Health index estimation through integration of general knowledge with unsupervised learning

Bajarunas, Kristupas; Baptista, Marcia L.; Goebel, Kai; Chao, Manuel Arias

DOI
10.1016/j.ress.2024.110352

Publication date
2024

Document Version
Final published version

Published in
Reliability Engineering and System Safety

Reliability Engineering and System Safety 251 (2024) 110352

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Health index estimation through integration of general knowledge with unsupervised learning

Kristupas Bajarunas [a,b,*], Marcia L. Baptista [a], Kai Goebel [c,d], Manuel Arias Chao [a,b]

[a] Faculty of Aerospace Engineering, Delft University of Technology, HS 2926 Delft, The Netherlands
[b] Institute of Data Analysis and Process Design, Zurich University of Applied Sciences, 8401 Winterthur, Switzerland
[c] SRI International, 3333 Coyote Hill Rd, CA 94304 Palo Alto, United States
[d] Luleå University of Technology, 971 87 Luleå, Sweden

ARTICLE INFO

Keywords:
Prognostics
Health index
Hybrid model
Unsupervised learning
Convolutional autoencoder

ABSTRACT

Accurately estimating a Health Index (HI) from condition monitoring data (CM) is essential for reliable and interpretable prognostics and health management (PHM) in complex systems. In most scenarios, complex systems operate under varying operating conditions and can exhibit different fault modes, making unsupervised inference of an HI from CM data a significant challenge. Hybrid models combining prior knowledge about degradation with deep learning models have been proposed to overcome this challenge. However, previously suggested hybrid models for HI estimation usually rely heavily on system-specific information, limiting their transferability to other systems. In this work, we propose an unsupervised hybrid method for HI estimation that integrates general knowledge about degradation into the convolutional autoencoder's model architecture and learning algorithm, enhancing its applicability across various systems. The effectiveness of the proposed method is demonstrated in two case studies from different domains: turbofan engines and lithium batteries. The results show that the proposed method outperforms other competitive alternatives, including residual-based methods, in terms of HI quality and their utility for Remaining Useful Life (RUL) predictions. The case studies also highlight the comparable performance of our proposed method with a supervised model trained with HI labels.

1. Introduction

Understanding the health condition of complex systems is an important step in prognostics and health management (PHM) [1,2]. A Health Index (HI) represents the system's health state over time or usage on a scale from 1 (perfect health) to 0 (failure) and, therefore, provides a clear and interpretable measure of degradation. HIs are also instrumental in predicting remaining useful life (RUL). For instance, HIs can be integrated into prognostic models by matching HI patterns with known failure times [3–6], or extrapolated until the failure threshold [7–9] for RUL predictions.

Different data-driven approaches have been proposed for estimating HI from condition monitoring (CM) data, but many of these approaches rely extensively on labeled data. For instance, when dealing with datasets containing HI labels, the utilization of supervised models is prevalent [10]. Another common strategy is the residual technique, where models are trained to recognize a system's normal behavior using health state labels, subsequently identifying the HI by analyzing reconstruction errors [11–13]. However, for complex systems, the challenge lies in obtaining representative labeled data, which can be costly or unfeasible in industrial contexts. This limitation has motivated a growing interest in unsupervised learning methods for HI estimation, circumventing the need for labeled datasets.

To address the difficulty of dealing with unlabeled data, researchers have proposed hybrid unsupervised models combining data-driven models with prior knowledge about the system for HI estimation. For instance, Biggio et al. [14] leverage a battery simulator for training a transformer architecture with synthetic data of degraded system dynamics, allowing their model to uncover degradation patterns from real-world experiments. Alternatively, Guo et al. [15] combine a data-driven HI with knowledge-based HI to model power transformer degradation. Nonetheless, a characteristic of many current hybrid models is their dependence on system-specific knowledge (e.g., detailed simulators of degraded system dynamics), limiting their applicability to other systems exhibiting diverse degradation patterns. Moreover, as pointed out in recent review work [16], most models rely on a single strategy to integrate data and prior knowledge, potentially limiting

* Corresponding author at: Faculty of Aerospace Engineering, Delft University of Technology, HS 2926 Delft, The Netherlands.
E-mail addresses: baja@zhaw.ch, k.v.b.bajarunas@tudelft.nl (K. Bajarunas), m.lbaptista@tudelft.nl (M.L. Baptista), kai.goebel@sri.com (K. Goebel), aria@zhaw.ch, m.a.c.ariaschao@tudelft.nl (M.A. Chao).

https://doi.org/10.1016/j.ress.2024.110352

Available online 14 July 2024

Nomenclature

HI	Health index
PHM	Prognostics and health management
CM	Condition monitoring
RUL	Remaining useful life
SL	Supervised learning
UL	Unsupervised learning
RM	Residual method
AE	AutoEncoder
PCA	Principal component analysis
CNN	Convolutional neural network
MAE	Mean absolute error
RMSE	Root mean squared error
MAPE	Mean absolute percentage error
Mon	Monotonicity
Tren	Trendability
Prog	Prognosability
MutInf	Mutual information score
CMAPSS	Commercial modular aero-propulsion system simulation
ANM	Additive noise model
SCM	Structural causal model
DAG	Directed acyclic graph

Symbols in the equations

X	Sensor readings
W	Operating conditions
Z	Degradation (representation)
T	Cycle number
u	Unit of a fleet
m	Observations
p	Number of sensors
k	Number of operating conditions
C	Correlation constraint
NG	Negative gradient constraint
F	Functional constraint

their effectiveness. These gaps hinder the broader application of hybrid models and emphasize the need for more general hybrid models that accommodate a wide array of complex systems [17].

In this paper, we build upon our initial study [18], where we showcased the feasibility of inferring HIs for turbofan engines with a hybrid unsupervised method. In this current research, we expand the methodology with the primary objective of demonstrating the generalization of our method across various complex systems. To this end, we address the following research question: *How can knowledge about degradation be incorporated into an unsupervised hybrid method for HI estimation applicable to diverse complex systems?*

To achieve the intended generalization, we propose to integrate general domain knowledge about the HI problem into the method using multiple hybridization strategies. Specifically, we postulate that there are common fundamental degradation characteristics at a certain level of abstraction that are informative and, hence, transferable across a range of complex systems. For instance, degradation characteristics of multiple complex systems exhibit a fast wear period, followed by a period of almost steady decline, which is ultimately followed by another period of faster wear towards the end of life. Additionally, we hypothesize that an understanding of known causal relationships can be leveraged for more accurate HI estimation

In line with these hypotheses, we propose an unsupervised hybrid method for HI estimation with two distinct design features: (1) a novel network architecture of a convolutional AutoEncoder (AE) preserving the causal relationships among sensor readings, operating conditions, and degradation within complex systems, and (2) the incorporation of soft constraints within the loss function derived from general knowledge of the degradation process, guiding the AE to infer degradation in its latent space. Fig. 1 provides an overview of our proposed HI estimation methodology.

To demonstrate the intended generalization, our experimental analysis investigates two distinct case studies, turbofan engines and Li-ion batteries. Our proposed method is thoroughly compared against two alternative methodologies: a residual-based method and a supervised model. This comparison encompasses scenarios both within and out of distribution. For reproducibility and future research purposes, the complete code reproducing this study is released in open-source at https://github.com/KBaja/UnsupervisedHI.

The main contributions of this study are as follows:

1. We propose a novel hybrid unsupervised method for HI estimation that relies on general knowledge about degradation and combines multiple hybridization techniques. We demonstrate that our model can accurately estimate the HI of various systems (turbofan engines and batteries) with distinct degradation patterns.
2. We provide an extensive comparative analysis involving supervised, residual, and unsupervised methods for HI estimation, enabling a quantified assessment of their respective performances. The outcomes highlight the superiority of our proposed unsupervised method over the residual method, positioning it at par with the supervised model.
3. We evaluate different forms of general knowledge options for integration into our model to be used as constraints in the latent space of an AE: monotonicity, negative gradient, and functional HI.

The remainder of this paper is organized as follows: Section 2 presents background information about general degradation dynamics. Section 3 presents the problem formulation and related work. Section 4 proposes the unsupervised hybrid model, while Section 5 presents the case studies and the training set-up. In Section 6, the results are presented, followed by a discussion in Section 7, and the conclusion in Section 8.

2. General domain knowledge about degradation

While system-specific knowledge can be valuable for HI estimation, it often limits the generalizability of the approach. This section explores the concept of **general knowledge** about degradation, referring to domain knowledge that applies broadly to complex systems exhibiting degradation. In particular, we introduce two key examples of general knowledge utilized in this paper: the causal structure of condition monitoring data and degradation dynamics.

In the context of a complex system, the causal structure describes the underlying network of cause-and-effect relationships that link the various components together. Uncovering the causal structure is crucial for understanding how a complex system functions, responds to external influences and evolves over time. It involves identifying the key components, their interactions, and the mechanisms through which they influence each other, as well as the potential feedback loops and non-linear dynamics that can emerge from these interactions.

Knowledge of the underlying cause-and-effect relationships within a complex system can provide valuable insights into degradation processes.[1] This knowledge, often regarded as general knowledge [19–21],

[1] Under the hypothesis that signatures of faults are present in the condition monitoring data.

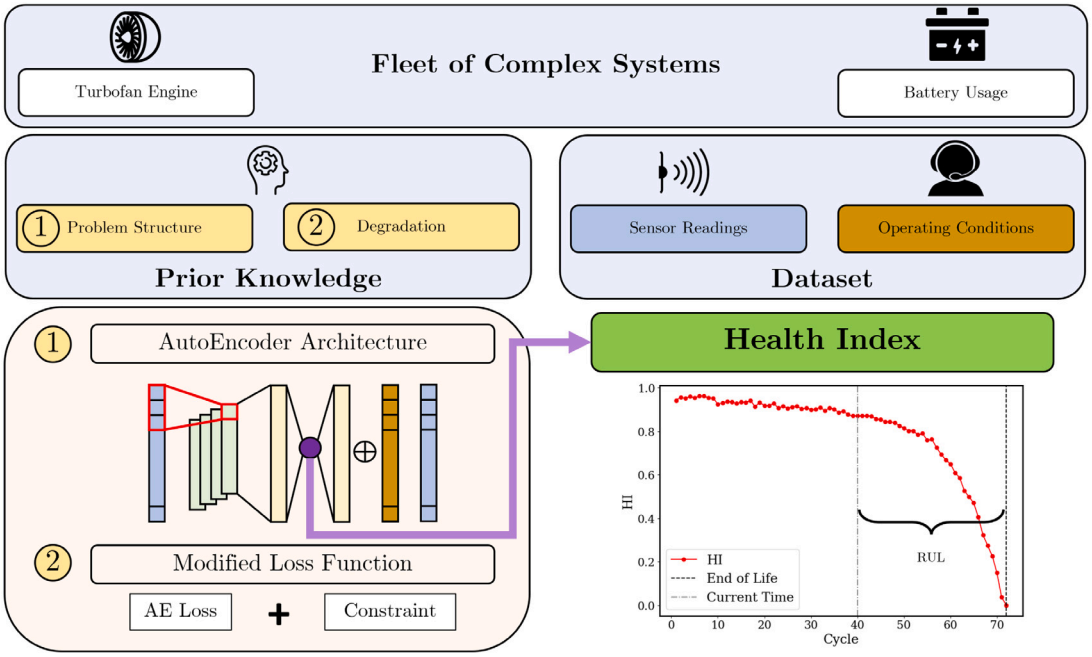


Fig. 1. Overview of the unsupervised hybrid method for HI estimation that relies primarily on general knowledge about degradation.

is based on a general understanding of which variables (causes) directly influence degradation and which variables (effects) are themselves affected by degradation. For instance, in a battery system, increasing the ambient temperature (cause) can accelerate the degradation of battery capacity (effect). From another perspective, a decrease in capacity (cause) can lead to faster constant current charging times (effect).

Beyond understanding the cause-and-effect relationships within a system, general knowledge also extends to the concept of degradation dynamics. In complex systems, degradation often unfolds gradually, with failures evolving over time rather than occurring abruptly [22]. Most failure modes stem from an underlying degradation process, where gradual deterioration eventually reveals weaknesses that can lead to system failure [23]. Degradation in complex systems manifests itself in various forms, ranging from observable changes in physical components like crack growth to subtler alterations affecting system dynamics and performance degradation, such as changes in battery output voltage. Despite the diverse manifestations of degradation, consistent patterns emerge across different systems and their degradation dynamics.

Fig. 2 illustrates three common temporal evolutions of degradation: linear, convex, and concave. While real-world systems may exhibit a combination of these patterns, this figure showcases each in isolation for clarity.. The horizontal line on the degradation scale represents the failure threshold. Linear degradation involves a steady increase in degradation over time, such as the wear of automobile tire treads, which appears linear over a certain time. Convex degradation entails an accelerating rate of degradation increase, as seen in crack growth scenarios. Conversely, concave degradation entails an increase in degradation over time at a diminishing rate, such as the growth of chlorine–copper compounds in printed circuit boards [23].

At a certain level of abstraction, fundamental degradation characteristics remain consistent across various system dynamics [24]. For instance, degradation is typically monotonic or non-decreasing, as depicted in Fig. 2, which may be expressed mathematically through positive first differences between observations or positive gradients with respect to time. Even though some of these commonalities may not universally apply to all real-world systems (e.g., the apparent short-term capacity recovery of batteries during no-load periods), we anticipate their persistence in the long-term degradation process.

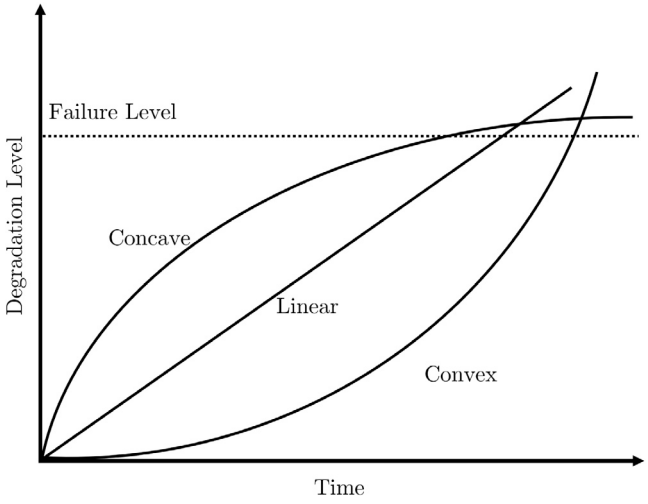


Fig. 2. Possible shapes for univariate health degradation curves.
Source: Adapted from [23].

