the results from the VAE_1 model are shown in Figure 10. It is clear from Figures 10(a) and (b) that although there is a fault present from record 100, the ROC curves for the $HI^{(1)}$ and $LHI^{(2)}$ measures are insensitive to the fault. The ROC curves for the records with damage never obtain a good TPR at low FPRs, highlighting the shortcoming of the conventional approach as damage is not identifiable. In Figures 10(c) and (d), it is clear that there is some increase in the HI and LHI measures from record 100, but this is not significant enough to show a clear deviation from the learnt healthy state. The ROC curve responses in Figures 10(a) and (b) are explained by Figures 10(c) and (d), as there is clear histogram overlaps and no distinguishable histogram deviation



(a) ROC curve - $HI^{(1)}$

(b) ROC curve - $LHI^{(2)}$

(c) Record histogram - $HI^{(1)}$

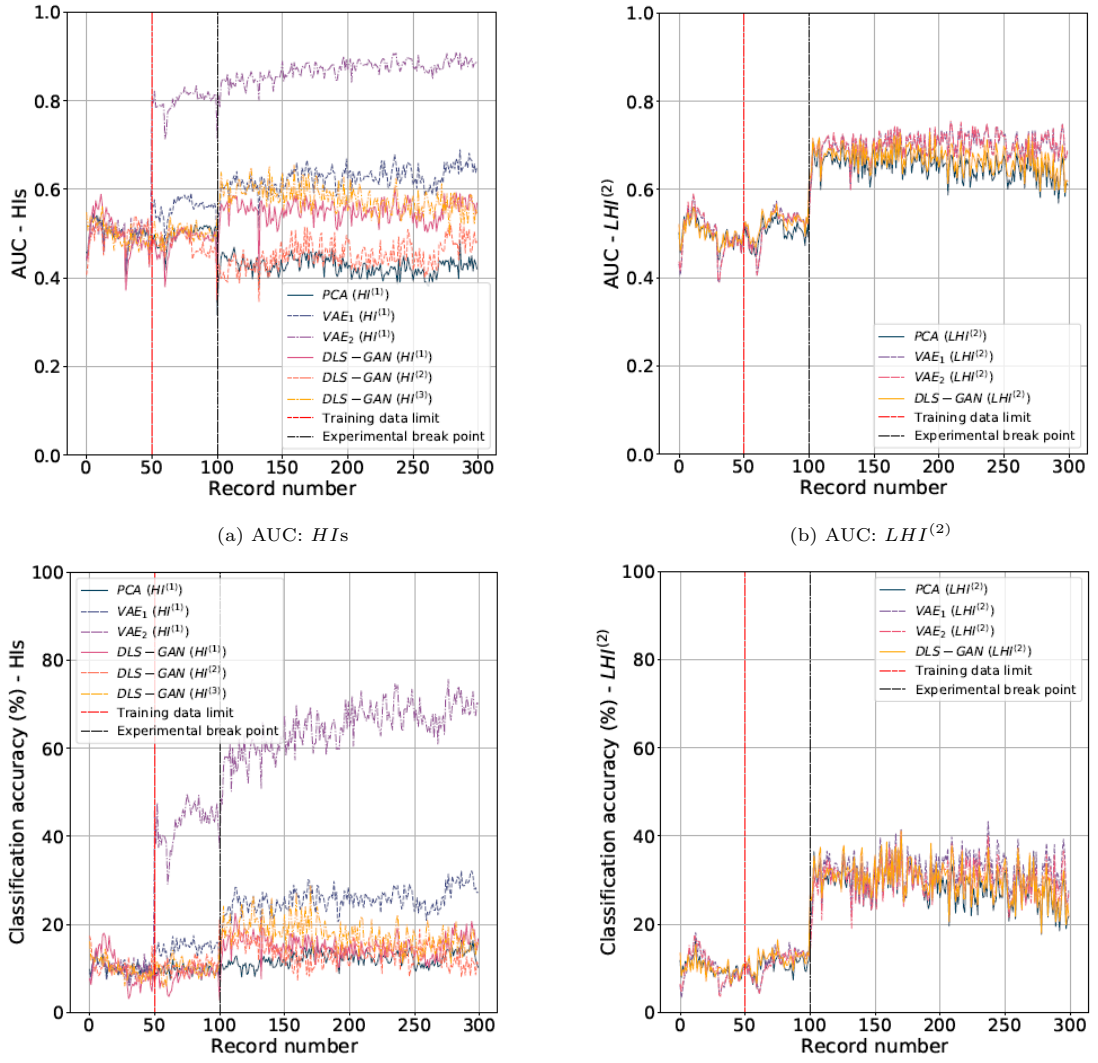
(d) Record histogram - $LHI^{(2)}$

Figure 10: The ROC curves and record histograms for the VAE_1 model. Note the clear inability to detect the damage present in the conventional analysis setting. In Figures (a) and (b), the ROC curves are only developed for the records with damage. The ROC curves shown are purely demonstrative, and the AUC is used in this work to summarise the ROC curve information. In Figures (c) and (d), the record histograms are shown from record 50, as the healthy density contains the HI measures for the time-series data used for model training.

from the healthy state in the damaged records. In Figure 10(d) the shift in the histogram for the unhealthy test data is attributed to the experimental breakpoint.

In Figure 11, the AUC and classification accuracies for the models considered in this work are shown. The AUC and classification accuracy indicate poor performance from $HI^{(1)}$ and $LHI^{(2)}$ as there is no growth from record 100, except for the VAE_2 model. The VAE_2 response growth is attributed to reconstruction loss overfitting to the training data. However, it is clear that the latent manifold AUC and classification accuracy has no overfitting phenomenon present. This

indicates the duality of the learnt variance reconstruction loss and the learnt manifold, as the model has exploited the additional variance flexibility to reconstruct the training data better. In contrast, the manifold indicates that the data between records 50 and 100 are from the same healthy distribution. Figure 11 demonstrates that limited insight can be drawn from the models for CM condition inference under the conventional analysis setting. The conventional analysis approach is limited by the reformulation of the inference objective to a classification setting, which relies on strong and consistent HI or LHI measure deviations. Furthermore, the latent manifold cannot be further investigated or explored as there is no temporal structure.



(a) AUC: HIs

(b) AUC: $LHI^{(2)}$

(c) Classification accuracy: HIs

(d) Classification accuracy: $LHI^{(2)}$

Figure 11: The AUC and classification accuracies for the models considered. It is expected in Figures (a) and (b) that records 0 to 100 have an AUC measure of around 0.5, and in Figures (c) and (d), these records are expected to have a classification accuracy of around 10% as these records are healthy. Note in Figures (a) and (c) that the $HI^{(1)}$ response for the VAE_2 exhibits a jump at record 50 in the $HI^{(1)}$ measures, which is attributed to model reconstruction sensitivity to data, not from the training set. This sensitivity is not present in the $LHI^{(2)}$ response in Figures (b) and (d).

3.2.3. Temporal preservation analysis

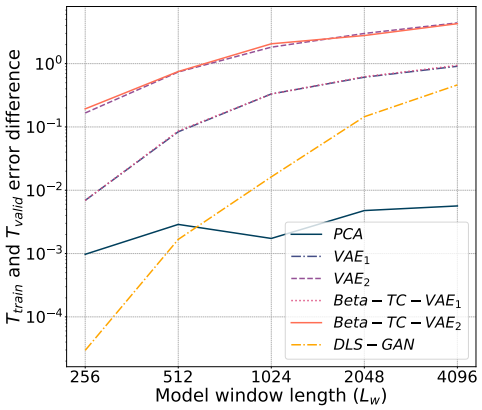
In contrast to the conventional approach, it is possible to use conventional signal analysis methods to process the health indicators under temporal preservation. In this section, the time-synchronous average (TSA) of the HI and LHI signals is used for condition inference and obtained from

$$\overline{\mathbf{x}}_{TSA}(n) = \frac{1}{N_r}\sum_{i=0}^{N_r-1}\mathbf{x}(n+iN_s),\ \text{where } 1 \leq n \leq N_s, \tag{13}$$

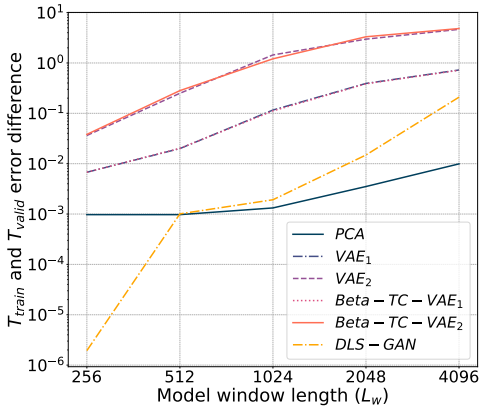
where the signal $\mathbf{x}(n)$ is processed to contain N_r rotations consisting of N_s points per rotation [41, 42]. The geometrical imperfection of the tachometer butt joint is used to align the different measurements. It is important to highlight that signal processing techniques are used to enrich the condition inference procedure, but the models were trained on the raw vibration data. This demonstrates the power of the *temporal preservation* approach, as it allows for the natural exploitation of signal processing techniques that can uncover the presence of fault covariates. Condition inference may be performed using the processed discrepancy signal's statistics. However, it was found that these statistics were insufficient to detect damage on this dataset. For the models used for this dataset, the window length was chosen as $L_w = 512$. Figure 12 shows the difference in average feature reconstruction error between the training and validation set to motivate this decision. The latent size and model window size affect the model's reconstruction ability, with a clear decreasing trend as the window length is lowered. This is attributed to the information in each signal segment, as shorter segments have less information that the model has to capture and decreases the emphasis of larger frequency components. The chosen window length was selected as it is near the region of error stability, and it offers a model with a larger number of parameters to improve model flexibility.

