Semi-supervised and unsupervised learning for health indicator extraction from guided waves in aerospace composite structures

Perry, James Josep; Garcia-Conde Ortiz, Pablo; Konstantinou, George; Vergouwen, Cornelie; Kumaran, Edlyn Santha; Moradi, Morteza

DOI
10.1016/j.jmsy.2025.12.014

Publication date
2026

Document Version
Final published version

Published in
Journal of Manufacturing Systems

Citation (APA)
Perry, J. J., Garcia-Conde Ortiz, P., Konstantinou, G., Vergouwen, C., Kumaran, E. S., & Moradi, M. (2026). Semi-supervised and unsupervised learning for health indicator extraction from guided waves in aerospace composite structures. *Journal of Manufacturing Systems*, *84*, 468-492. https://doi.org/10.1016/j.jmsy.2025.12.014

Journal of Manufacturing Systems 84 (2026) 468–492

Contents lists available at ScienceDirect

Journal of Manufacturing Systems

journal homepage: www.elsevier.com/locate/jmansys

Technical paper

Semi-supervised and unsupervised learning for health indicator extraction from guided waves in aerospace composite structures

James Josep Perry[1], Pablo Garcia-Conde Ortiz[1], George Konstantinou[1], Cornelie Vergouwen, Edlyn Santha Kumaran, Morteza Moradi *

Center of Excellence in Artificial Intelligence for Structures, Prognostics & Health Management, Aerospace Engineering Faculty, Delft University of Technology, Kluyverweg 1, Delft, 2629 HS, The Netherlands

ARTICLE INFO

Keywords:
Prognostics and health management
Health indicator
Constrained variational autoencoder
Semi-supervised learning
Aeronautical composite structures

ABSTRACT

Health indicators (HIs) are central to diagnosing and prognosing the condition of aerospace composite structures, enabling efficient maintenance and operational safety. However, extracting reliable HIs remains challenging due to variability in material properties, stochastic damage evolution, and diverse damage modes. Manufacturing defects (e.g. disbonds) and in-service incidents (e.g. bird strikes) further complicate this process. This study presents a comprehensive data-driven framework that learns HIs via two learning approaches integrated with multi-domain signal processing. Because ground-truth HIs are unavailable, a semi-supervised and an unsupervised approach are proposed: (i) a diversity deep semi-supervised anomaly detection (Diversity-DeepSAD) approach augmented with continuous auxiliary labels used as hypothetical damage proxies, which overcomes the limitation of prior binary labels and enables modelling of intermediate degradation, and (ii) a degradation-trend-constrained variational autoencoder (DTC-VAE), in which the monotonicity criterion is embedded via an explicit trend constraint. Guided waves with multiple excitation frequencies are used to monitor single-stiffener composite structures under fatigue loading. Time, frequency, and time–frequency representations are explored, and per-frequency HIs are fused via unsupervised ensemble learning to mitigate frequency dependence and reduce variance. Using fast Fourier transform features, the models achieved fitness scores of 81.6% (Diversity-DeepSAD) and 92.3% (DTC-VAE), indicating improved monotonicity and consistency over existing baselines. The proposed history-independent framework, supported by prognostic metrics–guided Bayesian optimisation and excitation frequency-agnostic HI fusion, enables the estimation of more robust HIs for aeronautical composite structures.

1. Introduction

The health of aerospace structures must be continuously monitored to ensure their safety and reliability. An effective monitoring and diagnostic system should detect damage early enough to issue warnings before it escalates into a dangerous situation [1]. Even earlier diagnostics using damage trend prediction, known as prognostics, can save significant time and cost by enabling timely repairs before a structure reaches an irreparable state [2,3], in addition to improving reliability [4–6]. In this context, developing a health indicator (HI) is essential for both diagnostics and prognostics [7–10]. An HI, being a value which quantifies the health of a structure, serves as a critical bridge between these two aspects toward condition-based maintenance (CBM) [11,12]. However, deriving a truly comprehensive and accurate HI is exceptionally challenging for most objects, especially engineering systems, due to their inherent complexity [13]. The comprehensiveness of an HI refers to its ability to account for all potential types of damage that could affect a system's health [13–15]. Tracking and quantifying all forms of degradation, even in simple objects made of isotropic materials, is often infeasible. Many degradation processes occur internally, beyond the reach of direct measurement. Although sensors may monitor the effects of different types of damage on the sensed signals, meticulously identifying the damage remains challenging due to numerous known and unknown factors, which depend on the object under monitoring, type of sensors, and structural health monitoring (SHM) techniques employed. This complexity is further exacerbated in the case of advanced systems and materials [14,16]. In summary, claiming to have completely true HI values for any object, especially a complex one like an engineering system, is an overstatement [13,17].

* Corresponding author.
E-mail address: m.moradi-1@tudelft.nl (M. Moradi).
[1] These authors contributed equally.

https://doi.org/10.1016/j.jmsy.2025.12.014
Received 21 October 2025; Received in revised form 14 December 2025; Accepted 14 December 2025
Available online 19 December 2025

List of Symbols

c	Hypersphere centre
F_{all}	Fitness function given all units
F_{test}	Test fitness function
f	Frequency
H	Number of eigenvalues
h	Eigenvalue number
i, j	Timestep
k	Prognostic criteria weighting
$\mathcal{L}$	Loss function
M	Total number of specimens
Mo	Monotonicity
m	specimen number
N	Number of measurements (timesteps)
N_f	Number of frequencies
n	Number of labelled samples
Pr	Prognosability
Q_ϕ	Encoder probability distribution
r	Rate of degradation
S	Frequency domain statistical feature
$s(f)$	Frequency domain signal
t_i	Time at timestep i
Tr	Trendability
u	Number of unlabelled samples
X	Time domain statistical feature
$x(i)$	Time domain signal
$x_i^{m,f}$	Features at timestep i of specimen m with measurement frequency f
$\hat{x}$	Reconstructed features
$y_i^{m,f}$	Health indicator at timestep i of specimen m with measurement frequency f
$\tilde{y}$	Auxiliary label
$\bar{y}$	Mean health indicator across folds
z	Autoencoder latent variables
α	Kullback-Leibler divergence weighting
β	Reconstruction loss weighting
γ	Monotonicity constraint loss weighting
ϵ	Small number to prevent zero errors
ε	Gaussian noise
ζ_h	h^{th} eigenvalue of Gram matrix
η	Labelled parameter loss weighting
θ	Neural network function
λ	Diversity loss weighting
μ	Mean
ν	L2 regularisation weighting
σ	Standard deviation
ϕ	Encoder parameters
ψ	Decoder parameters
ω_f	Weight for frequency f

While a perfectly accurate HI is impractical, certain principles about HIs hold universally true. For example, the end of life (EoL) represents the complete loss of health in any system. Similarly, a system's health inevitably decreases over time unless maintenance or self-healing mechanisms are applied. These principles form the foundation for evaluating an HI using key metrics. Prognostability (Pr), for instance, ensures that the failure state at EoL is represented by the same HI value across different systems [18,19]. Monotonicity (Mo) emphasises the consistency of health degradation over time [18,20–22]. Another desirable metric is trendability (Tr), which assesses whether the degradation trends of similar systems are consistent. Although uncertainties in manufacturing, loading conditions, and environmental factors may introduce variability, the general trends for similar systems should not diverge significantly. Therefore, ideal HIs for a group of similar systems should satisfy Mo, Pr, and Tr [13,19,23,24], all of which are considered in the fitness score.

As already mentioned, developing or identifying HIs is a challenging task due to the inherent complexity and uncertainty of material degradation processes. This difficulty is amplified in anisotropic materials like composites [25] which are widely used in aerospace applications thanks to their high strength-to-weight ratio and ability to withstand directional stresses [26]. Unlike isotropic materials, composite materials exhibit more complex failure modes arising from their anisotropic properties and the potential interactions between different types of damage [15,27,28]. These unique characteristics make designing effective HIs for composite structures particularly intricate [14,23]. Additionally, uncertainties arising during manufacturing (e.g. disbond defects) or operation (e.g. bird strikes) exacerbate this complexity [29,30]. To enable in situ monitoring and account for these uncertainties, SHM techniques are essential. An effective SHM technique must be capable of covering most structural areas, particularly critical regions. Among these techniques, guided wave (GW)-based SHM stands out as a practical approach, where multiple sensors are strategically attached to critical regions to monitor the structure's health. GWs are ultrasonic waves bounded by a structure's material [30] that produce data from which the location and severity of damage to a structure can be determined [27,31–33]. This enables SHM in otherwise inaccessible regions of complex composite structures [34]. As GW and most other SHM techniques rely on a network of sensors, model-based approaches face limitations in processing the large volume of complex, high-dimensional sensed signals. Furthermore, noise, multimodality, and the influence of environmental factors make it more challenging to generate HIs from raw GW data [15,35,36]. Consequently, data-driven and artificial intelligence (AI) approaches offer greater potential for handling such intricate datasets. Due to the unknown true health state of any engineering system, supervised models face limitations in constructing HIs. To address this challenge, semi-supervised or unsupervised AI models can be utilised for data fusion, enabling the generation of effective HIs [37,38].

Furthermore, the calculation of the typical fitness score, which includes Mo, Pr, and Tr, is prone to bias toward the training units. This bias can lead to potential issues, as the model may produce highly correlated HIs during training, resulting in misleadingly high fitness scores. However, when faced with an unmatched HI from a specific unit during testing, the performance may fall short. To address this, the rectified evaluation criteria [13,15] are adopted in this study, ensuring a more reliable assessment during the test phase. This adjustment provides a stronger foundation for comparison and enhances the practical reliability of the standard.

To construct a history-independent HI for aeronautical composite structures, GW-SHM emerges as a promising technique as mentioned earlier. However, each GW measurement requires data collection through a network of PZT sensors, resulting in an enormous volume of data to process. Feeding this data directly into deep learning models demands a high number of input layer nodes, which increases model complexity and computational requirements while reducing interpretability [36]. A more efficient alternative is to employ signal processing (SP) algorithms as a preprocessing step. SP algorithms often follow explicit solutions, producing globally applicable and faster outcomes. By extracting a set of statistical features from the processed signals, the size and complexity of the models can be significantly reduced. However, identifying which SP method yields the most effective features remains an open question. The proposed frameworks address this by first applying various SP algorithms to extract and compare features in terms of fitness criteria. The selected features are then fed into AI models to generate HIs, ensuring both efficiency and effectiveness in the process.

This work contributes to the field in the following ways:

(i) Extension of the Diversity-DeepSAD model into augmented Diversity-DeepSAD by embedding continuous degradation-progression labels instead of binary ±1 labels in its loss function, enabling modelling of intermediate fatigue states essential for composites.
(ii) Development of a degradation-trend-constrained variational autoencoder (DTC-VAE) architecture, in which a monotonicity constraint is embedded directly in the latent space to enforce consistent HI evolution, for guided wave monitoring of aerospace composites.
(iii) Use of Bayesian hyperparameter optimisation guided by a composite fitness score based on monotonicity, prognosability, and trendability, requiring construction of full run-to-failure HI trajectories for all training and validation units.
(iv) Implementation of a frequency-agnostic unsupervised HI fusion mechanism, weighted by composite fitness score, providing a principled alternative to selecting a single excitation frequency in guided wave SHM.
(v) Comprehensive comparison of semi- and unsupervised learning models for generating history-independent HIs, highlighting their behaviour, limitations, and applicability across guided wave SHM scenarios.

A review of fundamental literature follows in Section 2. The methodology employed in this study is described in detail in Section 3, with the corresponding results presented in Section 4. A discussion of these findings is provided in Section 5, culminating in the conclusions outlined in Section 6.

2. Literature study

In this section, we first review semi-supervised models relevant to HI generation, followed by a discussion of unsupervised approaches. Semi-supervised models begin with concepts developed from supervised methods. For example, the support vector machine (SVM), introduced by Cortes and Vapnik [39], is a fundamental supervised classification algorithm. However, its binary classification ability falls short of generating meaningful HIs, as it ignores the complex gradient between healthy and failure states. Progress toward HI generation came with support vector data description (SVDD) [40] and its deep learning extension, Deep SVDD, proposed by Kim et al. in 2015 [41], which captures more intricate behaviours. For SHM, a semi-supervised model is preferred to effectively handle unlabelled data. Deep semi-supervised anomaly detection (DeepSAD), proposed by Ruff et al. in 2019 [42], extends Deep SVDD by incorporating unlabelled data points. While DeepSAD has proven effective in various non-SHM contexts, its application to HI generation poses challenges. For instance, Han et al. [43] used DeepSAD to detect anomalies in vehicle emissions, by setting a threshold anomaly score. They evaluated the performance of models by the area under the receiver operating characteristic curve, thus without considering the distribution of scores on either side of the threshold. This approach, although beneficial for anomaly detection, is inadequate for HI extraction due to the complete neglect of prognostic criteria. Similarly, DeepSAD has been applied with consideration only to the classification of anomalies in hydraulic systems [44]. Other approaches have sought to enhance semi-supervised anomaly detection. Gao et al. [45] developed ConNet, a robust deep anomaly detection model for sparsely labelled data. However, ConNet only considers positive, anomalous, labelled samples, which is unsuitable for HI generation as the healthy state is not considered.

Recent advancements have specifically targeted HI generation. Frusque et al. [46] applied DeepSAD to HI generation for rotating machinery, and it was observed that the objective function considers only the norm of the embedding, therefore producing low-rank results with some dimensions not fully utilised. To diversify the representation of degradation trends, they incorporated a diversity term in the loss function, creating Diversity-DeepSAD. For the PHME2010 milling dataset, Diversity-DeepSAD achieved promising results, with performances of 99% for Mo, 99% for Tr, and 94% for Pr, indicating room for improvement in Pr. However, when applied to thermal spray coating monitoring data, the model's performance was lower, with 91% for Mo and 72% for Pr, while Tr was not reported. Based on the displayed HIs, the Tr score appears low. Notably, the coefficient of variation (CV), useful for comparing variability across datasets, was 24% for the thermal spray coating dataset. In contrast, aeronautical structures are significantly more challenging, with CV values of 51% for GW data from five single-stiffener composite panels [15] and 87% for acoustic emission data from 12 specimens [23,24]. Despite its potential, Diversity-DeepSAD has not yet been applied to aeronautical structures, which present unique challenges for HI construction and remaining useful life (RUL) prediction.

Moradi [13] developed various semi-supervised frameworks for single-stiffener composite structures. One approach utilised a multilayer long short-term memory (LSTM) network, employing an intrinsically semi-supervised inductive learning technique, trained on acoustic emission-based time and frequency domain features [23]. This method achieved a fitness score of 93% under a rigorous leave-one-out cross-validation (LOOCV) process. However, the fine-tuning process to optimise the number of epochs should be done using only the validation specimen (unit), and another unit, different from the training/validation units, should be investigated to confirm the generalisability of this fine-tuned hyperparameter—something that was not considered in their study. To address this, Moradi et al. [24] introduced CEEMDAN-driven semi-supervised ensemble deep learning (CEEMDAN-SSEDL), which combined complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) for feature extraction and ensemble learning techniques to reduce randomness. By incorporating bidirectional LSTM (BiLSTM) layers during ensemble learning, this approach achieved a fitness score of 91.3%, while ensuring separate test units were considered. Moradi et al.'s [23] semi-supervised inductive learning approach was later applied by Frusque et al. [46] to the PHME2010 milling dataset, achieving a fitness score of 96%. However, the details of the architecture and the hyperparameter fine-tuning process were not reported. Recognising the limitations of history-dependent models, Moradi et al. [13,37] noted that such models often perform poorly in the absence of historical data. To overcome this, they introduced the Hilbert transform semi-supervised convolutional neural network (HT-SSCNN), designed to generate HIs from history-independent GW data. Leveraging the GW monitoring technique, this approach achieved a 93% fitness score. Inspired by this, the methods in this paper are also designed to be history-independent, relying solely on data from the current state of the structure. Given the demonstrated capabilities of anomaly detection algorithms, this study seeks to extend the application of Diversity-DeepSAD to the SHM of aerospace structures.

Previous applications of DeepSAD, however, have been limited to binary auxiliary labels distinguishing only healthy and failed states. This binary formulation neglects the intermediate degradation states that make up the majority of a structure's lifetime, and in composites often leaves only a single label at failure, severely restricting the usefulness of labels. To address this gap, we introduce continuous auxiliary labels interpolated across the degradation process, applied only at early-life (e.g. 0%–25%) and late-life (e.g. 75%–100%) stages, to provide richer supervision and produce smoother HIs.

Unsupervised models, by design, rely on feedback from the model output to learn without labelled data. A common technique is the autoencoder (AE), which uses a neural network to encode input data into a reduced-dimensional latent space and then reconstruct the input from this representation. Yang et al. [47] generated HIs for rotating machinery using sparse AEs, evaluating them based on the Mann–Kendall (MK) monotonicity metric. Lin and Tao [48] extended this approach

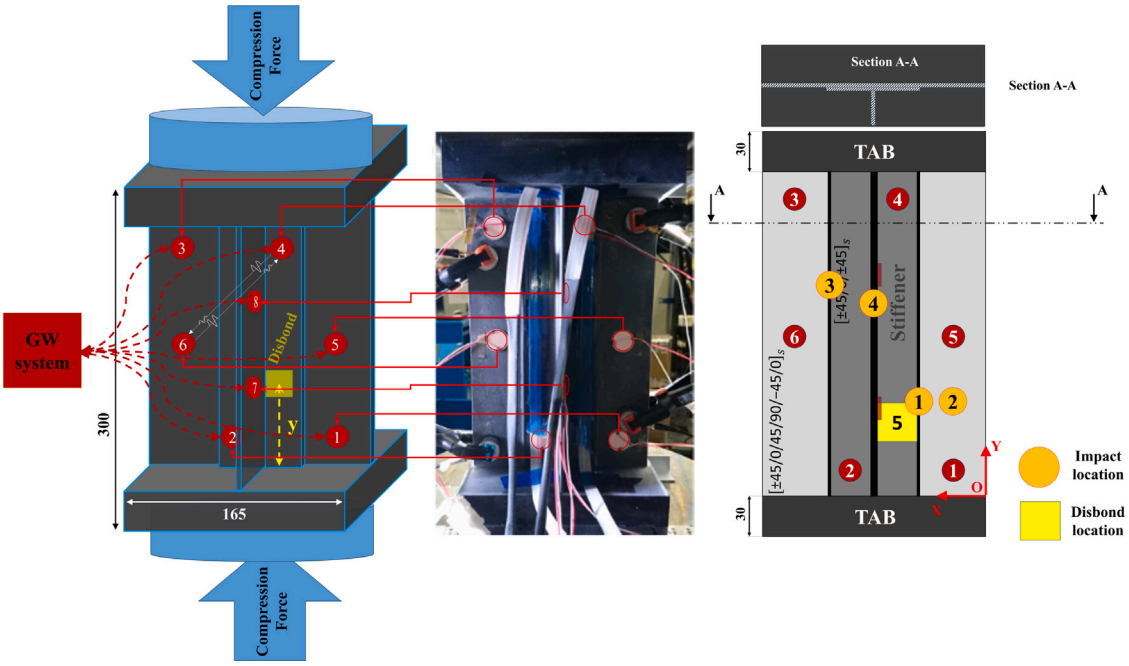


Fig. 1. Single T-stiffener CFRP panel under C–C fatigue loading monitored with PZT sensors (red circles).

by employing an ensemble of stacked AEs with linear targets, while Xu and Wang [49] enhanced stacked AEs with noise reduction using an exponential weight moving average model. In 2023, Xu et al. [50] implemented a deep convolutional AE and Mao et al. [51] applied tensor representation. These studies, however, focused on rotating machinery, where the vibrational behaviour is more self-exposing and sensitive to damage. In contrast, aerospace structures do not exhibit such clear damage effects on the monitored signals, making the application of these methods more challenging. Variational autoencoder (VAE) architecture is similar to AE, but makes use of a probabilistic latent variable distribution in order to increase generability to data beyond the training set. In 2019, Ping et al. [52] utilised a VAE with a logarithmic normal distribution to generate HIs for turbofan engines (CMAPSS simulated dataset, which is simpler compared to experimental dataset [23,53]). In 2020, Hemmer et al. [54] applied a VAE to rotating machinery at discrete damage levels. While generally increasing with damage level, the HIs do not conform to the prognostic criteria. Mitigating this, in 2022, Qin et al. [55] introduced a monotonic degradation constraint in the loss function, resulting in the degradation-trend-constrained VAE (DTC-VAE), which generates HIs that meet the Mo metric. Guo et al. [56] proposed a multiscale convolutional AE network for rotating machinery. Further studies in 2024 applied developments of VAE to the generation of HIs for bearings from vibrational data [57–59]. However, the application of VAEs for HI generation in aerospace structures remains unattempted. This paper seeks to address this gap by applying the DTC-VAE model to GW signals measured from aerospace structures, aiming to achieve high quality HIs. By leveraging the strengths of both Diversity-DeepSAD and DTC-VAE, the study aspires to advance HI generation for complex aeronautical structures.

