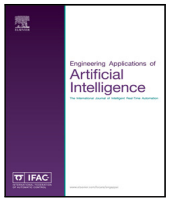
Engineering Applications of Artificial Intelligence 113 (2022) 104926

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Constructing robust health indicators from complex engineered systems via anticausal learning

Georgios Koutroulis [a,*], Belgin Mutlu [a], Roman Kern [b]

[a] Pro2Future GmbH, Inffeldgasse 25F, 8010 Graz, Austria
[b] Know Center, Inffeldgasse 16, 8010 Graz, Austria

ARTICLE INFO

Keywords:
Health indicator
Prognostics and health management (PHM)
Structural causal models
Robustness
Anticausal prediction

ABSTRACT

In prognostics and health management (PHM), the task of constructing comprehensive health indicators (HI) from huge amounts of condition monitoring data plays a crucial role. HIs may influence both the accuracy and reliability of remaining useful life (RUL) prediction, and ultimately the assessment of system's degradation status. Most of the existing methods assume apriori an oversimplified degradation law of the investigated machinery, which in practice may not appropriately reflect the reality. Especially for safety–critical engineered systems with a high level of complexity that operate under time-varying external conditions, degradation labels are not available, and hence, supervised approaches are not applicable. To address the above-mentioned challenges for extrapolating HI values, we propose a novel anticausal-based framework with reduced model complexity, by predicting the cause from the causal models' effects. Two heuristic methods are presented for inferring the structural causal models. First, the causal driver is identified from complexity estimate of the time series, and second, the set of the effect measuring parameters is inferred via Granger Causality. Once the causal models are known, off-line anticausal learning only with few healthy cycles ensures strong generalization capabilities that helps obtaining robust online predictions of HIs. We validate and compare our framework on the NASA's N-CMAPSS dataset with real-world operating conditions as recorded on board of a commercial jet, which are utilized to further enhance the CMAPSS simulation model. The proposed framework with anticausal learning outperforms existing deep learning architectures by reducing the average root-mean-square error (RMSE) across all investigated units by nearly 65%.

1. Introduction

All modern engineered systems inevitably go through a continuously evolving health degradation process, which finally may lead to their replacement, usually via conventional preventive maintenance policies (Shafiee, 2015). It is extremely important to obtain a transparent knowledge of the machinery's degradation levels so that unscheduled maintenance activities with great operational cost and possible reputational damage can be prevented (Rodrigues et al., 2012). According to Groenenboom (Groenenboom, 2018) in the domain of aviation, handling such issues by deploying intelligent condition monitoring approaches may enable airlines to save about $3 bn. per year. For developing such sophisticated solutions, it may be either impractical or infeasible to measure the exact health status of the machinery (Lei et al., 2018), let alone to unveil its hidden degradation trends. In this regard, Prognostics and Health Management (PHM) (Goebel et al., 2017) aims for the prediction of the remaining useful life (RUL) of the investigated machinery from comprehensive health indicator (HI) values that hopefully reflect the true health status. Such PHM schemes may not only improve the system's reliability and cost-efficiency, but further prevent major accidents with potential loss of human lives. Hence, the accurate estimation of HIs, in particular under time-varying operating conditions with different degradation effects, plays a critical role to the final assessment of the RUL prediction, and ultimately to the effectiveness of the entire PHM framework.

Generally, two main categories of HIs are widely known based on the direct association of the information carrier with the gradual deterioration of the machinery (Hu et al., 2012; Lei et al., 2018). First, Physical Health Indicators (PHIs) mostly utilize domain-driven features that characterize the system's degradation condition in a straightforward manner. In such PHI approaches (Javed et al., 2014; Medjaher et al., 2013; Benkedjouh et al., 2013), signal processing methods (e.g, wavelet transform) are usually employed with statistical-based ones so that the physical characteristics of the system are captured. PHIs are typically extracted from univariate raw vibration signals sampled at high frequencies, and measured in single mechanical components, such as bearings and gears (Ali et al., 2015; Hu et al., 2016). In

* Corresponding author.
E-mail address: georgios.koutroulis@pro2future.at (G. Koutroulis).

https://doi.org/10.1016/j.engappai.2022.104926
Received 26 November 2021; Received in revised form 15 April 2022; Accepted 28 April 2022
Available online xxxx
0952-1976/

more complex multi-component systems (e.g, jet engines) is much more difficult for PHIs to accurately extract the overall health status. Alternatively, Virtual Health Indicators (VHIs) are generally extracted by applying fusion and dimensionality reduction techniques either on multidimensional sensor readings or on individual physical features (Yang et al., 2016; Baraldi et al., 2018; Wang et al., 2008; Lei et al., 2018). In general, VHIs are one-dimensional unitless agents and they should clearly depict the health status of the machine regardless of any variations in the operating conditions. In this work, we focus on the later category of HIs, which is considered more challenging, since multifaceted degradation from complex systems must be extracted and summarized into a single highly representative and robust HI.