In Figure 13, the $HI^{(1)}$ and $LHI^{(1-3)}$ TSA responses are shown from the PCA model. In Figure 13(a), the gear tooth fault is around 140° but to prove that the fault has occurred is non-trivial due to apparent activity at other angular increments. It is clear from Figures 13(b-d) that there is damage present in the latent manifold and specifically emphasised by $LHI^{(2)}$. Figures 13(c) and (d) show a clear spike attributed to the tooth fault's presence. There is some noticeable deviation around record one hundred, which manifests from the experimental interruption to seed the fault.

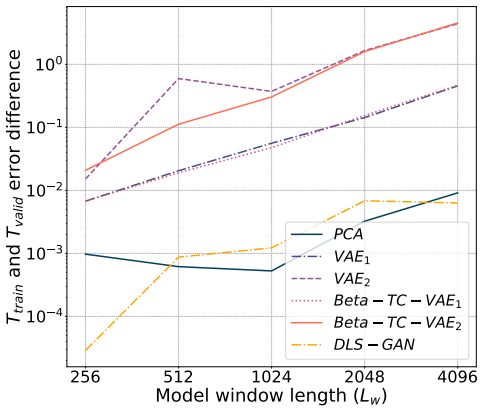
Figures 14 and 16 show the $HI^{(1)}$ and $LHI^{(1-3)}$ TSA responses for the VAE_1 and VAE_2 models. The benefit of transition function nonlinearity is evident, whereby the response from $HI^{(1)}$ clearly shows the presence and development of the fault. The effect of the VAE_2 model's learnt variance is evident through a greater response magnitude in Figure 14(b). To demonstrate the implicit nature of the model window length on discrepancy signals, consider the $HI^{(1)}$ TSA signal as shown in Figure 15 for window lengths of $L_w = 1024$ and $L_w = 2048$ using the VAE_2 model. Notice the decreased influence of the gear tooth fault. This phenomenon occurs as each segment captures more of the impulsive components that were not filtered out, explaining the deviations in Figure 12 as the window length increases. This degrades the fault covariate information captured by the HI discrepancy signal, and the fault inference procedure is therefore degraded. Under the examination of the latent manifold, shown in Figure 16, through latent metrics $LHI^{(1-3)}$, it is clear that the latent distance, $LHI^{(1)}$, is a poor metric for condition inference on this dataset. Both $LHI^{(2)}$ and $LHI^{(3)}$ show the presence of the tooth fault. Interestingly, the latent manifold is responding similarly to that seen from PCA in Figure 13. This is attributed to the shared L_2 loss that is used to develop these models. The L_2 loss is obtained from the explicit assumption that the generative distribution is Gaussian. This explicit choice of distribution may also explain why we cannot analyse simple discrepancy signal statistics or larger model windows, as the remaining impulsive signal



(a) $Z_{latent} = 25$, $CCR = 0.5$.



(b) $Z_{latent} = 50$, $CCR = 0.75$.

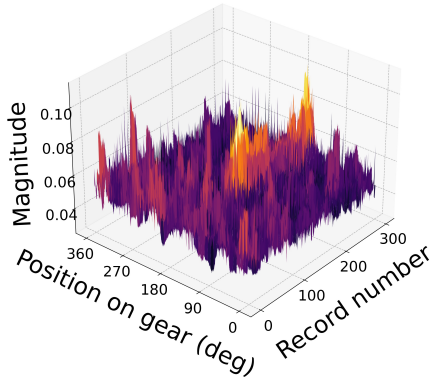


(c) $Z_{latent} = 100$, $CCR = 0.95$.

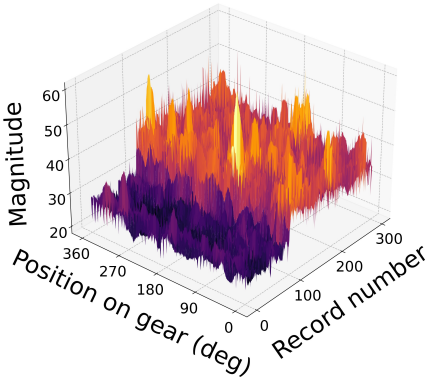
Figure 12: The effect of latent dimensionality and window length L_w on model reconstruction error. Note that the $x-$axis was $\log_2$ scaled, and the $y-$axis was $\log_{10}$ scaled. The reconstruction error is a proxy metric as it is a consistent metric indicative of training performance. For PCA, the CCR rate was investigated rather than the size of Z_{latent}, and for DLS-GAN, Z_{latent} refers to the dimensionality of the $\mathbf{n}$ space.

components that were not filtered out are known to be strongly non-Gaussian.

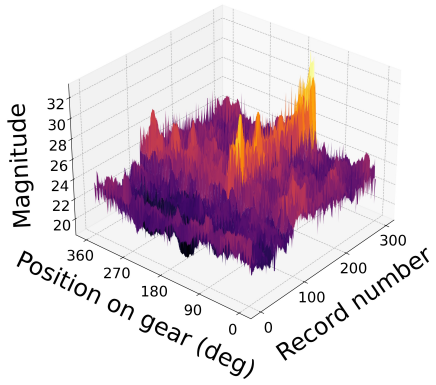
In Figure 17, the $HI^{(1-3)}$ and $LHI^{(1-3)}$ TSA responses are shown from the DLS-GAN model. This model offers two additional discrepancy metrics, and these metrics analyse both data samples and latent samples. It is clear that the parametric function nonlinearity improves the results obtained, with improved reconstruction metric response under the TSA. The data discriminator response in $HI^{(2)}$ is notably uninformative, with no detectable fault development. This is attributed to the conflicting training objective from the AAE and InfoGAN frameworks. This difference manifests in the learnt data distribution from the explicit and implicit training schemes for the AAE and GAN frameworks. This difference negatively favours training the data discriminator, as the MSE loss will naturally dominate the training as it is more stable. The latent critic response, $HI^{(3)}$, shown in Figure 17(c), indicates some damage, but there is some activity at other angular positions. This is attributed to the critic's design to use only latent samples and not exploit their time dependency. Thus, the manifestation of natural data perturbations in the latent manifold may corrupt the condition inference process. This is explained in Figures 17(d), (e) and (f), which shows the latent metric response. $LHI^{(2)}$ and $LHI^{(3)}$ are informative metrics for condition inference. This satisfactory metric response is shared with the other models considered, which indicates the impact of the L_2 loss on the learnt latent manifold. The consistently poor performance of



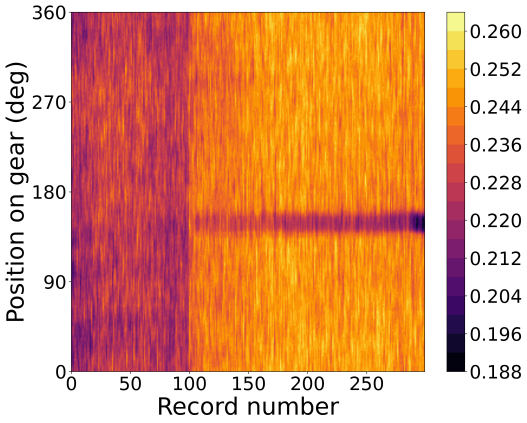
(a) $HI^{(1)}$



(b) $LHI^{(1)}$

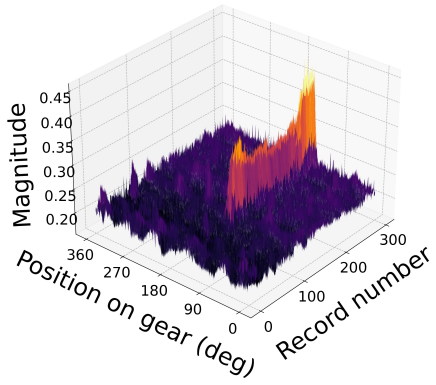


(c) $LHI^{(2)}$

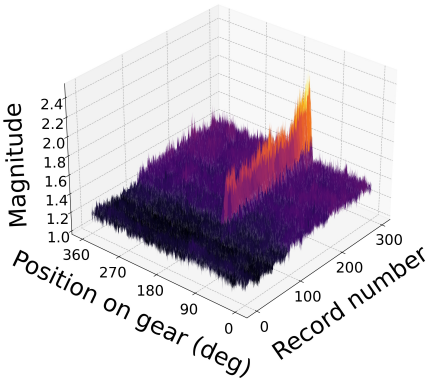


(d) $LHI^{(3)}$

Figure 13: The HI and three LHI metric responses under the synchronous average using PCA with the *temporal preservation* analysis approach.