3. Method

This section provides a detailed overview of the methodology adopted in this study, encompassing the experimental setup and the framework for HI generation. Subsection 3.1 introduces the experimental setup which forms the foundation of the analysis. Subsection 3.2 describes the overall proposed two-stage framework for HI generation, followed by Subsection 3.3, which outlines the criteria used to evaluate the generated HIs. Subsection 3.4 delves into the SP techniques applied to preprocess the data, while Subsection 3.5 focuses on the feature extraction process. Finally, Subsection 3.6 presents the feature fusion approach, including Diversity-DeepSAD, DTC-VAE, and ensemble learning models, followed by Subsection 3.7 on hyperparameter sensitivity analysis. Details on code are provided in "Code Availability".

3.1. Experimental setup and data preparation

To evaluate the proposed framework, five single-stiffener composite specimens were subjected to run-to-failure compressive fatigue loading under the ReMAP-H2020 project [60]. The specimens consist of IM7/8552 carbon fibre-reinforced epoxy in unidirectional prepreg, formed with layups of $[\pm 45/0/45/90/-45/0]_S$ and $[\pm 45/0/\pm 45]_S$ [15]. The specimens were monitored using different SHM techniques, from which only GWs are used to construct HIs.

The experimental setup is visualised in Fig. 1, which shows where an impact load (orange circles) or a disbond (yellow squares) was pre-applied to the corresponding numbered specimens. To carry out GW monitoring, intervals of 5000 cycles were considered [15]. Eight PZTs were attached to the structure to function as both actuators and sensors (see red circles in Fig. 1). GW data was collected at frequencies of 50 kHz, 100 kHz, 125 kHz, 150 kHz, 200 kHz, and 250 kHz. At a given moment, one of the PZTs acted as an actuator, while the rest acted as sensors, and this process rotated through all eight PZTs. This created a total of (8 × 7) 56 actuator–sensor paths. It should be highlighted that the proposed models in this study do not receive indication of uncertainties that may have occurred during testing, such as disbond defects, broken sensors, and impact loading.

Fig. 2 presents example raw GW signals from the initial (left) and end-of-life (right) states of one composite skin–stiffener panel (composite specimen 5, 125 kHz excitation). Each GW measurement consists of 56 actuator–sensor paths recorded at 6 excitation frequencies, resulting in a 56 × 6 × 2000 data tensor per sample. Zoom-in views highlight changes in waveform characteristics across fatigue progression. The bottom plot shows the corresponding HIs predicted across the lifetimes of multiple units, demonstrating both unit-to-unit variability and complete run-to-failure coverage. In this example, unit 5 was held out

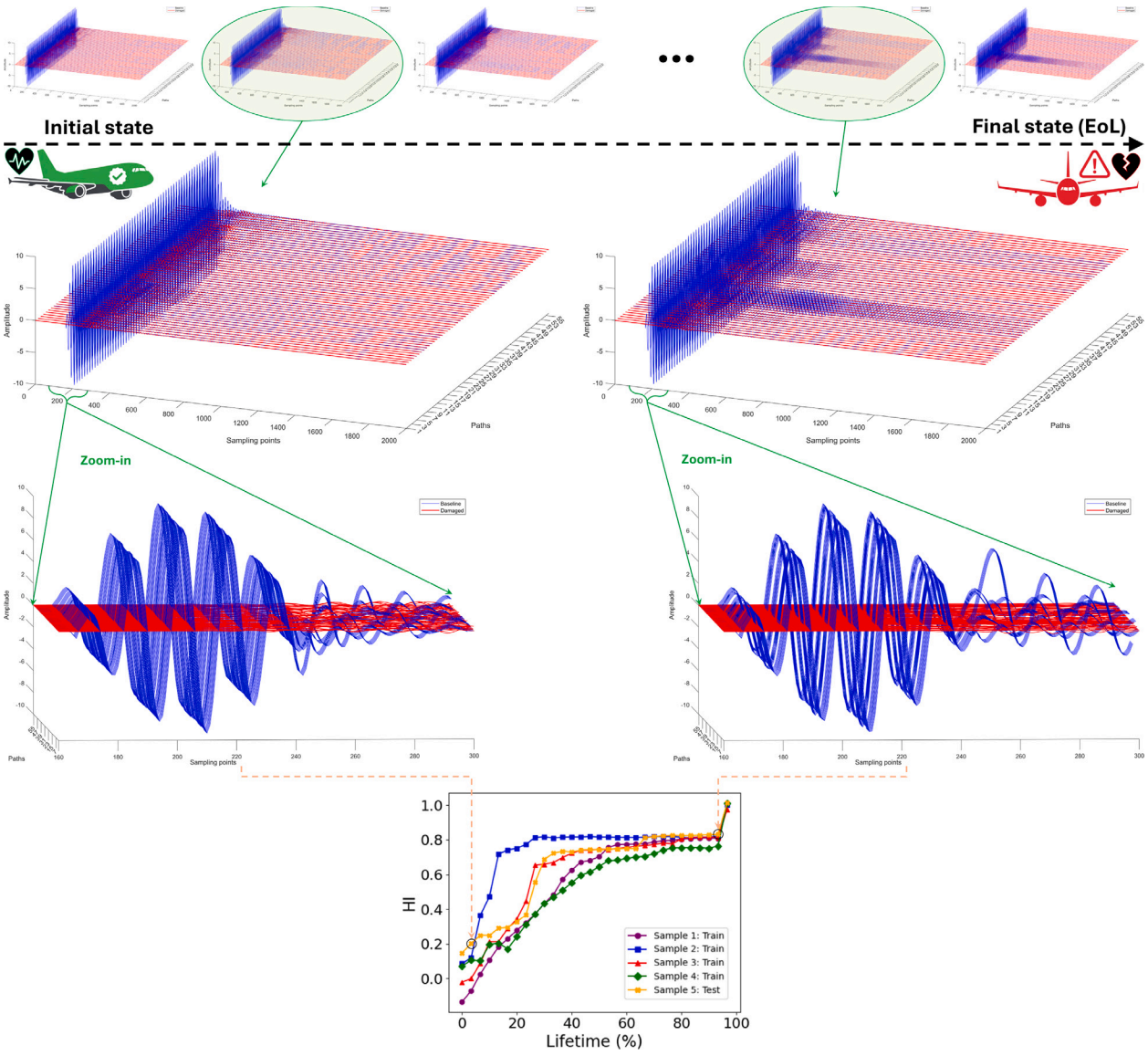


Fig. 2. Example initial and end-of-life guided wave signals and corresponding HIs illustrating the structure of the dataset and fatigue-induced waveform evolution.

entirely for testing, while the remaining units were used for training and validation.

The input feature data were arranged in 5 folds, in each of which a different specimen was used for testing while the remaining 4 were for training. This leave-one-unit-out strategy ensures that all measurements from the held-out panel, including its manufacturing tolerances and experimental uncertainties, remain completely unseen during training, providing a stringent cross-unit generalisation test and helping to mitigate overfitting despite the limited number of structural units. Training and test data were normalised relative to the training data only using Z-score scaling, in other words, subtracting by the mean and dividing by the standard deviation.

The model output was also normalised using the training data, with min–max normalisation, to ensure HIs had a value of approximately 0 in the first timestep (healthy state) and approximately 1 at the last (failure).

3.2. Four-stage framework for HI generation

This study followed a four-stage framework, as can be seen in Fig. 3, resulting in HI generation through the use of SP techniques and AI models. The first stage was pre-processing data to the correct format. In the second stage, time domain measurements were processed using different SP methods, including fast Fourier transform (FFT), short-time Fourier transform (STFT), empirical mode decomposition (EMD), and Hilbert transform (HT). Thirdly, statistical features were extracted from the processed signals for each path, actuator frequency f, and measurement timestep. As all input features should ideally be of the same SP method to ensure efficient processing, a single SP method is used in the workflow. With this in mind, the two SP methods with the highest performing features were selected for further individual use.

In the fourth stage, the features selected were fed into AI models, i.e. the Diversity-DeepSAD and DTC-VAE models, where hyperparameters were fine-tuned using Bayesian optimisation. This process was repeated for each combination of excitation frequency of PZTs and SP method. The extracted HIs (i.e. y) from each actuating frequency were then fused by an unsupervised ensemble learning model, enabling the evaluation of the remaining combinations of AI model and SP method against the HIs evaluation criteria to determine the optimal final combination.

3.3. HIs criteria

Given the lack of true HI labels, evaluating the quality of the developed HIs necessitates leveraging known physics-based principles

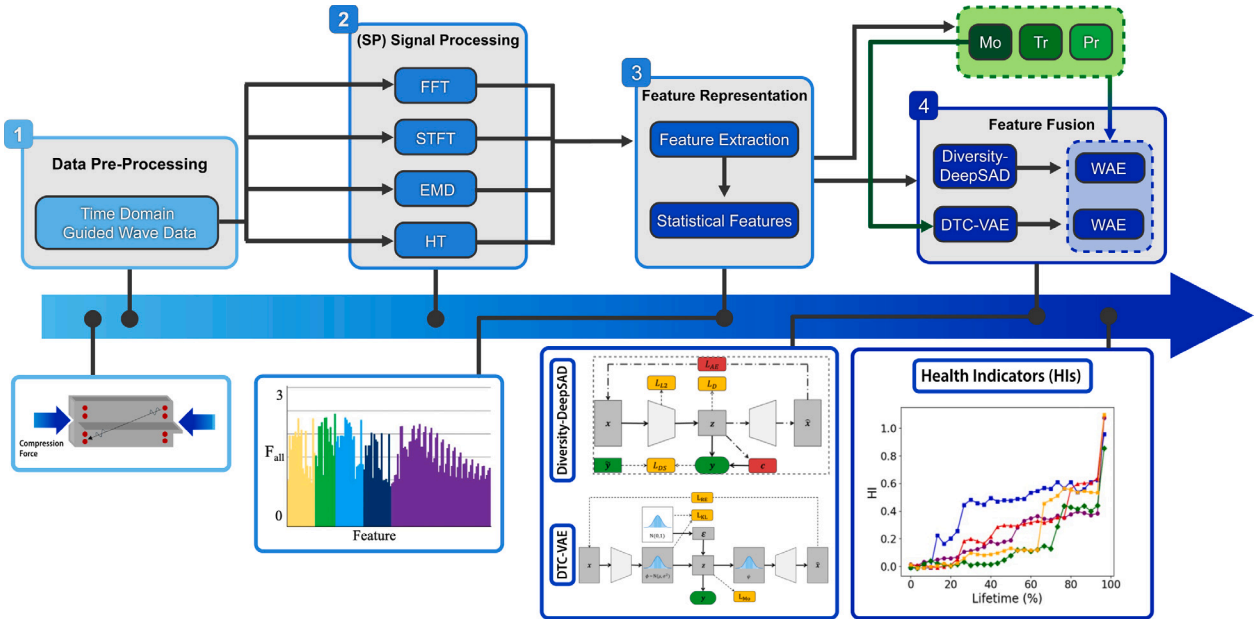


Fig. 3. Framework for HI generation and evaluation — Step 1: Data Pre-Processing, Step 2: Signal Processing, Step 3: Feature Representation, Step 4: Feature Fusion — with prognostic criteria (Mo, Pr, Tr).

and facts. The HI criteria include Mo, Pr, and Tr, and they are standard for this application [13,15,23,46,61]. While these metrics do not directly measure physical damage, they evaluate whether the learned HI behaves in a physically meaningful and degradation-consistent manner.

Mo measures the extent to which the HI shows only an increasing or decreasing trend. If maintenance is not performed, health should always decrease with an increasing number of cycles, and thus a monotonic HI should more closely model reality. The modified Mann–Kendall (MMK) formulation of Mo is presented in Eq. (1) [15]:

$$Mo = \frac{1}{M}\sum_{m=1}^{M}\left|\frac{1}{(N_m-1)}\sum_{i=1}^{N_m-1}\frac{\sum_{j=1,j>i}^{N_m}(t_j-t_i)\cdot sgn(y_j^m-y_i^m)}{\sum_{j=1,j>i}^{N_m}(t_j-t_i)}\right| \tag{1}$$

where M is the number of specimens, N is the number of GW measurements (timesteps), y_i^m denotes an HI at measurement i of specimen m, t_i is similarly the time at the ith measurement, while j can also represent a measurement number in this way. $sgn(\cdot)$ is the sign function. This formulation of Mo is preferable as it considers time gaps greater than one unit, thus reducing the effect of noise and favouring monotonicity through the whole trend, as opposed to between adjacent points [15].

Pr measures the HIs' deviation at failure between specimens. Preferably, all specimens should have the same value of HI at the end of life, such that they can be reliably used to predict failure. Eq. (2) expresses Pr [15], where $std(\cdot)$ refers to the standard deviation between specimens.

$$Pr = exp\left(-\frac{std(y_N^m)}{\frac{1}{M}\sum_{m=1}^{M}\left|y_1^m-y_N^m\right|}\right) \tag{2}$$

Finally, Tr is a measure of the correlation between HIs of different units [15]. The set of specimens should show the same HI pattern from beginning to end of life, as this makes their behaviour more predictable. Tr is given by Eq. (3), where $corr(\cdot)$ is the Pearson correlation function.

$$Tr = min_{a,b}\ corr(y^a, y^b), \quad a,b = 1,\ldots,M \tag{3}$$

In summary, Mo quantifies how consistently the HI evolves monotonically under the assumption that health should not improve without maintenance, Pr measures how closely specimens converge to a common HI value at failure, and Tr evaluates whether specimens follow similar degradation trends throughout their lifetimes. To effectively evaluate a model's performance on test specimens while mitigating bias toward highly matched training specimens and avoiding the obscuring of unmatched test specimen behaviour, alternative evaluation criteria must be established. These criteria should prioritise the behaviour of test specimens over those used during training, ensuring a more reliable assessment [15,24]. These criteria include Mo_{test}, which evaluates the monotonicity of only the test specimen(s), and Pr_{test}, which assesses the prognosability by emphasising the deviation of the test specimen(s) relative to those used for training. These metrics are defined in Eq. (4) and Eq. (5) [15]:

$$Mo_{test} = \left|\frac{1}{(N_{test}-1)}\sum_{i=1}^{N_{test}-1}\frac{\sum_{j=1,j>i}^{N_{test}}(t_j-t_i)\cdot sgn(y_j^m-y_i^m)}{\sum_{j=1,j>i}^{N_{test}}(t_j-t_i)}\right| \tag{4}$$

$$Pr_{test} = exp\left(-\frac{\left|y_N^{test}-\frac{1}{M^{train}}\sum_{m=1}^{M^{train}}y_N^m\right|}{\frac{1}{M}\sum_{j=1}^{M}\left|y_1^j-y_N^j\right|}\right) \tag{5}$$

where the index $train$ or $test$ represents train or test specimens, respectively, and thus M^{train} refers to the number of the training specimens.

The fitness score F_{all} is defined by Eq. (6), and measures the overall quality of HIs [15], while F_{test} is defined similarly in Eq. (7) with a focus on test units. In order to weight the criteria equally, the control constants k_{Mo}, k_{Pr}, and k_{Tr} are set to equal 1. Both fitness scores F_{all} and F_{test} are calculated and discussed in this study.

$$F_{all} = k_{Mo}\cdot Mo + k_{Pr}\cdot Pr + k_{Tr}\cdot Tr \tag{6}$$

$$F_{test} = k_{Mo}\cdot Mo_{test} + k_{Pr}\cdot Pr_{test} + k_{Tr}\cdot Tr \tag{7}$$

The prognostic criteria Mo, Pr, and Tr each take values in the range $[0,1]$, where 0 represents the worst behaviour and 1 represents the ideal expected behaviour of a degradation trajectory. The composite fitness scores F_{all} and F_{test} therefore lie in the range $[0,3]$, with 3 indicating the best overall performance across the three prognostic criteria. When reporting performance as a percentage (e.g. 92.3%), this corresponds to the fitness score normalised by its maximum value of 3.

3.4. Signal processing

SP techniques were applied to transform the measured signals into the frequency and time-frequency domains, in order to highlight relevant information, such as frequency components or noise isolation.

Table 1
Feature numbers extracted from the results of each SP method.

Method	Raw data	FFT	HT	EMD	STFT
Features domain	Time	Frequency	Time	Time	Time-Frequency
Features	1–19	20–33	34–52	53–71	72-139

These transformations enabled the extraction of relevant features for further analysis. The techniques used include FFT, STFT, EMD, and HT.

The FFT algorithm transforms the original temporal signal into the frequency domain by processing the entire time domain at once and outputting corresponding amplitudes for the underlying frequencies. However, it does not retain temporal variations in the signal, such as how these frequencies evolve over time, which may reveal additional information about the structure's health. In contrast, STFT preserves time-frequency information by dividing the data into smaller time intervals and analysing localised segments, making it advantageous for non-stationary signals where frequency content changes over time. The result of STFT is a 2D matrix, where time and frequency are the axes, and amplitude is the value at each position. Nevertheless, there is a nonlinear trade-off between time and frequency resolution: too many time segments reduces the available frequency samples, and vice versa. In this study, it was found that performing STFT with an FFT length of 250 maximised the number of data points, resulting in an overlap length of 125.

In the case of EMD, the sifting algorithm was implemented to iteratively decompose the signal into intrinsic mode functions (IMFs) in the time domain. The process involves constructing upper and lower envelopes of the signal using cubic splines to interpolate local maxima and minima, respectively. The mean of these envelopes is then subtracted from the signal at each iteration, refining the signal until it satisfies the conditions for an IMF. Suitable stopping conditions are necessary to avoid valuable information being lost [62]. Through trial and error, these criteria were set to a minimum standard deviation of 0.1 and a maximum of 10 iterations.

The HT algorithm produces a complex-valued signal, the analytic signal, from a real signal. The real component represents the original signal, while the imaginary component is created by phase-shifting each frequency component by $\frac{\pi}{2}$ [63]. The analytic signal therefore contains information concerning instantaneous amplitude and phase.

By employing these SP techniques, features can be extracted that capture essential signal characteristics across the time, frequency, and time-frequency domains, forming a robust foundation for subsequent analysis.

3.5. Feature extraction

Statistical features, as detailed in Appendix A, were extracted from the time, frequency, and time-frequency domains to characterise the signals comprehensively. The selected feature set aligns with established approaches in the literature [23]. They were chosen due to their proven capability to capture diverse and informative aspects of signal behaviour.

From the time domain, 19 statistical features were extracted, as shown in Table A.13. These features were computed for the original time-domain signals as well as the processed signals using HT and EMD. From the frequency domain, 14 statistical features were extracted from the FFT-transformed signals, as listed in Table A.14. For the time-frequency domain, the output of STFT was divided into 17 time windows, each with discrete frequency components. To reduce dimensionality, four statistical measures—mean, standard deviation, skewness, and kurtosis—were calculated across the frequency axis within each time window. This resulted in a total of 68 features (4 measures ×17 time windows), as summarised in Table A.15.

