Unsupervised Physics-Informed Health Indicator Discovery for Complex Systems

Kristupas Bajarunas[1,2], Marcia Baptista[1], Kai Goebel[3], and Manuel Arias Chao[1,2]

[1] *Air Transport and Operations, Faculty of Aerospace Engineering, Delft University of Technology, Delft, Netherlands*
{k.v.b.bajarunas, m.lbaptista, m.a.c.ariaschao}@tudelft.nl

[2] *Zurich University of Applied Sciences, Zurich, Switzerland*
{baja, aria}@zhaw.ch

[3] *Palo Alto Research Center, Palo Alto, California, USA*
kgoebel@parc.com

Abstract

Discovering health indicators (HI) is essential for prognostics and health management of complex systems, as an HI enables timely interventions and effective maintenance strategies. However, most of the existing methodologies for HI discovery rely on labeled data which is expensive and complicated to obtain in the real world. In this paper, we propose a novel, unsupervised physics-informed model structured after expert knowledge in the form of a graphical representation of the expected relationships between sensor readings, operating conditions, and degradation. In addition, a soft constraint is used to guide the representation of the HI according to generally available expert knowledge about degradation. We evaluated the model on a turbofan engine dataset and conducted four experiments by manipulating the original data to create realistic real-world scenarios. The proposed method discovers an HI that exhibits better intrinsic qualities than the current state-of-the-art methodologies, leading to enhanced prognostic performance. Notably, in situations where the initial health state of each system varies, the proposed method achieves an average prognostic performance improvement of approximately 20% compared to existing state-of-the-art methods.

1. Introduction

The ability to predict when a system will fail can provide significant benefits such as reducing maintenance costs, preventing unexpected downtime, and increasing safety. One essential tool in the realm of condition-based maintenance is the health indicator (HI), which offers an interpretable means to monitor a system's health over time. HIs find utility in various aspects, including fault diagnosis, anomaly detection, and prognostics. For example, in prognostics, HIs prove crucial in predicting the remaining useful life (RUL) by contrasting degradation patterns among different units (Wang, Yu, Siegel, & Lee, 2008). Moreover, alternative prognostic strategies seek to establish predictive mappings between HI and RUL (Yang et al., 2016).

However, HIs are rarely directly observed, which makes their discovery complex. Many existing methodologies hinge on labeled data for HI discovery. For instance, when dealing with data that contains labeled failures, insights into the degradation process can be extracted from the information about RUL, aiding in HI discovery (Guo, Lei, Li, Yan, & Li, 2018; Fu, Zhong, Lin, & Zhao, 2021; Cofre-Martel, Lopez Droguett, & Modarres, 2021). Another approach involves utilizing the residual technique, where a model learns the system's normal behavior and discovers the HI by calculating reconstruction errors. Studies often leverage health state labels to select healthy training data and then employ residuals to discover the HI (Ye & Yu, 2021).

Nevertheless, obtaining labeled data to train supervised learning algorithms for HI discovery is often expensive or impossible. Therefore, there is a growing interest in developing unsupervised learning methods. To address the difficulty of dealing with unlabeled data, a solution strategy within the PHM research community is leveraging additional knowledge about the degradation behavior of the system to construct the HI. For instance, one study used prior knowledge that degradation can be expressed as an exponential function and estimated the parameters of the function by solving an optimization task (Liu & Huang, 2014). However, relying on such specific knowledge can pose challenges when attempting to uncover unit-specific degradation patterns. By imposing predefined assumptions about the degradation patterns, the model may fail to capture the unique characteristics and complexities of individual units. Furthermore, such specific knowledge might not be applicable to different systems where the degradation patterns differ.

We attempt to answer the question of how to effectively discover the HI of a system without relying on labeled data. Our hypothesis is that an unsupervised learning method based on an Autoencoder (AE), coupled with expert knowledge, can discover HI patterns from condition monitoring (CM) data. Unlike current methods that rely on system-specific degradation knowledge, our approach aims to use general knowledge.

To achieve this, we first introduce a new graphical representation that illustrates the relationship between sensor readings, operating conditions, and degradation in a typical system. We demonstrate how this representation can inform the design of an AE's architecture for the purpose HI discovery. Finally, we incorporate an extra constraint based on expert knowledge of the degradation process to guide the AE in generating results that are consistent with the knowledge.

Our approach's effectiveness is demonstrated by evaluating it under realistic data scenarios commonly seen in the industry. We examine uncertainties in the amount of healthy data during training, significant variations in the initial health state of each unit, disparities in the distribution of operating conditions between training and test datasets leading to out-of-distribution scenarios, and cases where the majority of the data is healthy. Our results show that the proposed method outperforms the residual approach in most situations, especially when the initial health state of each unit differs or when performing out-of-distribution tests. These findings highlight the effectiveness of integrating expert knowledge into a learning algorithm, which can lead to more accurate and robust health indicators for prognostic models.

The paper is organized as follows: Section 2 presents the problem of HI discovery, followed by Section 3, which introduces the background knowledge. Section 4 proposes the unsupervised physics-informed model, while Section 5 presents the case study. In Section 6, the results are presented, and Section 7 provides the conclusion and future work. Section 8 discusses the current limitations of the proposed approach.

2. UNSUPERVISED HEALTH CONDITION DISCOVERY OF A FLEET OF TURBOFAN ENGINES

We consider the challenging problem of discovering the health condition of a fleet of turbofan engines from CM data. Our focus lies on the challenging scenario where the engines operate under a wide range of flight conditions, and the direct observation of component degradation is not possible. However, the effects of degradation manifest in the sensor readings distributed throughout the engine. Specifically, we emphasize failure modes driven by cycle loading resulting from flight cycles of varying duration, including take-off, cruise, and descent operations.

The degradation trajectories of individual engines exhibit inherent differences attributed to various factors. Firstly, variations in the manufacturing process lead to different initial health states for each engine. Secondly, the diverse operating conditions experienced by each engine contribute to varying levels of degradation. Lastly, maintenance activities carried out at random intervals can improve the engine's health state.

In this context, the goal is to discover the degradation patterns for each engine given CM data. In particular, given the importance of obtaining an interpretable degradation measure for maintenance decision-making, we aim to discover an HI. In fact, HI is a widely used metric in the literature, providing a standardized and comprehensible representation of degradation through a single numerical value (Zhou et al., 2022).