VHIs can be constructed by supervised or unsupervised learning methods, depending on the availability of the degradation (RUL) labels. For instance, Guo et al. (2018) proposed a supervised method to construct HIs via deep convolutional neural networks (CNN) by utilizing the cumulative service life in percentage from rolling element bearings as the target label. Although the authors used a non-linear mapping to construct the HI, they assumed a linear degradation trajectory for RUL labeling that adds a major restriction to the method, as different operating conditions may vary within the service life of the asset, and accordingly, they might accelerate or distort the degradation process. Similarly, the authors in Chen et al. (2020) employed in a non-linear way CNNs with long–short term memory neural networks (LSTMs) to capture long term dependencies in the time series which they ultimately used for bearing HI construction. Even though deep learning architectures are designed to overcome the tedious hand-crafting effort of manual feature extraction (Qin et al., 2016), the former study require sufficiently large amounts of run-to-failure vibration data to map the automatically extracted features to the target value that represents the health status of the investigated bearings. Both previous methods are basically supervised and might work well for individual components. However, in complex mechanical modules or systems, such as jet engines, it is often extremely costly or even impossible to quantitatively capture the holistic degradation state with a high accuracy that might be further used as the label at a specific service time. Finally, the dependence of these methods on run-to-failure training data is obviously an additional factor of limited applicability in real-operating conditions of safety–critical systems. Such limitations pave the way for unsupervised learning methods, in which only data from healthy conditions are utilized to learn the prediction algorithms.

Data-driven approaches that employ machine learning models, and especially deep neural network architectures, are currently proposed for PHM applications under nonlinear and multidimensional settings (Fink et al., 2020; Thoppil et al., 2021). Besides these fundamental challenges, changing operating conditions in engineered systems may rapidly deteriorate the model's HI predictions due to sensitivity issues to perturbations of the input independent variables, which is attributed to the lack of robustness (Khan et al., 2021). Uncertainties that emanate either from measurement errors or from any stochasticity of the degradation process further impact the robustness of the model (Lei et al., 2018) in such a way that it is infeasible for HIs to be later used for RUL prediction purposes. On the other hand, another source of error might originate from model complexity itself. For example, deep neural network architectures that are trained with vast amounts of data, they consist of millions of parameters in order to learn the mapping functions between the input variables and the output. Such models usually suffer from increased complexity, and at the end they might yield poor generalization performance on future unseen data. Developing data-driven solutions with low model complexity that are able to robustly account for data variations is the key for ensuring the safety and reliability of engineered systems.

Most of the existing works rely on methods that learn horizontal dependencies in the data without the underlying causal structure under consideration. Such approaches usually lack generalization abilities and robustness, since they may collapse in non-i.i.d. (independent and identically distributed) regime, in which the probability distribution may strongly vary between the source and the target domain (e.g., different operating conditions). Since causal relationships are inherently invariant and stable over different domains (Bühlmann, 2020), models that are built on this principle can effectively address the challenges from non-i.i.d. settings. Such useful properties are actually entailed in structural causal models (Pearl, 2009; Peters et al., 2017) that mathematically describe the underlying causal relationships between cause and effect via deterministic functions, and noise variables to account any randomness in the model. Intuitively, SCMs represent the mechanism that is responsible for the data generation, and they represent a trade-off between physical and statistical models, as it is summarized in Table 1. In the seminal work of Schölkopf et al. (2012), the notion of structural causal models is utilized to investigate their implications on machine learning problems, like covariate shift and efficient usage of data in semi-supervised learning. In particular, the authors showed that in case of alignment of the causal direction $X \rightarrow Y$ with the predictive one (predicting Y from X), where the input of the model X is the cause and the target Y is the effect, robustness to covariate shift is easier to achieve. On the other hand, when the causal direction $X \leftarrow Y$ is the opposite with the predictive one, where we are trying to predict the cause Y from the effect X, this is said to be *anticausal prediction* and semi-supervised learning can interestingly work. Experiments in Schölkopf et al. (2012) with linear models for semi-supervised learning demonstrate the effectiveness of the approach, which we also adopt in the proposed framework due to availability of few data from healthy state.