The extracted features are outlined in Table 1. By systematically extracting statistical features across multiple domains, the framework ensures a concise yet rich representation of the signal information.

For the first stage, features across different actuator–sensor paths and frequencies are averaged, resulting in 139 features per timestep (GW measurement) for each specimen. This averaging is intended to capture a robust representation of the specimen's response, to emphasise consistent trends across multiple paths and frequencies whilst reducing the influence of local variations or noise. After obtaining all timesteps and specimens, the HI prognostic criteria defined in Section 3.3 are applied to evaluate each feature. In order to avoid anomalously poor features disproportionately impacting the SP method selected, the mean fitness score across all features is calculated to serve as a benchmark. Features with fitness scores exceeding this benchmark are retained for further analysis. This benchmark based filtering process acts as a feature selection step, comparable to Baraldi et al. [64] method of identifying the most informative set of features to ensure that only consistently relevant features contribute to HI development. The SP method yielding features with the highest remaining mean fitness scores is identified as optimal for developing an efficient framework to extract HI, providing a systematic approach for feature selection rather then relying on arbitrary choices.

3.6. Feature fusion

Given the extracted and selected features as inputs, two deep learning models were developed to generate HIs, i.e. semi-supervised Diversity-DeepSAD and unsupervised DTC-VAE, which will be presented in this section.

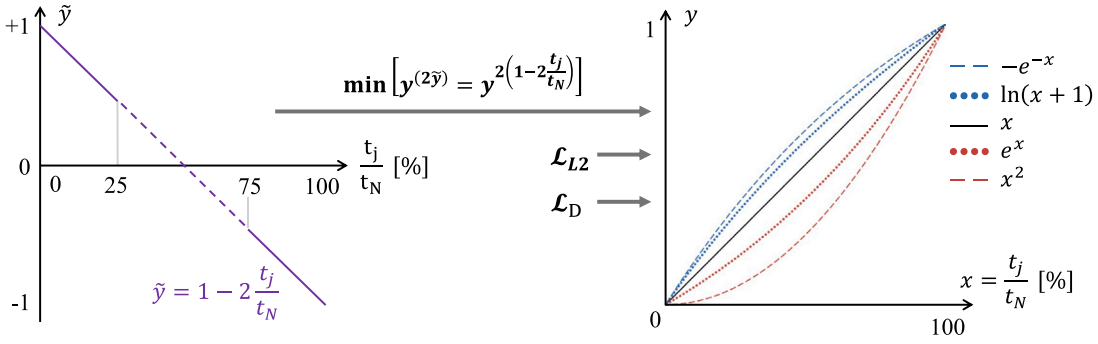
3.6.1. Diversity-DeepSAD

The original DeepSAD model incorporates labelled training data in the loss function of the otherwise unsupervised Deep SVDD model [42]. This approach transforms input data into a hyperspace of reduced dimensions, in which the distance to a point, the hypersphere centre $\mathbf{c}$, represents the extent to which the datapoint is an anomaly. Frusque et al. [46] modified the loss function further to include a term to increase diversity, resulting in Diversity-DeepSAD for HI generation. This was demonstrated on a rotational machinery dataset with hundreds of measurement timesteps, a condition not often met in many other applications. For example, the ReMAP dataset may contain as few as 30 timesteps before failure. Consequently, it is necessary to explore the impact of alternative functions for generating auxiliary artificial labels and to evaluate the influence of the number of labelled measurements.

In prior studies [42,46], a binary set of auxiliary labels $\tilde{y}_j \in \{1, -1\}$ was employed, where a label of 1 was assigned to measurements representing a very healthy condition, and -1 corresponded to those reflecting a severely damaged state. This approach focuses solely on two extreme conditions. However, these two phases typically occupy only a small portion of the system's entire lifetime, limiting the potential to fully leverage auxiliary labels. This issue is particularly pronounced in the severely damaged phase, where measurements are rare, often with only a single observation available at the point of failure for assigning the label -1. To address these limitations, this study generated continuous auxiliary labels $\tilde{y}_j \in [1, -1]$. Although many functions were considered, best performance was achieved through a simpler labelling scheme according to Eq. (8):

$$\tilde{y}_j = 1 - 2\frac{t_j}{t_N} \tag{8}$$

where a linear function was applied for the first and last quarters of the lifetime. The use of a linear function avoids imposing any assumed physical degradation which could reduce the flexibility of the



(a) Auxiliary labels generated for training as a function of lifetime. Data collected in the dashed region is left unlabelled.

(b) Examples of possible HI embeddings, where each curve represents a basic function capturing the main trend, either individually or in combination. Note that 0 corresponds to healthy and 1 to failure, unlike auxiliary labels (1 = healthy, -1 = failure).

Fig. 4. Relationship between new auxiliary labels and embedding output expected for the proposed Diversity-DeepSAD.

model, given this is an extension from a pure binary ±1 formulation. The quarter ratio was established experimentally to produce improved results. As visualised in Fig. 4a, the remaining data points do not leverage auxiliary labels, providing flexibility to the model, given the uncertainty of the true health status for these intermediate conditions.

It should be noted that auxiliary labels $\tilde{y}_j$ differ from the expected HI, denoted by y_j. Eq. (8) therefore need not bear resemblance to the physical behaviour of the sample. Assuming the hypersphere centre $\mathbf{c}$ represents the healthy state, the HI is defined as the distance from that centre:

$$\text{HI} = y_j = \|z_j - \mathbf{c}\| \tag{9}$$

where z_j denotes the latent space representation obtained from the autoencoder. When z_j is close to the healthy state $\mathbf{c}$, y_j approaches 0, indicating minimal deviation from the healthy condition. Conversely, as z_j moves farther from $\mathbf{c}$, y_j increases, reflecting a more degraded condition. In this framework, the maximum value of 1 is assigned to y_j in the fully damaged state. Consequently, the HI can freely exhibit linear or non-linear behaviour—with different main trends, such as exponential or polynomial behaviour—between these bounds, as illustrated in Fig. 4b.

Denoting the encoder function as ϕ, the HI can be expressed as:

$$y_j = \|\phi_\theta(\mathbf{x}_j) - \mathbf{c}\| \tag{10}$$

where $\phi_\theta(\mathbf{x}_j)$ represents the latent space embedding of the input data $\mathbf{x}_j$ at timestep j, and θ denotes the learnable parameters of the encoder. The sum of squares of these embeddings, raised to the power of the auxiliary label $\tilde{y}_j$ where applicable, is minimised using the following loss function:

$$\mathcal{L}_{DS} = \frac{1}{n+u}\sum_{i=1}^{u}\left(y_i\right)^2 + \frac{\eta}{n+u}\sum_{j=1}^{n}\left(y_j+\epsilon\right)^{2\tilde{y}_j} \tag{11}$$

in which n represents the number of labelled samples, u is the number of unlabelled samples, and η is a hyperparameter that controls the contribution of the auxiliary labels. To prevent numerical zero errors when $\tilde{y}_j$ is negative, a small term ϵ is added to labelled HIs. This loss component is the first term in the overall Diversity-DeepSAD loss function:

$$\mathcal{L}_{Diversity-DeepSAD} = \mathcal{L}_{DS} + \nu\mathcal{L}_{L2} + \lambda\mathcal{L}_D \tag{12}$$

where $\mathcal{L}_{L2}$ represents L_2 regularisation [65], minimising the sum of squares of node weights to encourage a greater distribution of smaller weights and prevent overfitting. ν and λ are hyperparameters controlling the weights of the regularisation terms $\mathcal{L}_{L2}$ and $\mathcal{L}_D$, respectively. $\mathcal{L}_D$ is defined as follows:

$$\mathcal{L}_D = \sum_{h=1}^{H}(\zeta_h - \ln(\zeta_h)) \tag{13}$$

where ζ_h represents the h^{th} out of H eigenvalues of the Gram matrix $G = Z^T Z$, in which Z is the DeepSAD embedding. $\mathcal{L}_D$ promotes a greater distribution of training data in the model output [46]. These three terms therefore work together to ensure that the model effectively learns the latent representations, while the regularisation terms help prevent overfitting and promote diversity within the learned embeddings.

A standard AE is always used to pretrain the network weights (learnable parameters θ). This AE is formed by adding a decoder ψ which inverts the Diversity-DeepSAD network layers, considering the mean squared difference between input and output data as a loss function, $\mathcal{L}_{AE}$. This loss function is used to pre-train the autoencoder, but is disregarded in the training of Diversity-DeepSAD in Eq. (12). The hypersphere centre $\mathbf{c}$ is initialised by computing the average output from the model during a forward pass on the training data. Initially, $\mathbf{c}$ is set to a zero matrix. In the pretraining phase, denoted by $\phi_\theta^{[0]}$, the model performs a forward pass on the training data, and the outputs for each sample are accumulated. The number of samples is incremented with each data point, and the cumulative outputs are added to $\mathbf{c}$. After processing all the data, the matrix $\mathbf{c}$ is normalised by the total number of samples to obtain the average. To avoid numerical issues, any component of $\mathbf{c}$ with an absolute value smaller than ϵ is adjusted as follows:

$$\mathbf{c} = \begin{cases} \phi_\theta^{[0]} + \epsilon & \text{if} \quad 0 < \phi_\theta^{[0]} < \epsilon \\ \phi_\theta^{[0]} - \epsilon & \text{if} \quad -\epsilon < \phi_\theta^{[0]} < 0 \end{cases} \tag{14}$$

This adjustment ensures stability in the hypersphere centre values, helping the model to establish a reference point for the healthy state. The hypersphere centre $\mathbf{c}$ serves as a critical foundation for anomaly detection and HI estimation. The model architecture is illustrated in Fig. 5, where red, yellow, and green represent pretraining, training, and model output, respectively.

Diversity-DeepSAD architecture and hyperparameter optimisation: The Diversity-DeepSAD model was implemented with 6 hidden layers, each containing half the number of neurons as the previous layer, and leaky rectified linear units (Leaky ReLU) used as activation functions between layers. The hyperspace was set to be 16-dimensional (resulting in $\mathbf{c}$ with dimensions 4×4), which was found to enhance the performance of AE during pretraining.

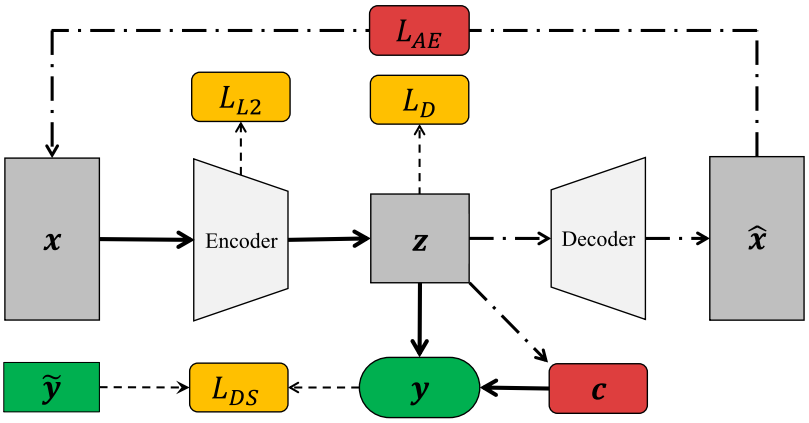


Fig. 5. Architecture of Diversity-DeepSAD. Dash-dotted lines indicate the pretraining autoencoder and dashed lines indicate the loss function.

Table 2
Diversity-DeepSAD hyperparameters space.

Hyperparameter	Min value	Max value
Batch Size	50	150
Learning Rate (Pretraining)	0.0001	0.001
Learning Rate	0.0001	0.001
Number of Epochs (Pretraining)	5	20
Number of Epochs	50	200

Hyperparameters with the strongest influence on model performance, i.e. batch size, learning rate (AE and DeepSAD stages) and the number of training epochs, were optimised via Bayesian optimisation (see Table 2). Optimisation was performed independently for every pair of frequency and feature type. Each run comprised 20 objective evaluations: the first 10 corresponded to uniform random sampling of the search space, and the remaining 10 were selected by a Gaussian-process surrogate model. The surrogate used a Matérn kernel with automatic relevance determination of the length scales, and a default acquisition strategy, which adaptively balances expected improvement, probability of improvement and lower confidence bound. The scalar optimisation objective was the F_{all} score.

At every Bayesian optimisation iteration, a DeepSAD model was trained on the four training specimens of the fold, and F_{all} was computed using only these units. The held-out specimen was excluded entirely during optimisation and used solely for testing. Across frequencies, optimisation trajectories typically stabilised after 15–18 evaluations, with flat incumbent curves and repeated sampling of similar hyperparameters, indicating convergence. For each (frequency, method), the hyperparameter set achieving the maximum observed F_{all} was selected and retrained once on the full training set prior to evaluation on the test specimen.

All other hyperparameters were fixed based on literature [46]: $\nu = 10$, $\eta = 10$, $\lambda = 0.001$ and $\epsilon = 1 \cdot 10^{-6}$. Section 3.7, Section 4.6 and Appendix B discuss the sensitivity analysis to demonstrate that these fixed weights lie within broad regions of high performance, confirming their robustness.

3.6.2. DTC-VAE

VAEs are probabilistic generative models that encode an input vector x_j into a lower-dimensional latent space z_j. While traditional autoencoders map x_j to z_j directly, VAEs instead learn a probabilistic mapping $\phi(z_j|x_j)$ in the encoder. This is parametrised by a mean μ and variance σ^2, allowing the model to represent uncertainty in the latent space. The latent variable, which in this paper is considered as the HI, is sampled according to Eq. (15), where ϵ is Gaussian noise added for variability.

$$z = \mu + \sigma \odot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1) \tag{15}$$

Finally, the decoder produces a reconstruction of the input $\hat{x}_j$ from the latent variable z_j, using the decoder probability distribution $\psi(x_j|z_j)$. This mechanism is illustrated in Fig. 6, where yellow and green represent training and model outputs, respectively.

The DTC-VAE model is optimised by minimising the loss function, $\mathcal{L}_{DTC-VAE}$, which is a summation of the Kullback–Leibler divergence ($\mathcal{L}_{KL}$) [66], reconstruction loss ($\mathcal{L}_{RE}$) and monotonicity constraint loss ($\mathcal{L}_{Mo}$):

$$\mathcal{L}_{DTC-VAE} = \alpha\mathcal{L}_{KL}(\phi,\psi) + \beta\mathcal{L}_{RE}(\phi,\psi) + \gamma\mathcal{L}_{Mo}(z) \tag{16}$$

where α, β, and γ are their respective hyperparameter weights.

Secondly, $\mathcal{L}_{RE}$, portrayed in Eq. (17), characterises the difference between the encoder input $\mathbf{x}_j^m$ and reconstructed decoder output $\hat{x}_j^m$ for data of a specimen m and timestep j, ensuring that these two are as similar as possible.

$$\mathcal{L}_{RE}(\phi,\psi) = \sum_{m=1}^{M}\sum_{j=1}^{N_m}(\mathbf{x}_j^m - \hat{\mathbf{x}}_j^m)^2 \tag{17}$$

Finally, the latent variable is prompted to follow a monotonic trend by adding the monotonicity constraint loss, $\mathcal{L}_{Mo}$. This was proposed by Qin et al. [55], ensuring the extracted HIs followed a monotonic trend by adding a degradation-trend constraint, hence the name DTC-VAE. While the other loss terms are common for VAE architectures, this one is unique to the DTC-VAE model and has a direct effect on the fitness of the output. The monotonicity constraint loss is expressed as follows:

$$\mathcal{L}_{Mo}(z) = \sum_{j=2}^{N}(z_j^m - z_{j-1}^m - r)^2 \tag{18}$$

where z_j^m represents the latent variable for specimen m at timestep j, while the rate of degradation r is a hyperparameter ensuring that $z_j^m > \left(z_{j-1}^m + r\right)$, which controls the magnitude of change of an HI between timesteps. Its value was determined from the original DTC-VAE study [55], in essence, a randomly selected constant in the range $(9, 10)$.

The original HIs criteria functions could not be adapted for gradient descent, given the fact that they are not differentiable and require comparison of latent outputs across specimens, inaccessible during model training. In the case of monotonicity, originally presented in Eq. (1), an alternate formulation $\mathcal{L}_{Mo}$ has been implemented, which ensures high monotonicity scores and is compatible with the training process. Ideally, all prognostic criteria (Mo, Pr, Tr) shall be implemented into the loss function in future to maximise fitness.

A common issue in VAEs is posterior collapse, where the latent variables fail to encode meaningful information because the decoder learns to reconstruct inputs directly and the posterior distribution collapses to the prior. This typically occurs when the KL divergence term dominates the loss [67]. In this work, the risk of collapse is mitigated through optimisation of the loss weights α, β, and γ, following Qin et al. [55],

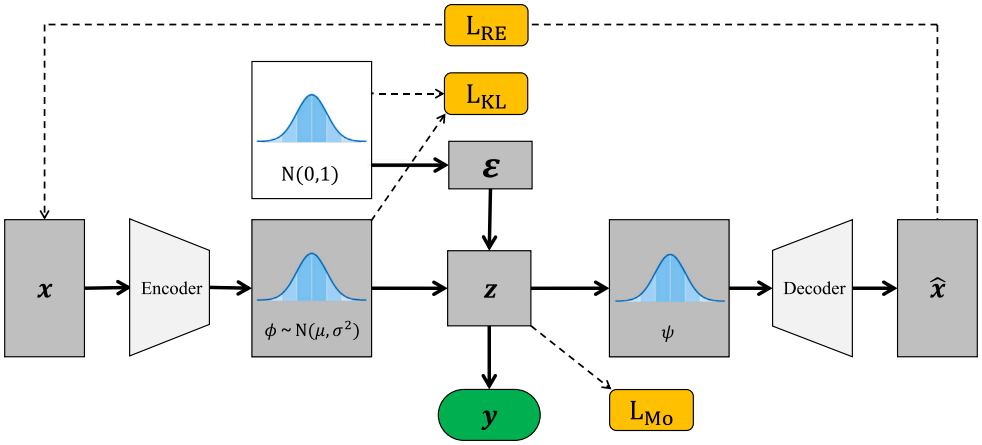


Fig. 6. Architecture of degradation-trend-constrained variational autoencoder (DTC-VAE).

and by the inclusion of the monotonicity constraint $\mathcal{L}_{Mo}$, which forces the latent variables to capture degradation trends instead of defaulting to the prior.

DTC-VAE architecture and hyperparameter optimisation: The DTC-VAE model consists of 3 components: the encoder, latent space and decoder. The encoder's only hidden layer uses a sigmoid activation function and has a varying number of neurons, kept as a hyperparameter. This is followed by two output layers of size 1, representing the mean and log variance of the latent space. As explained by Eq. (15), the reparameterisation trick is used to allow for gradient calculation and thus backpropagation. The decoder acts as a mirror to the encoder, it takes the latent variable as input to a hidden layer with sigmoid activation functions. This hidden layer has the same number of neurons as in the encoder. It is followed by an output layer the same size as the input with linear activation function, which returns the reconstructed input. All weights are initialised with uniform Glorot (Xavier) initialisation [68] due to its state-of-the-art performance with sigmoid activation functions and improved stability. Adam is used as the optimiser.

The hyperparameters were optimised over the space detailed in Table 3, again using Bayesian optimisation with Gaussian processes and F_{all} as the objective. For each frequency and feature type (SP method), the optimiser was run for 40 evaluations: the first 10 corresponded to random samples from the hyperparameter space, and the remaining 30 were guided by the Gaussian-process surrogate. At each evaluation, DTC-VAE was trained on the four training specimens of the current fold, and F_{all} was computed from the resulting HIs of these four specimens only. The held-out panel did not contribute to the optimisation objective.