2.1. Problem formulation

Formally, we are given multi-variate time-series of CM sensor readings $X_u = [x_u^1, ..., x_u^m]$ of a fleet of N units $(u = 1, ..., N)$ each with m observations. Each observation $x_u^i \in \mathbb{R}^p$ is a vector of p raw measurements. We are also given the history of operating conditions $W_u = [w_u^1, ..., w_u^m]$ for each unit, where each $w_u^i \in \mathbb{R}^s$. The goal is to discover the unobservable state of degradation Z of each unit at each point in time z_u^i. For interpretation purposes, the history of degradation will be transformed into a sequence of health indexes h, such that $\{h_u^i \in \mathbb{R}^1 | 0 \leq h_u^i \leq 1\}$.

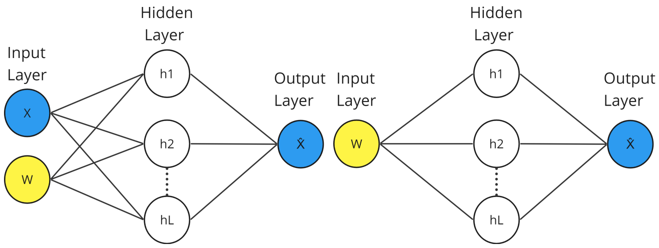
3. BACKGROUND

Under specific circumstances, it is feasible to acquire system health labels. For instance, healthy system condition labels can be obtained through inspections conducted by maintenance engineers who carefully assess and verify the system's health. In such a scenario, the residual approach (i.e., modeling of the healthy system) can be employed to discover hidden anomalies in the system's health.

However, obtaining health labels for HI discovery is often impractical, as it can be complex and expensive. Consequently, in practice, unsupervised learning methods are usually needed. In the following section, both approaches for HI discovery, i.e., residual and unsupervised, are further introduced.

3.1. Residual approach

The residual approach is a semi-supervised method for HI discovery that has been widely used in previous literature (Zhai, Gehring, & Reinhart, 2021; Lee, Lim, & Chattopadhyay, 2021; Koutroulis, Mutlu, & Kern, 2022; Zgraggen,



(a) AE (b) Regression

Figure 1. Residual model options

Pizza, & Huber, 2022). During training, a learning algorithm is trained to reconstruct healthy data. At prediction time, the trained model is evaluated by measuring the reconstruction error for new data inputs. Under the hypothesis that the training dataset is fully representative of a healthy system, small reconstruction errors are typically indicative of healthy inputs, while large reconstruction errors are typically indicative of faulty operation that was not observed during training.

The residual approach typically employs the AE as the preferred learning algorithm. However, it is crucial to account for changing operating conditions in order for the model to distinguish the effects of degradation and operating conditions. Figure 1 (a) illustrates the residual approach utilizing the AE. The model reconstructs healthy sensor readings while utilizing operating conditions as additional input (de Pater & Mitici, 2023).

An alternative approach to using an AE is to directly predict the sensor readings based on the operating conditions, without an intermediate representation (Lövberg, 2021). The approach is illustrated in Figure 1(b). This approach may be preferable in situations where data collection is limited, but may not perform as well in cases where the sensor readings contain outliers (Chalapathy & Chawla, 2019). Ultimately, the choice of approach should depend on the specific characteristics of the data and the problem being addressed.

In this study, the assumption is made that the health state of the system is uncertain, and no health state labels are available for HI discovery. To address this challenge, the residual model employed in this study relies on a specific assumption that a certain number of the initial flight cycles of each engine are healthy. This assumption is valid for brand-new engines, but might not be good in a different situation. In our experiments, we will demonstrate that the selection of the number of healthy cycles has a significant impact on the quality of HI produced by the residual approach.

3.2. Unsupervised approaches

In the absence of health state labels, unsupervised learning methods become crucial as an alternative to semi-supervised approaches. Notably, in systems where variations in sensor readings are predominantly driven by degradation, employing an AE with reduced dimensions in the latent space has demonstrated the ability to identify meaningful degradation patterns. For instance, this was demonstrated in an experiment using a subset of turbofan datasets where the operating conditions are kept constant (de Beaulieu, Jha, Garnier, & Cerbah, 2022). However, it is important to recognize that in systems where the impact of degradation is masked by varying operating conditions, fully unsupervised HI discovery methods are challenging. Therefore, alternative approaches have been proposed that emphasize the integration of additional knowledge with unsupervised methods.

For instance, in (Magadán et al., 2023), an AE was trained using features extracted by considering interesting frequencies given by prior expert knowledge. Through this process, the AE was able to uncover health indicators from the low-dimensional latent space.

An alternative approach (Qin et al., 2023) leverages knowledge about the shape of degradation and imposes constraints on the functional form of degradation. By incorporating this constraint, an HI can be effectively extracted from the latent space representation of an AE.

These methods showcase the utilization of domain-specific knowledge to guide the unsupervised learning process and enhance the accuracy of the extracted health indicators.

Despite these advancements, there are still challenges with the current unsupervised approaches which use additional expert knowledge. Namely, relying on specific knowledge can hinder the discovery of unit-specific degradation patterns, limiting the model's ability to capture unique characteristics. Additionally, such knowledge may not apply to diverse systems with different degradation patterns.

4. METHODOLOGY

To discover an HI in situations where no information about the health state is available, we propose an unsupervised physics-informed model. In alignment with (Karniadakis et al., 2021), the term "physics-informed" encompasses additional knowledge that becomes integrated within a machine learning model. In our proposed methodology, we utilize expert knowledge about the degradation process of complex systems. This expert knowledge, inherently informal and non-specific to a particular system, serves as the basis of our approach. To be more precise, we use two pieces of expert knowledge: we impose an inductive bias on a model's architecture and a learning bias on the objective function.

The introduction of an inductive bias stems from a novel graphical representation of the expected relationship between sensor readings, operating conditions, and degradation in a typical complex system. The graphical representation encapsulates the expected dependencies between observable and hidden factors involved in the degradation of turbofan engines; thus corresponding to a causal graph (Pearl et al., 2000). We embed the causal graph within the AE network's architecture as an inductive bias.