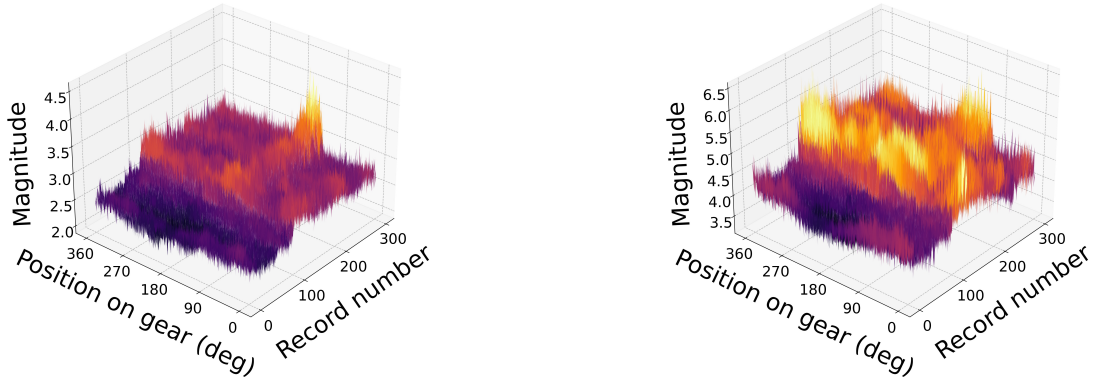


(a) $HI^{(1)}$: VAE_1



(b) $HI^{(1)}$: VAE_2

Figure 14: The $HI^{(1)}$ discrepancy signal synchronous average for VAE_1 and VAE_2 models for the gearbox dataset.



(a) $HI^{(1)} : L_w = 1024$

(b) $HI^{(1)} : L_w = 2048$

Figure 15: The $HI^{(1)}$ signal content degradation as the window length is increased for the VAE_2 model.

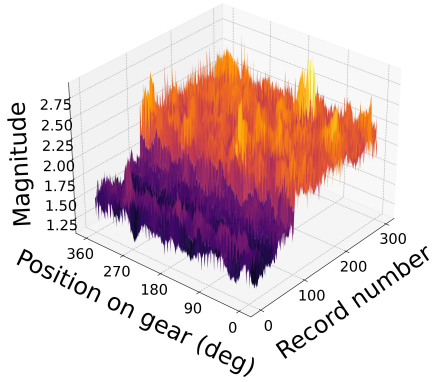
$LHI^{(1)}$ may indicate that the encoding transition function attempts to preserve the traversal rate. In doing so, anomalous instances are mapped into unknown regions of the manifold. This is another benefit of the *temporal preservation* approach, as it allows one to interpret the dynamics of the latent manifold.

3.2.4. Discussion

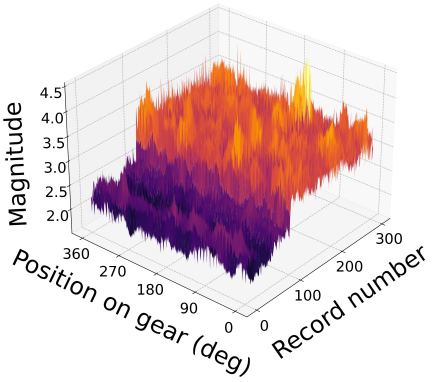
From the presented results in Sections 3.2.2 and 3.2.3, the benefits of the proposed method are clear: by preserving time and using the latent information, it is possible to find powerful health indicators that can be analysed using established signal analysis methods. By utilising signal analysis methods, it is possible to enhance the anomalous components and to identify their source. This section compares the performance of the health indicators obtained using the *temporal preservation* approach and the conventional approach. To quantify the temporally preserved HI and LHI response performance, it is required that a condition deviance point be prescribed to the various TSA responses. This is non-trivial for the current TSA representations. This requires that a technique be used that transforms the individual HI and LHI signals into an interpretable result that can be trended through time. Schmidt et al. [43] introduced a method that uses $k-$means clustering to assign two centroids μ_{max} and μ_{min}, to a single discrepancy signal. These means are trended through time, and the larger of the two is used for condition inference. This approach may be used should one wish to identify a condition deviance point, but this is beyond the scope of this work.

The objective of this discussion is to further highlight the benefit of the *temporal preservation* approach over the conventional approach through direct comparison. To ensure comparison consistency, the VAE_1 model records histograms for the TSA of the HIs and LHIs are shown in Figure 18. The histograms are centred by removing the median from the TSA for each record. This step is not integral to the proposed approach. Rather, it emphasises the damage sensitivity in the TSA and where damage manifests in the record histograms. Removing the median is not universally applicable to all fault cases. However, as the localised fault has been identified this has no impact and this step simply removes any data nonstationarity that is not due to the tooth fault.

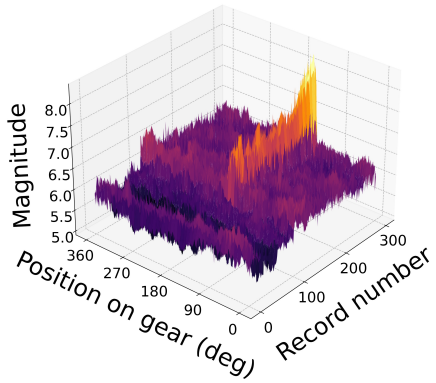
In Figures 18(a) (c) and (d), the damage sensitivity regions manifest in the histogram outliers. These outliers are attributed to the gear tooth fault damage. The damage response manifests in the variance of the histograms, and there



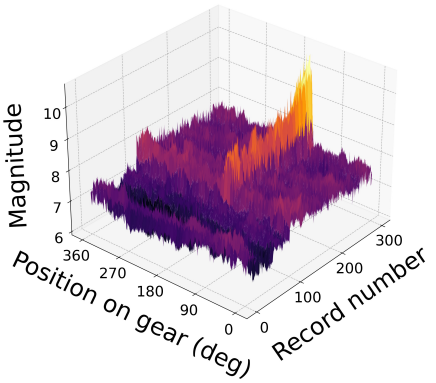
(a) $LHI^{(1)}$: VAE_1



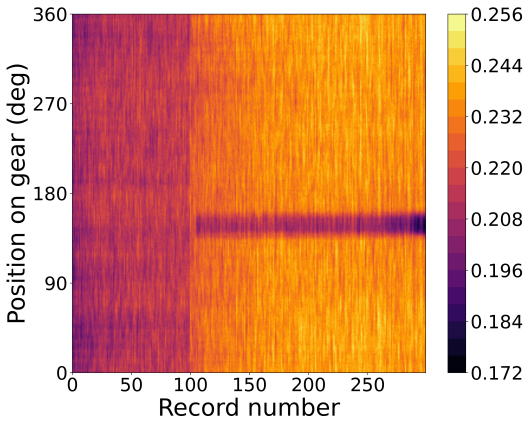
(b) $LHI^{(1)}$: VAE_2



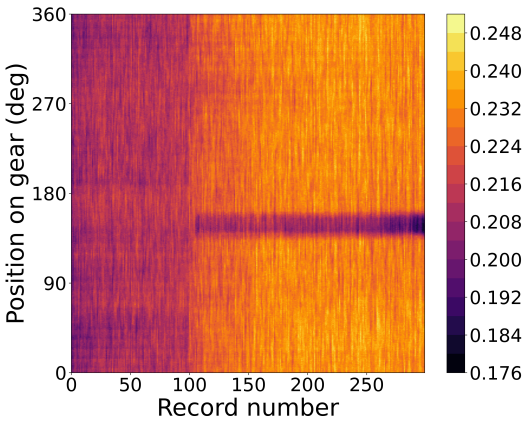
(c) $LHI^{(2)}$: VAE_1



(d) $LHI^{(2)}$: VAE_2



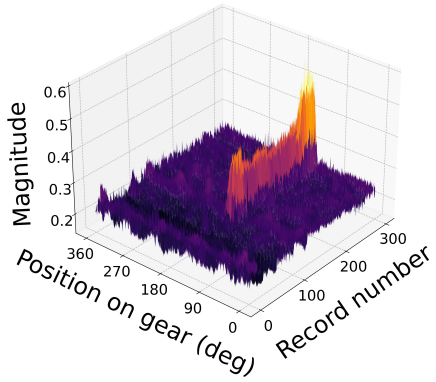
(e) $LHI^{(3)}$: VAE_1



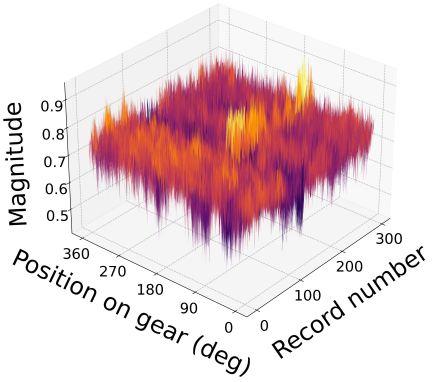
(f) $LHI^{(3)}$: VAE_2

Figure 16: The LHI synchronous average responses from the VAE_1 and VAE_2 models for the gearbox dataset.