Convergence behaviour was similar to that of Diversity-DeepSAD: the best observed F_{all} typically stabilised after approximately 15 iterations, with subsequent iterations exploring hyperparameters within a narrow band of performance. The final configuration (per frequency and method) was selected as the hyperparameter set achieving the highest F_{all} across all iterations and retrained on the full training set before computing the test metrics. The sensitivity analysis in Section 3.7, Section Section 4.6 and Appendix B confirms that the selected loss-weight combinations lie inside broad plateaus of high performance, rather than isolated optima.

3.6.3. Ensemble learning model

To ensure the reproducibility of results, it is essential to initialise the models with predetermined random seeds. After hyperparameter optimisation with a given seed, the models were each trained for 5 different random seeds to evaluate the stability of the models. The HIs generated by each model were then mean-averaged, and the HIs criteria applied to obtain the fitness scores reported in Section 4.

Table 3
DTC-VAE hyperparameters space.

Hyperparameter	Min value	Max value
Hidden Layer Size	40	60
Batch Size	75	95
Learning Rate	0.001	0.01
Number of Epochs	500	600
α	1.4	1.8
β	0.05	0.1
γ	2.6	3

The weighted averaging ensemble (WAE) model was implemented to fuse the HI results of the N_f different excitation GW frequencies. This is done to improve performance by reducing instability, and removing the need to select and rely on only one frequency. It was chosen to use the normalised fitness scores of the averaged HIs as weights $\tilde{\omega}_f$ for the same frequency, as this method was most successfully implemented previously [15]. This method is described by Section 3.6.3, with the constants ω_f adjusted depending on their fitness, producing a new weighted-average HI value $\bar{y}$.

$$\bar{y} = \sum_{f=1}^{N_f} \tilde{\omega}_f \bar{y}^f \tag{19a}$$

$$\tilde{\omega}_f = \frac{\omega_f}{\Sigma_{f=1}^{F} \omega_f} \tag{19b}$$

$$\omega_f = F_{all}\left(\bar{y}^f\right) \tag{19c}$$

To summarise the framework, Algorithm 1 presents the full HI generation workflow. It links SP and feature selection with per-frequency model training under leave-one-unit-out cross-validation, prognostic metric computation across random initialisation seeds, and the final weighted frequency-fusion stage. The information pipeline during this process is also shown visually in Fig. 7. Multi-frequency GW measurements are first processed via feature extraction methods, including the HT and FFT, to obtain informative signal representations. These features are then used as input to the proposed models, augmented Diversity-DeepSAD and DTC-VAE, to generate per-frequency HIs. Frequency fusion, performed using WAE, integrates the per-frequency HIs into a unified HI for each model-transform combination.

3.7. Hyperparameter sensitivity analysis

In addition to Bayesian optimisation, a sensitivity analysis was performed to assess the robustness of Diversity-DeepSAD and DTC-VAE to their hyperparameters, and quantify how perturbations around the nominal hyperparameters affect the resulting HIs.

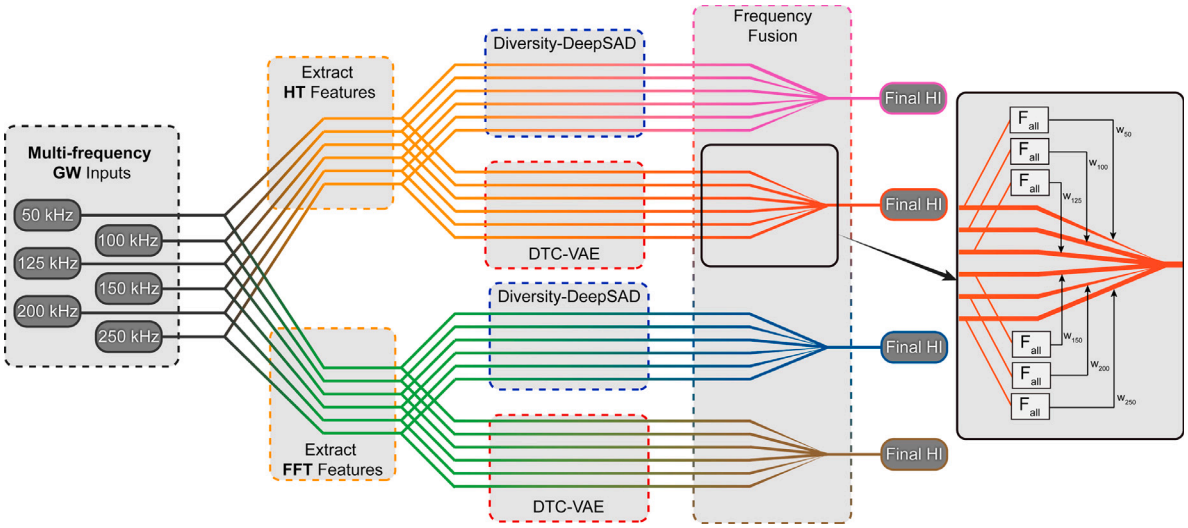


Fig. 7. Schematic of the frequency-fusion pipeline, where multi-frequency guided waves are transformed into FFT- and HT-based features to generate per-frequency HIs, which are fused using prognostic-criteria–optimised weights.

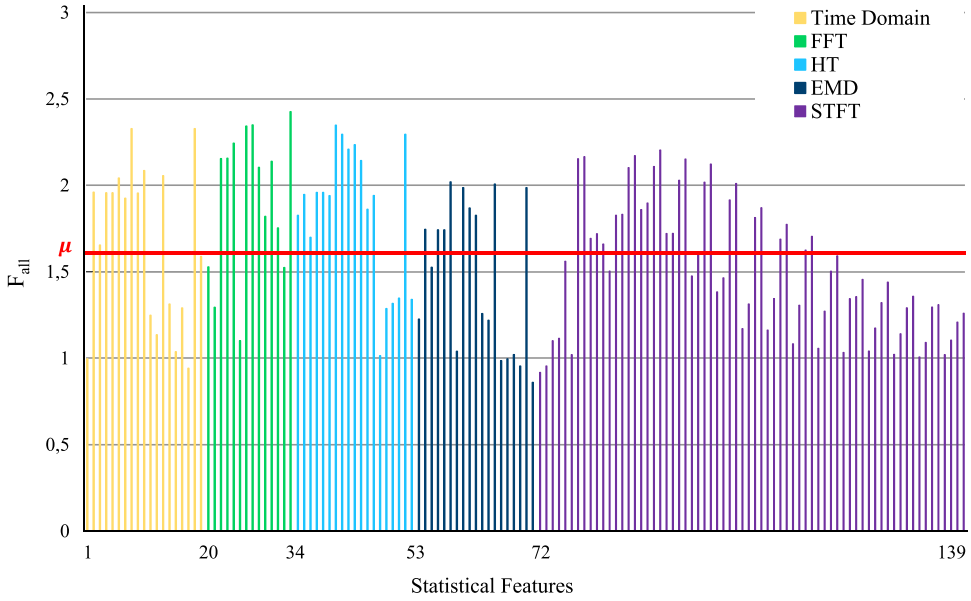


Fig. 8. Fitness scores of statistical features extracted from the time domain and other domains derived from various signal processing methods.

For Diversity-DeepSAD, the analysis focused on the three weighting hyperparameters in Eq. (11) and Eq. (12): the L_2 regularisation weight ν, the auxiliary label weighting η, and the diversity term weight λ. Other hyperparameters were fixed to their per-frequency optimal values as baseline. A three-dimensional grid search was then performed over ν, η and λ on a logarithmic scale, $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ for each parameter. For every triplet (ν, η, λ) and each excitation frequency, the model was retrained using the same cross-validation folds as in the main experiments, and the HIs were recomputed for all specimens. The prognostic criteria in Section 3.3 were then evaluated to obtain both F_{all} and F_{test}, and the scores were aggregated by averaging across folds and excitation frequencies. This allowed response surfaces $F_{all}(\eta, \nu|\lambda)$ and $F_{test}(\eta, \nu|\lambda)$ to be constructed for each value of λ, allowing the identification of regions of stable high performance and the verification of the configuration used in the main study.

As for DTC-VAE, a similar procedure was followed for the loss weights α, β and γ in Eq. (16). Similarly, other hyperparameters were fixed to optimal parameters found during Bayesian optimisation to create a per-frequency baseline, allowing a comparison where only the loss weights were altered. A regular grid was then sampled over (α, β, γ) beyond the ranges reported in Table 3. Again, response surfaces $F_{all}(\alpha, \beta|\gamma)$ and $F_{test}(\alpha, \beta|\gamma)$ were plotted for each γ, demonstrating the effect of each hyperparameter and justifying the convergence and stability in the given hyperparameter space.

4. Results

In this section, SP methods are first compared based on the HIs criteria calculated from the statistical features extracted for each method. Following this, the results of the semi-supervised Diversity-DeepSAD model and the unsupervised DTC-VAE model are presented and compared using the HIs criteria across all specimens F_{all} (training, validation, and test) as well as for test specimens alone F_{test}. Beyond comparing the two deep learning models, the analysis also examines the impact of different excitation GW frequencies, the fusion of all excitation GW frequencies, and the selection of SP methods (focusing on the two top-performing ones).

Algorithm 1: Overview of HI generation and frequency-fusion framework

Input: Guided wave measurements for all specimens and excitation frequencies
Output: Per-frequency and fused HI trajectories with associated prognostic metrics

Stage 2-3: Feature extraction and selection
foreach *GW measurement (specimen, frequency, path, timestep)* **do**
  apply SP methods (FFT, STFT, EMD, HT)
  extract statistical features
foreach *candidate feature* **do**
  aggregate per timestep and specimen; compute Mo, Pr, Tr and F_{all} for feature
retain features with fitness above the mean
select the best-performing SP methods (i.e. FFT, HT)

Stage 4a: Per-frequency HI generation
foreach *selected SP method* **do**
  foreach *GW excitation frequency* f **do**
    foreach *LOOCV fold (choose one specimen as test, remaining specimens as train)* **do**
      build training and test feature sets for all timesteps
      foreach *model* $\in$ *{Diversity-DeepSAD, DTC-VAE}* **do**
        fine-tune hyperparameters by Bayesian optimisation, maximising F_{all} given only training specimens
        foreach *random seed number* **do**
          train the fine-tuned model using the random seed number and generate per-frequency HIs $y^{m,f}_{(\text{seed})}(t)$ for all specimens
          compute Mo, Pr, Tr, F_{all} and Mo_{test}, Pr_{test}, F_{test} for this combination (model, SP, f, fold, seed) and store the non-fused HIs and metrics
        aggregate prognostic metrics across seeds (mean and standard deviation) for this combination (model, SP, f, fold) for reporting

Stage 4b: Frequency fusion (WAE)
foreach *selected SP method* **do**
  foreach *model* $\in$ *{Diversity-DeepSAD, DTC-VAE}* **do**
    foreach *LOOCV fold* **do**
      foreach *random seed number* **do**
        apply WAE to per-frequency HIs $y^{m,f}_{(\text{seed})}(t)$ across all GW excitation frequencies f to obtain fused HIs $\bar{y}^{m}_{(\text{seed})}(t)$ for all specimens
        compute Mo, Pr, Tr, F_{all} and Mo_{test}, Pr_{test}, F_{test} for this fused combination (model, SP, fold, seed) and store fused HIs and metrics
      aggregate fused prognostic metrics across seeds (mean and standard deviation) for this combination (model, SP, fold)
return *Per-frequency and fused HIs together with their prognostic metrics*

4.1. Signal processing

Fig. 8 shows the F_{all} scores averaged across all paths and frequencies for each of the extracted statistical features. It is clear that all SP methods contain some features that perform poorly when evaluated for fitness. In order to discard the low-performing features, the set is reduced to only those performing above average ($\mu = 1.614$), which is indicated by the red horizontal line in Fig. 8. The mean score of features extracted from each SP method is presented in Table 4. It can be seen that after the feature reduction, FFT and HT yield the highest-performing features, which are used in the subsequent AI models.

4.2. Diversity-DeepSAD

The Diversity-DeepSAD model was trained across all combinations of frequencies and data folds (i.e. training and test divisions), with WAE implemented as described. The HIs constructed by Diversity-DeepSAD upon FFT and HT features are shown in Figs. 9 and 10, respectively. For better comparison, the HIs are displayed over the normalised lifetime of specimens, ranging from 0% to 100%. The resulting fitness scores of the HIs are reported in Table 5 and Table 6. These tables include F_{all} scores, from Eq. (6), and F_{test} scores, from Eq. (7), for HIs generated from both FFT and HT features.

Diversity-DeepSAD results show high fitness scores for both FFT and HT, with mean F_{test} scores of 2.27 and 2.13 respectively, with F_{all} scores averaging slightly higher at 2.35 and 2.22. This implies a general higher performance of features from the FFT than those from the HT. No fold performed consistently worse than the others, although average results from test Sample 1 were consistently highest. Inspecting the graph, this may be due to the invariably conforming behaviour of this sample, aiding in high F_{test} scores. All frequencies performed similarly, with 50 kHz producing marginally better average F_{all} results for both SP methods.

While the range of scores is similar for both sets of input data, FFT demonstrates a greater robustness due to its low standard deviation between random seeds, with an F_{test} mean $\bar{\sigma} = 0.10$ compared to 0.14 for HT. This indicates a stronger robustness in this model using the FFT features than using those of the HT.

4.3. DTC-VAE

The DTC-VAE model was similarly trained and tested on all combinations of frequencies and folds for 5 random seeds. The HIs extracted by DTC-VAE from both FFT and HT features are displayed in Figs. 11 and 12, respectively. The resulting fitness scores of the HIs generated by the DTC-VAE are thus provided in Tables 7 and 8. These tables cover both FFT and HT features, detailing scores for the general fitness F_{all} and test fitness F_{test}.

FFT features achieved consistently high results, with a mean F_{test} scores of 2.55 and 2.20 for the FFT and HT respectively, and similar F_{all} scores of 2.56 and 2.25. These imply features from the FFT generate again generally better performing HIs than the HT. No fold produced consistently different scores to the others; inspection of the graphs reveals distributed anomalies for all. However, for DTC-VAE, HIs generated using features from the 100 kHz signal tended to perform poorly, with an average F_{test} score of 2.06 compared to the overall mean score of 2.38.

FFT features exhibited higher stability than those of the HT, with an overall lowest F_{test} score of 2.24 compared to 1.42. This is further highlighted by the low standard deviation values, with an F_{test} mean $\bar{\sigma} = 0.09$ compared to 0.12 for HT. The stronger robustness in both models using the FFT features than those of the HT indicates that the former include more reliable information on the structure's health, consistent with the results in Table 4.

4.4. Weighted average ensemble

Overall, the fused models show improved results compared to single frequencies, with generally higher scores consistently outperforming

Table 4
Mean fitness scores for each signal processing method, calculated before and after feature reduction.

	Time domain	FFT	HT	EMD	STFT
Mean F_{all} before reduction	1.673 (± 0.45)	1.924 (± 0.41)	1.840 (± 0.39)	1.474 (± 0.42)	1.511 (± 0.37)
Mean F_{all} after reduction	2.022 (± 0.18)	**2.149** (± 0.21)	**2.046** (± 0.20)	1.880 (± 0.11)	1.910 (± 0.19)

Table 5
Fitness scores across all units (F_{all}) for the Diversity-DeepSAD model using FFT and HT features over 5 iterations.

f [kHz]	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
	FFT	HT	FFT	HT	FFT	HT	FFT	HT	FFT	HT
50	2.55 (± 0.04)	2.29 (± 0.28)	2.47 (± 0.03)	2.49 (± 0.03)	2.40 (± 0.05)	1.85 (± 0.06)	2.50 (± 0.06)	2.47 (± 0.07)	2.47 (± 0.06)	2.67 (± 0.08)
100	2.34 (± 0.07)	2.30 (± 0.05)	2.22 (± 0.05)	2.19 (± 0.14)	2.36 (± 0.04)	2.31 (± 0.04)	1.92 (± 0.19)	2.22 (± 0.10)	2.47 (± 0.03)	1.99 (± 0.11)
125	2.43 (± 0.06)	2.02 (± 0.22)	2.20 (± 0.12)	2.42 (± 0.04)	2.38 (± 0.03)	2.37 (± 0.08)	2.42 (± 0.10)	2.07 (± 0.23)	2.47 (± 0.05)	2.18 (± 0.33)
150	2.38 (± 0.12)	2.56 (± 0.11)	2.41 (± 0.04)	2.42 (± 0.06)	2.16 (± 0.14)	2.28 (± 0.08)	2.06 (± 0.11)	2.37 (± 0.14)	2.49 (± 0.06)	2.06 (± 0.14)
200	2.39 (± 0.05)	2.35 (± 0.04)	2.32 (± 0.03)	2.04 (± 0.17)	2.34 (± 0.06)	1.82 (± 0.09)	2.51 (± 0.05)	2.33 (± 0.13)	2.32 (± 0.11)	2.51 (± 0.03)
250	2.36 (± 0.05)	2.26 (± 0.04)	2.30 (± 0.05)	1.95 (± 0.13)	2.32 (± 0.05)	1.92 (± 0.32)	2.14 (± 0.07)	1.67 (± 0.19)	2.28 (± 0.10)	2.25 (± 0.15)
Fusion	2.50 (± 0.03)	2.51 (± 0.02)	2.42 (± 0.03)	2.37 (± 0.07)	2.44 (± 0.02)	2.01 (± 0.08)	2.38 (± 0.05)	2.22 (± 0.12)	2.51 (± 0.05)	2.44 (± 0.06)

Table 6
Test fitness scores (F_{test}) for the Diversity-DeepSAD model using FFT and HT features over 5 iterations.

f [kHz]	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
	FFT	HT	FFT	HT	FFT	HT	FFT	HT	FFT	HT
50	2.63 (± 0.06)	2.25 (± 0.24)	2.44 (± 0.04)	2.48 (± 0.02)	2.30 (± 0.08)	1.57 (± 0.09)	2.41 (± 0.08)	2.48 (± 0.07)	2.39 (± 0.13)	2.53 (± 0.10)
100	2.37 (± 0.08)	2.26 (± 0.05)	2.17 (± 0.12)	2.00 (± 0.15)	2.21 (± 0.06)	2.12 (± 0.06)	1.67 (± 0.22)	2.13 (± 0.11)	2.40 (± 0.08)	1.64 (± 0.10)
125	2.42 (± 0.12)	1.99 (± 0.18)	2.14 (± 0.17)	2.28 (± 0.08)	2.35 (± 0.07)	2.43 (± 0.11)	2.30 (± 0.15)	2.03 (± 0.30)	2.36 (± 0.05)	2.21 (± 0.27)
150	2.29 (± 0.11)	2.58 (± 0.14)	2.42 (± 0.04)	2.27 (± 0.10)	2.31 (± 0.10)	2.32 (± 0.06)	1.75 (± 0.08)	2.30 (± 0.17)	2.35 (± 0.07)	1.93 (± 0.12)
200	2.40 (± 0.05)	2.31 (± 0.08)	2.28 (± 0.06)	2.02 (± 0.23)	2.31 (± 0.05)	1.63 (± 0.11)	2.46 (± 0.04)	2.38 (± 0.15)	2.07 (± 0.25)	2.47 (± 0.03)
250	2.36 (± 0.08)	2.18 (± 0.11)	2.17 (± 0.06)	1.59 (± 0.15)	2.23 (± 0.13)	1.77 (± 0.44)	1.88 (± 0.11)	1.56 (± 0.13)	2.17 (± 0.12)	2.15 (± 0.20)
Fusion	2.56 (± 0.03)	2.50 (± 0.03)	2.44 (± 0.04)	2.31 (± 0.07)	2.35 (± 0.05)	1.74 (± 0.06)	2.25 (± 0.08)	2.20 (± 0.09)	2.37 (± 0.09)	2.18 (± 0.09)

Table 7
Fitness scores across all units (F_{all}) for the DTC-VAE model using FFT and HT features over 5 iterations.