Furthermore, we extend the causal graph by introducing additional expected dependencies related to the system's failure modes of interest. Specifically, we illustrate how knowledge about failure mechanisms primarily influenced by operational cycle loading can be integrated into the model's objective function. This knowledge serves as a learning bias modifying the training objective function.

4.1. Causal graph of a degrading system

We present a graphical representation of the health discovery problem for degrading turbofan engines. From the problem description, it follows that the sensor measurements X are influenced by the operating conditions W and the degradation of the system Z.

Therefore, the different types of variables of the system can be graphically represented in Figure 2. Links between observable variables are shown in a solid line, while links between hidden variables are shown in dashed lines. Although the graphical representation was developed for the turbofan dataset, the representation is generic enough to cover multiple complex systems. For example, similar variable interactions are expected to be found when modeling batteries, since the evolution of voltage over time given the current demands are changing based on the degradation of the battery.

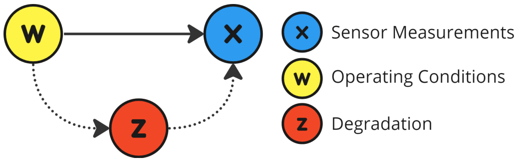


Figure 2. Graphical representation of variables in a turbofan engine

The graphical representation is explained by the expected behavior of the turbofan system. The association $W \rightarrow X$ can be justified by observing that sensor readings for a unit with a certain level of degradation vary significantly under different operational conditions. For instance, an engine during take-off experiences higher power input than during cruising, which impacts the sensor readings. The association between $W \rightarrow Z$ is included to reflect the way engines operate influences the level of degradation. For example, short flights, dominated by take-off and landing operations can lead to accelerated degradation. Finally, the association between $Z \rightarrow X$ is implicit in the definition of degradation. When comparing two engines operating under the same conditions but at different degradation states, a significant difference in sensor readings is observed. Engine performance worsens with degradation, leading to observable differences in sensor readings.

When certain observations are assumed to be healthy and not subject to degradation, the system can be represented graphically with only the association of $W \rightarrow X$. If a model is trained solely with healthy data, then the difference between the model prediction and the actual result will be exactly the degradation of the system, as it is the only unaccounted variable. Thus, the graphical representation also fits in the context of the residual approach.

4.2. Inductive bias: Derived model architecture

To detect degradation and discover the HI, a data-driven model can be created by incorporating the graphical knowledge presented in Figure 2. One approach to incorporate dependencies between different variables is to use an AE, as depicted in Figure 3. The model is trained to not only reconstruct sensor readings X in a supervised manner but, also to predict W. This step involves partitioning the latent space into two sections that represent W and Z, respectively, implicating the associations $W \rightarrow X$ and $Z \rightarrow X$. The effect of degradation on sensor readings is essentially decoupled from the effect of operating conditions. The overall training objective is given by:

$$L_{MSE} = ||X - \hat{X}||_2 + ||W - \hat{W}||_2 \tag{1}$$

where $||\cdot||_2$ represents the mean absolute error.

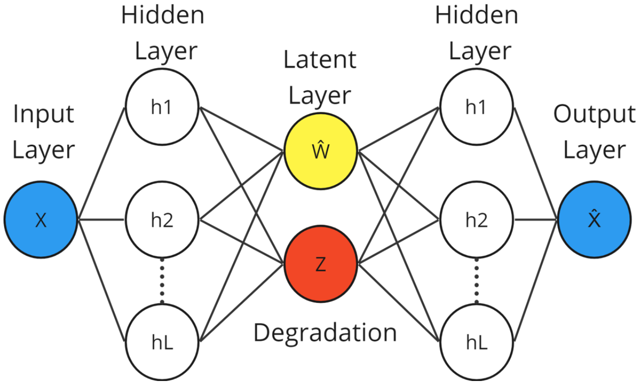


Figure 3. AE structure derived from the graphical representation

For the majority of systems, the effect of the association $W \rightarrow X$ is much stronger than $Z \rightarrow X$. As a result, given a sufficiently strong decoder, the sensor readings can be reconstructed regardless of the value of the unsupervised portion of the latent space. Therefore, without explicit guidance of the latent space Z, there is no guarantee that the latent space will show a clear degradation pattern.

4.3. Learning Bias: Modified objective function

In this paper, we propose to use additional knowledge about degradation to enhance the graphical representation shown in Figure 2. We argue that important degradation mechanisms in turbofan engines such as friction, erosion, and fouling of the rotating components are dominated by cycle operation. As such, the change in degradation within an operational cycle t is minor.

To capture the influence of the operation cycle time on the degradation process, we propose that the association $t \rightarrow Z$ must be included in the graphical representation, as shown in Figure 4.

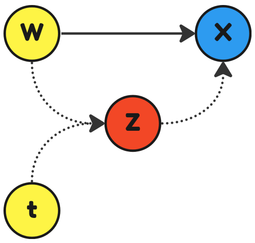


Figure 4. Modified graphical representation with additional knowledge

In order to incorporate information about the operational cycle time in the model, we implement a soft constraint on the latent space of degradation. Specifically, this soft constraint aims to minimize the correlation between operational cycle time and the latent space Z, and is defined as follows:

$$L_{corr} = \frac{\sum(t_i - \bar{t})(Z_i - \overline{Z})}{\sqrt{\sum(t_i - \bar{t})^2 \sum(Z_i - \overline{Z})^2}} \tag{2}$$

Our proposed approach is trained with the following objective function:

$$L = L_{MSE} + \lambda L_{corr} \tag{3}$$

By incorporating the soft constraint, we introduce additional expert knowledge about degradation to the model. The soft constraint serves as a guide to the latent space, allowing it to uncover degradation without becoming overly restrictive and compromising the model's ability to function as an AE. In essence, the constraint assists in shaping the latent space into a more meaningful representation that can better capture the hidden degradation of the system. The parameter λ controls the importance of the constraint and is used to reduce the risk of overfitting.

5. CASE STUDY

5.1. Dataset: A small fleet of turbofan engines

We demonstrate the proposed method for unsupervised HI discovery on the new Commercial Modular Aero-Propulsion System Simulation (N-CMAPSS) dataset (Arias Chao, Kulkarni, Goebel, & Fink, 2021). The N-CMAPSS dataset was created using a high-fidelity simulation model, which was given real flight conditions as recorded on board a commercial jet.