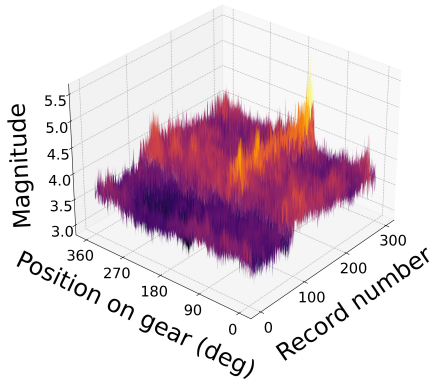
are clear healthy and unhealthy regions. The LHI_1 response seen in Figure 18(b) is expected as there is no apparent damage seen in Figure 16(a). This variance manifestation is caused by the interaction between L_w and the fault frequency,



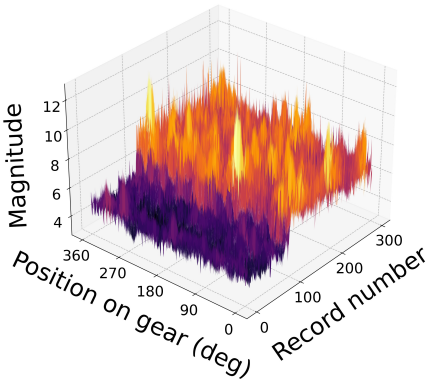
(a) $HI^{(1)}$: DLS-GAN



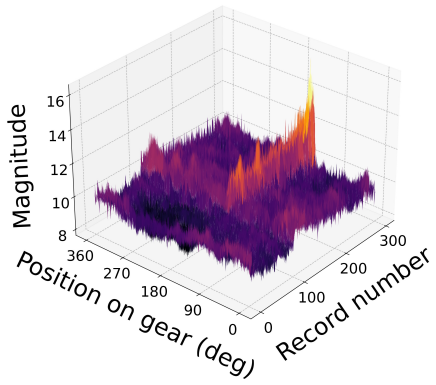
(b) $HI^{(2)}$: DLS-GAN



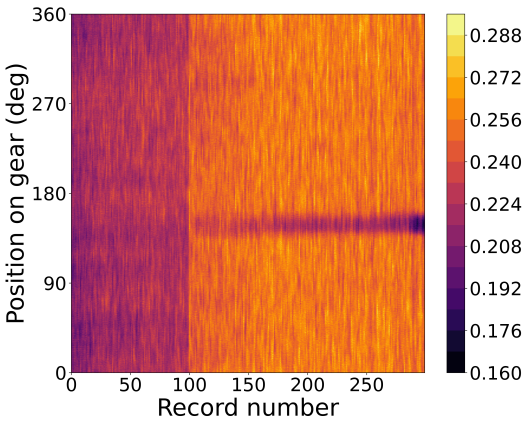
(c) $HI^{(3)}$: DLS-GAN



(d) $LHI^{(1)}$: DLS-GAN



(e) $LHI^{(2)}$: DLS-GAN



(f) $LHI^{(3)}$: DLS-GAN

Figure 17: The three HI and LHI metrics obtained from the DLS-GAN model analysed using the synchronous average.

which indicates that there are many signal segments that are healthy, with only a few segments that contain the localised damage. Hence, metrics that monitor deviations in the variance of the TSA are suited to identify condition deviance

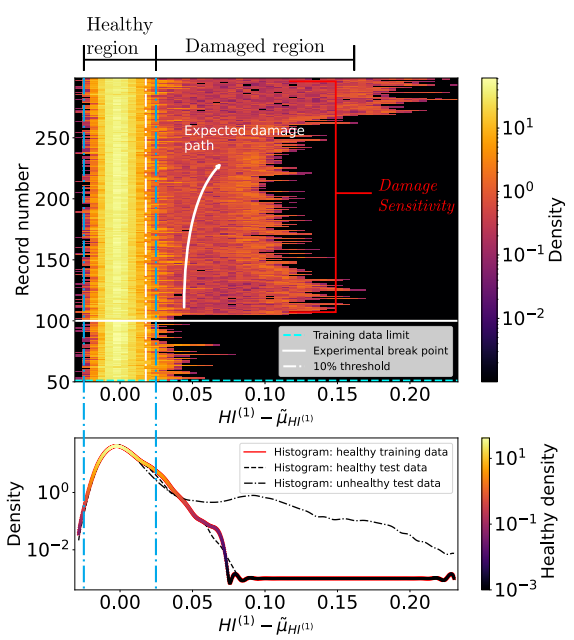
points, such as the method proposed by [43]. This result is crucial to highlight the shortcoming of the conventional approach, as the interaction between L_w and the fault may cause the ROC, AUC and classification accuracy metrics from the conventional approach to be biased towards poor performance. Naturally, L_w may be increased to reduce this bias, but, as shown in Figure 15, this decreases the ability to perform fault localisation. For this reason, the AUC and the classification accuracy are not shown for the *temporal preservation* approach, as this information is not a true representation of the model responses for the chosen L_w. However, this level of insight is only available through the *temporal preservation* approach, which enriches the condition inference procedure and deepens analysis inference.

When comparing Figures 10 and 18, it is clear that the temporal preservation approach offers numerous benefits during the condition inference stage. Damage was detected using the HIs and the proposed LHIs, highlighting that preserving temporal structure is key to manifold interpretation in LVM-based CM. The results of the HIs and LHIs under the *temporal preservation* approach indicates that the decision to obtain more informative health indicators is beneficial to condition monitoring and will enable future research to improve the methods used by using an improved analysis methodology. The use of simple signal processing techniques, such as the TSA, can enhance the condition inference stage by exploiting the temporal structure that is preserved during model evaluation.

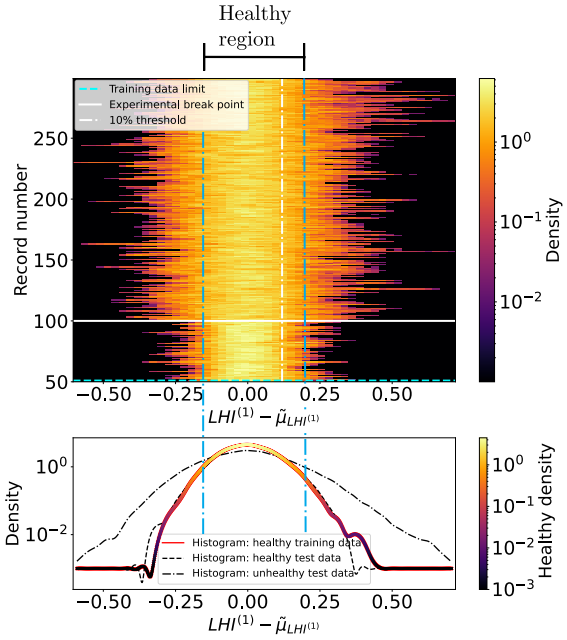
3.3. IMS dataset

The Intelligent Maintenance Systems (IMS) dataset is a well-used bearing failure dataset used for benchmarking condition monitoring methodologies [44]. This dataset consists of three run-to-failure sub-datasets containing accelerometer data from an experimental setup, shown in Figure 19, consisting of four double row Rexnord ZA-2115 bearings. The experimental setup ran under a constant load of $6000lb$s and a constant shaft speed of $2000rpm$. In each dataset, a $1s$ vibration sample was obtained under a sampling rate of $F_s = 20.48kHz$. For this work, the second IMS dataset will be investigated, and this dataset contained 984 measurements throughout the experimental lifespan. The second IMS dataset had an outer race bearing fault that occurred in the first bearing. The IMS dataset is investigated to ensure that the investigated models and proposed LHIs apply to a dataset with constant operating conditions. This investigation ensures that the results obtained are compared to state-of-the-art methods in the literature. By ensuring that our analysis methodology is appropriate, competitive performance results are obtained using vastly different LVM methods.

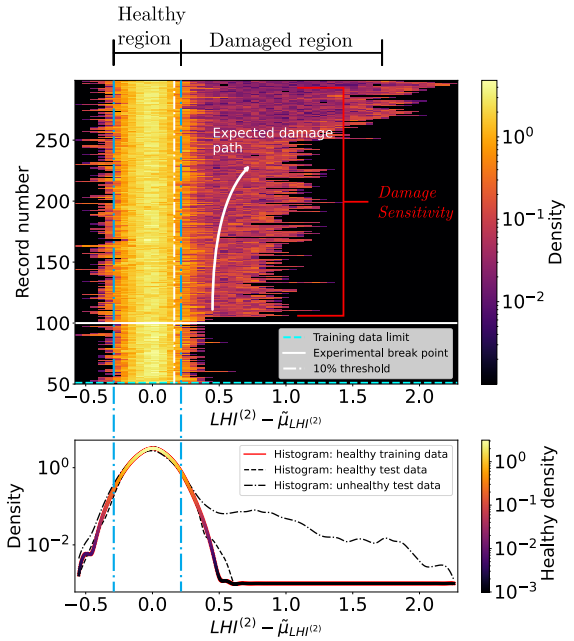
To develop the training and validation datasets for model optimisation, the first 10% (98 records) of the available data were assumed to be healthy data. The processing and model optimisation strategy used for the Gearbox dataset in Section 3.2 was followed, and the *temporal preservation* approach was employed during the development of the HI and LHI signals during model evaluation. The authors chose to investigate two model window lengths as $L_w = 512 \leq \frac{F_s}{f_s} = \frac{20480}{33.3} = 615.02 \leq L_w = 4096$ bound the implicit window ratio $\frac{F_s}{f_s}$. This dataset will demonstrate how simple HI and LHI signal statistics can be used for novelty detection. The HI and LHI signal statistic used for condition inference is the signal mean. To detect a condition deviance point, a condition threshold was determined using the median $\tilde{\mu}$ and standard deviation σ of the HI and LHI signal means that correspond to the time-series data used for model training, and this threshold is given as $thres = \tilde{\mu} \pm 6\sigma$. This approach was chosen to ensure result comparison consistency and comparison flexibility as the threshold is conservative by design. This selection is not integral to the proposed framework, however, to ensure that the healthy data satisfies the Gaussian distribution assumption for the threshold, the authors performed a hypothesis test analysis using the D'Agostino k-squared test [45], Anderson-Darling test [46] and Shapiro-Wilk test [47].