f [kHz]	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
	FFT	HT	FFT	HT	FFT	HT	FFT	HT	FFT	HT
50	2.66 (± 0.03)	2.16 (± 0.11)	2.45 (± 0.05)	2.55 (± 0.07)	2.57 (± 0.04)	2.42 (± 0.06)	2.52 (± 0.11)	2.40 (± 0.06)	2.50 (± 0.05)	2.09 (± 0.14)
100	2.37 (± 0.11)	1.96 (± 0.11)	2.41 (± 0.19)	1.88 (± 0.13)	2.55 (± 0.09)	1.71 (± 0.19)	2.52 (± 0.14)	1.61 (± 0.20)	2.45 (± 0.08)	1.50 (± 0.15)
125	2.43 (± 0.07)	2.00 (± 0.13)	2.49 (± 0.11)	2.11 (± 0.11)	2.54 (± 0.07)	2.20 (± 0.13)	2.57 (± 0.13)	2.30 (± 0.10)	2.57 (± 0.05)	2.27 (± 0.04)
150	2.69 (± 0.05)	2.63 (± 0.10)	2.54 (± 0.09)	2.59 (± 0.05)	2.53 (± 0.11)	2.60 (± 0.07)	2.51 (± 0.08)	2.62 (± 0.03)	2.75 (± 0.02)	2.65 (± 0.05)
200	2.81 (± 0.03)	2.50 (± 0.07)	2.55 (± 0.10)	2.51 (± 0.02)	2.73 (± 0.06)	2.62 (± 0.08)	2.77 (± 0.03)	2.64 (± 0.07)	2.79 (± 0.04)	2.52 (± 0.08)
250	2.35 (± 0.11)	2.06 (± 0.06)	2.66 (± 0.05)	2.19 (± 0.07)	2.44 (± 0.21)	1.97 (± 0.04)	2.51 (± 0.13)	1.99 (± 0.25)	2.67 (± 0.06)	2.16 (± 0.05)
Fusion	2.75 (± 0.01)	2.55 (± 0.02)	2.72 (± 0.05)	2.51 (± 0.07)	2.77 (± 0.07)	2.56 (± 0.05)	2.80 (± 0.03)	2.61 (± 0.04)	2.80 (± 0.03)	2.49 (± 0.02)

Table 8
Test fitness scores (F_{test}) for the DTC-VAE model using FFT and HT features over 5 iterations.

f [kHz]	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
	FFT	HT	FFT	HT	FFT	HT	FFT	HT	FFT	HT
50	2.68 (± 0.02)	2.20 (± 0.13)	2.41 (± 0.07)	2.50 (± 0.08)	2.57 (± 0.07)	2.35 (± 0.10)	2.52 (± 0.08)	2.39 (± 0.08)	2.53 (± 0.04)	2.03 (± 0.15)
100	2.34 (± 0.21)	1.92 (± 0.08)	2.36 (± 0.15)	1.98 (± 0.11)	2.49 (± 0.13)	1.42 (± 0.30)	2.41 (± 0.21)	1.54 (± 0.19)	2.52 (± 0.07)	1.61 (± 0.21)
125	2.43 (± 0.06)	1.86 (± 0.19)	2.43 (± 0.14)	2.00 (± 0.13)	2.58 (± 0.08)	2.32 (± 0.08)	2.65 (± 0.10)	2.22 (± 0.14)	2.62 (± 0.08)	2.27 (± 0.06)
150	2.72 (± 0.05)	2.66 (± 0.12)	2.35 (± 0.16)	2.44 (± 0.12)	2.58 (± 0.11)	2.65 (± 0.05)	2.51 (± 0.08)	2.53 (± 0.05)	2.75 (± 0.04)	2.62 (± 0.06)
200	2.83 (± 0.02)	2.58 (± 0.09)	2.51 (± 0.15)	2.26 (± 0.10)	2.73 (± 0.06)	2.57 (± 0.07)	2.78 (± 0.02)	2.65 (± 0.07)	2.76 (± 0.05)	2.61 (± 0.05)
250	2.24 (± 0.07)	1.90 (± 0.11)	2.60 (± 0.11)	2.13 (± 0.08)	2.44 (± 0.27)	1.55 (± 0.06)	2.58 (± 0.12)	2.05 (± 0.33)	2.65 (± 0.05)	2.35 (± 0.09)
Fusion	2.76 (± 0.02)	2.60 (± 0.03)	2.71 (± 0.05)	2.43 (± 0.09)	2.76 (± 0.08)	2.41 (± 0.08)	2.77 (± 0.02)	2.61 (± 0.04)	2.81 (± 0.03)	2.61 (± 0.03)

the corresponding average fitness score of a particular fold. Across both FFT and HT, the average unfused F_{all} and F_{test} scores for Diversity-DeepSAD are 2.29 and 2.20, while for DTC-VAE they are 2.41 and 2.38 respectively. The same fused scores average 2.38, 2.29, 2.66 and 2.65. This is an average score increase of 3.2% for Diversity-DeepSAD, but 10.9% for DTC-VAE, with both F_{all} and F_{test} scores showing similar improvement. The high performance of WAE, fusing different GW excitation frequencies, combined with DTC-VAE is also evident from the graphs, where the fused HIs consistently behave more cleanly than their constituent frequencies.

The stability of fused results is also greatly improved, with deviations between folds smaller than within individual frequencies. For both models, but still most noticeably DTC-VAE, the fused graphs are visually more consistent between folds than unfused. Between seeds, the standard deviations of fitness for the fusion are always below 0.1, with one exception in Diversity-DeepSAD fold 4. The average standard deviation for Diversity-DeepSAD-Fusion scores is half that of unfused, 0.05 compared to 0.10 for F_{all} and 0.06 to 0.12 for F_{test}. For DTC-VAE, the reduction is better than half at 0.04 from 0.09 and 0.05 from 0.11 respectively. These low standard deviations indicate that HIs fused using WAE of multiple GW excitation frequencies are more robust. This is because different GW excitation frequencies can capture various damage modes in structures, and the ensemble approach reduces susceptibility to errors in individual constituent models.

Across all results, there are only three occurrences where the standard deviation of fused results for HIs from the FFT exceeds that of the HT, confirming that the features extracted from the output of FFT also tend to produce more consistently performing HIs. The F_{test} scores for Diversity-DeepSAD with WAE average 2.39 and 2.19 for FFT and HT respectively, while DTC-VAE shows a large increase to 2.76 and 2.53. Every DTC-VAE WAE fitness score is higher than its Diversity-DeepSAD

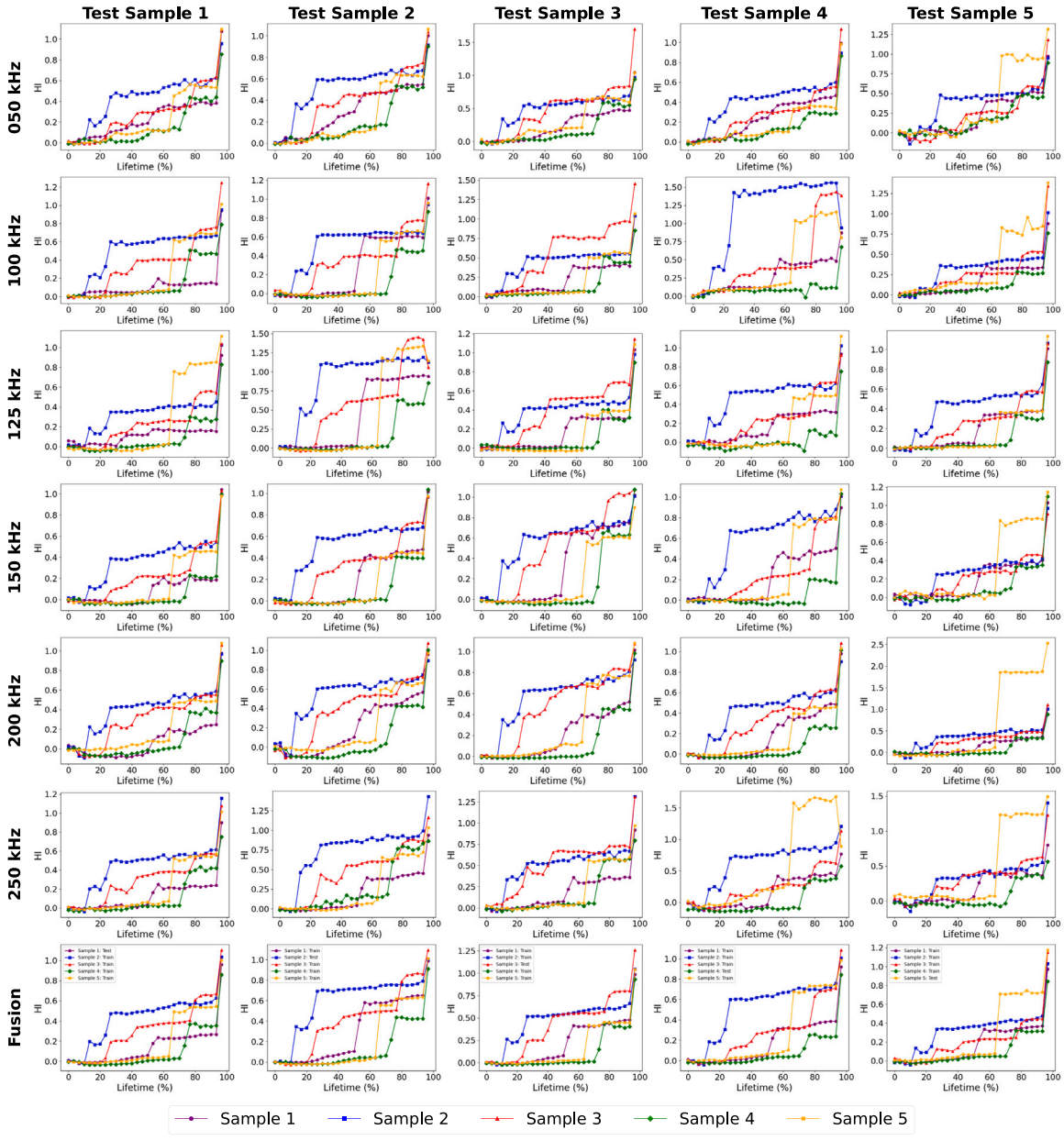


Fig. 9. HIs constructed by Diversity-DeepSAD using FFT features.

counterpart, showing exceptionally high performance at a peak of 2.81 or 94%.

4.5. Model comparison

The prognostic criteria and fitness scores (reported in the previous subsections and tables) were averaged across all folds for both models to comprehensively perform a comparison, including the WAE models. The resulting averages for F_{all}, along with the prognostic criteria scores, are displayed in Table 9, while those for F_{test} are included in Table 10.

Both models showed promising monotonicity scores, with Diversity-DeepSAD achieving an average Mo score of 0.87 and 0.87 for FFT and HT features respectively, while its average Mo_{test} scores were 0.83 and 0.82. DTC-VAE achieved average Mo scores of 0.92 and 0.85 and average Mo_{test} scores 0.91 and 0.81.

The average prognosability scores were also high, particularly for DTC-VAE. Diversity-DeepSAD achieved an average Pr score of 0.89 and 0.85 for FFT and HT features respectively. Its average Pr_{test} scores were 0.85 and 0.81. On the other hand, DTC-VAE achieved average Pr scores of 0.97 and 0.87, demonstrating exceptional performance on FFT features. Similarly, its average Pr_{test} scores were 0.91 and 0.81.

The results reflect the fact that trendability was the lowest performing prognostic criterion for both models, as can be seen by the inconsistency of shapes between specimens in the HI graphs. Both models achieved moderate scores, with Diversity-DeepSAD scoring an average Tr score of 0.59 and 0.50, and DTC-VAE scoring 0.68 and 0.52, for FFT and HT features respectively. As explained in Section 3.1, each specimen had differences in manufacturing and the presence or absence of manufacturing disbonds or impact events. These factors significantly impact GW propagation characteristics, and it is therefore physically justified that the Tr score produced was lower than that for Mo or Pr.

Overall, HIs extracted by Diversity-DeepSAD using FFT features performed slightly better than using HT features. Similarly, DTC-VAE also generated better HIs when trained on FFT features, although the difference in performance was greater. However, both models showed significant improvements when combined with WAE, with all prognostic criteria showing improved scores and stability for both FFT and

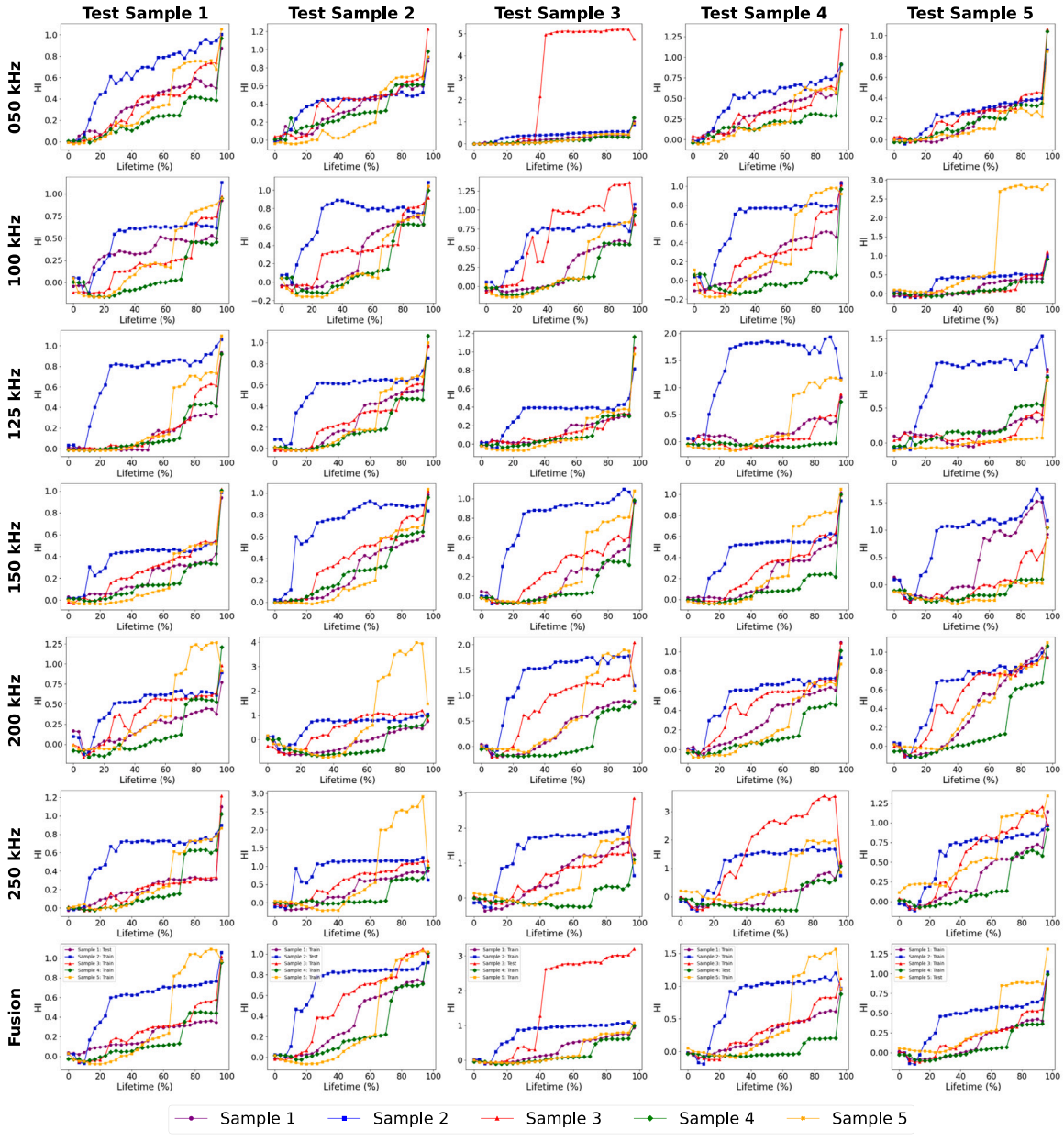


Fig. 10. HIs constructed by Diversity-DeepSAD using HT features.

Table 9
Prognostic criteria and fitness scores (F_{all}) averaged across all folds.

Score	Diversity-DeepSAD		Diversity-DeepSAD WAE		DTC-VAE		DTC-VAE WAE	
	FFT	HT	FFT	HT	FFT	HT	FFT	HT
Mo	0.87 (± 0 .04)	0.87 (± 0 .06)	0.94 (± 0 .01)	0.93 (± 0 .03)	0.92 (± 0 .06)	0.85 (± 0 .07)	0.99 (± 0 .01)	0.93 (± 0 .03)
Pr	0.89 (± 0 .07)	0.85 (± 0 .16)	0.91 (± 0 .03)	0.84 (± 0 .18)	0.97 (± 0 .04)	0.87 (± 0 .13)	0.98 (± 0 .01)	0.90 (± 0 .02)
Tr	0.59 (± 0 .13)	0.50 (± 0 .17)	0.60 (± 0 .08)	0.54 (± 0 .11)	0.68 (± 0 .13)	0.52 (± 0 .22)	0.80 (± 0 .05)	0.71 (± 0 .05)
F_{all}	2.35 (± 0 .16)	2.22 (± 0 .28)	2.45 (± 0 .06)	2.31 (± 0 .20)	2.56 (± 0 .16)	2.25 (± 0 .34)	2.77 (± 0 .05)	2.54 (± 0 .06)

HT features. Diversity-DeepSAD WAE achieved average F_{all} scores of 2.45 and 2.31, while its F_{test} scores were 2.39 and 2.19. In addition, DTC-VAE WAE scored 2.77 and 2.54 in F_{all}, and 2.76 and 2.53 in F_{test}.

4.6. Sensitivity analysis of hyperparameters

The hyperparameter sensitivity analysis confirmed that the conclusions drawn in the previous subsections are not the result of fragile or excessively fine-tuned configurations. For both Diversity-DeepSAD and DTC-VAE, the response surfaces constructed for F_{all} and F_{test} show broad regions of consistently high performance around the nominal hyperparameters obtained via Bayesian optimisation.