3. Problem formulation and related work

In this section, we formally introduce the problem of HI estimation from CM data (Section 3.1). We also review related work and discuss three established solution strategies: residual methods (Section 3.2), unsupervised (Section 3.3), and supervised methods (Section 3.5). These methods serve as a benchmark against which we evaluate the methodology proposed in this work. In Fig. 3, we provide an overview of the model functional mappings and data requirements for various HI estimation methods.

3.1. Problem formulation

We are given multi-variate time series of sensor readings

$$X_u = [x_u^1, \dots, x_u^m] \tag{1}$$

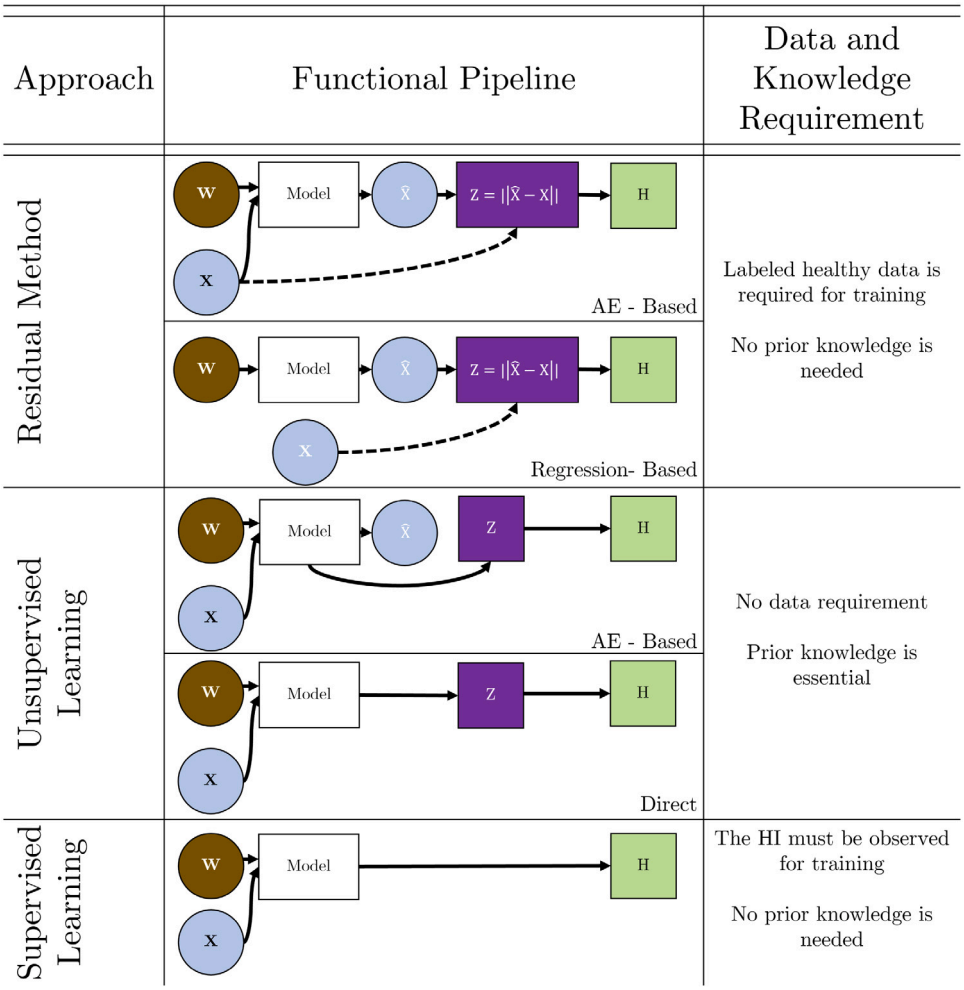


Fig. 3. Overview of different HI estimation methods.

of a fleet of N units ($u = 1, \ldots, N$), 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 corresponding scenario–descriptor operating conditions

$$W_u = [w_u^1, \ldots, w_u^m] \tag{2}$$

for each unit, where each $w_u^i \in \mathbb{R}^k$. The goal is to estimate the state of degradation Z of each unit at each point in time $z_u^i \in \mathbb{R}^z$. The HI of each unit at each point in time h_u^i is then a normalized 1-D representation of the state of degradation z_u^i, such that

$$\{z_u^i \in \mathbb{Z}^z\} \rightarrow \{h_u^i \in \mathbb{R} | 0 \leq h_u^i \leq 1\} \tag{3}$$

3.2. Residual method HI estimation

Safety critical systems undergo comprehensive health monitoring and inspections, allowing the accurate labeling of subsets of CM data as healthy. Alternatively, in certain conditions, as in the case of new units, CM data can be labeled healthy, assuming minimal degradation during the initial operational cycles. Therefore, one of the most common methods for HI estimation in the literature is the residual method [25–33].

The residual method uses CM data which is labeled as healthy to train a model $f(\bullet)$ emulating healthy system responses X. Once the model is trained, an HI is estimated from the reconstruction residual r of current CM data X, which is given by:

$$r = |f(\bullet) - X| \tag{4}$$

In the final step, the HI is found by reducing the dimensionality of r and normalizing the resulting one-dimensional projection in the range [0, 1]

$$r \in \mathbb{R}^p \rightarrow h \in \mathbb{R} \tag{5}$$

The residual method typically works well under the hypothesis that the training dataset is representative of a healthy system, meaning that small reconstruction errors are typically indicative of healthy inputs, while large reconstruction errors are typically indicative of faulty operations that were not observed during training. For the residual model to function effectively, it is essential to consider variations in operating conditions. This ensures that the model can discern alterations in sensor readings that are unrelated to degradation.

Two types of residual methods are often proposed depending on the inputs of the model $f(\bullet)$. The residual method has been used with an asymmetric AE to reconstruct the sensor readings based on information about the operating conditions and the historical sensor readings [12,34]. Another approach is the regression residual method that maps the sensor readings based on the operating conditions [13,34]. The difference between the two methods is visualized in Fig. 3. Previous research indicates that the regression-based residual method may be preferable in situations where data collection is limited, but may not perform as well in cases where the sensor readings contain outliers [35].

3.3. Unsupervised HI estimation

In the general scenario of complex systems, the acquisition of labeled data indicating whether or not a system is healthy or the extent of degradation can often prove challenging. In such a scenario, unsupervised learning methods become essential. AE models are a popular unsupervised technique for HI estimation, aiming to learn a representation of unlabeled CM data encompassing both healthy and degraded conditions. For instance, de Beaulieu et al. [36] showed that, in certain scenarios, AEs can reveal degradation patterns in their latent space in a case study involving turbofan engines.

Another commonly employed technique is Principal Component Analysis (PCA). Schwartz et al. [37] achieved successful HI estimation for turbofan engines by leveraging the first principal components extracted from sensor reading data using Kernel PCA. Both PCA and its various extensions have been shown to perform comparably to AEs in certain scenarios.

While the mentioned methods showed success in specific cases using subsets of the CMAPSS turbofan dataset with constant operating parameters, as we reveal in Section 6.3 of our case study, their effectiveness diminishes in scenarios where operational conditions mask degradation's impact on sensor readings.

3.4. Unsupervised hybrid HI estimation

Hybrid methods integrating prior knowledge with unsupervised models have emerged as a promising solution strategy for HI estimation [16]. Following the taxonomy of hybrid models proposed in [38,39], these methods can be classified based on **where prior knowledge** is integrated into the machine learning pipeline and **what type of knowledge** is integrated.

Regarding **where prior knowledge** is integrated, we refer to three main hybridization strategies. Observational bias involves augmenting the training data with synthetic data or derived features that reflect underlying prior knowledge, serving as a weak mechanism to embed knowledge into machine learning models. Inductive bias focuses on crafting specialized model architectures that implicitly incorporate the additional knowledge. Learning bias seeks to infuse prior knowledge by modifying the model's learning algorithm, ensuring that the model simultaneously fits the observed data and approximately adheres to a set of specified constraints.

Beyond hybridization strategies, **the type of knowledge** integrated plays a pivotal role in the performance of hybrid methodologies. Two primary sources of knowledge stand out: system-specific high-fidelity knowledge, such as simulators, while another group leverages general low-fidelity knowledge.

System-specific knowledge. The application of *observational bias* in combination with system-specific knowledge for HI estimation is demonstrated in the research conducted by Magadan et al. [40]. In their work, a model was trained using features that were extracted by considering interesting bearing frequencies, which were identified based on prior knowledge about the system. Another instance of observational bias is seen in the work of Biggio et al. [14]. The authors used a battery simulator to train a model with multiple degradation parameters. The resulting model was then capable of extracting health information from actual data. An *inductive bias* strategy is proposed in the work of Guo et al. [15]. The authors combine a data-driven HI with an HI based on knowledge about the specific system. A similar methodology for inductive bias can also be found in [41,42].

General knowledge. Despite these advancements, there are still challenges with methods that use system-specific prior knowledge. Namely, specific knowledge may not apply to diverse systems with different degradation patterns. Additionally, relying on specific knowledge such as predefined linear or nonlinear functions for degradation can bias the estimation of unit-specific degradation patterns, limiting the model's ability to capture unique characteristics.

A few studies have leveraged general knowledge that is not specific to a single system. This prior knowledge is typically based on the expectation of the HI in these complex systems. The most commonly used piece of knowledge is the understanding that the estimated HI should exhibit certain characteristics, such as being monotonic, correlating with the operational cycle time, and having a consistent threshold for failure across a fleet of units.

Researchers have used the knowledge about the expectation for the HI in two ways. The first is by using it to guide the selection of data features that mirror the HI's desired characteristics, thereby introducing an observational bias. This method is evident in the works in [43–45]. The second method constructs the learning objective function of the model tasked with HI estimation. This introduces a learning bias into the model, steering the learning process towards solutions that are consistent with our prior understanding of the HI. This method is showcased in the research conducted in [2,46–52].

While the use of general knowledge can lead to the development of a more universal model for HI estimation, it is not always clear whether this general knowledge is adequate to enhance model performance. This uncertainty arises from the limited informativeness of the knowledge employed.

3.5. Supervised HI estimation

Supervised learning methods can be used for HI estimation when the state of degradation of a system is directly observable. For instance, in controlled laboratory experiments with battery usage, one indication of degradation is declining performance, often reflected in phenomena like capacity fade. Following analytical calculations of capacity fade, it can function as a proxy for HI and be utilized to train supervised models [10,53–55]. Similarly, in the field of fracture mechanics, where observations of crack length serve as a degradation proxy [56,57], and in the field of machining tool wear, where wear can be measured directly [58].

However, for most complex systems, the degradation of a system is complex, affecting multiple components, and is unobservable without a detailed inspection. In these cases, supervised learning methods are not applicable since no desired output labels exist.

3.6. Overview of related work

Our analysis of the related work, summarized in Table 1, reveals prevailing trends in HI estimation. Labeled data approaches are notably dominant, as seen in the widespread use of residual methods (RM) and supervised learning (SL) techniques. Although these categories typically do not involve hybrid approaches, there are notable exceptions. For example, in [32], the authors employ a residual approach augmented with physics-informed fault signatures for gearbox degradation discovery, introducing an observational bias. Similarly, in the supervised learning approach of [58], the authors integrate a physical model of tool cutting, fitted from observed data, to augment observed tool wear data with synthetic features generated by the model.

However, a specific focus on purely hybrid unsupervised methods reveals a significant research gap: the scarcity of models that employ multiple hybridization strategies. As discussed in [16], a combined approach could significantly enhance the flexibility and robustness of hybrid methods. Additionally, the table underscores a critical issue of limited generalizability; only the work by Chen et al. [48] has demonstrated their methodology across multiple case studies. This restriction highlights concerns about the broad applicability of existing methods. For wider utility, HI estimation methods need to be adaptable and effective across various systems.

Table 1
Overview of recent works on HI estimation. The column explanations are as follows: *Approach* indicates HI estimation method (RM — residual method, UL — unsupervised learning, SL — supervised learning), *Hybrid Strategy* shows which hybridization strategy was used (O — observational bias, I — inductive bias, L — learning bias, X — none), *General Knowledge* denotes whether general knowledge was used for a hybrid model (✓) or not (X), and *Case study* specifies the systems on which the method was tested.

Reference	Approach	Hybrid strategy	General knowledge	Case study
[25]	RM	–	–	Turbofan (CMAPSS)
[26]	RM	I	–	Bearings
[27]	RM	–	–	Bearings
[28]	RM	–	–	Turbofan (CMAPSS)
[29]	RM	–	–	Turbofan (DASHlink)
[30]	RM	I	–	Turbofan (N-CMAPSS)
[31]	RM	–	–	Wind turbine
[32]	RM	O	–	Gearbox
[33]	RM	–	–	Gearbox
[12]	RM	–	–	Turbofan (N-CMAPSS)
[34]	RM	–	–	Turbofan (N-CMAPSS)
[13]	RM	–	–	Turbofan (N-CMAPSS)
[36]	UL	–	–	Turbofan (CMAPSS)
[37]	UL	–	–	Turbofan (CMAPSS)
[40]	UL	O	–	Bearings
[14]	UL	O	–	Battery
[15]	UL	I	–	Electric transformer
[41]	UL	I	–	Electric transformer
[42]	UL	I	–	Electric transformer
[43]	UL	O	✓	Bearings
[44]	UL	O	✓	Bearings
[45]	UL	O	✓	Bearings
[2]	UL	L	✓	Turbofan (CMAPSS)
[46]	UL	L	✓	Turbofan (CMAPSS)
[47]	UL	L	✓	Turbofan (CMAPSS)
[48]	UL	L	✓	Battery + Bearings
[49]	UL	L	✓	Turbofan (CMAPSS)
[50]	UL	L	✓	Turbofan (CMAPSS)
[51]	UL	L	✓	Turbofan (CMAPSS)
[52]	UL	L	✓	Turbofan (CMAPSS)
[10]	SL	–	–	Battery
[54]	SL	–	–	Battery
[55]	SL	–	–	Battery
[56]	SL	–	–	Materials
[57]	SL	–	–	Materials
[58]	SL	O+L	–	Materials
Proposed method	UL	I+L	✓	Turbofan (N-CMAPSS) +Battery

4. Methodology

To address the challenge of HI estimation from CM data of various complex systems, we propose a novel unsupervised hybrid method based on general knowledge about degradation. To compensate for the potential lack of information of such general knowledge, we propose combining multiple hybridization strategies. Specifically, the method incorporates two key aspects visualized in Fig. 4. We introduce an inductive bias directly into the model architecture of the method. This is achieved by utilizing an AutoEncoder whose structure is informed by the causal structure between variables involved in HI estimation. In addition, we apply a learning bias to modify the objective function used to train the model, reflecting expected degradation dynamics in complex systems. In the following section, we present more details about the inductive bias in Section 4.1 and the learning bias in Section 4.2.