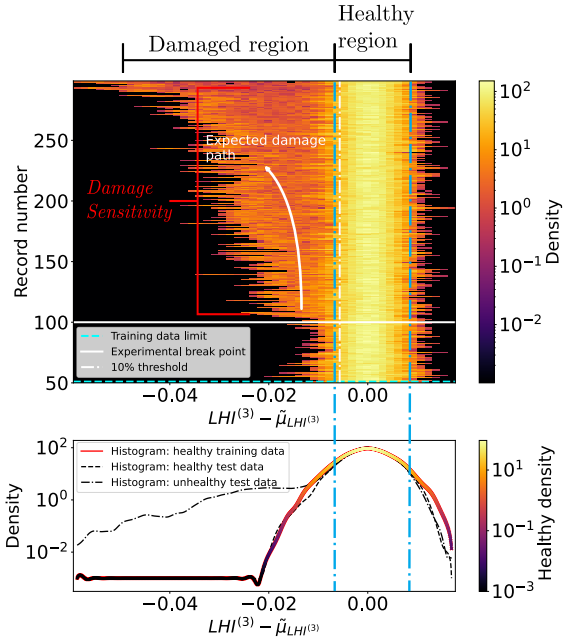
(a) Record histogram - $HI^{(1)}$.



(b) Record histogram - $LHI^{(1)}$.



(c) Record histogram - $LHI^{(2)}$.



(d) Record histogram - $LHI^{(3)}$

Figure 18: The record histogram for the TSA of the $HI^{(1)}$ and $LHI^{(1-3)}$ responses for the VAE_1 model. Notice the growth of the histogram variance in Figures (a), (c) and (d). This variance/outlier growth is due to the gear tooth fault. Note that the record histograms are shown from record 50 as the healthy histograms summarise records 0 to 50.

These tests validated the Gaussian assumption. An example of the test results is given in Appendix B. To detect a change in condition, the mean of the five succeeding points is used. If this mean is greater than the threshold, the record is identified as a deviance point. As the model performance has been adequately demonstrated in Section 3.2, a single set of results is shown to the reader in Figure 20. The results from the literature are shown in Table 1, and the results for the considered LVMs are tabulated in Table 2.

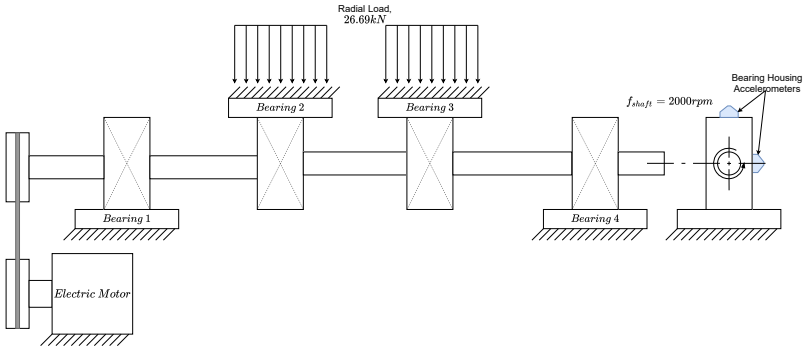


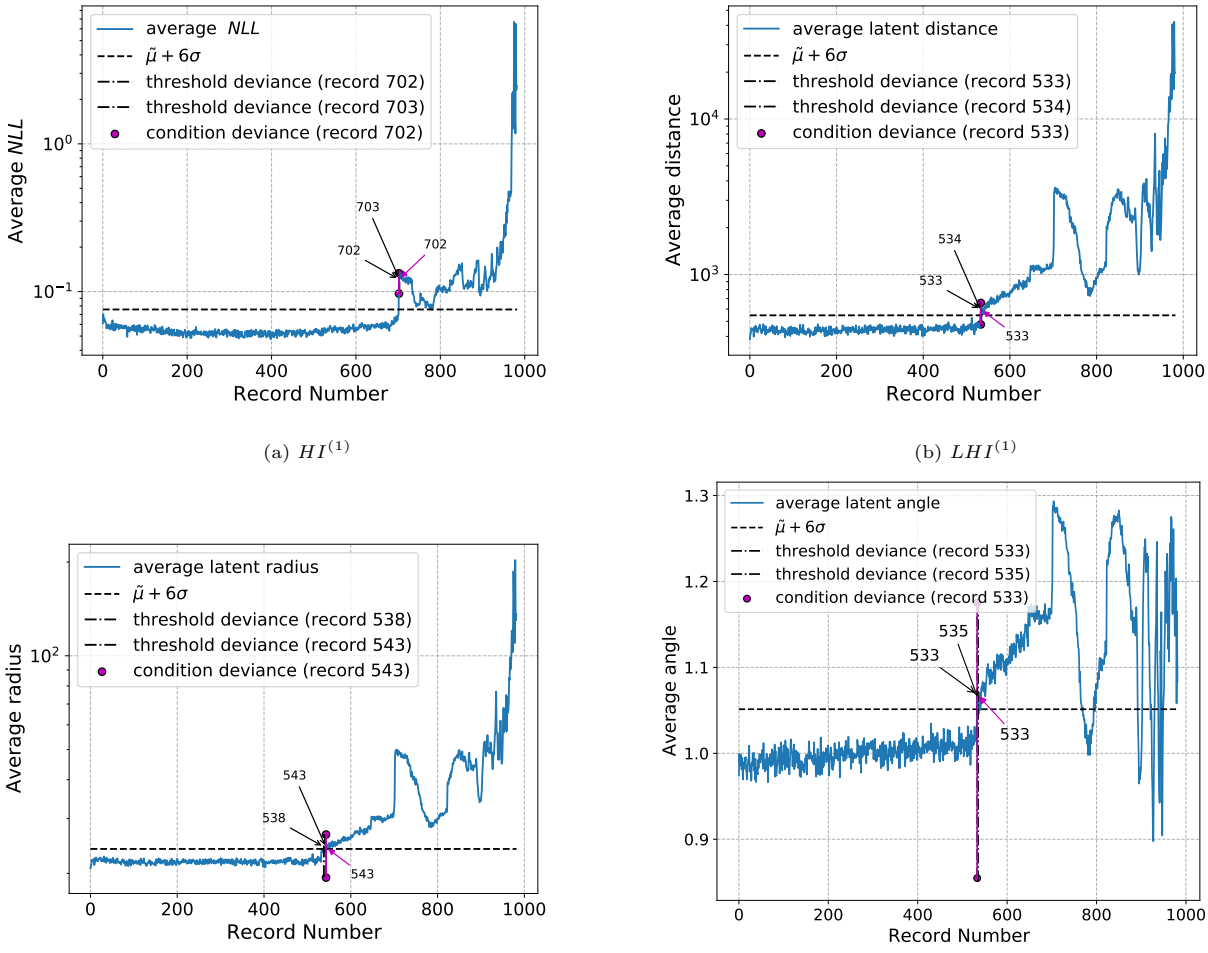
Figure 19: The IMS experimental setup [44]. Notice the presence of four bearings with two accelerometers per bearing.

In Figure 20, it is clear that the linear PCA LVM is able to sufficiently capture and describe the data. While the reconstruction HI appears to detect damage late, the proposed LHIs all detect damage earlier than the reconstruction HI. The PCA model's performance indicates that this dataset is simple to capture, as a linear LVM can sufficiently capture the healthy data distribution and the HIs and LHIs are sensitive to anomalous instances. To quantify whether the deviations are due to the development of a fault, Figure 21 was developed by taking the discrete Fourier transform of each HI and LHI signal. In Figure 21(a), the $HI^{(1)}$ record-frequency plot is shown, and in Figure 21(b) the frequency content at the known BPFO component through time is tracked from each available HI and LHI from the PCA model. This result indicates that the fault development can now be connected to an outer race fault, highlighting how the *temporal preservation* approach bridges techniques between signal processing and deep learning to enhance the condition inference procedure.

In the analysis of Table 1, it is clear that the results from signal processing literature appear more consistent in identifying a condition deviance point around record 533. This is also attributed to the current focus of deep learning literature, which is often concerned with semi-supervised learning or transfer learning techniques [5]. Signal processing methodologies often require more effort to design a suitable framework. Most supervised learning or semi-supervised learning methods used in the literature require historical fault data that is non-trivial to obtain in practice. The use of LVMs as a novelty detection methodology can overcome both these issues, and the performance of the LVMs shown in Table 2 indicates that the *temporal preservation* approach is a suitable model evaluation process that deep learning models should consider when being applied as a novelty detection mechanism. The performance of the PCA model also indicates the benefit of considering different classes of LVMs for learning-based condition monitoring techniques, as this model requires minimal computational resources and has noticeable benefits in its ease of implementation.