From the eight available data subsets, we consider the set DS003 which is characterized by a single fault mode that affects the low-pressure turbine efficiency and flows in combination with the high-pressure turbine efficiency degradation. The training set contains 9 units, and the test set contains 6 units. Each unit in the training set and testing set contains 14 observable sensor measurements, denoted as X, which are recorded from the time of engine installation until engine failure (run-to-failure data). In addition to the sensor measurements, four operating conditions W are available. The operating conditions include altitude, Mach number, throttle-resolved angle, and total temperature at the fan inlet. The units are divided into three flight classes depending on whether the unit is operating short-length flights (i.e., flight class 1), medium-length flights (i.e., flight class 2), or long-length flights (i.e., flight class 3). The sensor signals and operating conditions are sampled once per second (1Hz).

The N-CMAPSS dataset models degradation at the component level through initial, normal, and abnormal degradation stages. The ground truth HI was derived from a non-linear mapping of multiple degradation variables and was used to declare system failure when its value reached 0. More details about degradation modeling in the N-CMAPSS dataset can be found in the dataset description paper (Arias Chao et al., 2021). The ground truth HI (h_{gt}) will be used for evaluation purposes only. An overview of the used dataset is provided in Table 1.

Table 1. Summary of DS03 N-CMAPSS dataset. The table provides information on the average number of healthy cycles and end-of-life cycles for each flight class.

Flight Class	Average of Healthy Cycles	Average of Total Cycles
Short	36.3	82.6
Medium	23.5	68.4
Long	17	66.2

5.2. Pre-processing

For all experiments, data were first normalized to the range [0,1] using min-max normalization. Following the pre-processing methodology in (Lövberg, 2021), the data sampling frequency was reduced to 0.1Hz, and the float format was changed to a half-float format.

5.3. Constructing a health indicator with the reconstruction error

In this work, we resort to Principal Component Analysis to convert the degradation patterns identified by the baseline residual and our proposed methods into HIs. Specifically, the dimensionality of the degradation signal is reduced using the first principal component. Subsequently, the training sets' 2.5% percentiles are employed to remove the smallest and largest outliers. The final step involves normalizing the HIs through min-max normalization and averaging observations per cycle.

5.4. Network Configurations

Residual model - AE. The asymmetric-AE residual model is shown in Figure 1(a). The model is trained to reconstruct sensor readings when operating conditions and the sensor readings are concatenated. The architecture of the asymmetric-AE residual model used here comprises 4 feed-forward layers with ReLU activation functions. The input layer has a dimension size of 18, the hidden layers have a dimension size of 128, and the final layer has a size of 14.

Residual model - Regression. The regression type residual model implemented in this study is shown in Figure 1(b). We train a model to predict sensor readings given operating conditions. The model contains 4 feed-forward layers with ReLU activation functions. The input layer has a dimension size of 4, the hidden layers have a dimension size of 128, and the final layer has a size of 14. The implementation is identical to previous work in (Lövberg, 2021).

Proposed model - Physics-informed AE. The structure of the proposed algorithm is shown in Figure 3. The model is composed of two parts: an encoder and a decoder. Both of these parts are built using feed-forward neural networks. The encoder and decoder have the same structure but with reversed layer dimensions. Specifically, each model comprises two hidden layers with dimensions of size 128 with ReLU activation functions used throughout. After the input is processed through the encoder, the output is then passed through two fully connected layers with linear activation functions. The first fully connected layer has the same dimensionality as W, while the second has a dimensionality of 1. These two layers are then concatenated before being fed into the decoder. The parameter λ is set to 1.

5.5. Training set-up

The optimization of the network's weights is carried out with mini-batch stochastic gradient descent (SGD) and with the Adam algorithm. The batch size is set to 248 and the learning rate to 1e-5. The maximum number of epochs was set to 50. Early stopping was implemented for all models to stop training once the AE loss was below 1e-5 for 5 epochs. Early stopping was implemented to reduce the risk of overfitting.

5.6. Evaluation

Following the evaluation philosophy in (Nguyen & Medjaher, 2021), we compare and analyze the performance of the proposed method for HI discovery based on two evaluation aspects: quality of the HI and impact on the prognostic performance when the HI is used for an RUL estimation task. For each of the two aspects, we consider evaluation metrics that are defined in the following sections.

5.6.1. HI Criteria

There are several desirable properties that an HI should exhibit to represent the degradation of a system accurately. Although initial health conditions and operational modes can cause some variability in the discovered HIs, it is still desirable for them to demonstrate consistent behavior.

In this work, we employ the following criteria for HI evaluation:

Monotonicity measures the tendency for the HI to consistently increase or decrease (Coble, 2010). The monotonicity M of health index h_u of unit u with m observations is expressed as

$$M = \frac{1}{m-1} \sum_{j=1}^{m-1} |I(h_u^{j+1} - h_u^j) - I(h_u^j - h_u^{j+1}))| \quad (4)$$

$$I(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Trendability is used to evaluate the degree to which the HIs of a fleet of systems have a similar shape and underlying form (Coble, 2010). Trendability T of health index h_u of unit u with cycles t_u is expressed as

$$T = |\text{corr}(t_u, h_u)| \quad (5)$$

Prognosability is used to evaluate consistent HI behavior towards the end of life of units (Coble, 2010). Prognosability P of all health indexes in a set E^d is given by,

$$P = exp(-\frac{\sigma(h_u^{end})}{\mu(|h_u^{end} - h_u^0|)}) \quad u \in E^{\text{d}} \quad (6)$$

Where the starting and ending HI values of unit u are denoted as h_u^0 and h_u^{end}, respectively, while σ and μ refer to the standard deviation and mean operators.