Regarding Diversity-DeepSAD, the surfaces $F_{all}(\eta, \nu|\lambda)$ and $F_{test}(\eta, \nu|\lambda)$ for λ of 0.001 and 0.01 in Fig. 13, with full plots in Appendix B, revealed that performance for both features is governed primarily by the diversity weight λ, with some sensitivity to the auxiliary label

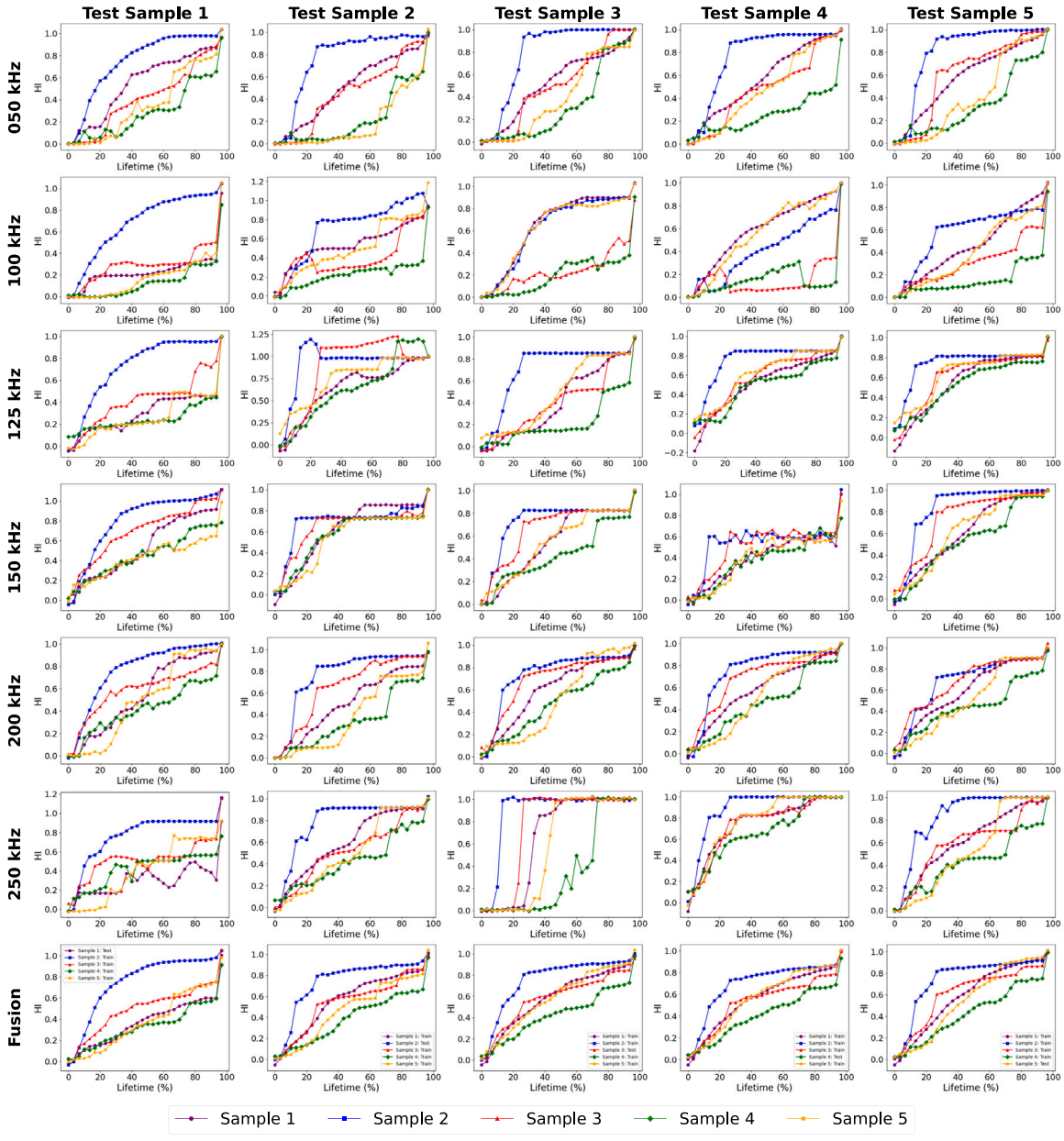


Fig. 11. HIs constructed by DTC-VAE using FFT features.

Table 10
Test prognostic criteria and test fitness scores (F_{test}) averaged across all folds.

Score	Diversity-DeepSAD		Diversity-DeepSAD WAE		DTC-VAE		DTC-VAE WAE	
	FFT	HT	FFT	HT	FFT	HT	FFT	HT
Mo_{test}	0.83 (± 0 .04)	0.82 (± 0 .12)	0.93 (± 0 .04)	0.91 (± 0 .05)	0.91 (± 0 .10)	0.81 (± 0 .13)	0.98 (± 0 .02)	0.91 (± 0 .10)
Pr_{test}	0.85 (± 0 .13)	0.81 (± 0 .24)	0.86 (± 0 .10)	0.74 (± 0 .29)	0.97 (± 0 .05)	0.88 (± 0 .16)	0.97 (± 0 .02)	0.91 (± 0 .09)
Tr	0.59 (± 0 .13)	0.50 (± 0 .17)	0.60 (± 0 .08)	0.54 (± 0 .11)	0.68 (± 0 .13)	0.52 (± 0 .22)	0.80 (± 0 .05)	0.71 (± 0 .05)
F_{test}	2.27 (± 0 .23)	2.13 (± 0 .34)	2.39 (± 0 .12)	2.19 (± 0 .26)	2.55 (± 0 .18)	2.20 (± 0 .38)	2.76 (± 0 .06)	2.53 (± 0 .11)

weight η and weaker dependence on the regularisation weight ν. For small values of λ, a plateau of high fitness values was observed, with F_{all} and F_{test} remaining close to their optima. Only very large ν values systematically degraded performance, showing that the regularisation term can have a significant effect. The large dependence on λ confirmed that the diversity term is essential for producing smooth, monotonic HIs, while the regularisation refined performance. Overall, the configuration used ($\eta = 10$, $\nu = 10$, $\lambda = 0.001$) lies well within the stable high-performance region and is therefore fully justified.

Regarding DTC-VAE, the response surfaces $F_{all}(\alpha, \beta|\gamma)$ and $F_{test}(\alpha, \beta|\gamma)$ exhibited a stronger stability across all hyperparameters, with two examples shown in Fig. 14 and all figures being included in Appendix B. Increases beyond the optimised regions did lead to gradual

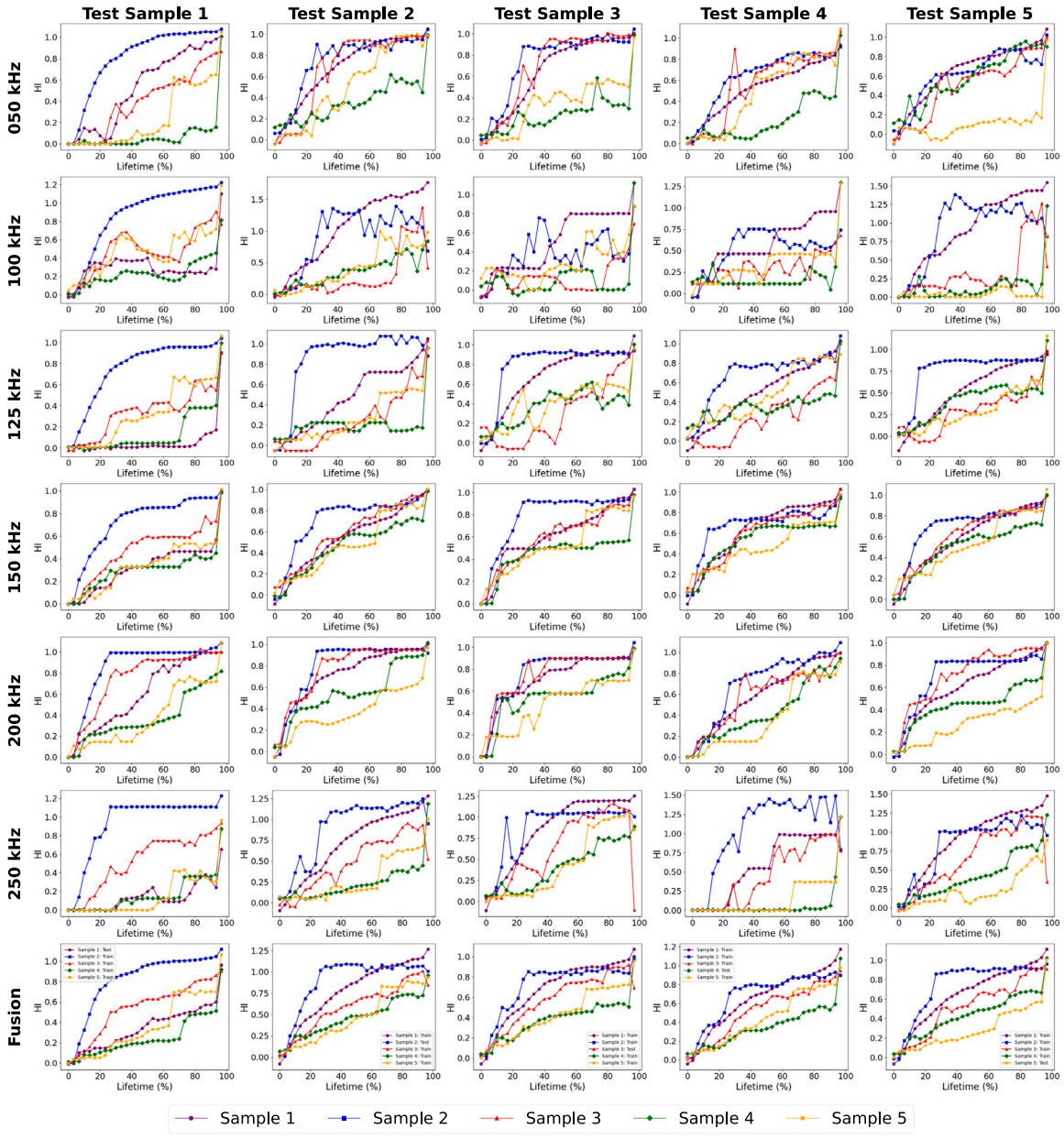


Fig. 12. HIs constructed by DTC-VAE using HT features.

changes in fitness. Very small γ values saw the greatest decrease in fitness, highlighting the importance of the monotonicity constraint in constructing reliable HIs. Furthermore, for small γ, increases in β beyond the optimised region led to reduced fitness, consistent with an overly strong reconstruction loss term dominating the objective, reducing the effective influence of the KL term and thus degrading the learned latent structure. However, this was not observed when the monotonicity constraint was dominant, showing improved stability. In all cases, the ranges optimised in the main experiments fell inside a stable high-fitness region, rather than at a sharp optimum.

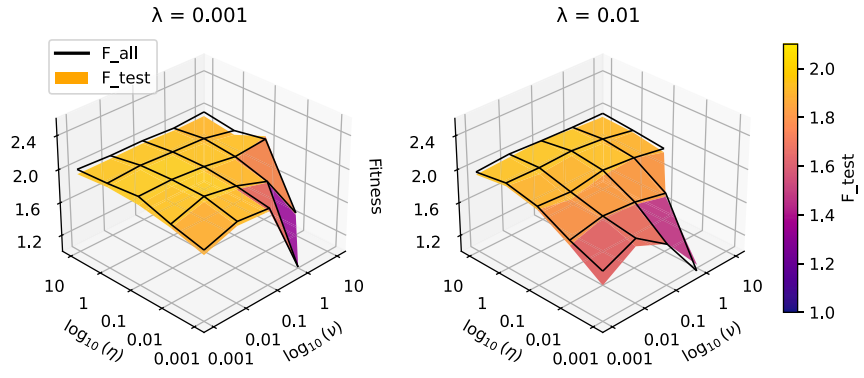
Overall, the sensitivity analysis shows that both models maintain high F_{all} and F_{test} scores across a non-trivial neighbourhood of their nominal hyperparameters. This demonstrates that the comparative advantages of FFT over HT, and of DTC-VAE over Diversity-DeepSAD, arise from genuinely robust model behaviour rather than from narrowly tuned hyperparameter settings.

5. Discussion

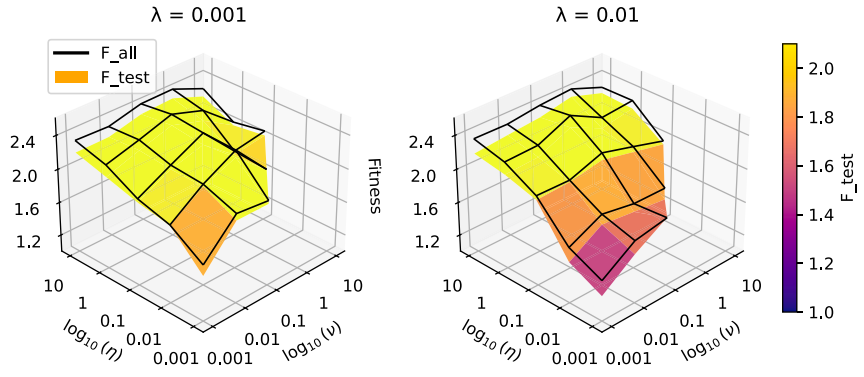
This section contextualises the results by analysing the behaviour of the constructed HIs, assessing their robustness and computational feasibility, and comparing the proposed models with state-of-the-art methods. The discussion links model behaviour to underlying physical mechanisms and highlights practical implications for GW SHM.

5.1. Interpretation of constructed HIs

Developing robust and reliable HIs for composite structures remains a challenging task, as these indicators must capture meaningful



(a) Diversity-DeepSAD using FFT features.



(b) Diversity-DeepSAD using HT features.

Fig. 13. Hyperparameter sensitivity analysis of augmented Diversity-DeepSAD over (ν, η, λ), for the two SP methods.

trends while remaining interpretable and suitable for prognostic applications. This analysis evaluates the augmented Diversity-DeepSAD and DTC-VAE, highlighting key differences in their performances and behaviours, which reveal their respective strengths and limitations in extracting meaningful HIs for composite structure prognostics.

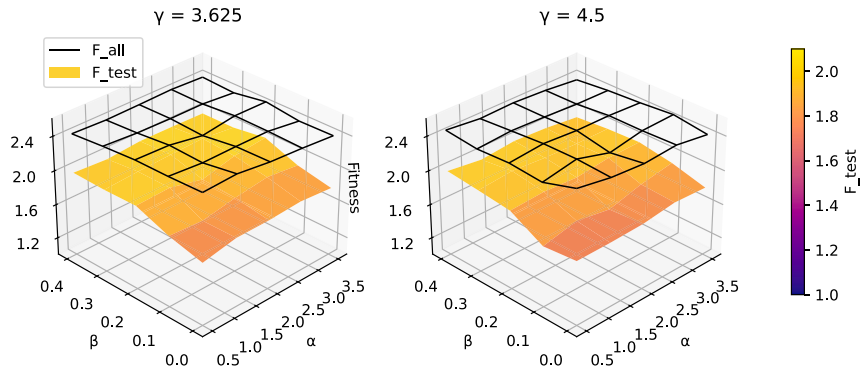
Diversity-DeepSAD graphs show a tendency for HIs to increase in sharp steps, which could indicate different damage states during the fatigue life of composite structures. However, these increases were sometimes erroneous or abrupt; this is most visible in Fig. 10 in test Sample 3 at 50 kHz and test Sample 4 at 125 kHz, showing the model may be sensitive to sudden changes in test sample structural health or sensor damage. On the other hand, DTC-VAE produced smoother HIs, indicating improved performance in this regard. This is particularly the case with FFT features, with some exceptions in the case of HT features.

The incremental changes in HIs produced by Diversity-DeepSAD could be seen as a more explainable result, potentially illustrating the distinct damage states of composite structures. This could provide valuable insights for prognostic models, especially state-based ones, in predicting RUL. However, the increases in HIs should not be too abrupt, else using HIs for RUL prediction would be challenging, as discussed previously. Further studies and testing are needed to establish a stronger connection between these phases and actual physical damage states. On the other hand, HIs produced by DTC-VAE resulted in slightly higher fitness scores and less sharp jumps, that could improve RUL prediction.

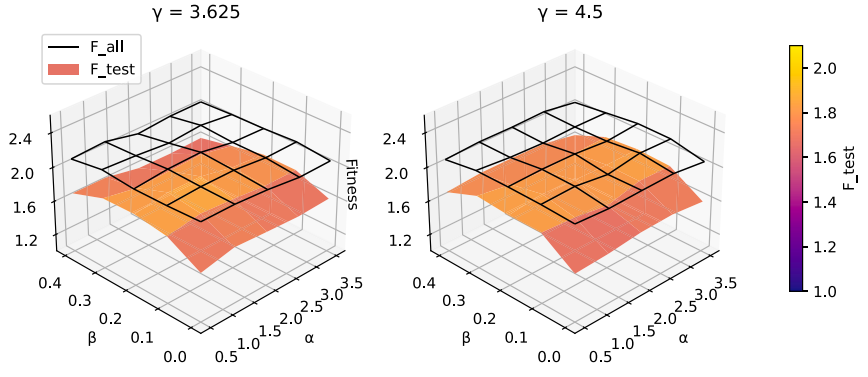
In order for further physical interpretation, Fig. 15 shows the stiffness of each specimen, which is estimated by the slope of the linear region of the load–displacement curves during post-buckling (i.e. ranging from 40 kN to 60 kN) [69]. The stiffness reduction is due to complex damage mechanisms and therefore also represents the physical health of the structure, similarly to the generated HIs. It should be noted that stiffness was not included in the data fed to the models, as it cannot be easily measured in operation, particularly in an aerospace application.

The HIs produced by augmented Diversity-DeepSAD and DTC-VAE can be meaningfully linked to plausible damage mechanisms and measured stiffness trends, while remaining cautious about causality (see Table 11). Stepwise HIs produced by Diversity-DeepSAD highlight discrete changes in GW scattering—consistent with localised events such as delamination initiation, disbond propagation, or sudden coalescence of cracks—whereas the DTC-VAE yields smoother HIs that reflect the gradual accumulation of distributed microdamage (e.g. matrix cracking). Considering the impact/disbond locations, that of Sample 2 lies off the stiffener and shows the earliest HI transition but the smoothest stiffness decline, suggesting local damage detected early by GW but not yet affecting global stiffness; Sample 4's is located on the stiffener and exhibits long stability followed by late HI jumps and a sharper stiffness loss, consistent with critical stiffener-related propagation; Sample 5's pre-existing disbond explains its early, abrupt stiffness reduction (~20% lifetime) while GW responses are path-dependent and thus more model-sensitive. Where both models and feature folds consistently indicate transitions (for instance Sample 2's early change), confidence is increased that the HI reflects a genuine physical change. Differences in timing between HI transitions and stiffness drops therefore provide diagnostic value: early HI changes with late stiffness loss point to local damage not yet structural, while early stiffness drops with variable GW response point to gross load-path alterations (e.g. a disbond).

Furthermore, results from both DTC-VAE and in particular Diversity-DeepSAD showed a tendency for HIs to spike at EoL, such as test



(a) DTC-VAE using FFT features.



(b) DTC-VAE using HT features.

Fig. 14. Hyperparameter sensitivity analysis of DTC-VAE over (α, β, γ), for the two SP methods.

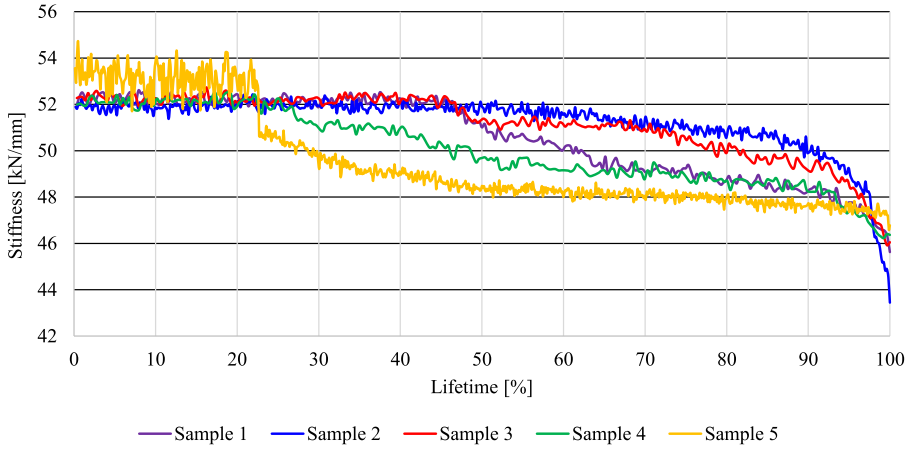


Fig. 15. Post-buckling stiffness against number of compression cycles for each specimen.

Sample 3 at 125 kHz. This is due to the model learning the failure state as a separate category to the continuous range of degradation until that point, an effect which can be emphasised with alternative hyperparameters. This suggests that the deviation in the structural health condition after failure is significantly greater and distinct from the condition before failure, demonstrating the damage accumulation process in composite structures. Although high prognosability scores may indicate that extracted HIs are more suitable for RUL prediction, it is crucial to ensure that changes in HI values are not too abrupt. For example, a sharp increase near the end of life may leave insufficient time to take maintenance action. In this regard, DTC-VAE could potentially provide HIs that are more useful than Diversity-DeepSAD.