When the results detailed in Table 2 are compared to Table 1, the performance obtained from the considered LVMs is on par with the current state-of-the-art techniques from both learning-based literature and signal processing literature. Interestingly, in Table 2, the effect of a larger model window length appears to hinder the nonlinear models, with worsened performance in the $HI^{(1)}$ indicators when compared to their shorter window length alternative. This is attributed to the signal information present in segments from larger window length models and the relationship to model parametric flexibility, where more flexible models may over-fit slightly to the training data. The $LHI^{(2)}$ performance appears to weaken for models of larger window length to the point where, for some models, a response is obtained that is not indicative of any damage, as indicated by table elements containing IC. This highlights the requirement to consider multiple LHI metrics, as the latent manifold is a function of the dataset. The DLS-GAN model provides satisfactory

performance from all elements except the data discriminator D_χ. This is attributed to the conflicting training objective and L_2 dominance, which causes poor discriminator training as the generator is less flexible.



(a) $HI^{(1)}$

(b) $LHI^{(1)}$

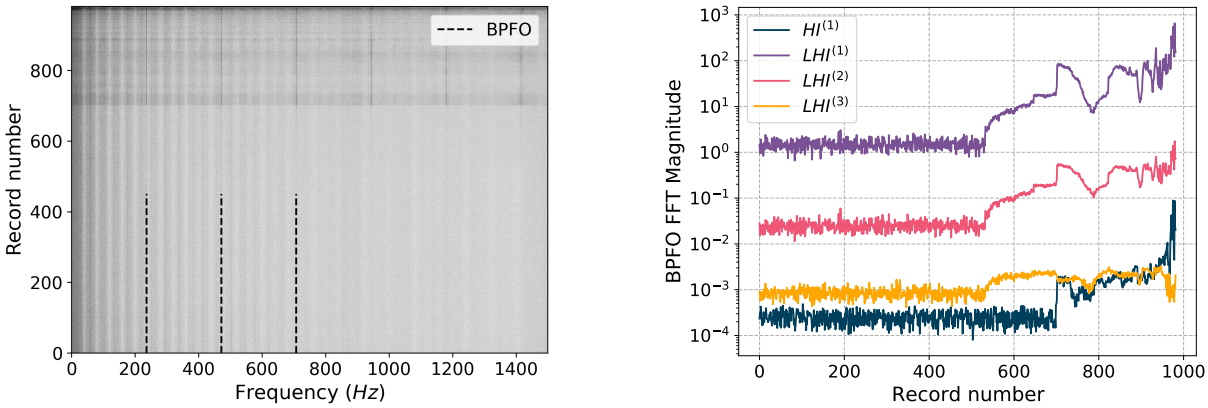
(c) $LHI^{(2)}$

(d) $LHI^{(3)}$

Figure 20: The three LHIs obtained using a PCA model with $L_w = 512$ for the first bearing from the second IMS dataset. Notice the strong response to damage that is identifiable through the latent manifold under the *temporal preservation* approach.

4. Conclusion and recommendations

In this work, a *temporal preservation* approach and a set of model agnostic LHIs are proposed and evaluated on experimental data consisting of stationary and time-varying operating conditions. This work highlights the requirement to improve the procedures to analyse LVMs before an optimal model is proposed. This contribution contrasts current techniques, whereby the focus is placed on how time-series data is analysed rather than on improving the training or the LVM. The use of various LVMs, trained only on raw time-series asset data in a healthy condition, as a condition monitoring technique is explored, and the responses from the data space and latent space are investigated. It is shown that both components of the model are indicative of damage. However, the temporal structure of the data must be retained to detect anomalous instances in the latent manifold. The detection of anomalous instances in the latent manifold improves model interpretability and allows for physical intuition to be built into the model by investigating the dynamics of time-series



(a) $HI^{(1)}$ record-frequency plot

(b) $BPFO$ frequency content in the HI and LHI signals

Figure 21: The $HI^{(1)}$ discrepancy signal Fourier transform plot and the $BPFO$ frequency component for each available HI and LHIs using the PCA model with $L_w = 512$. Notice how the threshold deviations in Figure 20 can now be confirmed by analysing the discrepancy signals' frequency content and using bearing characteristic fault frequency information.

Table 1: A summary of the certain results from the literature for the IMS dataset two for bearing one.

Literature type	**Method used**	**Article author**	**Deviance record**
Signal processing	Cyclostationary indicators	Antoni and Borghesani [7]	534
	$MED - SK - NES$, IES	Abboud et al. [48]	533, 542
	Frequency spectral coherence indicator	Kass et al. [49]	540 (3.4 days)
	QAM	Hou et al. [50]	533
	SAM impulsivity	Moshrefzadeh [51]	533
	WPT-based discrepancy analysis	Schmidt et al. [20]	535
	Weighted normalised square envelope	Wang et al. [52]	533
Learning-based	EMD-ANN	Ben Ali et al. [53]	553
	PCA LS-SVM	Dong and Luo [54]	700
	RUL using T-S FIS	Huang et al. [55]	630 (approximated)
	RUL using MLGTCN	Li et al. [56]	653
	NSVDD	Liu and Gryllias [57]	532 (best)
	SDIAE	Mao et al. [12]	Classification
	MoGAN	Zareapoor et al. [14]	Classification

data as it traverses through the latent space. The benefit of using LVMs to capture the asset's healthy state allows one to bypass the requirement for extensive expert knowledge and any labelled failure data. This naturally overcomes the limitations of supervised learning methodologies that have been developed for condition monitoring. The benefit of the *temporal preservation* approach is the ability to treat the HI or LHI signal as a processed representation of the original time-series signal, which can then be further analysed using simple signal statistics or using signal processing techniques. This result introduces a natural unification between the two fields. It allows for improved condition inference as the type

Table 2: The obtained threshold condition deviance point from the second IMS dataset for bearing one when investigating the available HIs and LHIs. Note that IC is the abbreviation used for results deemed inconclusive by the authors.

Model type and characteristics		**Health Indicator condition deviance point**					
Model used	**Window length**	**HI**$^{(1)}$	**HI**$^{(2)}$	**HI**$^{(3)}$	**LHI**$^{(1)}$	**LHI**$^{(2)}$	**LHI**$^{(3)}$
PCA	$L_w = 512$	702	N/A	N/A	**533**	**538**	**533**
	$L_w = 4096$	**543**	N/A	N/A	**533**	579	**533**
VAE_1	$L_w = 512$	**533**	N/A	N/A	578	648	**554**
	$L_w = 4096$	579	N/A	N/A	702	IC	622
VAE_2	$L_w = 512$	**545**	N/A	N/A	550	702	556
	$L_w = 4096$	647	N/A	N/A	586	IC	621
DLS-GAN	$L_w = 512$	**537**	621	**544**	**533**	**542**	**533**
	$L_w = 4096$	567	578	571	549	647	**542**

of fault can be uncovered if external information such as tachometer information or fault frequency information is used in the analysis framework.

It is proposed that future research must consider implicit density estimation techniques that can perform model inference. This will allow the generative distribution to capture more complex data distributions and allow the latent manifold to be explored and interpreted for condition inference purposes. Furthermore, alternative techniques such as normalising and auto-regressive flows can be investigated to determine whether the ability to evaluate $p(\mathbf{x})$ exactly is beneficial over generative likelihood HIs. Additionally, as the flows are designed to be invertible, it is expected that the LHIs proposed in this work can be readily applied.

Acknowledgements

The authors gratefully acknowledge the support that was received from the Eskom Power Plant Engineering Institute (EPPEI) in the execution of the research.

Appendix A. Network architectures and parameters

In this work, Pytorch was used to optimise and train the models [58]. The system used to train and evaluate the models had an Intel i7-8750H processor and a Geforce RTX 2070 graphics card. The decision was made to generalise the network architecture design based on the window length L_w to simplify the model development process for consistent network design referencing. A fixed stride and kernel size was defined for any network that used a convolutional layer, with the padding used to ensure that equal feature map division could be found from one layer to the next. The convolutional layer design used in this work consists of: $L_{stride} = 4$, $L_{kernel} = 32$, $L_{padding} = \frac{L_{kernel} - L_{stride}}{2} = 14$. Under these properties, the network convolutional or de-convolutional layer input size is enforced to undergo a reduction or expansion of a factor of four. A fully connected layer of size of $L_{FC_1} = 800$ is connected at the intermediary level to move from a convolutional layer. A second fully connected layer transitions to the latent space dimensionality or to a single node in a discriminator

Table 3: A table showing the basic network architecture for an encoder network for $L_{stride} = 4, L_{kernel} = 32$ and $L_{pad} = 14$. Note that N is the batch size and if one wishes to design a decoder network, steps $1 - 5$ need to be reversed.

Network depth level	**Layer Operator**	**Output layer dimensionality**
0	Input layer	$\mathbb{R}^{N\times 1\times L_w}$
1	Convolution	$\mathbb{R}^{N\times 32\times \frac{L_w}{4}}$
2	Convolution	$\mathbb{R}^{N\times 64\times \frac{L_w}{16}}$
3	Convolution	$\mathbb{R}^{N\times 128\times \frac{L_w}{64}}$
4	Fully-connected	$\mathbb{R}^{N\times 800}$
5	Fully-connected (output layer)	$\mathbb{R}^{N\times Z_{latent}}$

or critic network. For a tabular visualisation of what these architectures may look like for E_ϕ, G_θ or D_χ, please refer to Table 3.