4.1. Inductive bias: Derived model architecture

In the context of PHM, it is widely accepted [19–21] that the performance of systems, as expressed in the sensor readings (X), is typically influenced by both operating conditions (W) and by degradation (Z). In this study, we show this empirically by leveraging elements from causal theory. Concretely, we use the additive noise model (ANM) [59] to establish the causal relationships between the variables of interest in a directed acyclic graph (DAG) [60] (see Fig. 5). The empirical study is presented in more detail in Appendix A.

The causal direction $W \rightarrow X$ can be justified by observing that sensor readings (X) vary significantly under different operational conditions (W). For example, a commercial aircraft will go through a series of flight stages (e.g., taxiing, take-off, cruise, descend) that affect its sensor recordings. This effect is usually easily observed or recognized. The causal direction $W \rightarrow Z$ represents the influence that operational conditions have on degradation. The stresses and conditions to which a system is subject over its lifetime will have a long-term impact on the system's degradation. The third causal direction $Z \rightarrow X$ is a central assumption in prognostics: sensor readings (X), which reflect the performance of a system, are subject to changes due to degradation (Z), even though this effect may not be as pronounced as the influence of operating conditions (W). Formally, we can express the previous causal graph as a structural causal model (SCM) with assignments given as follows:

$$\text{Operational conditions } W := f_1(\epsilon_1) \tag{6}$$

$$\text{Degradation } Z := f_2(W, \epsilon_2) \tag{7}$$

$$\text{Sensor readings } X := f_3(W, Z, \epsilon_3) \tag{8}$$

where $\epsilon_1, \epsilon_2, \epsilon_3$ are jointly independent noise variables and f_1, f_2, f_3 are deterministic causal functions. It is important to recognize that these assignments (operator :=) are unidirectional, with causal variables on the right and dependent variables on the left.

The SCM implies that operational conditions (W) are an independent process. Degradation (Z) is dependent and caused by the operational conditions (W). The sensor readings (X) are caused by both operational conditions (W) and degradation (Z). As described previously, this structure can be derived empirically from observational data using Algorithm 1 in Appendix A.

Understanding the causal relationships between the variables W, Z, and X is crucial to our work because it enables the development of a more appropriate unsupervised learning architecture [61]. Even though it may appear intuitive from Assignment (7) to estimate the degradation Z from the operational conditions W, this approach is misleading. Formally, the principle of causal conditional independence dictates that knowing the distribution of a cause (W) does not provide additional insights into how W influences the effect (Z). In simpler terms, even with extensive data about W, we cannot directly infer how W affects Z because causality flows in one direction.

On the other hand, the scenario becomes more nuanced when considering "anticausal" learning. This is, if the variable previously considered as an effect (Z) becomes a cause for another variable (X), then information about the latter (X) can be informative about the former (Z). In essence, by studying the effect of Z on X, we gain indirect insights into the nature of Z itself. Note, however, that $f_3(W, Z)$ captures the combined influence of both Z and W on X. Simply estimating Z from X would inherit the confounding effect of W, making it difficult to isolate the true effect of Z on X.

To address this, we propose a specific model architecture for isolating Z. This architecture leverages an AE with an encoder $\mathcal{G}_\theta$ and decoder $\mathcal{F}_\phi$. The encoder is trained to encode X to an estimated representation of Z, denoted as $\hat{Z}$.

$$\mathcal{G}_\theta(X) \rightarrow \hat{Z} \tag{9}$$

Which is the anticausal direction permitted by the SCM structure. Thereafter, the decoder $\mathcal{F}_\phi$ uses W and $\hat{Z}$ to reconstruct X.

$$\mathcal{F}_\phi(W, \hat{Z}) \rightarrow \hat{X} \tag{10}$$

This is consistent with the expression presented in Assignment (8). The proposed AE model is given by:

$$\mathcal{F}_\phi(W, \mathcal{G}_\theta(X)) = \hat{X} \tag{11}$$

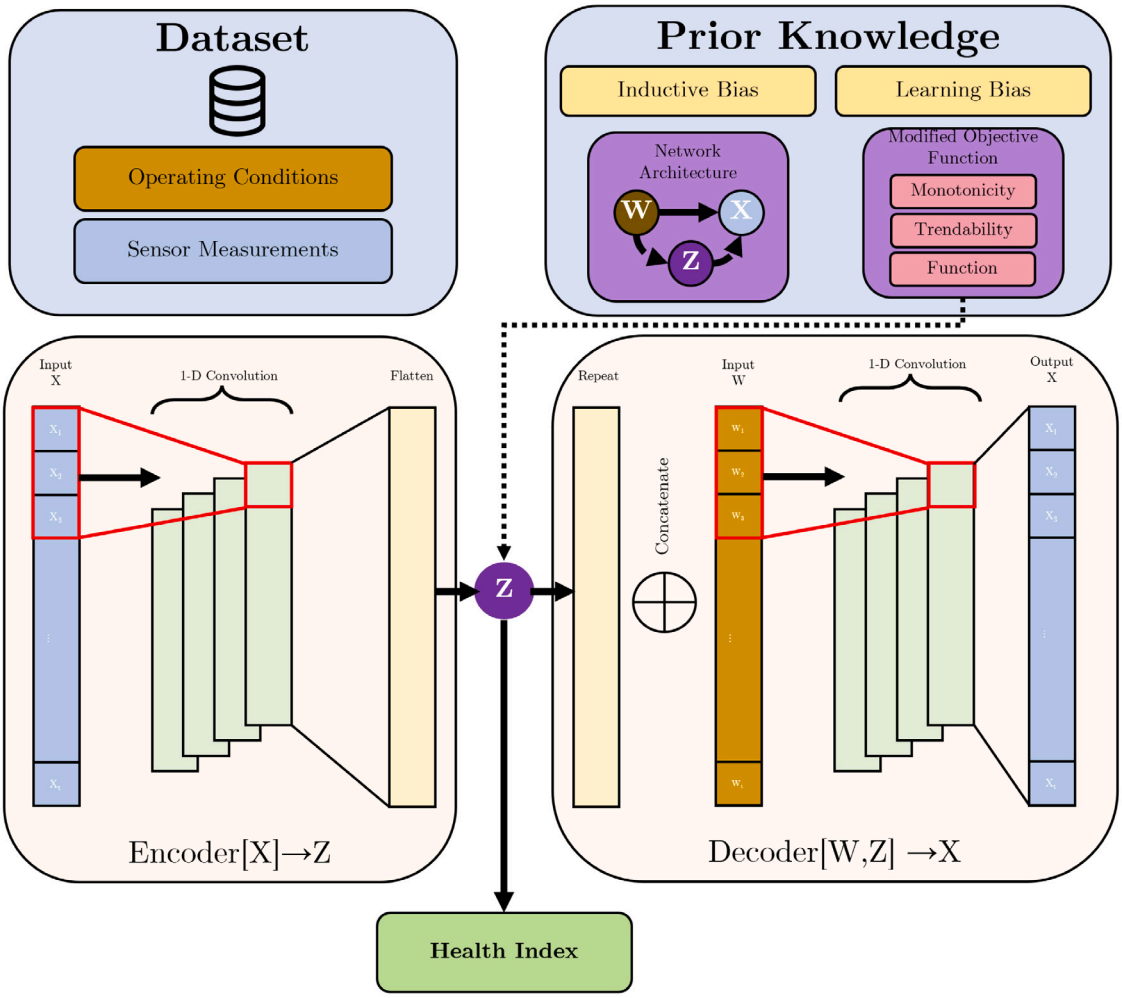


Fig. 4. Overview of the proposed unsupervised hybrid method. The proposed method utilizes an AutoEncoder whose architecture is informed by prior knowledge regarding the structure of the HI estimation problem. The encoder processes sensor readings (X) to estimate degradation (Z), while the decoder reconstructs (X) using operating conditions (W) and estimated degradation (Z). The loss function incorporates an additional constraint derived from prior knowledge about degradation's temporal evolution. Finally, the HI is derived from degradation through subsequent post-processing steps.

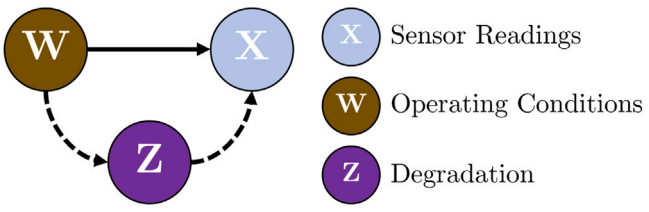


Fig. 5. Causal representation of variables in a degrading system. X denote the sensor readings, W the operating conditions and Z the degradation. Observable variables are depicted by solid lines, while dashed lines illustrate hidden variables.

It is trained with the following objective function

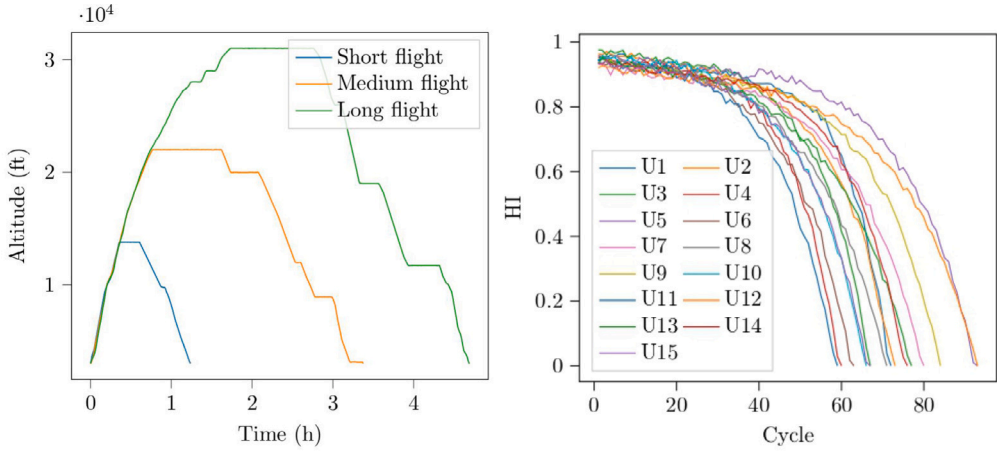
$$L_{MAE} = \frac{1}{m}\sum |X_i - \hat{X}_i| \tag{12}$$

The decoder effectively captures changes in X attributed to variations in W and Z. Because the encoder solely relies on X to derive Z, the network is compelled to learn crucial information unrelated to W. In the ablation study detailed in Section 6.3, we show that our proposed architecture performs better than a model that inputs operating conditions into both the encoder and decoder.

4.2. Learning bias: Embedding cycle information

Important degradation mechanisms in complex systems are typically dominated by operation time. For instance, in turbofan engines, degradation mechanisms such as friction, erosion, and fouling of rotating components, are dominated by cycle operation. Similarly, degradation mechanisms in batteries, such as solid–electrolyte interphase layer growth, lithium plating, or particle fracture, are also dominated by cycling [62]. Therefore, in this subsection, we present how general knowledge about the temporal dependence of the degradation can be embedded into the data-driven pipeline as a learning bias modifying the loss function of the proposed AE with a soft constraint.

We show three potential methods to incorporate the influence of operational cycles on degradation in the architecture: (1) trendability, (2) negative gradient, and (3) HI function derived from reliability theory. Each method consists of a different soft constraint that is imposed on the latent space of the AE architecture. Notably, these hybridization techniques are referred to as inductive bias because the



(a) Flight altitude conditions

(b) Health Index.

Fig. 6. Typical flight conditions and HI for units in the N-CMAPSS dataset.

soft constraints guide the latent space, enabling it to unveil degradation without overly restricting the AE's functionality. The soft constraints are implemented as an additional term in the objective function of the model with the parameter λ used to control the significance of the constraint and mitigate the risk of overfitting.