Recently, the authors in Khan et al. (2021) highlight the importance of integrating causal models towards achieving robust AI-based solutions for PHM applications. Furthermore, they assert that the black-box problem in deep learning hinders any transparency of how input variables are interrelated with each other and with the outcome. This problem can be eliminated by the inherent interpretability of the structural causal models. We employ causality to seamlessly model the degradation process of the system and build multiple structural causal models from the time series of the monitoring signals and the inferred causal driver. Hence, we exploit the remarkable invariance properties of the causal mechanisms (Bühlmann, 2020; Schölkopf, 2019) and learn our models upon these mechanisms for better generalization and robustness.

The main contributions of this work are summarized as follows:

1. A complexity-estimate metric for time series data to rank operational and environmental parameters (potential causes) by computing the largest variability in relation with the time scale. The intuition behind this metric is that within same time periods *having more, and larger peaks and valleys* will yield higher complexity estimate. In addition to limited background knowledge from the application domain, we finally select the causal driver from the overall parameter set.
2. A causal feature selection for time series is developed based on non-linear *Granger Causality* to capture complex dependencies between the inferred causal driver and the measuring parameters. The computed causal indices are then used for selecting the dependent measuring parameters, from which a set of structural causal models is yielded.
3. A novel HI is proposed by introducing anticausal learning for robust predictions of the holistic health status of complex engineered systems under time-varying operating conditions. To the best of our knowledge, the proposed framework is the first that integrates such powerful techniques from the field of causal inference for PHM applications.
4. Our anticausal-based framework is capable of employing any kind of regression method. However, we show empirically that indeed the linear regression outperforms all other methods as it inherently disentangles the derived structural causal models

Table 1
Taxonomy of modeling approaches (Peters et al., 2017). Starting from top, physical models are the most detailed ones and they are in principle described by ordinary differential equations. At the bottom, statistical models can be only learned by observational data, and learning is not feasible under distribution shift (non i.i.d.). Striking a balance between two former models, structural causal models, on the one hand, move beyond the i.i.d. settings by enabling specific graphical interventions, while, on the other hand only limited physical knowledge of the system is necessary.

Model type	Predict in i.i.d. regime	Predict under distribution shift	Require domain knowledge
Physical model	Yes	Yes	Complete
Structural causal model	Yes	Yes	Limited
Statistical model	Yes	No	None

in the anticausal direction. Last but not least, the proposed approach requires no assumption (e.g., bilinear) regarding the underlying degradation law, which most of methods do.

The rest of the paper is organized as follows: Section 2 reviews the related work of data-driven methods employed for VHI construction in different PHM applications. Section 3 presents the required theoretical background. In Section 4, the framework and details of the proposed anticausal learning approach are briefly introduced. In Section 5, details of the experimental results with the analysis and discussion on the effectiveness of the proposed method is presented. Finally, Section 6 concludes the paper and presents the potential of future work.