The improved performance of DTC-VAE can be linked to the connection between its loss function and the fitness of the extracted HIs. The incorporation of the monotonicity constraint into its loss function makes part of the fitness metric differentiable, and thus, optimisable. This provides a direct relationship between training and output fitness,

Table 11
Summary of data-driven HI behaviour (constructed by augmented Diversity-DeepSAD vs. DTC-VAE, both upon FFT features), stiffness evolution, impact/disbond location, and plausible physical interpretation.

Specimen	Impact/disbond & relative position	DeepSAD's HI (damage states before final jump to EoL & timing)	DTC-VAE's HI (shape & timing)	Stiffness behaviour	Plausible physical interpretation (linking HI→stiffness)
Sample 1	Impact near stiffener edge	2 damage states; first state longer than second	Very smooth HI; gradual monotonic rise	Smoothed sudden reduction, later in life	Impact near stiffener edge causes early local cracking/small delaminations detected by GW as two discrete stages (DeepSAD). Global stiffness declines later when local damage coalesces. DTC-VAE's smooth HI captures overall monotonic accumulation.
Sample 2	Impact at off-stiffener region	3 damage states; earliest first transition among samples; long final state	Fast HI rise early (first ~30% lifetime), then slow growth	Smoothest stiffness decline; no early jump (stiffness accelerates near EoL)	Off-stiffener impact produces local scattering changes that GW detects very early (DeepSAD). Because damage is local and not initially in primary load paths, global stiffness remains unaffected until late accumulation—consistent with DTC-VAE plateau after early rise.
Sample 3	Impact near opposite stiffener edge	4 damage states; 3rd state longest	Gradual multi-stage trend, intermediate between extremes	Smooth reduction; late acceleration toward EoL	Multiple HI states suggest sequential damage mechanisms (matrix cracking→interface debonding→delamination) at/around the stiffener edge. Stiffness reduces progressively and accelerates as damage becomes global.
Sample 4	Impact on the stiffener, centre	2 damage states; fewer pre-final stages	Slow progression until ~70% lifetime, then jump(s) toward EoL	Early-to-mid life smoother drop then sharper loss	Impact on stiffener produces damage that remains subcritical for long time (little HI change) then propagates rapidly along the stiffener causing HI jumps and stiffness loss (loss of load-transfer at stiffener). Timing matches DTC-VAE late jumps.
Sample 5	Manufacturing disbond between stiffener and skin (right side)	2 damage states (sometimes 3 depending on fold); early transition visible	Smooth monotonic HI but early rise relative to some units	Earliest stiffness decrease (~20% lifetime); sudden early drop	Pre-existing disbond alters global load path→early measurable stiffness loss. GW HIs detect local scattering but effectiveness depends on which paths intersect the disbond; model variability reflects path sensitivity. Overall, early stiffness drop is consistent with a structural discontinuity (disbond) rather than purely fatigue accumulation.

by penalizing the model for producing HIs with lower fitness scores, ensuring that these follow a smooth, monotonic trend. This trend is easier to predict, making the extracted HIs more suitable for prognostics.

5.2. Uncertainty and dataset size

The suitability of incorporating uncertainty quantification into the evaluation of the constructed HIs was examined, given its common application for prognostic modelling. Although it is widely used in RUL prediction [70], its use in HI construction remains limited. Studies on HIs typically rely on the use of prognostic criteria and do not report confidence intervals [15,36]. This is partly because HIs represent abstract latent variables rather than directly measurable physical quantities. As a result, interval-based uncertainty estimates for fitness scores may be misleading, especially in relation to smaller datasets where statistical variance cannot be reliably defined. Nevertheless, as reported in this study, evaluating the standard deviation of Mo, Pr, Tr, and the fitness score—across (i) multiple random initialisations of the deep learning models, which reflects model uncertainty, and (ii) different composite specimens, which reflects variability due to GW sensing, manufacturing imperfections, and minor differences in sensor placement and test setup—is necessary. These statistics provide a quantitative assessment of the robustness of the learned HIs and their sensitivity to both model and experimental variability.

Although only five composite specimens were employed in this study, the dataset is highly information-rich: each unit provides dozens of GW measurements, each comprising 56 actuator–sensor paths across six frequencies. Since the proposed models operate in a history-independent manner, every GW measurement timestep forms an independent sample, yielding approximately 230 machine learning measurements in total. Furthermore, the leave-one-unit-out evaluation, where an entire specimen is unseen during training, provides a stringent generalisation test and helps mitigate overfitting despite the limited number of structural units.

5.3. Computational feasibility

Both Diversity-DeepSAD and DTC-VAE remain computationally lightweight. All experiments were conducted on a laptop equipped with a 12th-Gen Intel Core i7-12700H CPU (14 cores, 20 threads), 16 GB RAM and an NVIDIA RTX A1000 Laptop GPU with 4 GB VRAM. Training and inference were performed using GPU acceleration (CUDA 12.8). Training a single model instance, in other words one frequency and feature type for one validation fold, required on average 35.68 s for Diversity-DeepSAD and 2.27 s for DTC-VAE, with a maximum of 62.34 s and 3.01 s respectively. The corresponding parameter counts averaged 519,664 trainable parameters for Diversity-DeepSAD and 71,309 for DTC-VAE (1.982 MB and 0.272 MB respectively in memory), depending mainly on the input dimensionality and hidden layer width.

Inference for a new specimen, i.e. generating a complete HI trajectory for a single specimen at one frequency, took on average 17.47 ms per forward pass for Diversity DeepSAD and 1.03 ms for DTC-VAE. Since training and Bayesian optimisation are performed offline, only this inference step is relevant for online SHM. The sub-second inference times leave a large margin with respect to typical GW measurement intervals, indicating that both methods are suitable for real-time deployment, for example, for in-service aircraft structures in the industrial context. Moreover, the training time scales approximately linearly with the number of specimens and epochs, and model size scales linearly with the hidden layer width, suggesting that the framework can be extended to larger datasets and additional frequencies without prohibitive computational overhead.

5.4. Comparison with state-of-the-art

Based on the average prognostic criteria and fitness scores across all units, Diversity-DeepSAD and DTC-VAE demonstrate competitive performance compared to other proposed models, as shown in Table 12. In this table, two SHM techniques — GW and acoustic emission — are also compared. It should be noted that the constructed HIs using acoustic emission are inherently time-dependent due to the nature of the sensing technology, whereas the GW technique has the potential to provide data for designing history-independent HIs, as demonstrated by the proposed models in this study and HT-SSCNN [15].

While SSLSTM using acoustic emission data [23] achieves the highest fitness score across all units (F_{all}) with a performance of 93%, it relies on test units for validation during the LOOCV process to stop training and fine-tune hyperparameters. This reliance potentially limits its applicability to real-world scenarios where test units are new and unknown. However, this limitation has been addressed in subsequent work [24], where CEEMDAN-SSLSTM followed by a BiLSTM ensemble

Table 12
Performance of various methods in the ReMAP dataset collected using guided wave (GW) and acoustic emission techniques.

Prognostic criteria	Diversity-DeepSAD on FFT (WAE)	DTC-VAE on FFT (WAE)	HT-SSCNN [15] (WAE)	SSLSTM [23] (simple ensemble)	CEEMDAN-SSLSTM [24] (BiLSTM ensemble)
Mo	0.94 (± 0.01)	0.99 (± 0.01)	–	0.99 (± 0.01)	–
Pr	0.91 (± 0.03)	0.98 (± 0.01)	–	0.86 (± 0.14)	–
Tr	0.60 (± 0.08)	0.80 (± 0.05)	–	0.93 (± 0.03)	–
F_{all}	2.45 (± 0.06)	2.77 (± 0.05)	2.78 (± 0.15)	2.79 (± 0.14)	2.74 (± 0.19)
Mo_{test}	0.93 (± 0.04)	0.98 (± 0.02)	–	–	–
Pr_{test}	0.86 (± 0.10)	0.97 (± 0.02)	–	–	–
Tr	0.60 (± 0.08)	0.80 (± 0.05)	–	–	–
F_{test}	2.39 (± 0.12)	2.76 (± 0.06)	2.67 (± 0.20)	–	2.59 (± 0.24)
Number of units	*5 composite specimens*			*12 composite specimens*	
SHM technique	*GW*			*Acoustic Emission*	

achieved the best performance for acoustic emission data, with 91.3% for F_{all} and 86.3% for test units (F_{test}).

On the other hand, using GW data, the proposed history-independent HIs in this study result in the lowest deviation between fitness scores obtained for all units (F_{all}) and test units (F_{test}), demonstrating reduced overfitting. Both the semi-supervised Diversity-DeepSAD model and the unsupervised DTC-VAE model are significantly more stable, with much lower standard deviations compared to existing state-of-the-art models. The DTC-VAE model, followed by WAE fusion, achieved the highest performance of 92% for test units (F_{test}) and 92.3% for all units (F_{all}). Notably, the deviation between fitness scores for all units and test units is minimal (0.3%).

Although this study focuses on the model families previously explored for the ReMAP GW dataset, the results highlight opportunities for evaluating more advanced architectures. Models such as transformers, graph neural networks, and physics-informed neural networks may further exploit spatial–temporal dependencies in GW signals. A systematic assessment of these architectures is left for future work.

5.5. Limitations and future work

Although the proposed framework demonstrates strong performance and robustness, some limitations identified in the discussion are summarised explicitly:

(i) The dataset contains only five composite specimens. The data richness and leave-one-unit-out evaluation provide sufficient validation coverage, however, future work should aim to include more specimens to improve statistical generalisation.
(ii) Diversity-DeepSAD and DTC-VAE occasionally produce abrupt transitions at EoL, somewhat influencing the prognosability criterion and their use for prognostics. Further work should aim to incorporate further constraints into the learning process and loss functions.
(iii) The framework relies on multi-frequency guided wave measurements with dense actuator–sensor paths, which may not be available in all SHM deployments.
(iv) Only two model families were explored. More advanced architectures, such as transformers, graph neural networks, and physics-informed neural networks, may further improve degradation representation.

6. Conclusion

This study proposed a novel plenary data-driven framework for constructing health indicators (HIs) for aeronautical composite structures by developing two advanced machine learning models: Diversity-DeepSAD and DTC-VAE. First, statistical features from the best-performing signal processing techniques, fast Fourier transform (FFT) and Hilbert transform (HT), were extracted as inputs for the AI models. This step enabled dimensionality reduction of GW signals, each 2000 data points in length, recorded by PZTs across 56 actuator–sensor inspection paths at six excitation frequencies (2000 × 56 ×6 data points per measurement), to just a few tens of features per excitation frequency. This allows for simpler deep learning architectures while preserving critical information. Feature selection was based on fitness scores, calculated as the sum of the prognostic criteria of monotonicity, prognosability, and trendability. FFT features demonstrated superior performance, achieving a performance of 71.0%, i.e. an average fitness score of 2.13 (out of 3.00), compared to 68.2% for HT features.

Diversity-DeepSAD and DTC-VAE exhibited distinct behaviours in HI construction. Diversity-DeepSAD produced incremental changes in HIs, potentially reflecting distinct damage states, but occasionally showed abrupt increases, particularly near the end of life (EoL). In contrast, DTC-VAE generated smoother HIs with higher fitness scores, benefiting from its monotonicity-constrained loss function, which directly optimised the fitness metric. This smoother trend makes DTC-VAE more suitable for RUL prediction, as it avoids abrupt changes that could limit actionable maintenance time. Therefore, the augmented Diversity-DeepSAD is recommended as a closer representative of true health state, while DTC-VAE may be more useful for prognostic applications.

The proposed framework outperformed state-of-the-art models in constructing history-independent HIs. DTC-VAE, followed by weighted averaging ensemble fusion, achieved the highest test fitness score of 2.76 (92% performance) with minimal deviation (0.3%) from the fitness score across all units, indicating reduced overfitting and improved generalisation. Moreover, the models demonstrate greater stability in performance compared to existing approaches. Unlike some existing models, which rely on test units for validation, the proposed framework ensures applicability to real-world scenarios where test units are unknown.

Overall, this study addresses the critical challenge of constructing robust HIs for aeronautical composite structures, where reliable monitoring and prognostics are essential for operational safety and effective maintenance planning. By leveraging the strengths of augmented Diversity-DeepSAD and DTC-VAE, the proposed frameworks provide stable HIs while supporting more reliable RUL predictions. The frameworks are designed to be computationally efficient, scalable and feasible for monitoring, making them suitable for SHM applications such as in-service aircraft structures. Future work will focus on extending the DTC-VAE loss function to incorporate additional prognostic criteria, exploring larger datasets with more units to validate generalisability and assessing the implementation in operational environments to further enhance the frameworks practical applicability. Additional work will also explore the applicability of graph neural networks and physics-informed neural networks to GW SHM.

Table A.13
Time domain features [23].

No	Feature	Equation	No	Feature	Equation
1	Mean value	$X_m = \frac{\sum_{i=1}^{N} x(i)}{N}$	11	Shape factor	$X_{shape} = \frac{X_{rms}}{\frac{1}{N}\sum_{i=1}^{N}\lvert x(i)\rvert}$
2	Standard deviation	$X_{sd} = \sqrt{\frac{\sum_{i=1}^{N}(x(i)-X_m)^2}{N-1}}$	12	Impulse factor	$X_{impulse} = \frac{X_{peak}}{\frac{1}{N}\sum_{i=1}^{N}\lvert x(i)\rvert}$
3	Root amplitude	$X_{root} = \left(\frac{\sum_{i=1}^{N}\sqrt{\lvert x(i)\rvert}}{N}\right)^2$	13	Max–min difference	$X_{p2p} = max\,(x(i))$
4	Root mean square	$X_{rms} = \sqrt{\frac{\sum_{i=1}^{N}(x(i))^2}{N}}$	14–17	k^{th} central moment ($k = 3, 4, 5, 6$)	$X_{k\_cm} = \frac{\sum_{i=1}^{N}(x(i)-X_m)^k}{N}$
5	Residual sum of squares	$X_{rss} = \sqrt{\sum_{i=1}^{N}\lvert x(i)\rvert^2}$			
6	Peak maximum	$X_{peak} = max(\lvert x(i)\rvert)$	18	FM4	$X_{FM4} = \frac{X_{4\_cm}}{X_{sd}^4}$
7	Skewness	$X_{skewness} = \frac{\sum_{i=1}^{N}(x(i)-X_m)^3}{(N-1)X_{sd}^3}$	19	Median	$X_{med} = \frac{\sum_{i=1}^{N} t(i)}{N}$
8	Kurtosis	$X_{kurtosis} = \frac{\sum_{i=1}^{N}(x(i)-X_m)^4}{(N-1)X_{sd}^4}$			
9	Crest factor	$X_{crest} = \frac{X_{peak}}{X_{rms}}$			
10	Clearance factor	$X_{clearance} = \frac{X_{peak}}{X_{root}}$			

Table A.14
Frequency domain features S [23].

No	Feature	Equation	No	Feature	Equation
1	Mean frequency	$S_1 = X_{mf} = \frac{\sum_{f=1}^{N_f} s(f)}{N_f}$	8	–	$S_8 = \sqrt{\frac{\sum_{f=1}^{N_f} f_k^4 s(f)}{\sum_{f=1}^{N_f} f_k^2 s(f)}}$
2	*(same as variance)*	$S_2 = \frac{\sum_{f=1}^{N_f}(s(f)-S_1)^2}{N_f-1}$	9	–	$S_9 = \frac{\sum_{f=1}^{N_f} f_k^2 s(f)}{\sqrt{\sum_{f=1}^{N_f} s(f)\sum_{f=1}^{N_f} f_k^4}}$
3	*(same as skewness)*	$S_3 = \frac{\sum_{f=1}^{N_f}(s(f)-S_1)^3}{N_f\left(\sqrt{S_2}\right)^3}$	10	–	$S_{10} = \frac{S_6}{S_5}$
4	*(same as kurtosis)*	$S_4 = \frac{\sum_{f=1}^{N_f}(s(f)-S_1)^4}{N_f S_2^2}$	11	–	$S_{11} = \frac{\sum_{f=1}^{N_f}(f_k-S_5)^3 s(f)}{N_f S_6^3}$
5	–	$S_5 = X_{fc} = \frac{\sum_{f=1}^{N_f} f_k s(f)}{\sum_{f=1}^{N_f} s(f)}$	12	–	$S_{12} = \frac{\sum_{f=1}^{N_f}(f_k-S_5)^4 s(f)}{N_f S_6^4}$
6	–	$S_6 = \sqrt{\frac{\sum_{f=1}^{N_f}(f_k-S_5)^2 s(f)}{N_f}}$	13	–	$S_{13} = \frac{\sum_{f=1}^{N_f}\sqrt{(f_k-S_5)}s(f)}{N_f\sqrt{S_6}}$
7	–	$S_7 = X_{rmsf} = \sqrt{\frac{\sum_{f=1}^{N_f} f_k^2 s(f)}{\sum_{f=1}^{N_f} s(f)}}$	14	–	$S_{14} = \sqrt{\frac{\sum_{f=1}^{N_f}(f_k-S_5)^2 s(f)}{\sum_{f=1}^{N_f} s(f)}}$

Table A.15
Time-frequency domain features X [71].

No	Feature	Equation
1	Mean value	$X_m = \frac{\sum_{i=1}^{N} x(i)}{N}$
2	Standard deviation	$X_{sd} = \sqrt{\frac{\sum_{i=1}^{N}(x(i)-X_m)^2}{N-1}}$
3	Skewness	$X_{skewness} = \frac{\sum_{i=1}^{N}(x(i)-X_m)^3}{(N-1)X_{sd}^3}$
4	Kurtosis	$X_{kurtosis} = \frac{\sum_{i=1}^{N}(x(i)-X_m)^4}{(N-1)X_{sd}^4}$

CRediT authorship contribution statement

James Josep Perry: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Pablo Garcia-Conde Ortiz:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **George Konstantinou:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Cornelie Vergouwen:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Edlyn Santha Kumaran:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Morteza Moradi:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Formal analysis, Conceptualization.

Code availability

The source code used for all experiments in this study is available on GitHub at the following repository: https://github.com/mortezamkh1992/Diversity-DeepSAD-vs-DTC-VAE

Declaration of competing interest

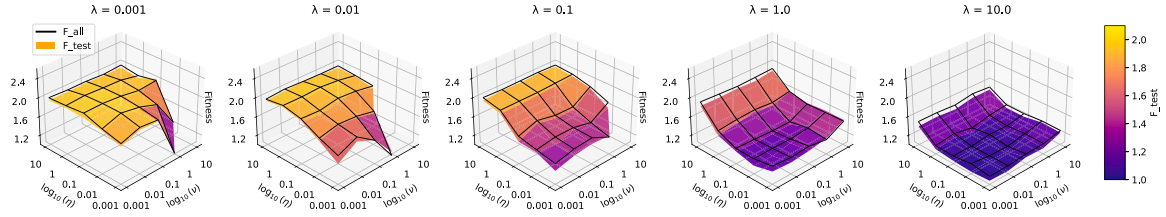
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. List of features by domain

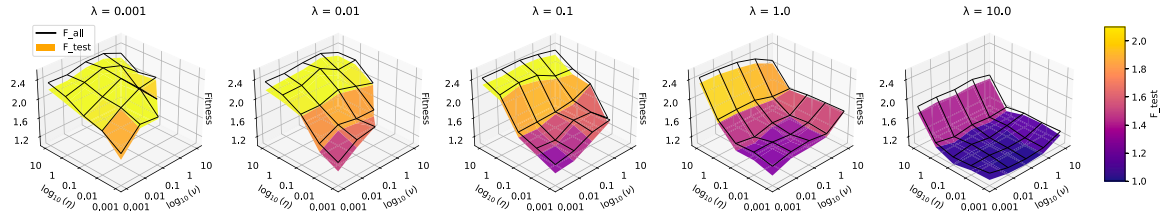
Features extracted from time, frequency, and time-frequency domains are listed in Tables Table A.13, Table A.14, and Table A.15, respectively. Each time or time-frequency domain feature X is given in terms of the time or time-frequency domain signal samples x, while frequency domain features S are given in terms of the frequency domain signal samples $s(f)$,

Appendix B. Sensitivity analysis

All plots for the sensitivity analysis described in Section 3.7 are included below in Fig. B.16 and Fig. B.17.