Trendability. The first method involves imposing a restriction based on the Spearman correlation between t and Z. The motivation for imposing a correlation between operation cycles and degradation lies in the necessity to establish a relationship between the increasing age of the equipment and the decreasing HI. The constraint is defined as follows:

$$L_C = \frac{\sum(t_i - \bar{t})(Z_i - \overline{Z})}{\sqrt{\sum(t_i - \bar{t})^2 \sum(Z_i - \overline{Z})^2}} \tag{13}$$

Our proposed method is trained with the following objective function:

$$L = L_{MAE} + \lambda L_C \tag{14}$$

Monotonicity. The second method is similar to the first one, but instead, we aim to constrain the gradient of Z with respect to t to be negative, that is $\frac{\partial Z}{\partial t} \leq 0$. The constraint is motivated by the understanding that the rate of Z change concerning t must be negative to derive a monotonically decreasing HI. Compared to the previously mentioned correlation constraint, the negative gradient constraint is more flexible because it does not implicitly impose a linear constraint. We define the negative gradient constraint as:

$$L_{NG} = \max\{0, \frac{\partial Z}{\partial t}\} \tag{15}$$

Our proposed method is trained with the following objective function:

$$L = L_{MAE} + \lambda L_{NG} \tag{16}$$

HI function derived from reliability theory. The third method can be used if one knows a function of the expected health index for a given cycle, i.e. $h = g(t)$. The function $g(t)$ might be known from prior knowledge or could be derived from the history of HI. We propose a method to derive $g(t)$ from the history of HIs inspired by reliability theory. For more information see Appendix B. Compared to the two previous constraints, the functional constraint is the most restrictive and system-specific, since it is individual to the specific system being investigated. We define a function $g(t)$ given by:

$$g(t) = C - ((t(log(1-P))^{-1/\beta})A)^B \tag{17}$$

where the parameters A, B, C, β are estimated from data. The constraint is then given by:

$$L_F = \frac{1}{m}\sum |g(t_i) - Z_i| \tag{18}$$

Our proposed method is trained with the following objective function:

$$L = L_{MAE} + \lambda L_F \tag{19}$$

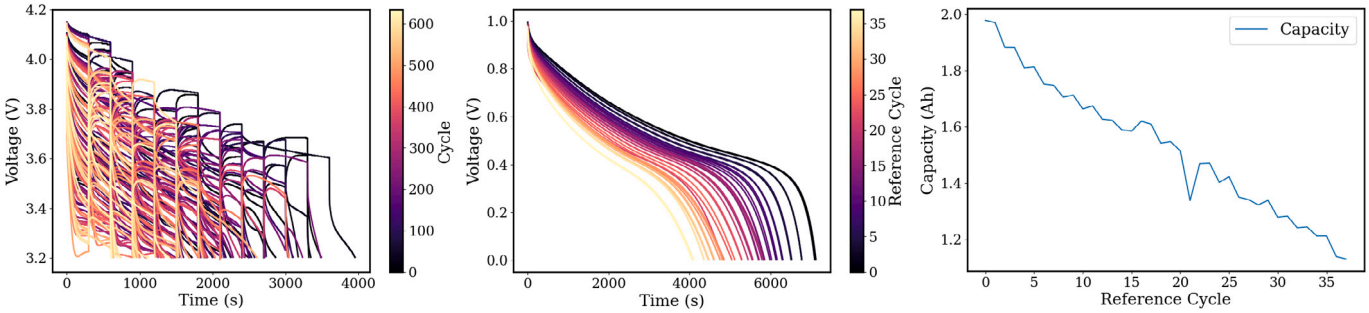
5. Case studies

The proposed method is demonstrated and evaluated in two case studies featuring distinct complex systems. These case studies vary in multiple aspects, including the obvious difference between an electrochemical process and a thermo-kinetic process. but also the rate of degradation with respect to operational cycles and the manifestation of the impact of degradation. By examining these diverse scenarios, we aim to illustrate the robustness and applicability of our proposed method across multiple complex systems.

Section 5.1 presents the airplane turbofan engine case study, and Section 5.2 introduces the lithium battery case study. In Section 5.3, we describe the preprocessing of the datasets. Section 5.4 describes how the HI is extracted for the considered methods. The implemented network architectures are briefly presented in Section 5.5. Finally, we discuss how to evaluate HI estimation methods in Section 5.6.

5.1. Dataset: Aircraft turbofan engines

The new Commercial Modular Aero-Propulsion System Simulation (N-CMAPSS) dataset [63] provides full run-to-failure degradation trajectories of turbofan engines. From the eight available data subsets within the N-CMAPSS dataset, we consider the set DS003, which is characterized by a failure mode that affects the low-pressure turbine efficiency and flows in combination with the high-pressure turbine efficiency. Each unit within the fleet contains 14 observable sensor measurements, denoted as X, which are recorded from an initial health condition until engine failure (run-to-failure data). In addition to the sensor measurements, the corresponding 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, medium-length flights, or long-length flights. Fig. 6(a) displays the typical altitude conditions for randomly selected flight cycles and units.



(a) Random discharge cycles. (b) Reference discharge cycles. (c) Calculated capacity.

Fig. 7. Example of random discharge curves available for model training, and reference discharge curves used to calculate capacity values and establishing ground truth HI.

The N-CMAPSS dataset models degradation at the component level through initial, normal, and abnormal degradation stages. As a result of degradation, a HI is computed in the form of a non-linear mapping of multiple operational margins taken at reference conditions. The resulting HI was used to declare system failure when its value reached 0. The dataset also incorporates between-flight maintenance by permitting improvements in engine health parameters within allowable limits. This ground truth HI (i.e., h_{GT}) is available for DS03, and will be used for evaluation purposes only. Fig. 6(b) illustrates the ground truth Health Index (HI) for the fleet of units.

5.2. Dataset: Li-ion batteries

The randomized battery usage dataset from the NASA Ames Prognostics Center of Excellence repository [64] was considered to further evaluate the proposed methodology. The dataset encompasses data from individual 18650 LCO cells, each subjected to cycles of charging and discharging under randomized protocols.

The most common physical aging mechanisms observed in batteries are graphite exfoliation, loss of electrolyte, solid–electrolyte interface layer formation, continuous thickening, lithium plating, etc. [65]. Consequently, the battery aging process gives rise to two primary changes to the battery electrodynamics, which impact its performance: capacity fade and increase in internal resistance. In this particular case study, we will focus on capacity fade as the selected HI for the batteries under examination.

To simulate real-world battery usage scenarios, we only consider the randomized battery discharge cycle data. For each battery in the dataset, voltage and temperature measurements (X) were recorded during different operational conditions (W), defined by the applied current during the discharge process. Random discharge cycles are illustrated in Fig. 7(a).

To estimate the ground truth HI, we analytically calculate capacity values from reference discharge cycles with constant current. The current capacity Q of the battery is calculated as $Q = \frac{1}{3600}\int W(t)\,dt$. Illustrated in Fig. 7(b) are the constant current reference discharge cycles, and in Fig. 7(c) the calculated capacity values for each corresponding cycle.

In battery-related contexts, it is more common to use the notation of State of Health (SOH) instead of the HI. The SOH is defined as the ratio between the present capacity and the nominal capacity ($SOH = Q/Q_{nominal}$). In this paper, a failure of a battery will be defined once SOH is less than 60%. Since HI and SOH have similar meanings, the terms will be used interchangeably in this paper.

5.3. Pre-processing

In all our experiments, we initially performed min–max normalization on both X and W to scale their values within the range [0,1]. To improve computational efficiency, we also reduced the data sampling frequency. Specifically, for the turbofan dataset, we decreased the frequency from 1 Hz to 0.1 Hz and from 1 Hz to 0.5 Hz for the battery dataset.

Following these pre-processing steps, the next stage involved segmenting the data into fixed-length windows of size S. In the context of optimizing both the residual and supervised methods, we conducted a grid search to identify the optimal window size for each. Consequently, for the residual method and the supervised model, we employed a sliding window with a length of $S = 50$ for the turbofan dataset and $S = 200$ for the battery dataset. In contrast, our proposed method involved windowing the data based on individual operational cycles. Since operational cycles can vary in duration, we employed padding with zeros at the end of each cycle to equalize their lengths. The window size was determined by selecting the minimum integer that aligned with the length of the longest cycle. For the turbofan dataset $S = 2030$, and the battery dataset $S = 2160$.

Windowing whole operational cycles is the preferred methodology since it makes post-processing of the HI straightforward. For the residual method and supervised model windowing whole cycles was not possible since it created unwanted artifacts due to the zero-padding. It is also worth noting that our proposed method also works with shorter window sizes, but as previously mentioned due to the nature of easier HI construction, we have opted to window whole cycles. This will be discussed in more detail in Section 5.4.

5.4. Constructing the HI

In the context of our research, the extraction of HIs depends on the chosen methodology and the specific case study at hand. In the supervised model, the HI is directly estimated from the model's output. In contrast, the residual method requires a multi-step process. We first compute the residual error for each prediction window. Next, we flatten each window and employ the Principal Component dimensionality reduction technique, effectively transforming the residual vector, denoted as $r \in \mathbb{R}^p$, into a $\mathbb{R}$. Since the supervised and the residual methods use short window sizes, which also have overlapping observations, we also average the resulting HI per cycle.

Meanwhile, our proposed method offers a more streamlined HI extraction process. Degradation is extracted directly from the output of the encoder, i.e., the latent layer corresponding to Z in Fig. 4. Since whole operational cycles are used as input, there is no need to smooth the resulting HI.

Table 2
Hyperparameters of investigated methods.

Dataset	Model	Window size	Epochs	Batch size	Learning rate
CMPAPSS	Supervised	50	20	512	1e−4
	Residual	50	20	512	1e−4
	Proposal	2030	20	20	1e−4
Battery	Supervised	200	20	1024	1e−4
	Residual	200	20	1024	1e−4
	Proposal	2160	20	128	1e−4

For the turbofan dataset, we normalize the resulting HI to be within $[0, 1]$. For the battery dataset, we normalize the resulting HI to be within $[0\%, 100\%]$. As mentioned earlier, the resulting HI for the battery dataset corresponds to SOH, a more frequently used metric for expressing the health of the battery in literature.

5.5. Network architectures

In this section, the network architectures of the considered HI estimation methods are described.

Residual method — AE. The asymmetric-AE residual model is shown in Fig. 3(a). The model is trained to reconstruct X when W and X are used as input. The architecture of the asymmetric-AE residual model used here comprises four identical 1-D CNN layers with 64 filters of size 11 and with ReLU activation functions.

Residual method — Regression. The regression type residual model implemented in this study is shown in Fig. 3(b). We train a model to predict X given W as input. The model contains four identical 1-D CNN layers with 64 filters of size 11 and with ReLU activation functions.

Proposed method — Unsupervised Hybrid AE. The structure of the proposed model is shown in Fig. 4. The model is composed of two parts: an encoder and a decoder. Both of these parts are built using 1D-convolution layers. Specifically, the encoder takes as input X and passes it through three 1-D convolution layers with a number of filters equal to [128,64,16]. Afterward, the output of the encoder is flattened and passed through a fully connected layer with one neuron. The output of the fully connected layer is Z. The decoder concatenates Z and W as input and passes it through three 1-D CNN layers with a number of filters equal to [16,64,128]. The last layer of the decoder is a fully connected layer with a number of neurons equal to the dimensionality of X. The loss function of the model is modified based on the chosen general knowledge. We do not perform any fine-tuning of the constraint parameter λ and set it to 1.

Supervised model. We train a supervised model for HI estimation inspired by the architecture of [55]. The model uses as input X and W to predict the HI in a supervised manner. The model contains four identical 1-D convolution layers with the number of filters set to 64 and a kernel width of 11. Each convolution layer is followed by batch normalization and a max pool layer of size 2. The output of the final convolution layer is flattened and then passed through a dense layer with the output size of 1, corresponding to the HI.

The optimization of the network's weights is carried out with mini-batch stochastic gradient descent (SGD) and with the Adam algorithm. The training hyper-parameters for each model are given in Table 2.

5.6. Evaluation

Based on the state-of-the-art HI evaluation methodology [66], we compare and analyze the performance of the proposed method for HI estimation based on two evaluation aspects: quality of the HI and impact on prognostic performance when the HI is used for RUL prediction task. For each of the two aspects, we consider evaluation metrics that are defined in the following sections. An overview of the evaluation methodology is given in Fig. 8.

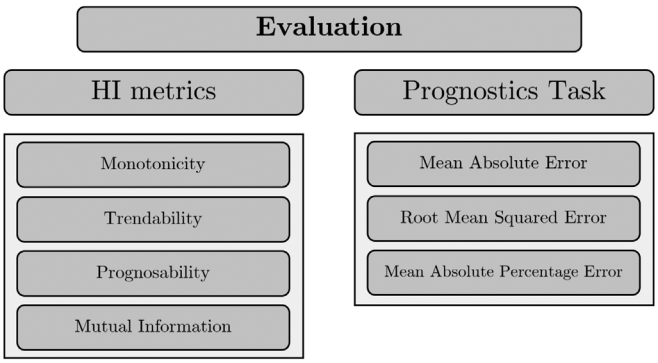


Fig. 8. Evaluation methodology overview.

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 estimated HIs, it is still desirable for them to demonstrate consistent behavior.

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

- **Monotonicity (Mon)** measures the tendency for the HI to consistently increase or decrease [67,68].
- **Trendability (Tren)** is used to evaluate the degree to which the HIs of a fleet of systems have a similar shape and underlying form [67,68].
- **Prognosability (Prog)** is used to evaluate consistent HI behavior towards the end of life of units [67,68].
- **Mutual Information (MutInf)** score quantifies the information obtained about RUL by observing HI [66].

For more information about each of the HI criteria see Appendix C.

5.6.2. Prognostic performance

A key objective of HI estimation is to enhance the performance of prognostic models. To validate the effectiveness of the proposed HI estimation techniques, a baseline prognostic model is needed. The sensor signals, operating conditions, and cycles are used as inputs to predict RUL. The model is given by:

$$G(X, W, t) = RUL \tag{20}$$

To test whether the estimated HIs increase prognostic performance, HIs are used to augment the input space.

$$G(X, W, t, h) = RUL \tag{21}$$

The chosen RUL model for both case studies is based on a 1D-CNN architecture, as outlined in the work by Chao et al. [69]. The parameters of the model are kept fixed for all experiments.

In particular, for the turbofan dataset, the CNN architecture includes five layers. Three initial convolutional layers utilize filters of size 10. The first two convolutions have ten channels, and the last convolution has only one channel. Zero padding is applied to maintain the feature map throughout the network. The resulting 2D feature map is flattened, leading to a 50-way fully connected layer followed by a linear output neuron. ReLu serves as the activation function for the network. The window length matches that of the residual and supervised HI

Table 3
Overview of the results section.

Experiment	Turbofan	Battery
In distribution	Section 6.1.1	Section 6.1.2
Out of distribution	Section 6.2.1	Section 6.2.2

Table 4
In-distribution training set-up for N-CMAPSS dataset.

Flight class	Training units	Testing units
Short	U1, U5, U9	U12, U14
Medium	U2, U3, U4, U7	U15
Long	U6, U8	U10, U11, U13

Table 5
In-distribution training set-up for NASA battery dataset.[a]

Load profile	Training batteries	Testing batteries
Uniform	RW4, RW5	RW6
Uniform	RW1, RW7	RW8
Skewed high	RW17, RW19	RW20
Skewed low	RW13, RW14, RW15	RW16

[a] Certain subsets of batteries were omitted from consideration: one due to a complex charging process, two because experiments were conducted at 40 °C external temperature while the rest were at room temperature, and three (RW2, RW3, and RW18) due to corrupted temperature readings.

estimation models (s = 50). Similarly, the battery dataset adopts a comparable CNN architecture to the turbofan dataset. The distinction lies in all three convolution layers having ten channels, and the last fully-connected layer has a size of 200 (matching the window length).