2. Related work

To date, a large body of literature aimed for high robustness to noisy data in their degradation modeling approaches (Lei et al., 2018; Guo et al., 2019). Noise in sensor data is ubiquitous in every real-world PHM application of complex systems that may largely distort the quality of the predictions. Other sources of noise, such as stochasticity of the degradation and random fluctuations due to changeable operating conditions, further exacerbate the prediction error of the HI. Towards the goal of constructing robust HIs in noisy data, Gugulothu et al. (2017) used originally autoencoders (AE) (Malhotra et al., 2016, 2017) for computing time series embeddings. In the hidden layers of the architecture, Recurrent Neural Networks (RNNs) are trained in an unsupervised manner with data from both normal and faulty operations of the machines. To alleviate vanishing and exploding gradient issues during back-propagation (Hochreiter and Schmidhuber, 1997) in traditional RNNs, LSTM cells are employed that are capable of learning long term dependencies from non-linear and noisy time series (Karasu et al., 2020; Karasu and Altan, 2022). HI curves in Gugulothu et al. (2017) are initially derived from embedding distances that are computed with reference to normal instances. Although the authors clearly demonstrate the robustness of the embeddings-based HI to noisy data, the approach is highly dependent on run-to-failure data that is difficult to obtain in safety–critical systems. Similar to our framework, the embeddings-based approach does not assume any specific degradation trend, and it also considers sensor data as time series. Fu et al. (2021) introduced a novel LSTM architecture with sequential updated reconstructions for constructing robust HIs, and subsequently the RUL of turbofan engines. Although both model-based approaches generate HIs with robust attributes, they mainly lack of interpretability, since extracted features usually have no physical meaning regarding the degradation process. Moreover, data abundance is usually required for training such deep learning architectures, which is rather difficult due to limited amount of degradation free data. Nguyen and Medjaher (2021) developed an end-to-end HI construction framework from automated low level feature extraction that accounts various performance criteria via genetic programming. Experiments on the C-MAPSS turbofan engine dataset (Saxena et al., 2008b) demonstrated the effectiveness of the extracted features over the raw sensor measurements in terms of monotonicity, smoothness and robustness, which we similarly integrate for evaluation of our approach.

In PHM, most of the signal reconstruction methods with deep learning are based on autoencoder schemes, as they considered a mainstream network architecture for learning meaningful and compressed representations of the data (Fink et al., 2020). In the beginning of the machine's operation, normal behavior is mostly dominant and it is expected that accurate reconstruction of this behavior is feasible. On the other hand, gradual abnormal behavior, due to system's long-term degradation, leads to poor reconstruction of the input data, and hence higher prediction error. In alignment with this principle, Malhotra et al. (2016) first introduced Long–Short Term Memory (LSTM) based Encoder–Decoder architecture (LSTM-ED) for HI estimation, and subsequently for RUL prediction. The authors utilized the reconstruction error on test time series as the HI that represents the long term degradation trend. Once the HI curve is constructed, a similarity-based technique (Yu et al., 2019) is employed to estimate the RUL, which is evaluated on C-MAPSS turbofan engine dataset and a real-world dataset from a milling machine. In a recent study, Liu et al. (2020) compared several generative methods on a real-world dataset from an aircraft air conditioning system, including variational autoencoder (VAE) that are able to reconstruct and denoise the input by learning a latent Gaussian representation. Experimental results demonstrated the superiority of HIs that are constructed from LSTM-AE and AE with the reconstruction error approach. We also employ exactly these architectures as baseline methods for comparison. Although deep learning methods exhibit high potential for PHM applications, they still learn purely associational relationships between set of features (Marcus, 2018) that might be fragile to time-varying operating conditions and noisy environments.

Mechanical systems operate always at constantly varying and heterogeneous conditions (different speed, temperature, load, etc.), which may to a considerable degree distort their canonical degradation trajectory. As the authors in Li et al. (2019) pointed out, operational variations may have great impact both on the degradation speed as well as on the distortion of the sensor readings. In this regard, several studies (Lei et al., 2018; Hu et al., 2012; Baraldi et al., 2018; Arias Chao et al., 2022; Li et al., 2019) so far considered time-varying conditions for degradation modeling. Luo et al. (2018) proposed a supervised deep learning framework for health estimation of CNC machine tools under non-stationary working conditions. The authors built a deep neural network for classification of dynamical impulse/non-impulse modes which they later used to compute the natural frequencies that remain invariant in different operational schemes. However, the method is strongly dependent on the intermediate classification step which introduces a kind of aggregation scheme, that may influence the final estimation of the HI. Recently, Zhai et al. (2021) introduced an unsupervised deep learning approach for deriving HIs from large-scale industrial data that it is mainly comprised of two stages. First, the authors employed K-Means to cluster the different operating regimes from the expert-selected operational parameters. Second, the labeling information from the clustering step is passed to a conditional VAE architecture so that a conditional probability distribution from the healthy instances is learned. HIs are then computed as the distance (Manhattan, Euclidean) between the reconstructed and the original data. Although the generative-based model captures the machinery's degradation under various working conditions, the clustering approach for obtaining the operational regime partitions may influence the performance of the predictions in case of strong non-stationarities in the time series. This can be easily confirmed since the approach in Zhai et al. (2021) is evaluated with the C-MAPSS dataset from Saxena et al.

(2008b) as well, in which all operational parameters are synthetic and aggregated at the cycle level.