Mutual Information score quantifies the information obtained about RUL by observing HI (Nguyen & Medjaher, 2021). Mutual information is a non-negative measure of dependency between two random variables. In this study, the negative exponential of mutual information $I(h_u, RUL_u)$ is used to obtain a score ranging from 0 to 1. Mutual Information score MI between h_u and RUL_u for unit u can be expressed as:

$$MI = \frac{1}{m}\sum_{i=1}^{m}[1 - exp(-I(h_u, RUL_u)] \qquad (7)$$

5.6.2. Prognostic Performance

A key objective of discovering HIs is to enhance the performance of prognostic models. In order to validate the effectiveness of the proposed HI discovery techniques, a baseline prognostic model is needed. From the existing methods for the turbofan dataset, a 1D-CNN- based model (Arias Chao, Kulkarni, Goebel, & Fink, 2022) is chosen due to its good performance. For more details about the implementation see the original paper.

The sensor signals, operating conditions, and cycle numbers are used as inputs to predict RUL. The model is given by:

$$G(X, W, t) = RUL$$

To test whether the constructed HI's help prognostic performance, HIs are used to augment the input space.

$$G(X, W, h, t) = RUL$$

Typical prognostic metrics, such as mean absolute error (MAE), root mean squared error (RMSE), Mean absolute percentage error (MAPE), and the NASA score function (s) are evaluated.

6. NUMERICAL EXPERIMENTS

To demonstrate the capabilities of the proposed method in comparison to the baseline residual approach, we consider four experiments performed by manipulating the original N-CMAPSS data to reflect realistic data scenarios. Firstly, we evaluate the effect of assuming varying amounts of healthy observations for training the residual models. Secondly, we represent a scenario where each unit of the fleet has a clearly different initial health condition. Thirdly, we train and test the model on non-overlapping flight classes to check the robustness of our method to out-of-distribution testing. Lastly, we consider the case where most CM data is healthy.

The selection of experiments is based on two main objectives. The first two experiments aim to highlight the shortcomings of the baseline residual approach, aiming to emphasize issues that our proposed method addresses more effectively. The subsequent two experiments are focused on evaluating the robustness of both models in addressing typical challenges encountered in practical prognostic scenarios.

Finally, to evaluate the efficacy of our proposed approach, we demonstrate the sensitivity of the model performance for different λ parameters.

6.1. Unknown Initial Health State

In real-world scenarios, it is often not possible to have access to the full health state of the system. Therefore, it is necessary to make assumptions about the amount of healthy data available for training. The proposed approach is insensitive to the amount of healthy data, but the performance of the residual approaches highly depends on this assumption. To demonstrate this impact, we will assume that the first [5, 10, 20, 40, 60] cycles of each unit are healthy, and evaluate the performance accordingly. The true number of healthy cycles is shown in Table 1.

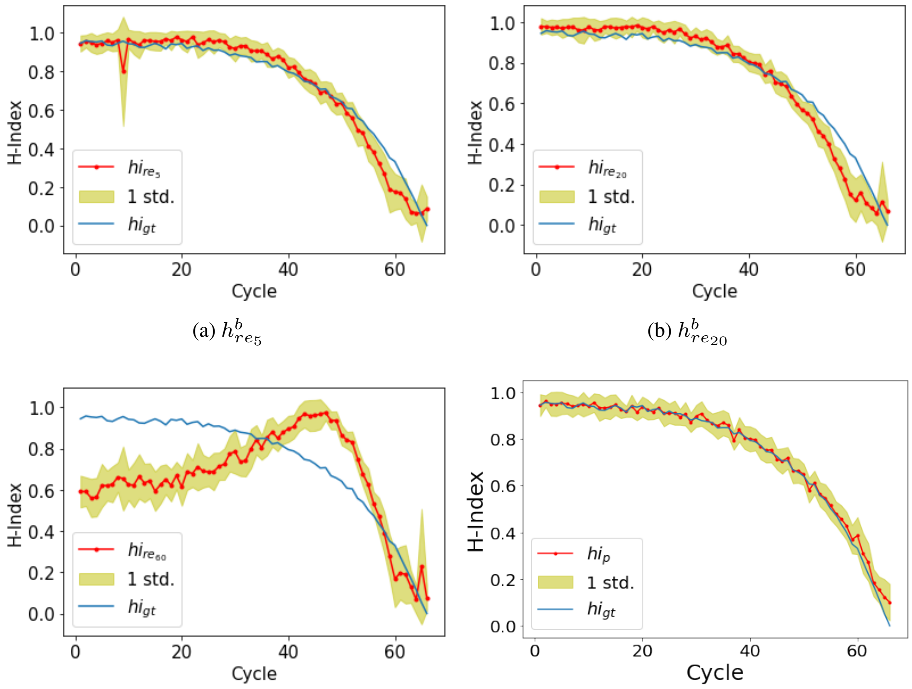
Table 2 presents the quantitative analysis results of the HI criteria. The table indicates the mean and standard deviation of the criteria values from 5 runs, along with an evaluation of the true HIs used to generate the N-CMAPSS dataset.

The regression-based residual model outperformed the AE-based residual model for all choices of the number of healthy observations. Therefore, the evaluation will only consider the regression-based residual approach going forward.

Moreover, the results indicate that the number of assumed healthy cycles significantly affects the performance of the residual approach. The best performance is achieved when 20 cycles are assumed to be healthy, which corresponds to the true number of healthy observations in the training dataset (on average 26 cycles). Assuming more than 40 healthy cycles leads to a significant drop in the model's performance.

Table 2. Quantitative results of the HI criteria under a varying amount of healthy observations. $h^a_{re(H)}$ health index of AE residual approach assuming H healthy observations, $h^b_{re(H)}$ health index of regression residual approach assuming H healthy observations, h_p health index of the proposed approach, h_{gt} ground truth health index.