Prognostic assessments are commonly based on metrics such as mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). Furthermore, we compute '% Average Improvement' which denotes the performance improvement over the baseline model upon inclusion of HI, with the same consideration being given to MAE, RMSE, and MAPE.

6. Results

In this section, we analyze the performance of the proposed model on two case studies: the turbofan dataset and the battery dataset. We evaluate a situation where the training and test datasets are of the same distribution in Section 6.1. Furthermore, we investigate HI estimation performance in the context of out-of-distribution testing in Section 6.2. The resulting structure of this section is shown in Table 3.

6.1. In-distribution testing

The initial set of experiments will focus on the in-distribution case, where the term "in-distribution" refers to training and testing data originating from similar data distribution. The N-CMAPSS dataset encompasses engines classified into three distinct flight classes. The operational conditions of engines are influenced by their assigned flight class, subsequently impacting degradation patterns. To create a balanced dataset for training and testing we use the data split given in Table 4.

Similarly, for the battery dataset, batteries undergo discharging with randomized loads selected from a given uniform distribution. These batteries are categorized into three load classes based on the selected load distribution: uniform, skewed high, and skewed low. The load class significantly influences the operational conditions of the batteries, consequently affecting degradation patterns. We utilize four subsets for training and testing our proposed method. Table 5 provides a summary of the data subsets, along with the division into training and test sets.

Table 6
Results of the HI criteria for the turbofan dataset. Mean and standard deviation values over 5 runs are provided. h_r^a — health index of AE residual method, h_r^b — health index of regression residual method, h_p^C — health index of the proposed method using correlation constraint, h_p^{NG} — health index of the proposed method using negative gradient constraint, h_p^F — health index of the proposed method using functional constraint, h_s — health index of supervised model, h_{gt} — ground truth health index.

HI	Mon	Tren	Prog	MutInf	MAPE
Residual method					
h_r^a	0.18(0.04)	0.79(0.08)	0.85(0.05)	0.50(0.06)	24.4(4.3)
h_r^b	0.33(0.03)	0.91(0.03)	**0.98(0.01)**	0.62(0.02)	10.9(1.5)
Proposed method					
h_p^C	0.33(0.02)	**0.98(0.00)**	0.93(0.01)	**0.81(0.01)**	8.8(0.7)
h_p^{NG}	**0.36(0.05)**	**0.98(0.00)**	0.95(0.01)	**0.81(0.01)**	**8.5(1.6)**
h_p^F	0.28(0.01)	0.97(0.01)	0.85(0.03)	0.66(0.02)	9.5(1.0)
Supervised					
h_s	0.40(0.03)	0.99(0.00)	0.94(0.01)	0.79(0.01)	8.2(1.4)
Ground truth					
h_{gt}	0.50	0.99	1.0	0.84	–

6.1.1. In-distribution testing — Turbofan

The HI metrics obtained from the six evaluated models are shown in Table 6. We also report the mean absolute percentage error (MAPE) between the estimated HI and the ground truth HI.

Comparing residual-based methods, both trained with CM data from the initial 20 flight cycles of each training unit, we observe that the regression-based residual method (h_r^b) outperforms the AE-based method (h_r^a) in all the metrics.

Analyzing the proposed method, we find that the correlation constraint (h_p^C) and the negative gradient constraint (h_p^{NG}) demonstrate comparable performance across various HI metrics. In contrast, the function-based constraint (h_p^F) exhibits relatively poorer performance in terms of mutual information and MAPE.

It is worth noticing that the proposed methods consistently outperform the residual methods across all metrics, except for prognosability. This trend is further supported by the MAPE score, indicating that the proposed methods yield HIs more closely aligned with the ground truth compared to the residual methods. Furthermore, when considering a supervised model (h_s) for HI estimation as an alternative to our proposed unsupervised method, we observe only marginal performance enhancements.

In addition to the quantitative evaluation of the HI metric, Fig. 9 depicts the estimated HI for test unit 10 in the considered methods. This unit is randomly selected for visualization purposes. The proposed method utilizing the negative gradient constraint (i.e., h_p^{NG}) results in an HI showing a closer match to the ground truth HI (h_{gt}) than the other methods considered.

The prognostic prediction performance obtained when the prognostics model trained with the HI estimated with six evaluated methods is shown in Table 7. The results show improved prognostics when integrating ground truth HI into the training data set. On average, the model shows a 32% improvement compared to the baseline model, which underlines the importance of HI as a valuable source for the prediction of RUL. Subsequently, the second most effective option is the supervised model, resulting in a 29% performance enhancement. The proposed method yields comparable results, with the proposed negative gradient constraint showing a notable improvement of 28%. It is worth noting that the inclusion of HI estimated by the residual method presents the least favorable scenario, yielding an improvement of 24%.

6.1.2. In-distribution testing — Batteries

The HI metrics obtained from the six evaluated models are shown in Table 8. Comparing residual-based methods, both trained with CM data from the initial 100 random walk cycles of each training unit,

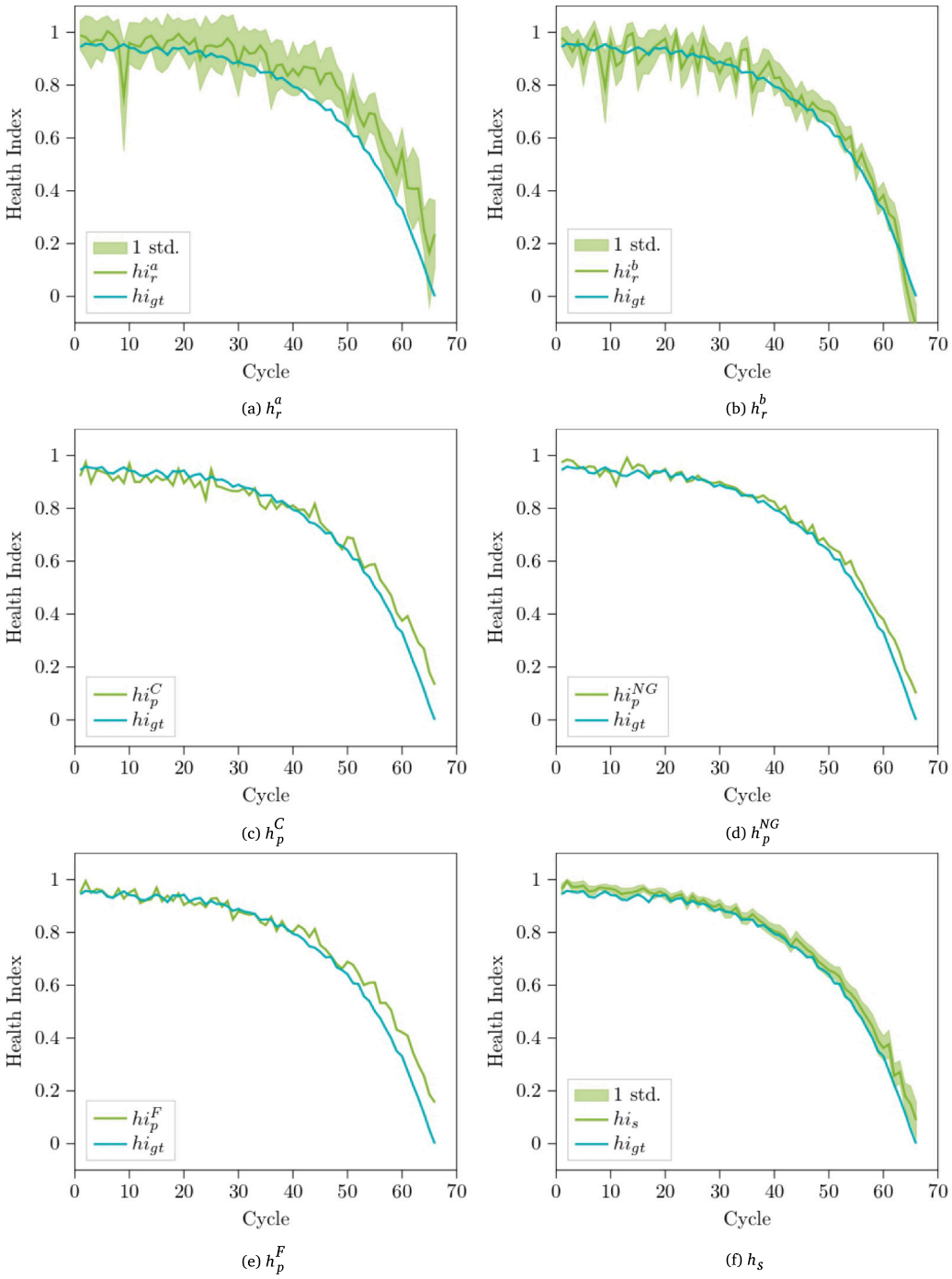


Fig. 9. Estimated HI of test unit 10 for the turbofan dataset. Residual methods (a) and (b). Proposed methods (c), (d), and (e). Supervised model (f).

we observe that the AE-based method outperforms the regression-based residual method in all the metrics.

Among the proposed methods, both the correlation constraint and the functional constraint exhibit superior performance, showcasing equivalent results. Conversely, the negative gradient constraint displays comparatively inferior performance across the HI criteria.

Compared to residual methods, the proposed methods incorporating correlation and functional constraints exhibit superior performance. Additionally, when comparing the proposed methods to the supervised model, an overall equivalency in performance is observed, albeit with notable distinctions. While the supervised model shows significantly improved results in terms of MAPE, the other criteria demonstrate a nearly identical performance.

Fig. 10 depicts the estimated HI for test battery 20 in the considered methods. This battery is randomly selected for visualization purposes. The HI estimated by the supervised model seems to be a closer match to the ground truth HI than HIs estimated by the other methods. The proposed methods employing the correlation or functional constraint are the next best choice.

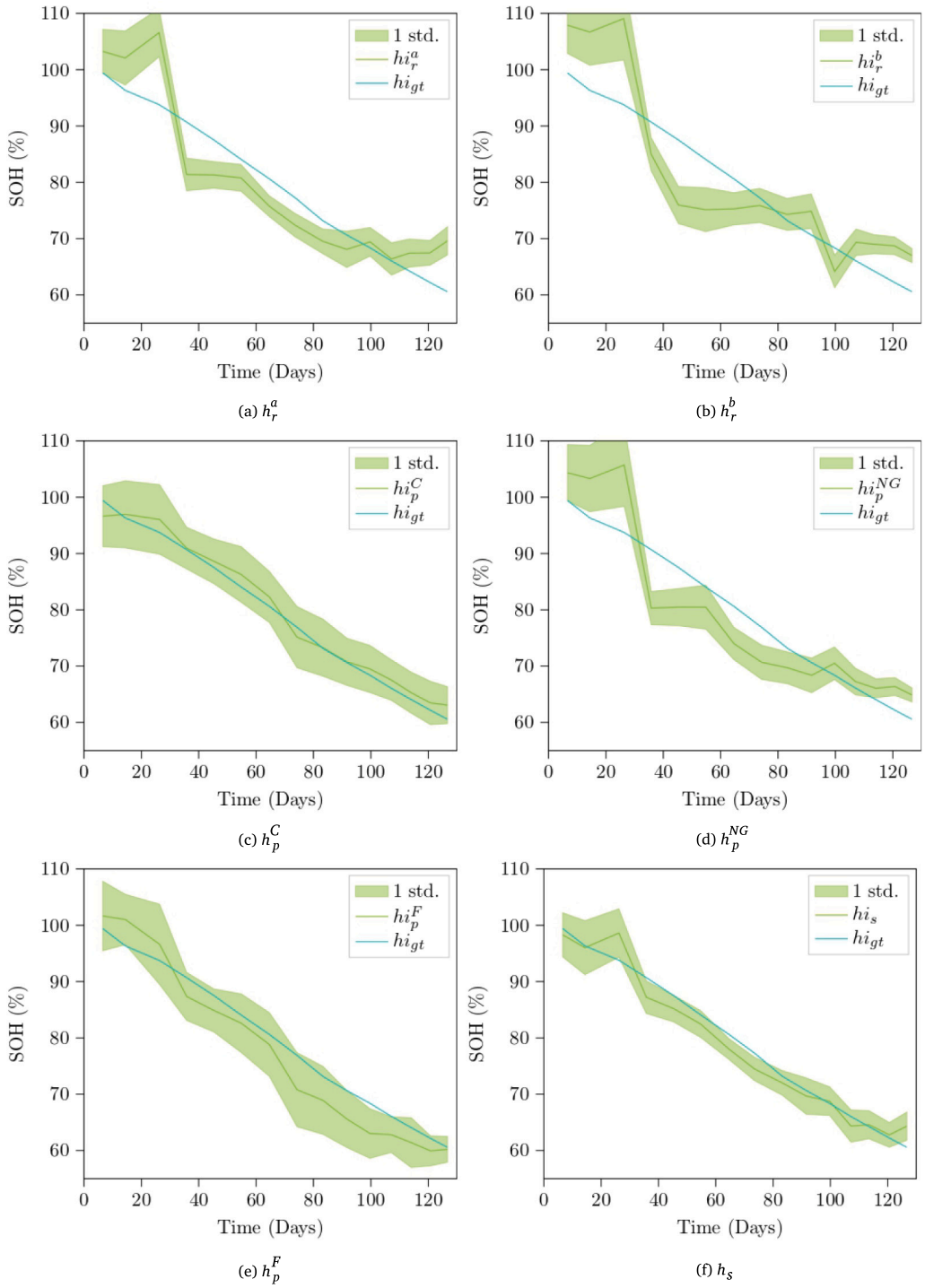


Fig. 10. Estimated HI of test unit 20 for the battery dataset. Residual methods (a) and (b). Proposed methods (c), (d), and (e). Supervised model (f).

The prognostic prediction performance obtained when the prognostics model trained with the HI estimated with six evaluated methods is shown in Table 9. The results show improved prognostics when integrating ground truth HI into the training data set. On average, the model shows a 54% improvement compared to the baseline model. Subsequently, the second most effective option is the supervised model, resulting in a 26% performance enhancement. The results of the proposed method yield comparable results, with the proposed functional constraint showing a notable improvement of 25%. The other two constraints, namely correlation and negative gradient, lead to 20% and 11%, respectively. Using the residual method leads to 14% improvement.