Very few studies have been proposed in the vast sphere of PHM literature using notions from causality. Nevertheless, very recently Baptista et al. (2022) investigated the causal influence between the prognostic model's input and its output via the game-theoretical Shapley (SHAP) values (Lundberg and Lee, 2017). The constructed SHAP values are evaluated in the C-MAPSS dataset by computing monotonicity, trendability, and prognosability metrics. Although most complex models yielded the best performance in RUL prediction, the authors showed that linear regression with higher interpretability outperforms other methods with increased complexity in monotonicity scores. One of the most known frameworks for inferring causal relationships in time series data is the Granger causality, originally proposed by Granger (1969). Zhu (2021) utilized Conditional Granger Causality (Geweke, 1984), an extension of the original Granger's method, for identifying linear causal relationships between multivariate sensor signals, which is later used as a variable reduction technique solely for the RUL prediction. However, the constructed HIs follow a rather rigid assumption of an exponential degradation model, where both operational and measuring variables are included without causal hierarchy. For the HI construction, we integrate a sensor selection technique by estimating non-linear Granger causal indices (Ancona et al., 2004) between the causal driver and the measuring parameters that ultimately yields the structural causal model.

3. Theoretical background

In this section, we present the theories of nonlinear Granger causality, additive noise models and anticausal learning with their implications and assumptions, which constitute the foundations of the proposed approach.

3.1. Estimation of bivariate causal indices

Granger causality (Granger, 1969) has already been widely used with great success in many scientific domains, ranging from neuroscience (Roebroeck et al., 2005) and finance (Hong et al., 2009) to computer science (Qiu et al., 2012) and climate research (Lozano et al., 2009). Specifically, let X_t and Y_t be two time series and we say that *X Granger causes Y* at a specified lag τ, if the prediction error is significantly reduced by regressing Y_t on both $Y_{t-\tau}$ and $X_{t-\tau}$ than only by incorporating the past of Y_t itself up to the lag τ that also represents the order of the autoregressive model. In the context of linear bivariate time series, the following autoregressive models are compared:

$$Y_t = \sum_{i=1}^{\tau} \alpha_i Y_{t-i} + \epsilon_t^y \tag{1}$$

$$Y_t = \sum_{i=1}^{\tau} \alpha_i Y_{t-i} + \sum_{i=1}^{\tau} \beta_i X_{t-i} + \epsilon_t^{yx} \tag{2}$$

where the noise terms ϵ_t^y and ϵ_t^{yx} are assumed to be independent and identically distributed (i.i.d.) time series, respectively. Whenever the noise term ϵ_t^{yx} estimated by the residuals from the full model of Eq. (2) has significant smaller variance than the noise term ϵ_t^y obtained by the restricted model from Eq. (1), then it is inferred that X Granger causes Y. Since the initial approach is purely linear, nonlinear extensions have been already introduced (Ancona et al., 2004; Marinazzo et al., 2008). Here, we present the nonlinear Granger causality based on radial basis functions (RBF) (Ancona et al., 2004) that we use for the sensor selection step.

The authors in Ancona et al. (2004) proposed a method provided that the following property is satisfied: (P1) *if the past time series $\mathbf{Y}^-$ are statistically independent of X* and the past time series $\mathbf{X}^-$, then $var(\epsilon^x) = var(\epsilon^{xy})$; *if the past time series $\mathbf{X}^-$ are statistically independent of Y* and the past time series $\mathbf{Y}^-$, then $var(\epsilon^y) = var(\epsilon^{yx})$, where $var(\cdot)$ is the estimate of the variance of the noise terms (residuals). Linear models are quite limited to satisfy the property P1, as they require very large time series. On the other hand, finding the right classes of nonlinear models may overcome the size restriction of the time series. Hence, the authors in Ancona et al. (2004) suggested RBFs for bivariate models with a clustering procedure that are given by

$$X = \mathbf{w}_{11}\mathbf{\Phi}(\mathbf{X}^-) + \mathbf{w}_{12}\mathbf{\Psi}(\mathbf{Y}^-) \tag{3}$$

$$Y = \mathbf{w}_{21}\mathbf{\Phi}(\mathbf{X}^-) + \mathbf{w}_{22}\mathbf{\Psi}(\mathbf{Y}^-) \tag{4}$$