HI	**M** (Eq. 3)	**T** (Eq. 4)	**P** (Eq. 5)	**MI** (Eq. 6)
$h^a_{re_5}$	0.31(0.06)	0.90(0.10)	0.86(0.01)	0.65(0.05)
$h^a_{re_{10}}$	0.33(0.03)	0.91(0.13)	0.87(0.01)	0.64(0.03)
$h^a_{re_{20}}$	0.35(0.08)	0.90(0.08)	0.94(0.01)	0.67(0.09)
$h^a_{re_{40}}$	0.27(0.06)	0.85(0.19)	0.92(0.01)	0.64(0.03)
$h^a_{re_{60}}$	0.10(0.06)	0.79(0.40)	0.71(0.03)	0.45(0.01)
$h^b_{re_5}$	0.38(0.03)	0.96(0.00)	0.90(0.05)	0.68(0.01)
$h^b_{re_{10}}$	0.39(0.04)	0.97(0.01)	0.91(0.04)	0.68(0.01)
$h^b_{re_{20}}$	**0.40(0.03)**	0.97(0.01)	0.91(0.01)	0.68(0.02)
$h^b_{re_{40}}$	0.33(0.04)	0.83(0.15)	0.91(0.00)	0.66(0.03)
$h^b_{re_{60}}$	0.21(0.06)	0.71(0.31)	0.90(0.02)	0.64(0.03)
h_p	0.35(0.01)	**0.98(0.00)**	**0.96(0.01)**	**0.70(0.01)**
h_{gt}	0.50	0.99	1.0	0.70



(a) $h^b_{re_5}$

(b) $h^b_{re_{20}}$

(c) $h^b_{re_{60}}$

(d) h_p

Figure 5. Discovered HI of test unit 10 by residual approach (a), (b), and (c)) and proposed approach (d)

In contrast, the proposed approach consistently outperforms the residual approach in terms of all HI evaluation metrics except monotonicity, regardless of the chosen number of healthy observations. Furthermore, the proposed approach can accurately discover HI values that align closely with the ground truth (i.e. h_{gt}). The better performance of the proposed method can also be observed graphically in Figure 5 when comparing the discovered HIs for a single test unit. The proposed method results in an HI (i.e. hi_p) showing a closer match to the truth than the three residual models.

Table 3. Quantitative results of the prognostic prediction task under a varying amount of healthy observations. G - neural network, X - sensor readings, W -operating conditions, t - cycles, $hi_{re(H)}$ - health index of residual approach assuming H healthy observations, hi_p health index of the proposed approach, hi_{gt} ground truth health index.

Model	**MAE**	**RMSE**	**MAPE**	**s (1e5)**
$G(X, W, t)$	6.4(0.9)	8.4(1.3)	31.2(4.5)	1.5(1.0)
$G(X, W, t, h^b_{re_{20}})$	5.9(0.4)	7.9(0.5)	27.0(2.2)	1.3(0.4)
$G(X, W, t, h_p)$	**5.6(0.3)**	**7.5(0.2)**	**26.5(3.5)**	**1.2(0.2)**
$G(X, W, t, h_{gt})$	5.0(0.1)	7.2(0.1)	16.5(2.0)	1.1(0.2)

The results from a prognostic prediction task are given in Table 3. Compared to a model which predicts RUL directly, the model which utilized the ground truth HI improves prognostic performance on average by 43%. This demonstrates that the obtained HI is very informative for RUL prediction.

The proposed method's discovered HI leads to better prognostic performance than the best-performing residual method's HI. On average, the best-performing residual approach improves performance by 11% on average, while the proposed approach increases performance by 17% on average. These results suggest that the proposed approach is able to discover an HI which is more informative for RUL prediction.

6.2. Difference in the Initial Health State

A common situation in the real world is that the health state of each unit can vary significantly when the health management tool is first switched on. For instance, some engines may have been in use for longer periods (and sometimes very long periods) before sensor monitoring begins, creating uncertainty about their health state. However, the N-CMAPSS dataset does not reflect this fact, as each unit was generated to have only minimal differences in initial health state.

In this experiment, the aim is to simulate a scenario where the initial health state of each unit is significantly different due to varying usage patterns. To achieve this, a certain number of initial cycles are randomly removed from the training data for each unit in the dataset. The proportion of cycles to be removed is generated from a Uniform distribution with parameters $[0, 0.75]$. The resulting truncated cycles are [29, 54, 8, 42, 67, 43, 9, 25, 57] for each training unit, as illustrated in Figure 6. The test data remains unchanged.

The results in Table 4 show that the residual approach is unable to construct a reasonable HI, resulting in a significant drop in each HI metric's performance. On the other hand, the proposed approach is less sensitive to data truncation, as it does not assume a portion of data to be healthy like the residual approach. Table 5 provides additional evidence of the effectiveness of the proposed approach as it improves prognostic performance through the discovered HI values. The proposed method discovers an HI which has on average 20% better prognostic performance than the residual approach.

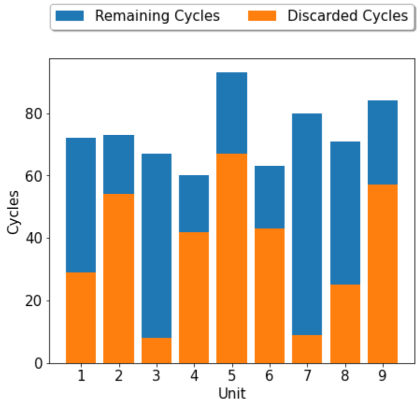


Figure 6. Illustration of data truncation

Table 4. Quantitative results of HI criteria for different initial health states of the system.

HI	**M** (Eq. 3)	**T** (Eq. 4)	**P** (Eq. 5)	**MI** (Eq. 6)
$h^b_{re_{20}}$	0.08(0.04)	0.36(0.21)	0.82(0.05)	0.50(0.03)
h_p	**0.25(0.01)**	**0.95(0.00)**	**0.96(0.01)**	**0.78(0.01)**
h_{gt}	0.49	0.99	1.0	0.75

Table 5. Quantitative results of prognostic prediction for different initial health states of the system.

Model	**MAE**	**RMSE**	**MAPE**	**s(1e7)**
$G(X, W, t)$	9.5(1.5)	17.4(5.8)	36.2(4.8)	3.3(2.3)
$G(X, W, t, h^b_{re_{20}})$	8.7(1.0)	13.9(2.8)	30.1(1.6)	1.2(0.9)
$G(X, W, t, h_p)$	**7.8(1.3)**	**12.7(2.8)**	**26.4(3.6)**	**0.8(0.9)**
$G(X, W, t, h_{gt})$	7.8(1.5)	12.2(2.6)	20.1(2.6)	0.8(1.2)

6.3. Out-of-Distribution Testing

The accuracy and reliability of prognostic prediction techniques depend highly on the quality and representativeness of the available time-to-failure data. Therefore, these methods may not perform well when applied to data from new units that operate under different conditions than those used during training (Nejjar, Geissmann, Zhao, Taal, & Fink, 2023).