For the latent critic $D_\omega(\mathbf{z_n})$ used in this work, a simple three-layer fully connected network is used where the dimensionality follows the process: $\mathbb{R}^{\mathbf{z_n}} \to \mathbb{R}^{3000} \to \mathbb{R}^{3000} \to \mathbb{R}^1$. In Table 4, the relevant activation functions for the different model components are given. In Tables 5 and 6, the latent dimensionality and hyper-parameters used for each dataset are given. To train the models in this work, the Adam and AdamW methods were used for the VAE and DLS-GAN methods, respectively, with parameters $\beta_1 = 0.6$, $\beta_2 = 0.999$ [59, 60]. Instance noise was also used for the first three thousand epochs of the DLS-GAN training [37].

Table 4: A table showing the basic network activation functions that were used alongside the general architectures in Table 3.

Method	**Hidden Layer Activations**	**Output Layer Activations**
VAE: $\mathbf{z} \in \mathbb{R}^{Z_{latent}}$	ReLU for all hidden layers	μ : linear, σ^2: softplus (for both encoder and decoder)
DLS-GAN: $\mathbf{z} \in \mathbb{R}^{Z_{latent}=[\mathbf{c},\mathbf{s},\mathbf{n}]}$	Encoder: leaky ReLU ($a = 0.2$) for all hidden layers Decoder: leaky ReLU ($a = 0.2$) for all hidden layers $D_\chi(\mathbf{x})$: Spectral Normalisation [61] on all hidden layers, ReLU for all hidden layers $D_n(\mathbf{n})/D_c(\mathbf{c})$: ReLU for all hidden layers	Encoder $[\mathbf{c}, \mathbf{s}, \mathbf{n}]$: softmax, linear, linear Decoder: linear $D_\chi(\mathbf{x})$: sigmoid $D_n(\mathbf{n})/D_c(\mathbf{c})$: Linear/Sigmoid

Table 5: The relevant latent dimensionality of the different models trained on different datasets.

LVM	**Gearbox Dataset**	**IMS dataset**
PCA	95% CCR	95% CCR
VAE_1 and VAE_2	$Z_{latent} = 50$	$Z_{latent} = 100$
DLS-GAN	$Z_{latent} = \mathbb{R}^{[10,10,150]}$	$Z_{latent} = \mathbb{R}^{[10,10,128]}$

Appendix B. Gaussian hypothesis test results

In Table 7, the results from three different Gaussian tests are shown with the corresponding p-values or test statistics, respectively.

Table 6: The learning rate η, reconstruction loss coefficient λ_{AE} and β parameters for the different models trained on different datasets.

LVM	Gearbox Dataset	IMS dataset
VAE_1 and VAE_2	$\eta = 1e-4$	$\eta = 1e-4$
DLS-GAN	$\eta = 1e^{-5}, \lambda_{AE} = 40, \beta_{1-6} = 1$	$\eta = 1e^{-5}, \lambda_{AE} = 80, \beta_{1-6} = 1$

Table 7: The Gaussian hypothesis test results for the VAE_1 model with $\alpha = 5\%$.

Model type	Health indicator metric	D'Agostino K-squared test	Anderson-Darling test	Shapiro-Wilk test	Acceptance total
VAE_1 ($L_w = 512$)	$HI^{(1)}$	Accept (p = 0.727)	Accept (0.154 < 0.758)	Accept (p = 0.900)	3
	$LHI^{(1)}$	Accept (p = 0.459)	Accept (0.283 < 0.758)	Accept (p = 0.289)	3
	$LHI^{(2)}$	Accept (p = 0.257)	Accept (0.300 < 0.758)	Accept (p = 0.378)	3
	$LHI^{(3)}$	Accept (p = 0.536)	Accept (0.210 < 0.758)	Accept (p = 0.605)	3

References

[1] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, D. Siegel, Prognostics and health management design for rotary machinery systems - Reviews, methodology and applications, Mechanical Systems and Signal Processing 42 (1-2) (2014) 314–334. doi:10.1016/j.ymssp.2013.06.004.
URL http://dx.doi.org/10.1016/j.ymssp.2013.06.004

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

[3] S. Schmidt, P. S. Heyns, Normalisation of the amplitude modulation caused by time-varying operating conditions for condition monitoring, Measurement: Journal of the International Measurement Confederation 149 (2020) 106964. doi:10.1016/j.measurement.2019.106964.
URL https://doi.org/10.1016/j.measurement.2019.106964

[4] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, R. X. Gao, Deep learning and its applications to machine health monitoring, Mechanical Systems and Signal Processing 115 (2019) 213–237. doi:10.1016/j.ymssp.2018.05.050.
URL https://doi.org/10.1016/j.ymssp.2018.05.050

[5] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, A. K. Nandi, Applications of machine learning to machine fault diagnosis: A review and roadmap, Mechanical Systems and Signal Processing 138 (2020) 106587. doi:10.1016/j.ymssp.2019.106587.
URL https://doi.org/10.1016/j.ymssp.2019.106587

[6] D. Wang, K. L. Tsui, Theoretical investigation of the upper and lower bounds of a generalized dimensionless bearing health indicator, Mechanical Systems and Signal Processing 98 (2018) 890–901. doi:10.1016/j.ymssp.2017.05.040.
URL http://dx.doi.org/10.1016/j.ymssp.2017.05.040

[7] J. Antoni, P. Borghesani, A statistical methodology for the design of condition indicators, Mechanical Systems and Signal Processing 114 (2019) 290–327. doi:10.1016/j.ymssp.2018.05.012.
URL https://doi.org/10.1016/j.ymssp.2018.05.012

[8] P. Večeř, M. Kreidl, R. Šmíd, Condition indicators for gearbox condition monitoring systems, Acta Polytechnica 45 (6) (2005) 35–43. doi:10.14311/782.
URL https://ojs.cvut.cz/ojs/index.php/ap/article/view/782

[9] N. Sawalhi, The application of spectral kurtosis to bearing diagnostics, in: Acoustics - Conference, no. November, 2004, pp. 393–398.
URL http://www.acoustics.asn.au/conference{_}proceedings/AAS2004/ACOUSTIC/PDF/AUTHOR/AC040115.PDF

[10] W. Booyse, D. N. Wilke, P. S. Heyns, Deep digital twins for detection, diagnostics and prognostics, Mechanical Systems and Signal Processing 140 (2020) 106612. doi:10.1016/j.ymssp.2019.106612.
URL https://doi.org/10.1016/j.ymssp.2019.106612

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

[12] W. Mao, W. Feng, Y. Liu, D. Zhang, X. Liang, A new deep auto-encoder method with fusing discriminant information for bearing fault diagnosis, Mechanical Systems and Signal Processing 150 (2021) 107233. doi:10.1016/j.ymssp.2020.107233.
URL https://doi.org/10.1016/j.ymssp.2020.107233

[13] X. Wu, Y. Zhang, C. Cheng, Z. Peng, A hybrid classification autoencoder for semi-supervised fault diagnosis in rotating machinery, Mechanical Systems and Signal Processing 149 (2021) 107327. doi:10.1016/j.ymssp.2020.107327.
URL https://linkinghub.elsevier.com/retrieve/pii/S0888327020307135

[14] M. Zareapoor, P. Shamsolmoali, J. Yang, Oversampling adversarial network for class-imbalanced fault diagnosis, Mechanical Systems and Signal Processing 149 (2021) 107175. doi:10.1016/j.ymssp.2020.107175.
URL https://doi.org/10.1016/j.ymssp.2020.107175

[15] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114arXiv:1312.6114.
URL http://arxiv.org/abs/1312.6114

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, arXiv:1406.2661arXiv:1406.2661.
URL http://arxiv.org/abs/1406.2661

[17] C. Fefferman, S. Mitter, H. Narayanan, Testing the manifold hypothesisarXiv:1310.0425, doi:10.1090/jams/852.
URL http://arxiv.org/abs/1310.0425

[18] H. Shao, H. Jiang, H. Zhao, F. Wang, A novel deep autoencoder feature learning method for rotating machinery fault diagnosis, Mechanical Systems and Signal Processing 95 (2017) 187–204. doi:10.1016/j.ymssp.2017.03.034.
URL http://dx.doi.org/10.1016/j.ymssp.2017.03.034

[19] S. Baggeröhr, A deep learning approach towards diagnostics of bearings operating under non-stationary conditions, Masters thesis, University of Pretoria (2019).

[20] S. Schmidt, P. S. Heyns, K. C. Gryllias, A discrepancy analysis methodology for rolling element bearing diagnostics under variable speed conditions, Mechanical Systems and Signal Processing 116 (2019) 40–61. doi:10.1016/j.ymssp.2018.06.026.
URL https://linkinghub.elsevier.com/retrieve/pii/S0888327018303583

[21] S. Schmidt, P. S. Heyns, Localised gear anomaly detection without historical data for reference density estimation, Mechanical Systems and Signal Processing 121 (2019) 615–635. doi:S0888327018307696.
URL https://www.sciencedirect.com/science/article/pii/S0888327018307696?dgcid=rss{_}sd{_}all

[22] T. Heyns, P. S. Heyns, J. P. De Villiers, Combining synchronous averaging with a Gaussian mixture model novelty detection scheme for vibration-based condition monitoring of a gearbox, Mechanical Systems and Signal Processing 32 (2012) 200–215. doi:10.1016/j.ymssp.2012.05.008.
URL http://dx.doi.org/10.1016/j.ymssp.2012.05.008

[23] F. Ding, F. Luo, Y. Yang, Clustering by directly disentangling latent space.
URL http://arxiv.org/abs/1911.05210

[24] D. Eisenbud, Linear sections of determinantal varieties, American Journal of Mathematics 110 (3) (1988) 541–575. doi:10.2307/2374622.
URL http://www.jstor.org.uplib.idm.oclc.org/stable/2374622

[25] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
URL http://www.deeplearningbook.org

[26] D. M. Blei, Build, compute, critique, repeat: data analysis with latent variable models, Annual Review of Statistics and Its Application 1 (2014) 203–232. doi:10.1146/annurev-statistics-022513-115657.