6.2. Out-of-distribution testing

The accuracy and reliability of prognostic prediction techniques hinge on the quality and representativeness of available time-to-failure data. As a result, these techniques may exhibit reduced performance when applied to data from new units operating under conditions

Table 7
Results of the prognostic prediction task for the turbofan dataset. Mean and standard deviation values over 5 runs are provided. G — neural network, X — sensor readings, W -operating conditions, t — cycles, h_r^b — health index of regression residual method, h_p^C — health index of the proposed method using correlation constraint, h_p^{NG} — health index of the proposed method using negative gradient constraint, h_p^F — health index of the proposed method using functional constraint, h_s — health index of supervised model, h_{gt} — ground truth health index.

Model	MAE	RMSE	MAPE	% Average improvement
Baseline model				
$G(X,W,t)$	6.0(0.4)	7.4(0.3)	30.5(4.8)	–
Residual method				
$G(X,W,t,h_{re}^b)$	4.9(0.2)	6.7(0.1)	16.1(1.2)	24%
Proposed method				
$G(X,W,t,h_p^C)$	**4.7(0.1)**	6.9(0.2)	15.0(0.4)	26%
$G(X,W,t,h_p^{NG})$	4.9(0.1)	**6.5(0.1)**	**14.4(0.6)**	**28%**
$G(X,W,t,h_p^F)$	4.8(0.1)	6.6(0.2)	15.0(0.4)	27%
Supervised				
$G(X,W,t,h_s)$	4.6(0.0)	6.4(0.1)	15.6(0.7)	29%
Ground truth				
$G(X,W,t,h_{gt})$	4.6(0.0)	6.3(0.1)	13.1(0.3)	32%

Table 8
Results of the HI criteria for the battery dataset.

HI	Mon	Tren	Prog	MutInf	MAPE
Residual method					
h_{re}^a	0.66(0.07)	0.96(0.02)	0.78(0.01)	0.53(0.04)	9.1(1.0)
h_{re}^b	0.50(0.07)	0.64(0.02)	0.66(0.01)	0.34(0.01)	11.0(0.2)
Proposed method					
h_p^C	**0.86(0.04)**	**0.99(0.00)**	**0.83(0.02)**	**0.63(0.00)**	7.3(0.6)
h_p^{NG}	0.39(0.08)	0.84(0.09)	0.83(0.11)	0.43(0.07)	9.6(1.0)
h_p^F	0.73(0.04)	0.98(0.00)	0.82(0.01)	0.62(0.00)	**7.0(0.5)**
Supervised					
h_s	0.94(0.08)	0.98(0.08)	0.88(0.08)	0.62(0.02)	3.1(0.2)
Ground truth					
h_{gt}	0.97	1.00	0.97	0.65	–

Table 9
Results of the prognostic prediction task for the battery dataset.

Model	MAE	RMSE	MAPE	% Average improvement
Baseline				
$G(X,W,t)$	165(15)	206(17)	256(16)	–
Residual method				
$G(X,W,t,h_{re}^a)$	144(11)	180(14)	215(19)	14%
Proposed method				
$G(X,W,t,h_p^C)$	134(5)	166(5)	202(13)	20%
$G(X,W,t,h_p^{NG})$	150(5)	185(5)	222(16)	11%
$G(X,W,t,h_p^F)$	**126(10)**	**161(13)**	**180(14)**	**25%**
Supervised				
$G(X,W,t,h_s)$	134(5)	167(8)	153(15)	26%
Ground truth				
$G(X,W,t,h_{gt})$	96(8)	113(9)	67(7)	54%

distinct from those in the training set [70], leading to an Out-of-Distribution (OOD) scenario. To assess the capabilities of the proposed method under such OOD scenarios, we intentionally designed challenging scenarios for each dataset.

For the turbofan dataset, we create an OOD scenario by considering different flight classes. In particular, we train models with short-flight class data and conduct tests on medium to long-flight classes. The specific training and testing units are detailed in Table 10.

Table 10
Out-of-distribution training set-up for N-CMAPSS dataset.

Flight class	Training units	Testing units
Short	U1, U5, U9, U12, U14	–
Medium	–	U2, U3, U4, U7, U15
Long	–	U6, U8, U10, U11, U13

Table 11
Out-of-distribution training set-up for NASA battery dataset.

Load profile	Training batteries	Testing batteries
Uniform	RW1, RW4–RW8	–
Skewed high	–	RW17, RW19, RW20
Skewed low	–	RW13–RW16

Table 12
Results of the HI criteria for the turbofan dataset under out-of-distribution scenario.

HI	Mon	Tren	Prog	MutInf	MAPE
Residual method					
h_r^b	0.12(0.03)	0.68(0.05)	0.86(0.03)	0.45(0.07)	25.1(3.6)
Proposed method					
h_p^C	0.10(0.03)	0.75(0.12)	0.74(0.13)	0.57(0.10)	34.2(5.9)
h_p^F	**0.16(0.03)**	**0.91(0.06)**	**0.89(0.05)**	**0.68(0.06)**	**16.6(3.5)**
Supervised					
h_s	0.11(0.04)	0.80(0.05)	0.88(0.04)	0.55(0.06)	22.8(4.4)
Ground truth					
h_{gt}	0.53	0.99	1.0	0.84	–

For the battery dataset, we create an OOD scenario by considering different load profiles. We train models using uniform load data and conduct tests using low-skew and high-skew data. For more information on the training/testing units see Table 11.

6.2.1. OOD testing — Turbofan

Table 12 illustrates the HI metrics in the out-of-distribution scenario. The proposed method, incorporating the functional constraint, surpasses the residual-based method across all metrics. However, the use of the correlation constraint does not yield improved performance compared to the residual method.

Moreover, when evaluating a supervised model for HI estimation, only marginal performance enhancements are observed in comparison to both the residual method and the proposed method with the correlation constraint. The supervised model demonstrates inferior performance across all metrics when compared to the proposed method utilizing the functional constraint.

For the prognostics prediction task in the out-of-distribution scenario, we decided to cap the maximum RUL value to 60 cycles to prevent large prediction errors for healthy units. The results are given in Table 13.

The results show the improvement in prognostics when integrating ground truth HI into the training data set. On average the model shows 47% improvement when incorporating the ground truth HI compared to the baseline model. The second most effective option was integrating the HI estimated by the supervised model. However, the inclusion of HI estimated by the proposed method, utilizing a functional constraint, showed a similar performance boost (41% compared to 37%). Comparatively, the inclusion of HI estimated by the residual method and the proposed method employing the correlation constraint result in marginal improvements of 5% and 9%, respectively.

Table 13
Results of the prognostics prediction task for the turbofan dataset under out-of-distribution scenario.

Model	MAE	RMSE	MAPE	% Average improvement
Baseline model				
$G(X, W, t)$	11.2(1.4)	13.9(1.4)	56.1(14.4)	–
Residual method				
$G(X, W, t, h_r^b)$	11.2(1.1)	14.4(1.2)	44.9(10.3)	5%
Proposed method				
$G(X, W, t, h_p^C)$	10.4(0.9)	12.8(1.1)	50.0(12.7)	9%
$G(X, W, t, h_p^F)$	**6.8(0.6)**	**9.1(1.0)**	**35.8(7.1)**	**37%**
Supervised				
$G(X, W, t, h_s)$	7.0(1.0)	9.1(1.4)	27.4(5.0)	41%
Ground truth				
$G(X, W, t, h_{gt})$	6.2(0.6)	8.5(0.6)	23.3(3.4)	47%

Table 14
Results of the HI criteria for the battery dataset under out-of-distribution scenario.

HI	Mon	Tren	Prog	MutInf	MAPE
Residual method					
h_r^a	**0.76(0.06)**	0.97(0.00)	**0.85(0.04)**	0.58(0.05)	9.5(0.5)
Proposed method					
h_p^C	0.41(0.11)	0.87(0.10)	0.71(0.04)	0.45(0.13)	9.0(0.9)
h_p^F	0.61(0.07)	**0.97(0.01)**	0.79(0.02)	**0.61(0.02)**	**8.0(0.3)**
Supervised					
h_s	0.71(0.04)	0.93(0.04)	0.77(0.13)	0.58(0.04)	6.4(1.3)
Ground truth					
h_{gt}	0.98	0.99	0.97	0.68	–

Table 15
Results of the prognostics prediction task for the battery dataset under out-of-distribution scenario.

Model	MAE	RMSE	MAPE	% Average improvement
Baseline model				
$G(X, W, t)$	52(8)	85(18)	142(37)	–
Residual method				
$G(X, W, t, h_r^a)$	52(5)	68(6)	89(12)	19%
Proposed method				
$G(X, W, t, h_p^C)$	46(3)	71(3)	83(6)	23%
$G(X, W, t, h_p^F)$	**41(6)**	**64(9)**	**77(6)**	**31%**
Supervised				
$G(X, W, t, h_s)$	43(5)	69(5)	74(4)	28%
Ground truth				
$G(X, W, t, h_{gt})$	21(1)	36(4)	26(7)	66%

Table 16
Results of the HI criteria for the turbofan dataset ablation study.

Mon	Tren	Prog	MutInf	MAPE
Proposed method				
0.36(0.05)	0.98(0.00)	0.95(0.01)	0.81(0.01)	8.5(1.6)
With inductive bias and without learning bias				
0.11(0.02)	0.77(0.06)	0.93(0.03)	0.41(0.08)	26.4(6.5)
Without inductive bias and with learning bias				
0.31(0.03)	0.98(0.00)	0.95(0.02)	0.69(0.01)	10.7(2.4)
Without inductive bias and learning bias				
0.05(0.01)	0.01(0.01)	0.03(0.03)	0.03(0.01)	89.6(6.0)

6.2.2. OOD testing — Batteries

Table 14 presents the HI metrics for the out-of-distribution scenario concerning the battery dataset. In comparison to the residual method, the proposed method using the functional constraint exhibits superior performance in trendability, mutual information, and MAPE. However, the proposed method using the correlation constraint performs worse than the residual method across all metrics except MAPE.

The HI metrics show a minimal performance gap between the supervised model and the proposed method utilizing the functional constraint. The proposed method with the functional constraint outperforms the supervised model in terms of trendability and mutual information, but scores lower in terms of monotonicity and MAPE.

For the prognostics prediction task in the out-of-distribution scenario, we decided to cap the maximum RUL value to 300 cycles to prevent large prediction errors. The results are given in Table 15.

Introducing the ground truth HI significantly enhances RUL prediction, achieving an average improvement of 66%. The next best approach involves integrating an HI estimated through the proposed method using the function constraint, resulting in a notable performance increase of 31%. Incorporating an HI estimated by the supervised model closely follows, showcasing an improvement of 28%. Conversely, integrating an HI estimated by the proposed method using the correlation constraint or the residual method yields more modest improvements, increasing performance by 23% and 19%, respectively.

6.3. Ablation study: Impact of hybridization techniques

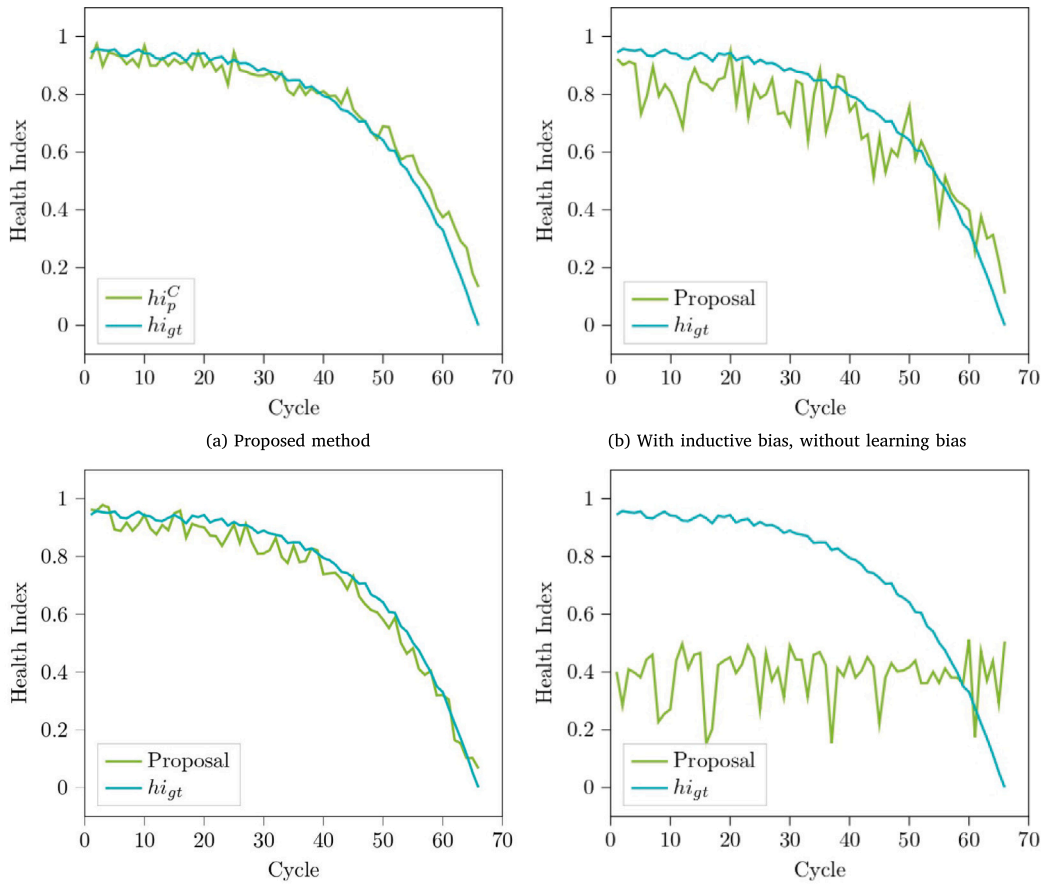
This section expands on the prior analysis by conducting an ablation study to show the advantages of incorporating multiple hybridization strategies in the proposed method. Specifically, the proposed method incorporates two hybridization strategies: inductive bias and learning bias. To assess the independent impact of each hybridization strategy on accurate HI estimation, we systematically eliminate each bias from the model.