As previously mentioned, engine degradation is influenced by operating conditions such as take-off and landing. The N-CMAPSS dataset consists of engines classified into three distinct flight classes, and it is reasonable to expect that the degradation patterns between engines of different flight classes may differ significantly. Furthermore, the sensor measurements of units from different flight classes may also be affected as the units reach different altitudes and speeds (Nejjar et al., 2023).

To assess the robustness of the baseline residual approach and the proposed approach to significant changes in operating conditions, we propose to train the models using short-flight class data and test them on medium to long-flight classes. The residual approach was trained assuming the first 20 cycles of each unit are healthy.

Table 6 shows the HI evaluation metrics. The results demonstrate that the proposed approach yields an HI that closely aligns with the ground truth HI, outperforming the HI discovered by the residual approach. Table 7 suggests that the proposed method is also able to discover an HI which is more beneficial for prognostic performance than the HI discovered by the residual approach.

Table 6. Quantitative results of the HI criteria out-of-distribution testing.

HI	**M** (Eq. 3)	**T** (Eq. 4)	**P** (Eq. 5)	**MI** (Eq. 6)
$h^b_{re_{20}}$	0.12(0.02)	0.81(0.01)	0.88(0.03)	0.62(0.01)
h_p	**0.21(0.06)**	**0.94(0.01)**	0.87(0.10)	**0.78(0.04)**
h_{gt}	0.44	0.99	1.0	0.85

Table 7. Quantitative results of the prognostic prediction out-of-distribution testing.

Model	**MAE**	**RMSE**	**MAPE**	**s(1e6)**
$G(X, W, t)$	11.4(3.2)	15.3(4.6)	43.7(9.2)	7.5(0.6)
$G(X, W, t, h^b_{re_{20}})$	11.4(2.9)	15.2(4.0)	37.4(4.7)	6.9(5.3)
$G(X, W, t, h_p)$	**10.8(2.9)**	**13.8(3.9)**	**33.5(2.8)**	**2.0(1.8)**
$G(X, W, t, h_{gt})$	8.5(1.2)	11.4(1.9)	31.4(1.2)	1.6(0.9)

6.4. Majority Healthy Data

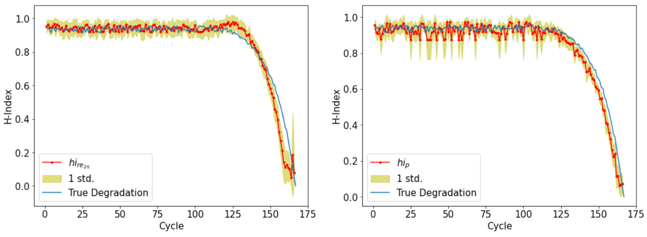
In certain scenarios, the majority of the operational data collected from an engine is healthy. This is because engines tend to be in good working condition for the majority of their lifetime, and only experience significant degradation towards the end of their useful life. We refer to this scenario as the “majority healthy" situation. In theory, the experiment should favor the residual model since there is an abundance of healthy data for training. The proposed model should have difficulties in this experiment, since degradation remains constant for a long period of time and does not correlate with cycle time.

To explore the performance of HI discovery models in the scenario where the majority of data is healthy, an experiment can be designed where the training and testing data are heavily skewed towards healthy observations. For each training and test unit, the amount of healthy data was augmented by sampling the first 20 cycles with repetition 100 times. To train the residual model, it was assumed that the first 100 cycles of each flight are healthy.

The results of the HI evaluation are presented in Table 8, and the discovered HIs are shown in Figure 7. The performance of the two methods is almost identical, and it is difficult to determine which model is preferable. Notably, the proposed method performs equally well as the residual approach, even in unfavorable conditions.

Table 8. Quantitative results of the HI criteria majority healthy data.

HI	**M** (Eq. 3)	**T** (Eq. 4)	**P** (Eq. 5)	**MI** (Eq. 6)
$h^b_{re_{100}}$	0.17(0.02)	0.62(0.03)	0.92(0.02)	0.48(0.02)
h_p	0.17(0.02)	0.65(0.00)	0.96(0.01)	0.50(0.01)
h_{gt}	0.21	0.65	1.0	0.62



(a) $h^b_{re_{100}}$

(b) h_p

Figure 7. Discovered HI of test unit 10. Majority healthy experiment results

6.5. Sensitivity of the learning bias

In a final ablation study, we investigate the sensitivity of the proposed approach to the choice of λ parameter. The λ parameter controls the trade-off between the importance of the additional constraint and the ability of the model to function as an AE. The results of the sensitivity analysis are shown in Table 9

When setting $\lambda = 0$, the introduced additional constraint concerning degradation within the latent space of the AE model is removed. We can observe that in the absence of this constraint, the model is no longer able to capture the degradation patterns. Hence, this constraint, informed by expert knowledge, serves as an effective guide for the discovery of degradation.

Table 9. Sensitivity of the proposed approach to given λ values

λ	**M** (Eq. 3)	**T** (Eq. 4)	**P** (Eq. 5)	**MI** (Eq. 6)
0	0.05(0.01)	0.05(0.02)	0.00(0.01)	0.05(0.02)
0.001	0.10(0.09)	0.21(0.37)	0.19(0.37)	0.18(0.25)
1	0.35(0.01)	0.98(0.00)	0.96(0.01)	0.70(0.01)
1000	0.33(0.02)	0.98(0.00)	0.96(0.01)	0.70(0.00)

With an increase of the λ parameter beyond 0.001, an enhancement in model performance becomes evident, as shown by the favorable HI evaluation metrics. It is worth highlighting that the model's responsiveness to the λ parameter is minimal since the model is able to reconstruct degradation patterns across a broad spectrum of λ parameter values. This insensitivity to λ variations stems from the model's training procedure, where early stopping is employed based solely on the AE loss.

7. CONCLUSION

In this paper, we introduce a novel physics-informed unsupervised model for HI discovery. We propose to incorporate general knowledge about the degradation process of complex systems as both an inductive bias on the network architecture and a learning bias on the objective function.