[27] M. E. Tipping, C. M. Bishop, Probabilistic principal component analysis, Journal of the Royal Statistical Society. Series B: Statistical Methodology 61 (3) (1999) 611–622. doi:10.1111/1467-9868.00196.

[28] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, B. Frey, Adversarial autoencoders, Arxiv:1511.05644arXiv:1511.05644.
URL http://arxiv.org/abs/1511.05644

[29] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, InfoGAN: interpretable representation learning by information maximizing generative adversarial nets.
URL http://arxiv.org/abs/1606.03657

[30] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, in: ICLR, Vol. 44, 2017, pp. 807–831.

[31] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International conference on machine learning, 2017, pp. 214—-223.
URL http://arxiv.org/abs/1701.07875

[32] S. Zhang, Y. Gao, Y. Jiao, J. Liu, Y. Wang, C. Yang, Wasserstein-Wasserstein auto-encoders, 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings (2019) 1–20.
URL http://arxiv.org/abs/1711.01558http://arxiv.org/abs/1902.09323

[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of Wasserstein GANs, Arxiv:1704.00028arXiv:1704.00028.
URL http://arxiv.org/abs/1704.00028

[34] S. Schmidt, P. S. Heyns, J. P. de Villiers, A novelty detection diagnostic methodology for gearboxes operating under fluctuating operating conditions using probabilistic techniques, Mechanical Systems and Signal Processing 100 (2018) 152–166. doi:10.1016/j.ymssp.2017.07.032.

[35] C. Peeters, J. Antoni, J. Helsen, Blind filters based on envelope spectrum sparsity indicators for bearing and gear vibration-based condition monitoring, Mechanical Systems and Signal Processing 138. doi:10.1016/j.ymssp.2019.106556.

[36] Y. Miao, B. Zhang, J. Lin, M. Zhao, H. Liu, Z. Liu, H. Li, A review on the application of blind deconvolution in machinery fault diagnosis, Mechanical Systems and Signal Processing 163 (May 2021) (2022) 108202. doi:10.1016/j.ymssp.2021.108202.
URL https://doi.org/10.1016/j.ymssp.2021.108202https://linkinghub.elsevier.com/retrieve/pii/S088832702100577X

[37] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, F. Huszár, Amortised MAP inference for image super-resolution (2016) 1–17.
URL http://arxiv.org/abs/1610.04490

[38] S. Schmidt, P. S. Heyns, K. C. Gryllias, A pre-processing methodology to enhance novel information for rotating machine diagnostics, Mechanical Systems and Signal Processing 124 (2019) 541–561. doi:10.1016/j.ymssp.2019.02.005.
URL https://doi.org/10.1016/j.ymssp.2019.02.005

[39] D. H. Diamond, P. S. Heyns, A. J. Oberholster, Online shaft encoder geometry compensation for arbitrary shaft speed profiles using Bayesian regression, Mechanical Systems and Signal Processing 81 (2016) 402–418. doi:10.1016/j.ymssp.2016.02.060.
URL https://www.sciencedirect.com/science/article/pii/S0888327016001114

[40] Y. Chen, S. Schmidt, P. S. Heyns, M. J. Zuo, A time series model-based method for gear tooth crack detection and severity assessment under random speed variation, Mechanical Systems and Signal Processing 156 (2021) 107605. doi:10.1016/j.ymssp.2020.107605.
URL https://doi.org/10.1016/j.ymssp.2020.107605

[41] P. D. McFadden, M. M. Toozhy, Application of synchronous averaging to vibration monitoring of rolling element bearings, Mechanical Systems and Signal Processing 14 (6) (2000) 891–906. doi:10.1006/mssp.2000.1290.

[42] S. Braun, The synchronous (time domain) average revisited, Mechanical Systems and Signal Processing 25 (4) (2011) 1087–1102. doi:10.1016/j.ymssp.2010.07.016.
URL http://dx.doi.org/10.1016/j.ymssp.2010.07.016

[43] S. Schmidt, P. S. Heyns, J. de Villiers, Discrepancy signal processing techniques for gearbox condition monitoring applications, Proceedings of the First World Congress on Condition Monitoring (WCCM 2017).

[44] H. Qiu, J. Lee, J. Lin, G. Yu, R. T. Services (2007), IMS, University of Cincinnati. Bearing Data Set, NASA Ames Prognostics Data Repository (2007).
URL http://ti.arc.nasa.gov/project/prognostic-data-repository

[45] R. D'Agostino, E. S. Pearson, Tests for departure from normality. Empirical results for the distributions of b2 and $\sqrt{\text{b1}}$, Biometrika 60 (3) (1973) 613–622.
URL http://www.jstor.org/stable/2335012

[46] M. A. Stephens, EDF statistics for goodness of fit and some comparisons, Journal of the American Statistical Association 69 (347) (1974) 730–737.
URL http://www.jstor.org/stable/2286009

[47] S. S. Shapiro, M. B. Wilk, An analysis of variance test for normality (complete samples), Biometrika 52 (3-4) (1965) 591–611. doi:10.1093/biomet/52.3-4.591.
URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/52.3-4.591

[48] D. Abboud, M. Elbadaoui, W. A. Smith, R. B. Randall, Advanced bearing diagnostics: a comparative study of two powerful approaches, Mechanical Systems and Signal Processing 114 (2019) 604–627. doi:10.1016/j.ymssp.2018.05.011.
URL https://doi.org/10.1016/j.ymssp.2018.05.011

[49] S. Kass, A. Raad, J. Antoni, Self-running bearing diagnosis based on scalar indicator using fast order frequency spectral coherence, Measurement: Journal of the International Measurement Confederation 138 (2019) 467–484. doi:10.1016/j.measurement.2019.02.046.
URL https://doi.org/10.1016/j.measurement.2019.02.046

[50] B. Hou, D. Wang, T. Xia, Y. Wang, Y. Zhao, K. L. Tsui, Investigations on quasi-arithmetic means for machine condition monitoring, Mechanical Systems and Signal Processing 151 (2021) 107451. doi:10.1016/j.ymssp.2020.

107451.
URL https://doi.org/10.1016/j.ymssp.2020.107451

[51] A. Moshrefzadeh, Condition monitoring and intelligent diagnosis of rolling element bearings under constant/variable load and speed conditions, Mechanical Systems and Signal Processing 149 (2021) 107153. doi:10.1016/j.ymssp.2020.107153.
URL https://doi.org/10.1016/j.ymssp.2020.107153

[52] D. Wang, Z. Peng, L. Xi, The sum of weighted normalized square envelope: A unified framework for kurtosis, negative entropy, Gini index and smoothness index for machine health monitoring, Mechanical Systems and Signal Processing 140 (2020) 106725. doi:10.1016/j.ymssp.2020.106725.
URL https://doi.org/10.1016/j.ymssp.2020.106725

[53] J. Ben Ali, N. Fnaiech, L. Saidi, B. Chebel-Morello, F. Fnaiech, Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals, Applied Acoustics 89 (2015) 16–27. doi:10.1016/j.apacoust.2014.08.016.
URL http://dx.doi.org/10.1016/j.apacoust.2014.08.016

[54] S. Dong, T. Luo, Bearing degradation process prediction based on the PCA and optimized LS-SVM model, Measurement: Journal of the International Measurement Confederation 46 (9) (2013) 3143–3152. doi:10.1016/j.measurement.2013.06.038.
URL http://dx.doi.org/10.1016/j.measurement.2013.06.038

[55] F. Huang, A. Sava, K. H. Adjallah, Z. Wang, Fuzzy model identification based on mixture distribution analysis for bearings remaining useful life estimation using small training data set, Mechanical Systems and Signal Processing 148 (2021) 107173. doi:10.1016/j.ymssp.2020.107173.
URL https://doi.org/10.1016/j.ymssp.2020.107173

[56] F. Li, Y. Cheng, B. Tang, X. Zhou, D. Tian, Multi-layer gated temporal convolution network for residual useful life prediction of rotating machinery, Mechanical Systems and Signal Processing 155 (2021) 107600. doi:10.1016/j.ymssp.2020.107600.
URL https://doi.org/10.1016/j.ymssp.2020.107600https://linkinghub.elsevier.com/retrieve/pii/S0888327020309869

[57] C. Liu, K. Gryllias, 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 (2020) 106682. doi:10.1016/j.ymssp.2020.106682.
URL https://doi.org/10.1016/j.ymssp.2020.106682https://linkinghub.elsevier.com/retrieve/pii/S0888327020300686

[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, Advances in Neural Information

Processing Systems 32 (NeurIPS).
URL http://arxiv.org/abs/1912.01703

[59] D. P. Kingma, J. Ba, Adam: a method for stochastic optimization.
URL http://arxiv.org/abs/1412.6980

[60] I. Loshchilov, F. Hutter, Decoupled weight decay regularization.
URL http://arxiv.org/abs/1711.05101

[61] T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings.
URL http://arxiv.org/abs/1802.05957