Initially, we examine the effect of removing the learning bias by setting the parameter λ, which controls the importance of the additional constraint, to 0. Subsequently, we investigate the effect of eliminating the inductive bias while preserving the learning bias by employing a convolutional AE where W and X serve as input to reconstruct X. The architecture mirrors that of the proposed method, with the distinction that the operating conditions are input to the encoder and the decoder (i.e., see UL method in Fig. 3). Lastly, we eliminate both learning bias and inductive bias, essentially creating a fully data-driven unsupervised model. This is achieved by employing the same architecture as in the second case but omitting the additional constraint term from the objective function.

The experiments focus on the in-distribution case of the turbofan dataset; the resulting HI metrics are shown in Table 16. In the initial scenario of eliminating the learning bias, a substantial decline in HI metrics is evident compared to the proposed method. In the subsequent case of removing the inductive bias, we present the outcomes of incorporating the correlation constraint. The results indicate a less pronounced decrease in HI metrics compared to the proposed method, with the most significant impact observed in Mutual Information and MAPE. Finally, we demonstrate that integrating no prior knowledge into the model results in the worst estimation of the HI. Fig. 11 visually represents the estimated HIs for each ablation experiment.

7. Discussion

This research aimed at achieving a reliable hybrid method for estimating the HI of diverse complex systems. In pursuit of this objective, we proposed a hybrid unsupervised method for HI estimation based on general knowledge. To compensate for the low informative nature of the prior knowledge, we opted to combine multiple hybridization strategies.



(a) Proposed method

(b) With inductive bias, without learning bias

(c) Without inductive bias, with learning bias

(d) Without inductive bias, without learning bias

Fig. 11. Estimated HI of test unit 10 for the turbofan dataset. Proposed method (a) Proposed method. (b) Proposed method without learning bias. (c) Proposed method without inductive bias. (d) Data-driven model.

To validate the generality of our proposed method, we conducted evaluations on two distinct case studies, each characterized by different degradation dynamics and their respective manifestations in observable sensor readings. The main findings are summarized as follows:

- The proposed method outperformed the industry standard residual method in both case studies. Furthermore, the performance of the proposed method was on par with that of a supervised model. This suggests that incorporating general knowledge contributes significantly to the method's performance across diverse systems.
- Among the tested soft constraints, the functional constraint emerged as the most effective choice for both case studies. The functional constraint uses the most system-specific knowledge and compels the latent space of the model to adhere to specific values.
- The advantage of the correlation and negative gradient constraints varied depending on the case study. In the case of turbofan engines characterized by highly non-linear degradation, the negative gradient constraint proved more effective. Conversely, for batteries exhibiting linear degradation and health recovery aspects, the correlation constraint demonstrated better performance.
- Ablation analysis (Section 6.3) shows that both hybridization strategies are effective for HI estimation, but the combination of the two is advantageous.

The aforementioned findings show that our proposed method has good generalization since it can accurately estimate the HI of various systems. However, it is important to acknowledge that the proposed method does have certain limitations. We outline these limitations below.

- The proposed method is tailored for systems characterized by failure modes predominantly driven by cycle loading. In instances where various factors drive the system's failure mechanism, the adaptability of the soft constraint employed in our method may require further consideration.
- The proposed method was only evaluated in case studies with continuous degradation. Exploring the suitability of our proposed method for systems exhibiting abrupt failures is a subject of future investigation.
- The evaluation of the proposed method involved case studies with run-to-failure data. Although our method applies to censored data, the interpreted meaning of the estimated HI differs. In instances where no failures are observed (HI = 0), the estimated HI is normalized with reference to the most degraded unit observed.

8. Conclusion

This work proposes an unsupervised hybrid method for HI estimation leveraging general knowledge and two hybridization strategies. The proposed method features two design features: (1) a novel network architecture of a convolutional AE preserving the causal relationships among sensor readings, operating conditions, and degradation within complex systems, and (2) the incorporation of soft constraints within the loss function derived from general knowledge of the degradation process, guiding the AE to infer degradation in its latent space.

In an extensive analysis involving turbofan engines and batteries, both in-distribution and out-of-distribution testing scenarios, we

demonstrated that this hybrid method, grounded in generalized knowledge, has wide applicability across diverse systems.

The effectiveness and generalization capabilities of the proposed method were demonstrated in comparative analysis involving alternative HI estimation methods. The evaluation encompassed both HI metrics and the utility of the HI for RUL prognostics. The proposed method consistently outperformed the industry standard residual method in all experimental setups. Notably, the performance gap between our approach and fully supervised models was minimal, particularly in RUL prediction tasks. For instance, in the turbofan dataset, both our method and the supervised model improved RUL predictions by approximately 28%. Similarly, in the battery dataset, both methods yielded approximately 25% improvement. The results emphasize the importance of integrating knowledge into neural networks, showcasing the informative potential embedded in such knowledge.

For future research, we plan to expand the application of our method beyond turbofan engines and batteries to include other critical systems, such as bearings. Furthermore, a key focus in future research will be the exploration of optimal strategies for effectively leveraging HI for RUL prognostics.

CRediT authorship contribution statement

Kristupas Bajarunas: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marcia L. Baptista:** Writing – review & editing, Supervision, Conceptualization. **Kai Goebel:** Writing – review & editing, Supervision, Conceptualization. **Manuel Arias Chao:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

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.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT (3.5) in order to improve the readability and language of the text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Appendix A. Additive causal model

This section is focused on the domain of causal inference and presents a common methodology for uncovering causal relationships from observational data. We demonstrate the results of applying such methodology for the turbofan case study.

To facilitate clarity, we begin by introducing causal notation. Initially, our attention is directed towards scenarios characterized by causal models involving only two variables. We then follow with a brief overview of the procedure for a multivariate case. In the context of the bivariate scenario a structural causal model (SCM) consists of two assignments:

$$C := f_1(\epsilon_1) \tag{A.1}$$

$$E := f_2(C, \epsilon_2) \tag{A.2}$$

where ϵ_1, ϵ_2 are jointly independent noise variables and f_1, f_2 are deterministic functions. In this model, we denote the random variables C as the cause and E as the effect. Furthermore, we refer to the causal graph $C \to E$ if C is a direct cause of E.

Determining the causal direction, even in a bivariate scenario, is challenging [71]. A recognized method for establishing causal direction is the non-linear additive noise model (ANM), as introduced in [59]. In general, if C is a direct cause of E, then it is intuitive to model the relationship as:

$$E = f(C) + \epsilon \quad C \perp \epsilon \tag{A.3}$$

where $f(\cdot)$ is an arbitrary nonlinear function and ϵ is the independent noise variable. The assertion that C is independent of ϵ ($C \perp \epsilon$) relies on several assumptions, including the absence of hidden common causes between C and E and no feedback loops between the two (i.e., no $C \leftrightarrow E$ interaction).

The non-linear ANM can be effectively applied to observational data in practical settings. Given two random variables C and E, the approach involves estimating the conditional expectation $\mathbb{E}(E|C)$ through regression analysis, followed by testing the independence of the residuals $E - \mathbb{E}(E|C)$ and C. Since the causal direction is typically unknown beforehand, it is necessary to test both possible causal directions.

The extension to the multivariate case is discussed in detail in [59]. Briefly, for each potential causal structure, represented by a directed acyclic graph (DAG) G_i, the procedure involves conducting a nonlinear regression for each variable against its parent variables. Subsequently, we test whether the resulting residuals are mutually independent. If any independence test is rejected, G_i is rejected. However, depending on the significance levels for rejecting and accepting independence, the ANM may indicate causality in both directions, no direction, or only one direction.

To address cases where these tests are inconclusive, [61] propose an alternative method based on the variance score of residuals. This method evaluates causality by assigning a higher score to models where the variance of the residuals is smaller, indicating a better fit. Such scores help in making definitive decisions about the causal direction. The procedure is outlined in [61] and summarized in Algorithm 1.

Algorithm 1 General procedure to find the optimal causal structure graph G_{opt}.

Input: Observational data $\mathcal{D}_N = \{V_j\}_j^N$, variables V_j

1: Construct all possible DAG G with d structural assignments $V_j = f_j(PA(V_j), \epsilon_j) \quad j = 1, ..., d$, where $PA(V_j)$ are the parents of V_j
2: For each graph structure G_k regresses each variable V_j on its parents $PA(V_j)$
3: Obtain residuals $R_j = V_j - \hat{f}_j(PA(V_j))$
4: Obtain a score $\log p(\mathcal{D}|G_k) = \sum_{j=1}^{d} -\log(\text{var}(R_j))$
5: Obtain most probable causal graph $G_{opt} = \arg\max_k \{\log p(\mathcal{D}|G_k)\}$

For the turbofan case study, we utilize the methodology given in Algorithm 1 specifically designed for the multivariate scenario. However, due to the large number of variables involved, including 14 sensor readings denoted as X, 4 operating conditions denoted as W, and the system health indicator Z, the exploration of all possible DAGs becomes impractically large.

To address this challenge, we introduce simplifying assumptions. Firstly, we assume mutual independence among the sensor readings X ($X_i \perp X_j, \forall i \neq j$) and mutual independence among the operating conditions W. This allows us to focus on analyzing smaller subsets of data, each containing a single sensor reading, a specific operating condition, and the health parameter (denotes as $D = \{X_i, W_j, Z\}$). While this simplification may not fully align with reality, it facilitates our goal of identifying causal relationships among X, W, and Z. Secondly, since the degradation effect is typically minor in the early cycles, we focus on data from later cycles (count greater than 45) where the effect of degradation is more noticeable.

We use DecisionTreeRegressor from sklearn with default parameters to perform regression. Since different combinations of chosen X_i and

Table A.17
All possible DAG with 3 variables. Each DAG is formatted as "variable ← [cause]", where "variable" represents the effect variable and "cause" denotes the set of potential causal factors or parents of the variable. The median and mean ranking over all choices of X_i and W_j are reported.

DAG structure	Median ranking	Mean ranking
$Z \leftarrow [X]$	0.0	0.75
$Z \leftarrow [W]$	1.0	1.14
$W \leftarrow [Z]$	2.0	2.05
$W \leftarrow [X]$	3.0	3.45
$X \leftarrow [Z]$	4.0	3.46
$X \leftarrow [W]$	5.0	5.29
$Z \leftarrow [X, W]$	6.0	5.54
$W \leftarrow [Z], Z \leftarrow [X]$	8.0	8.18
$W \leftarrow [Z, X]$	8.0	8.86
$W \leftarrow [X], Z \leftarrow [X]$	10.0	10.45
$W \leftarrow [X], Z \leftarrow [W]$	10.0	11.02
$X \leftarrow [Z], Z \leftarrow [W]$	11.0	10.57
$X \leftarrow [Z, W]$	11.0	12.04
$W \leftarrow [Z], X \leftarrow [Z]$	12.0	12.23
$X \leftarrow [W], Z \leftarrow [X]$	13.0	13.70
$X \leftarrow [W], Z \leftarrow [W]$	14.0	14.14
$W \leftarrow [X], X \leftarrow [Z]$	15.0	15.14
$W \leftarrow [Z], X \leftarrow [W]$	16.0	15.61
$W \leftarrow [X], Z \leftarrow [X, W]$	18.0	18.30
$W \leftarrow [Z, X], Z \leftarrow [X]$	19.0	19.70
$X \leftarrow [W], Z \leftarrow [X, W]$	20.0	19.71
$W \leftarrow [Z, X], X \leftarrow [Z]$	21.0	21.66
$X \leftarrow [Z, W], Z \leftarrow [W]$	**22.0**	21.11
$W \leftarrow [Z], X \leftarrow [Z, W]$	**22.0**	21.91

W_j lead to different optimal causal graphs, we report the median and mean ranking of all possible DAG structures given in Table A.17.

The two most frequently observed causal structures align with our earlier reasoning, indicating that X is influenced by both W and Z, represented as $X \leftarrow [Z, W]$. It is noteworthy that the illustration provided in Fig. 5, detailed in Section 4.1, corresponds to $X \leftarrow [Z, W], Z \leftarrow [W]$ and is in the top two ranking. The difference between the two most frequently found causal structures is the causal directions between W and Z. The ambiguity of the causal relationship between W and Z was anticipated in the C-MAPSS turbofan case study, given that the data generation process modeled degradation Z as an independent process.

Appendix B. Reliability type function

We propose a method to determine the function $g(t)$ representing the expected Health Indicator (HI) of a system for a given cycle t. Inspired by reliability theory, we hypothesize that the failure time of many complex systems (i.e., t for $HI(t) = 0$) follows a Weibull distribution with parameters β and η. Additionally, we hypothesize that the time to reach any intermediate HI threshold s is also Weibull distributed with parameters β_s and η_s.[2]

Under this hypothesis, the shape parameter β_s remains constant across different s thresholds (i.e., $\beta_s = \beta \quad \forall s$.) Meanwhile, the scale parameter η_s changes as a function of HI threshold (i.e., $\eta_s = h(s, \eta)$) [72]. This is because η_s is also known as the characteristic life and corresponds to the cycles at which 63% of the units have reached the threshold s. Since HI is decreasing as cycles increase, so does η_s. In this way, one can obtain the best-fit function $HI = h(\eta_s)$.

A good choice for this function (see [73]) is

$$HI = A - (B\eta_s)^C \tag{B.1}$$

The parameters A, B, and C are estimated from historical HI curves. Further, connecting cycle time t with HI involves the Weibull distribution's Cumulative Distribution Function, expressed as:

$$P = 1 - exp(-(t/\eta_s)^\beta) \tag{B.2}$$

[2] The validity of this hypothesis has been proved analytically in [72].

Rewriting (B.1) as $\eta_s = \sqrt[C]{(A - HI)}/B$ and substituting into (B.2), leads to

$$HI = C - ((t * (log(1 - P))^{-1/\beta}) * A)^B = g(t) \tag{B.3}$$

And thus we can obtain the best-fit HI for a fleet of units as a function of operational cycles. Note that adjusting P corresponds to adjusting the confidence of the HI. In our experiments, we have used P = 0.5.

Appendix C. HI criteria

This section presents the four HI criteria used for the evaluation of the proposed method.