We present the effectiveness of our proposed approach through a comparison with the residual approach using the N-CMAPSS turbofan dataset. It becomes apparent that the current HI discovery technique is sensitive to the availability of healthy training data and the uniformity of initial health states among units. In contrast, our novel approach displays resilience against these challenges, showcasing superior performance in detecting degradation patterns. Moreover, we highlight the robustness of our method in typical real-world scenarios where a substantial portion of the data is healthy or diverges from the distribution. Simultaneously, we illustrate the potential of the discovered HI in enhancing prognostic model performance. These outcomes emphasize the value of integrating expert knowledge into the learning algorithm, resulting in more precise and robust health indicators for prognostic models.

Future research will expand the proposed methodology by incorporating additional types of expert knowledge related to degradation. The methodology will be applied to other complex systems, such as batteries, where degradation is dependent on operational cycles.

8. LIMITATIONS

While we demonstrated excellent performance of the proposed model there are still some limitations. An inherent limitation of the proposed approach in its current form is that we only investigate one case study where the failure modes are dominated by cycle loading. In scenarios where the system's failure mechanism is governed by different factors, the soft constraint used by our proposed approach might not be suitable. We emphasise that an in-depth understanding of the precise physics of degradation is not mandatory; rather the main factors driving degradation need to be identified.

Furthermore, the proposed method inherits the general limitation of any HI estimation methodology for validation in certain real-world scenarios. In cases where degradation is observable the estimated HI can be compared to ground truth values. However, in many instances, the HI of the system can only be inferred from simulators or with performance tests, which makes validation challenging.

REFERENCES

Arias Chao, M., Kulkarni, C., Goebel, K., & Fink, O. (2021). Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data*, *6*(1), 5.

Arias Chao, M., Kulkarni, C., Goebel, K., & Fink, O. (2022). Fusing physics-based and deep learning models for prognostics. *Reliability Engineering & System Safety*, *217*, 107961.

Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*.

Coble, J. B. (2010). Merging data sources to predict remaining useful life–an automated method to identify prognostic parameters.

Cofre-Martel, S., Lopez Droguett, E., & Modarres, M. (2021). Remaining useful life estimation through deep learning partial differential equation models: A framework for degradation dynamics interpretation using latent variables. *Shock and Vibration*, *2021*, 1–15.

de Beaulieu, M. H., Jha, M. S., Garnier, H., & Cerbah, F. (2022). Unsupervised remaining useful life prediction through long range health index estimation based on encoders-decoders. *IFAC-PapersOnLine*, *55*(6), 718–723.

de Pater, I., & Mitici, M. (2023). Developing health indicators and rul prognostics for systems with few failure instances and varying operating conditions using a lstm autoencoder. *Engineering Applications of Artificial Intelligence*, *117*, 105582.

Fu, S., Zhong, S., Lin, L., & Zhao, M. (2021). A novel time-series memory auto-encoder with sequentially updated reconstructions for remaining useful life prediction. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 7114–7125.

Guo, L., Lei, Y., Li, N., Yan, T., & Li, N. (2018). Machinery health indicator construction based on convolutional neural networks considering trend burr. *Neurocomputing*, *292*, 142–150.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440.

Koutroulis, G., Mutlu, B., & Kern, R. (2022). Constructing robust health indicators from complex engineered systems via anticausal learning. *Engineering Applications of Artificial Intelligence*, *113*, 104926.

Lee, H., Lim, H. J., & Chattopadhyay, A. (2021). Data-driven system health monitoring technique using autoencoder for the safety management of commercial aircraft. *Neural Computing and Applications*, *33*, 3235–3250.

Liu, K., & Huang, S. (2014). Integration of data fusion methodology and degradation modeling process to improve prognostics. *IEEE Transactions on Automation Science and Engineering*, *13*(1), 344–354.

Lövberg, A. (2021). Remaining useful life prediction of aircraft engines with variable length input sequences. In *Annual conference of the phm society* (Vol. 13).

Magadán, L., Suárez, F. J., Granda, J. C., delaCalle, F. J., & García, D. F. (2023). A robust health prognostics technique for failure diagnosis and the remaining useful lifetime predictions of bearings in electric motors. *Applied Sciences*, *13*(4), 2220.

Nejjar, I., Geissmann, F., Zhao, M., Taal, C., & Fink, O. (2023). Domain adaptation via alignment of operation profile for remaining useful lifetime prediction. *arXiv preprint arXiv:2302.01704*.

Nguyen, K. T., & Medjaher, K. (2021). An automated health indicator construction methodology for prognostics based on multi-criteria optimization. *ISA transactions*, *113*, 81–96.

Pearl, J., et al. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, *19*(2).

Qin, Y., Yang, J., Zhou, J., Pu, H., & Mao, Y. (2023). A new supervised multi-head self-attention autoencoder for health indicator construction and similarity-based machinery rul prediction. *Advanced Engineering Informatics*, *56*, 101973. doi: https://doi.org/10.1016/j.aei.2023.101973

Wang, T., Yu, J., Siegel, D., & Lee, J. (2008). A similarity-based prognostics approach for remaining useful life estimation of engineered systems. In *2008 international conference on prognostics and health management* (pp. 1–6).

Yang, F., Habibullah, M. S., Zhang, T., Xu, Z., Lim, P., & Nadarajan, S. (2016). Health index-based prognostics for remaining useful life predictions in electrical machines. *IEEE Transactions on Industrial Electronics*, *63*(4), 2633–2644.

Ye, Z., & Yu, J. (2021). Health condition monitoring of machines based on long short-term memory convolutional autoencoder. *Applied Soft Computing*, *107*, 107379.

Zgraggen, J., Pizza, G., & Huber, L. G. (2022). Uncertainty informed anomaly scores with deep learning: Robust fault detection with limited data. In *Phm society european conference* (Vol. 7, pp. 530–540).

Zhai, S., Gehring, B., & Reinhart, G. (2021). Enabling predictive maintenance integrated production scheduling by operation-specific health prognostics with generative deep learning. *Journal of Manufacturing Systems*, *61*, 830–855.

Zhou, H., Huang, X., Wen, G., Lei, Z., Dong, S., Zhang, P., & Chen, X. (2022). Construction of health indicators for condition monitoring of rotating machinery: A review of the research. *Expert Systems with Applications*, *203*, 117297.