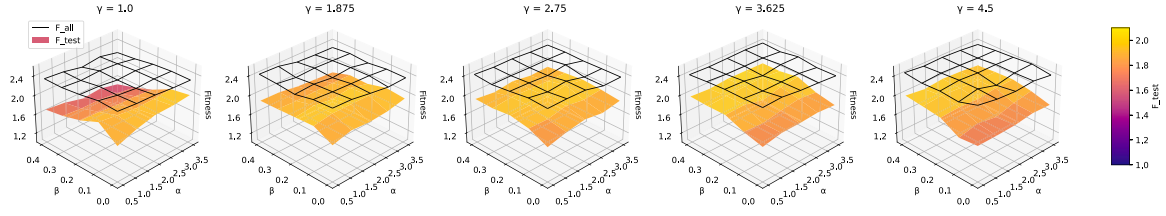


(a) Diversity-DeepSAD using FFT features.

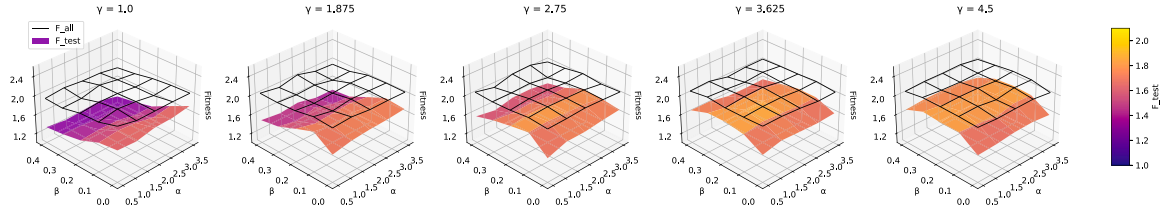


(b) Diversity-DeepSAD using HT features.

Fig. B.16. FUll hyperparameter sensitivity analysis of Diversity-DeepSAD over (ν, η, λ), showing response surfaces for the two SP methods.



(a) DTC-VAE using FFT features.



(b) DTC-VAE using HT features.

Fig. B.17. Full hyperparameter sensitivity analysis of DTC-VAE over (α, β, γ), showing response surfaces for the two SP methods.

References

[1] Jimenez JJM, Schwartz S, Vingerhoeds R, Grabot B, Salaün M. Towards multi-model approaches to predictive maintenance: A systematic literature survey on diagnostics and prognostics. J Manuf Syst 2020;56:539–57. http://dx.doi.org/10.1016/j.jmsy.2020.07.008.

[2] Sikorska JZ, Hodkiewicz M, Ma L. Prognostic modelling options for remaining useful life estimation by industry. Mech Syst Signal Process 2011;25(5):1803–36. http://dx.doi.org/10.1016/j.ymssp.2010.11.018.

[3] Badihi H, Zhang Y, Jiang B, Pillay P, Rakheja S. A comprehensive review on signal-based and model-based condition monitoring of wind turbines: Fault diagnosis and lifetime prognosis. Proc IEEE 2022;110(6):754–806. http://dx.doi.org/10.1109/JPROC.2021.3138128.

[4] Jia S, Ma B, Guo W, Li ZS. A sample entropy based prognostics method for lithium-ion batteries using relevance vector machine. J Manuf Syst 2021;61:773–81. http://dx.doi.org/10.1016/j.jmsy.2021.03.019.

[5] Guo J, Wang Z, Li H, Yang Y, Huang C-G, Yazdi M, Kang HS. A hybrid prognosis scheme for rolling bearings based on a novel health indicator and nonlinear Wiener process. Reliab Eng Syst Saf 2024;245:110014. http://dx.doi.org/10.1016/j.ress.2024.110014, URL https://www.sciencedirect.com/science/article/pii/S0951832024000899.

[6] Wu J, He D, Jin Z, Zhao M, Li X, Chen Y. Multi-view fully connected graph to fuse multi-sensor signals for mechanical equipment remaining useful life prediction. J Manuf Syst 2025;80:1029–52. http://dx.doi.org/10.1016/j.jmsy.2025.05.009.

[7] Qiu H, Lee J, Lin J, Yu G. Robust performance degradation assessment methods for enhanced rolling element bearing prognostics. Adv Eng Informatics 2003;17(3–4):127–40. http://dx.doi.org/10.1016/j.aei.2004.08.001.

[8] Guo L, Li N, Jia F, Lei Y, Lin J. A recurrent neural network based health indicator for remaining useful life prediction of bearings. Neurocomputing 2017;240:98–109. http://dx.doi.org/10.1016/j.neucom.2017.02.045.

[9] Khan A, Azad MM, Sohail M, Kim HS. A review of physics-based models in prognostics and health management of laminated composite structures. Int J Precis Eng Manufacturing-Green Technol 2023;10(6):1615–35. http://dx.doi.org/10.1007/s40684-023-00509-4.

[10] Wei J, Zhang F, Lu J. Health indicator construction based on double attribute feature deviation degree and its application into RUL prediction. Reliab Eng Syst Saf 2025;256:110785. http://dx.doi.org/10.1016/j.ress.2024.110785.

[11] Zhang P, Gao Z, Cao L, Dong F, Zou Y, Wang K, Zhang Y, Sun P. Marine systems and equipment prognostics and health management: a systematic review from health condition monitoring to maintenance strategy. Machines 2022;10(2):72. http://dx.doi.org/10.3390/machines10020072.

[12] Huang C, Bu S, Lee HH, Chan CH, Kong SW, Yung WK. Prognostics and health management for predictive maintenance: A review. J Manuf Syst 2024;75:78–101. http://dx.doi.org/10.1016/j.jmsy.2024.05.021.

[13] Moradi M. Designing health indicators for aerospace structures by intelligent information fusion [Dissertation (TU Delft)], Delft University of Technology; 2024, http://dx.doi.org/10.4233/uuid:7ac03701-b97a-427d-990c-e6c696d1254b.

[14] Beaumont PWR. The structural integrity of composite materials and long-life implementation of composite structures. Appl Compos Mater 2020;27(5):449–78. http://dx.doi.org/10.1007/s10443-020-09822-6.

[15] Moradi M, Gul FC, Zarouchas D. A novel machine learning model to design historical-independent health indicators for composite structures. Compos Part B: Eng 2024;275. http://dx.doi.org/10.1016/j.compositesb.2024.111328.
[16] Ferreira C, Gonçalves G. Remaining useful life prediction and challenges: A literature review on the use of machine learning methods. J Manuf Syst 2022;63:550–62. http://dx.doi.org/10.1016/j.jmsy.2022.05.010.
[17] Huang X, Peng Z, Tang D, Chen J, Zio E, Zheng Z. A physics-informed autoencoder for system health state assessment based on energy-oriented system performance. Reliab Eng Syst Saf 2024;242:109790. http://dx.doi.org/10.1016/j.ress.2023.109790.
[18] Coble J, Hines JW. Identifying optimal prognostic parameters from data: a genetic algorithms approach. In: Annual conference of the PHM society. vol. 1, (1). 2009.
[19] Coble JB. Merging data sources to predict remaining useful life–an automated method to identify prognostic parameters. 2010.
[20] Saxena A, Celaya J, Balaban E, Goebel K, Saha B, Saha S, Schwabacher M. Metrics for evaluating performance of prognostic techniques. In: 2008 international conference on prognostics and health management. IEEE; 2008, p. 1–17. http://dx.doi.org/10.1109/PHM.2008.4711436.
[21] González-Muñiz A, Díaz I, Cuadrado AA, García-Pérez D. Health indicator for machine condition monitoring built in the latent space of a deep autoencoder. Reliab Eng Syst Saf 2022;224:108482. http://dx.doi.org/10.1016/j.ress.2022.108482, URL https://www.sciencedirect.com/science/article/pii/S0951832022001417.
[22] Li X, Teng W, Peng D, Ma T, Wu X, Liu Y. Feature fusion model based health indicator construction and self-constraint state-space estimator for remaining useful life prediction of bearings in wind turbines. Reliab Eng Syst Saf 2023;233:109124. http://dx.doi.org/10.1016/j.ress.2023.109124.
[23] Moradi M, Broer A, Chiachío J, Benedictus R, Loutas TH, Zarouchas D. Intelligent health indicator construction for prognostics of composite structures utilizing a semi-supervised deep neural network and SHM data. Eng Appl Artif Intell 2023;117. http://dx.doi.org/10.1016/j.engappai.2022.105502.
[24] Moradi M, Galanopoulos G, Kuiters T, Zarouchas D. A novel intelligent health indicator using acoustic waves: CEEMDAN-driven semi-supervised ensemble deep learning. Mech Syst Signal Process 2025;224:112156. http://dx.doi.org/10.1016/j.ymssp.2024.112156.
[25] Senthilkumar M, Sreekanth T, Manikanta Reddy S. Nondestructive health monitoring techniques for composite materials: A review. Polym Polym Compos 2021;29(5):528–40. http://dx.doi.org/10.1177/0967391120921701.
[26] Hassani S, Mousavi M, Gandomi AH. Structural health monitoring in composite structures: A comprehensive review. Sensors 2021;22(1):153. http://dx.doi.org/10.3390/s22010153.
[27] Luca AD, Caputo F, Khodaei ZS, Aliabadi MH. Damage characterization of composite plates under low velocity impact using ultrasonic guided waves. Compos Part B: Eng 2018;138:168–80. http://dx.doi.org/10.1016/j.compositesb.2017.11.042.
[28] Saeedifar M, Zarouchas D. Damage characterization of laminated composites using acoustic emission: A review 8. 2020, http://dx.doi.org/10.1016/j.compositesb.2020.108039.
[29] Dienel CP, Meyer H, Werwer M, Willberg C. Estimation of airframe weight reduction by integration of piezoelectric and guided wave–based structural health monitoring. Struct Health Monit 2019;18:1778–88. http://dx.doi.org/10.1177/1475921718813279.
[30] Aujoux C, Mesnil O. Environmental impact assessment of guided wave-based structural health monitoring. Struct Health Monit 2023;22:913–26. http://dx.doi.org/10.1177/14759217221088774.
[31] Flynn EB, Todd MD, Wilcox PD, Drinkwater BW, Croxford AJ. Maximum-likelihood estimation of damage location in guided-wave structural health monitoring. In: Proceedings of the royal society a: mathematical, physical and engineering sciences. vol. 467, Royal Society; 2011, p. 2575–96. http://dx.doi.org/10.1098/rspa.2011.0095.
[32] Memmolo V, Monaco E, Boffa ND, Maio L, Ricci F. Guided wave propagation and scattering for structural health monitoring of stiffened composites. Compos Struct 2018;184:568–80. http://dx.doi.org/10.1016/j.compstruct.2017.09.067.
[33] Kralovec C, Schagerl M. Review of structural health monitoring methods regarding a multi-sensor approach for damage assessment of metal and composite structures 2. 2020, http://dx.doi.org/10.3390/s20030826.
[34] Janardhan PM, Balasubramaniam K. Lamb-wave-based structural health monitoring technique for inaccessible regions in complex composite structures. Struct Control Health Monit 2014;21:817–32. http://dx.doi.org/10.1002/stc.1603.
[35] Yang J, Xie G, Yang Y. An improved ensemble fusion autoencoder model for fault diagnosis from imbalanced and incomplete data. Control Eng Pract 2020;98. http://dx.doi.org/10.1016/j.conengprac.2020.104358.
[36] Yang K, Kim S, Harley JB. Guidelines for effective unsupervised guided wave compression and denoising in long-term guided wave structural health monitoring. Struct Health Monit 2023;22:2516–30. http://dx.doi.org/10.1177/14759217221124689.
[37] Moradi M, Chiachío J, Zarouchas D. Developing health indicators for composite structures based on a two-stage semi-supervised machine learning model using acoustic emission data. In: 10th ECCOMAS thematic conference on smart structures and materials. ECCOMAS; 2023, p. 923–34. http://dx.doi.org/10.7712/150123.9844.451295.
[38] Li H, Zhang Z, Li T, Si X. A review on physics-informed data-driven remaining useful life prediction: Challenges and opportunities. Mech Syst Signal Process 2024;209. http://dx.doi.org/10.1016/j.ymssp.2024.111120.
[39] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97. http://dx.doi.org/10.1007/BF00994018.
[40] Tax DMJ, Duin RPW. Support vector data description. Mach Learn 2004;54:45–66. http://dx.doi.org/10.1023/B:MACH.0000008084.60811.49.
[41] Kim S, Choi Y, Lee M. Deep learning with support vector data description. Neurocomputing 2015;165:111–7. http://dx.doi.org/10.1016/j.neucom.2014.09.086.
[42] Ruff L, Vandermeulen RA, Görnitz N, Binder A, Müller E, Müller K-R, Kloft M. Deep semi-supervised anomaly detection. In: International conference on learning representations. 2019, http://dx.doi.org/10.48550/arXiv.1906.02694.
[43] Han L, Zhang Y, He Y, You K, Liu W, Xie H. Rapid identification method for on-road high-emission vehicle based on deep semi-supervised anomaly detection. Meas: J Int Meas Confed 2025;239. http://dx.doi.org/10.1016/j.measurement.2024.115430.
[44] Dong Y, Chen K, Ma Z. Comparative study on semi-supervised learning applied for anomaly detection in hydraulic condition monitoring system. In: IEEE International Conference on Systems, Man and Cybernetics. 2023, http://dx.doi.org/10.1109/SMC53992.2023.10394193.
[45] Gao F, Li J, Cheng R, Zhou Y, Ye Y. Connet: Deep semi-supervised anomaly detection based on sparse positive samples. IEEE Access 2021;9:67249–58. http://dx.doi.org/10.1109/ACCESS.2021.3077014.
[46] Frusque G, Nejjar I, Nabavi M, Fink O. Semisupervised health index monitoring with feature generation and fusion. IEEE Trans Reliab 2024. http://dx.doi.org/10.1109/TR.2024.3496076.
[47] Yang Z, Baraldi P, Zio E. Automatic extraction of a health indicator from vibrational data by sparse autoencoders. In: 3rd International Conference on System Reliability and Science. 2018, p. 328–32. http://dx.doi.org/10.1109/ICSRS.2018.8688720.
[48] Lin P, Tao J. A novel bearing health indicator construction method based on ensemble stacked autoencoder. In: IEEE international conference on prognostics and health management. 2019, http://dx.doi.org/10.1109/ICPHM.2019.8819405.
[49] Xu F, Wang L. Constructing a health indicator for bearing degradation assessment via an unsupervised and enhanced stacked autoencoder. Adv Eng Informatics 2022;53. http://dx.doi.org/10.1016/j.aei.2022.101708.
[50] Xu Z, Bashir M, Liu Q, Miao Z, Wang X, Ekere N. A novel health indicator for intelligent prediction of rolling bearing remaining useful life based on unsupervised learning model. Comput Ind Eng 2023;176. http://dx.doi.org/10.1016/j.cie.2023.108999.
[51] Mao W, Wang Y, Kou L, Liang X. A new deep tensor autoencoder network for unsupervised health indicator construction and degradation state evaluation of metro wheel. IEEE Trans Instrum Meas 2023;72. http://dx.doi.org/10.1109/TIM.2023.3251399.
[52] Ping G, Chen J, Pan T, Pan J. Degradation feature extraction using multi-source monitoring data via logarithmic normal distribution based variational autoencoder. Comput Ind 2019;109:72–82. http://dx.doi.org/10.1016/j.compind.2019.04.013.
[53] Moradi M, Komninos P, Zarouchas D. Constructing explainable health indicators for aircraft engines by developing an interpretable neural network with discretized weights. Appl Intell 2025;55. http://dx.doi.org/10.1007/s10489-024-05981-2.
[54] Hemmer M, Klausen A, van Khang H, Robbersmyr KG, Waag TI. Health indicator for low-speed axial bearings using variational autoencoders. IEEE Access 2020;8:35842–52. http://dx.doi.org/10.1109/ACCESS.2020.2974942.
[55] Qin Y, Zhou J, Chen D. Unsupervised health indicator construction by a novel degradation-trend-constrained variational autoencoder and its applications. IEEE/ASME Trans Mechatronics 2022;27:1447–56. http://dx.doi.org/10.1109/TMECH.2021.3098737.
[56] Guo L, Yu Y, Duan A, Gao H, Zhang J. An unsupervised feature learning based health indicator construction method for performance assessment of machines. Mech Syst Signal Process 2022;167. http://dx.doi.org/10.1016/j.ymssp.2021.108573.
[57] Li X, Cheng C, Peng Z. Unsupervised construction of health indicator for rotating machinery via multi-criterion feature selection and attentive variational autoencoder. Sci China Technol Sci 2024;67(5):1524–37. http://dx.doi.org/10.1007/s11431-023-2610-4.
[58] Li S, Zhang C, Zhang X. A novel spatiotemporal enhanced convolutional autoencoder network for unsupervised health indicator construction. IEEE Trans Instrum Meas 2024;73:1–10. http://dx.doi.org/10.1109/TIM.2024.3383052.
[59] Ali Eftekhari Milani SJW. A hybrid convolutional autoencoder training algorithm for unsupervised bearing health indicator construction. Eng Appl Artif Intell 2025;139. http://dx.doi.org/10.1016/j.engappai.2024.109477.
[60] Zarouchas D, Broer A, Galanopoulos G, Briand W, Benedictus R, Loutas T. Compression fatigue tests on single stiffener aerospace structures. 2021, http://dx.doi.org/10.34894/QNURER.
[61] Baptista ML, Goebel K, Henriques EMP. Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. Artificial Intelligence 2022;306. http://dx.doi.org/10.1016/j.artint.2022.103667.

[62] Boudraa AO, Cexus JC, Benramdane S, Beghdadi A. Noise filtering using empirical mode decomposition. In: 2007 9th international symposium on signal processing and its applications. 2007, p. 1–4. http://dx.doi.org/10.1109/ISSPA.2007.4555624.

[63] Feldman M. Hilbert transform in vibration analysis. 2011, http://dx.doi.org/10.1016/j.ymssp.2010.07.018.

[64] Baraldi P, Bonfanti G, Zio E. Differential evolution-based multi-objective optimization for the definition of a health indicator for fault diagnostics and prognostics. Mech Syst Signal Process 2018;102:382–400. http://dx.doi.org/10.1016/j.ymssp.2017.09.013.

[65] Gauss CF. In: Perthes F, Besser IH, editors. Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Hamburg; 1809.

[66] Kingma DP, Welling M. Auto-encoding variational Bayes. In: Proceedings of the 2nd international conference on learning representations. ICLR, 2014.

[67] Lucas J, Tucker G, Grosse RB, Norouzi M. Understanding posterior collapse in generative latent variable models. In: Proceedings of the 36th international conference on machine learning. ICML, Proceedings of machine learning research, vol. 97, PMLR; 2019, p. 5502–11.

[68] Bengio Y, Glorot X. Understanding the difficulty of training deep feed forward neural networks. Int Conf Artif Intell Stat 2010;249–56.

[69] Yue N, Broer A, Briand W, Rébillat M, Loutas T, Zarouchas D. Assessing stiffness degradation of stiffened composite panels in post-buckling compression-compression fatigue using guided waves. Compos Struct 2022;293(115751). http://dx.doi.org/10.1016/j.compstruct.2022.115751.

[70] Wang R, Zhang Y, Hu C, Yang Z, Li H, Liu F, Bonfanti G, Zio E. A parallel prognostic method integrating uncertainty quantification for probabilistic remaining useful life prediction of aero-engine. Processes 2024;12(12):2925. http://dx.doi.org/10.3390/pr12122925.

[71] Buckley T, Ghosh B, Pakrashi V. A feature extraction & selection benchmark for structural health monitoring. Struct Health Monit 2023;22(3):2082–127. http://dx.doi.org/10.1177/14759217221111141.