- Monotonicity M of health index h_u of unit u with m observations is expressed as

$$M = \frac{1}{m-1}\sum_{j=1}^{m-1} |Ind(h_u^{j+1} - h_u^j) - Ind(h_u^j - h_u^{j+1})| \tag{C.1}$$

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

- Trendability T of health index h_u of unit u with cycles t_u is expressed as

$$T = |\text{corr}(t_u, h_u)| \tag{C.2}$$

where $corr(.)$ is the Spearman correlation coefficient.

- Prognosability P of all health indexes in a set E^d is given by,

$$P = exp(-\frac{\sigma(h_u^{end})}{\mu(|h_u^{end} - h_u^0|)}) \quad u \in E^{\mathrm{d}} \tag{C.3}$$

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 MI between h_u and RUL_u for unit u can be expressed as:

$$MI = \frac{1}{m}\sum_{i=1}^{m}[1 - exp(-I(h_u, RUL_u))] \tag{C.4}$$

where $I(.)$ is the mutual information measure.

References

[1] Lei Y, Li N, Guo L, Li N, Yan T, Lin J. Machinery health prognostics: A systematic review from data acquisition to rul prediction. Mech Syst Signal Process 2018;104:799–834. http://dx.doi.org/10.1016/j.ymssp.2017.11.016, URL https://www.sciencedirect.com/science/article/pii/S0888327017305988.

[2] Liu K, Gebraeel NZ, Shi J. A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis. IEEE Trans Autom Sci Eng 2013;10(3):652–64.

[3] Wang T, Yu J, Siegel D, Lee J. A similarity-based prognostics approach for remaining useful life estimation of engineered systems. In: 2008 international conference on prognostics and health management. IEEE; 2008, p. 1–6.

[4] Wang P, Youn BD, Hu C. A generic probabilistic framework for structural health prognostics and uncertainty management. Mech Syst Signal Process 2012;28:622–37.

[5] Yu W, Kim IY, Mechefske C. An improved similarity-based prognostic algorithm for rul estimation using an rnn autoencoder scheme. Reliab Eng Syst Saf 2020;199:106926.

[6] Liu Y, Hu X, Zhang W. Remaining useful life prediction based on health index similarity. Reliab Eng Syst Saf 2019;185:502–10.

[7] 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.

[8] Kumar A, Parkash C, Vashishtha G, Tang H, Kundu P, Xiang J. State-space modeling and novel entropy-based health indicator for dynamic degradation monitoring of rolling element bearing. Reliab Eng Syst Saf 2022;221:108356.

[9] Zhang Y, Zhang C, Wang S, Dui H, Chen R. Health indicators for remaining useful life prediction of complex systems based on long short-term memory network and improved particle filter. Reliab Eng Syst Saf 2024;241:109666.

[10] Roman D, Saxena S, Robu V, Pecht M, Flynn D. Machine learning pipeline for battery state-of-health estimation. Nat Mach Intell 2021;3(5):447–56.

[11] Ye Z, Yu J. Health condition monitoring of machines based on long short-term memory convolutional autoencoder. Appl Soft Comput 2021;107:107379.

[12] de Pater I, Mitici M. Developing health indicators and rul prognostics for systems with few failure instances and varying operating conditions using a lstm autoencoder. Eng Appl Artif Intell 2023;117:105582.

[13] Lövberg A. Remaining useful life prediction of aircraft engines with variable length input sequences. In: Annual conference of the PHM society. Vol. 13, 2021.

[14] Biggio L, Bendinelli T, Kulkarni C, Fink O. Ageing-aware battery discharge prediction with deep learning. Appl Energy 2023;346:121229.

[15] Guo H, Guo L. Health index for power transformer condition assessment based on operation history and test data. Energy Rep 2022;8:9038–45.

[16] 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;209:111120.

[17] Hagmeyer S, Zeiler P, Huber MF. On the integration of fundamental knowledge about degradation processes into data-driven diagnostics and prognostics using theory-guided data science. In: PHM Society European Conference. Vol. 7, 2022, p. 156–65. http://dx.doi.org/10.36001/PHME.2022.V7I1.3352, URL https://papers.phmsociety.org/index.php/phme/article/view/3352.

[18] Bajarunas K, Baptista M, Goebel K, Chao MA. Unsupervised physics-informed health indicator estimation for complex systems. In: Annual conference of the PHM society. Vol. 15, 2023.

[19] Si X-S. An adaptive prognostic approach via nonlinear degradation modeling: Application to battery data. IEEE Trans Ind Electron 2015;62(8):5082–96.

[20] Nouri Qarahasanlou A, ShakorShahabi R, Fallahnejad N. Assessment of spare parts requirement by reliability: A case study. Int J Reliab Risk Saf: Theory Appl 2022;5(1):9–19.

[21] Alves C. Group decision making approach for ranking and selecting maintenance tasks for joint scheduling with production orders. Int J Qual Res 18(1):235–58.

[22] Smith DJ. Reliability, maintainability and risk: practical methods for engineers. Butterworth-Heinemann; 2021.

[23] Meeker WQ, Escobar LA, Pascual FG. Statistical methods for reliability data. John Wiley & Sons; 2022.

[24] Saxena A, Goebel K, Simon D, Eklund N. Damage propagation modeling for aircraft engine run-to-failure simulation. In: 2008 international conference on prognostics and health management. 2008, p. 1–9. http://dx.doi.org/10.1109/PHM.2008.4711414.

[25] Malhotra P, Tv V, Ramakrishnan A, Anand G, Vig L, Agarwal P, Shroff G. Multi-sensor prognostics using an unsupervised health index based on lstm encoder–decoder. 2016, arXiv preprint arXiv:1608.06154.

[26] Xu F, Huang Z, Yang F, Wang D, Tsui KL. Constructing a health indicator for roller bearings by using a stacked auto-encoder with an exponential function to eliminate concussion. Appl Soft Comput 2020;89:106119.

[27] She D, Jia M, Pecht MG. Sparse auto-encoder with regularization method for health indicator construction and remaining useful life prediction of rolling bearing. Meas Sci Technol 2020;31(10):105005.

[28] Zhai S, Gehring B, Reinhart G. Enabling predictive maintenance integrated production scheduling by operation-specific health prognostics with generative deep learning. J Manuf Syst 2021;61:830–55.

[29] Lee H, Lim HJ, Chattopadhyay A. Data-driven system health monitoring technique using autoencoder for the safety management of commercial aircraft. Neural Comput Appl 2021;33:3235–50.

[30] Koutroulis G, Mutlu B, Kern R. Constructing robust health indicators from complex engineered systems via anticausal learning. Eng Appl Artif Intell 2022;113:104926.

[31] Zgraggen J, Pizza G, Huber LG. Uncertainty informed anomaly scores with deep learning: Robust fault detection with limited data. In: PHM society European conference. Vol. 7, 2022, p. 530–40.

[32] Chen Y, Rao M, Feng K, Zuo MJ. Physics-informed lstm hyperparameters selection for gearbox fault detection. Mech Syst Signal Process 2022;171:108907.

[33] Rao M, Zuo MJ, Tian Z. A speed normalized autoencoder for rotating machinery fault detection under varying speed conditions. Mech Syst Signal Process 2023;189:110109.

[34] Hsu C-C, Frusque G, Fink O. A comparison of residual-based methods on fault detection. In: Annual conference of the PHM society. Vol. 15, 2023.

[35] Chalapathy R, Chawla S. Deep learning for anomaly detection: A survey. 2019, arXiv preprint arXiv:1901.03407.

[36] de Beaulieu MH, Jha MS, Garnier H, Cerbah F. Unsupervised remaining useful life prediction through long range health index estimation based on encoders-decoders. IFAC-PapersOnLine 2022;55(6):718–23.

[37] Schwartz S, Jiménez JJM, Vingerhoeds R, Salaün M. An unsupervised approach for health index building and for similarity-based remaining useful life estimation. Comput Ind 2022;141:103716.

[38] Rueden LV, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, Kirsch B, Pfrommer J, Pick A, Ramamurthy R, Walczak M, Garcke J, Bauckhage C, Schuecker J. Informed machine learning-a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans Knowl Data Eng 2019. http://dx.doi.org/10.1109/TKDE.2021.3079836.

[39] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys 2021;1–19. http://dx.doi.org/10.1038/s42254-021-00314-5, URL www.nature.com/natrevphys.

[40] Magadán L, Suárez FJ, Granda JC, delaCalle FJ, García DF. A robust health prognostics technique for failure diagnosis and the remaining useful lifetime predictions of bearings in electric motors. Appl Sci 2023;13(4):2220.

[41] Jahromi A, Piercy R, Cress S, Service J, Fan W. An approach to power transformer asset management using health index. IEEE Electr Insul Mag 2009;25(2):20–34.

[42] Aizpurua JI, Stewart BG, McArthur SD, Lambert B, Cross JG, Catterson VM. Improved power transformer condition monitoring under uncertainty through soft computing and probabilistic health index. Appl Soft Comput 2019;85:105530.

[43] Bejaoui I, Bruneo D, Xibilia MG. A data-driven prognostics technique and rul prediction of rotating machines using an exponential degradation model. In: 2020 7th international conference on control, decision and information technologies (coDIT). Vol. 1, IEEE; 2020, p. 703–8.

[44] Wang P, Long Z, Wang G. A hybrid prognostics approach for estimating remaining useful life of wind turbine bearings. Energy Rep 2020;6:173–82.

[45] 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.

[46] Song C, Liu K. Statistical degradation modeling and prognostics of multiple sensor signals via data fusion: A composite health index approach. IISE Trans 2018;50(10):853–67.

[47] Qin Y, Yang J, Zhou J, Pu H, Mao Y. A new supervised multi-head self-attention autoencoder for health indicator construction and similarity-based machinery rul prediction. Adv Eng Inform 2023;56:101973. http://dx.doi.org/10.1016/j.aei.2023.101973.

[48] Chen Z, Xia T, Zhou D, Pan E. A health index construction framework for prognostics based on feature fusion and constrained optimization. IEEE Trans Instrum Meas 2021;70:1–15.

[49] Li Z, Wu J, Yue X. A shape-constrained neural data fusion network for health index construction and residual life prediction. IEEE Trans Neural Netw Learn Syst 2020;32(11):5022–33.

[50] Wang F, Du J, Zhao Y, Tang T, Shi J. A deep learning based data fusion method for degradation modeling and prognostics. IEEE Trans Reliab 2020;70(2):775–89.

[51] Wen P, Zhao S, Chen S, Li Y. A generalized remaining useful life prediction method for complex systems based on composite health indicator. Reliab Eng Syst Saf 2021;205:107241.

[52] Wang H, Li X, Zhang Z, Deng X, Jiang W. A deep learning based health index construction method with contrastive learning. Reliab Eng Syst Saf 2023;109799.

[53] Ng M-F, Zhao J, Yan Q, Conduit GJ, Seh ZW. Predicting the state of charge and health of batteries using data-driven machine learning. Nat Mach Intell 2020;2(3):161–70.

[54] Zhou D, Wang B, Zhu C, Zhou F, Wu H. A light-weight feature extractor for lithium-ion battery health prognosis. Reliab Eng Syst Saf 2023;237:109352.

[55] Fan Y, Xiao F, Li C, Yang G, Tang X. A novel deep learning framework for state of health estimation of lithium-ion battery. J Energy Storage 2020;32:101741.

[56] Liu CR, Choi Y. A new methodology for predicting crack initiation life for rolling contact fatigue based on dislocation and crack propagation. Int J Mech Sci 2008;50(2):117–23.

[57] Kunzelmann B, Rycerz P, Xu Y, Arakere NK, Kadiric A. Prediction of rolling contact fatigue crack propagation in bearing steels using experimental crack growth data and linear elastic fracture mechanics. Int J Fatigue 2023;168:107449.

[58] Wang J, Li Y, Zhao R, Gao RX. Physics guided neural network for machining tool wear prediction. J Manuf Syst 2020;57:298–310.

[59] Hoyer P, Janzing D, Mooij JM, Peters J, Schölkopf B. Nonlinear causal discovery with additive noise models. Adv Neural Inf Process Syst 2008;21.

[60] Pearl J, et al. Models, reasoning and inference, Vol. 19, Cambridge, UK: CambridgeUniversityPress; 2000, no. 2.

[61] Peters J, Janzing D, Schölkopf B. Elements of causal inference: foundations and learning algorithms. The MIT Press; 2017.

[62] Edge JS, O'Kane S, Prosser R, Kirkaldy ND, Patel AN, Hales A, Ghosh A, Ai W, Chen J, Yang J, et al. Lithium ion battery degradation: what you need to know. Phys Chem Chem Phys 2021;23(14):8200–21.

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

[64] Bole B, Kulkarni CS, Daigle M. Adaptation of an electrochemistry-based li-ion battery model to account for deterioration observed under randomized use. In: Annual conference of the PHM society. Vol. 6, 2014.

[65] Sui X, He S, Vilsen SB, Meng J, Teodorescu R, Stroe D-I. A review of non-probabilistic machine learning-based state of health estimation techniques for lithium-ion battery. Appl Energy 2021;300:117346.

[66] Nguyen KT, Medjaher K. An automated health indicator construction methodology for prognostics based on multi-criteria optimization. ISA Trans 2021;113:81–96.

[67] Coble JB. Merging data sources to predict remaining useful life–an automated method to identify prognostic parameters. TRACE; 2010.

[68] Mao W, Chen J, Chen Y, Afshari SS, Liang X. Construction of health indicators for rotating machinery using deep transfer learning with multiscale feature representation. IEEE Trans Instrum Meas 2021;70:1–13.
[69] Arias Chao M, Kulkarni C, Goebel K, Fink O. Fusing physics-based and deep learning models for prognostics. Reliab Eng Syst Saf 2022;217:107961.
[70] Nejjar I, Geissmann F, Zhao M, Taal C, Fink O. Domain adaptation via alignment of operation profile for remaining useful lifetime prediction. 2023, arXiv preprint arXiv:2302.01704.
[71] Mooij JM, Peters J, Janzing D, Zscheischler J, Schölkopf B. Distinguishing cause from effect using observational data: methods and benchmarks. J Mach Learn Res 2016;17(32):1–102.
[72] Diersin P, Chao MA, Bajarunas K. Analytical modeling of health indices for prognostics and health management. In: Annual conference of the PHM society. Unpublished results.
[73] Bagdonavicius V, Nikulin M. Accelerated life models: modeling and statistical analysis. CRC Press; 2001.