where $\{\mathbf{w}\}$ are four n-dimensional real vectors and $\mathbf{\Phi} = (\phi_1, \dots, \phi_n)$, $\mathbf{\Psi} = (\psi_1, \dots, \psi_n)$ are n given nonlinear RBFs with n the number of the clusters for estimating the prototypes of $\mathbf{X}^-$, $\mathbf{Y}^-$ respectively. In Lungarella et al. (2007) and Edinburgh et al. (2021) various nonlinear causal methods for bivariate time series are extensively compared and a value of 50 for the number of clusters n is suggested as an appropriate choice. Cluster centers $\{\hat{\mathbf{X}}_\rho^-\}_{\rho=1}^n$, $\{\hat{\mathbf{Y}}_\rho^-\}_{\rho=1}^n$ are derived from both vector spaces of $\mathbf{X}^-$, $\mathbf{Y}^-$ with k-means and are used to estimate the following vectors in the nonlinear feature space

$$\phi_\rho(\mathbf{X}^-) = exp(-\|\mathbf{X}^- - \hat{\mathbf{X}}_\rho^-\|^2/2\sigma^2), \quad \rho = 1, \dots, n, \tag{5}$$

$$\psi_\rho(\mathbf{Y}^-) = exp(-\|\mathbf{Y}^- - \hat{\mathbf{Y}}_\rho^-\|^2/2\sigma^2), \quad \rho = 1, \dots, n, \tag{6}$$

where σ is the scale parameter. The residuals are calculated by least squares fit from the Eqs. (3)–(6) and the causal index in the case of the unidirectional causality $x \to y$ is given by

$$ci_{x\to y} = var(\epsilon^y) - var(\epsilon^{yx}). \tag{7}$$

Such bivariate nonlinear models are accurate with high detection power, however they may not account indirect and common cause (confounded) links for multivariate time series models. Further, a critical assumption is that all observed variables are included, namely there are no *hidden confounders*. Even though in practice such assumption seems restrictive, background knowledge may alleviate to a large extent both previous limitations. In particular, *independent variables* in engineered systems can be relatively easy identified by domain experts, since either are tied with the physical laws of nature (environmental) or with direct interventional actions (operational). In the proposed approach, we initially introduce a complexity-based metric from the time series data mining field to rank all independent variables and finally select the causal driver that will be used for estimating all causal indices with the measuring parameters.

3.2. Anticausal learning

Before we rigorously present the notion of anticausal learning that was originally presented by Schölkopf et al. (2012), we briefly introduce structural causal models that set the foundation towards that direction.

Let us first assume the directed acyclic graph (DAG) $C \to E$, where C and E are two observed random variables. In contrast to Granger Causality, we say here that *C causes E* if we intervene on C, and the same time all the rest variables in the graph are held fixed (if any), then we should observe changes in the distribution of E. Note that, as in the previous section, we assume *Causal Sufficiency*, namely there are no hidden common causes of the variables in the observational data. To represent mathematically the causal knowledge from the underlying DAG, structural causal models (SCM) (Pearl, 2009; Peters et al., 2017) are proposed that consist of a set of structural equations of the form

$$C = f_1(\epsilon_1) \tag{8}$$

$$E = f_2(C, \epsilon_2) \tag{9}$$

where ϵ_1, ϵ_2 are jointly independent noise variables and f_1, f_2 are deterministic functions of the directed causes with the noise terms.

More specifically, the SCM from Eq. (9) represents an autonomous mechanism via the function f_2 from which the effect E is generated by the cause C and the noise term ϵ_2. The difference of the SCM with the equivalent algebraic equations is fundamental and should not be confused, since all variables appear on the left-hand side are the *dependent* variables and any change in this order breaks any connection with the underlying causal representation.

Inferring the causal direction even in the bivariate case by identifying the asymmetry of the association is far from trivial (Mooij et al., 2016). Certain assumptions must hold for the underlying SCM in order to have an identifiable causal direction. In this regard, Hoyer et al. (2008) first introduced the non-linear additive noise model (ANM) that utilizes the statistical independence of the underlying noise term with the cause variable for determining the true causal direction. Under this assumption, the SCM described by Eq. (9) can be rewritten as

$$E = f(C) + \epsilon, \quad C \perp\!\!\!\perp \epsilon \tag{10}$$

where $f(\cdot)$ is a arbitrary nonlinear regression method and ϵ the independent noise variable. The main assumption can be easily verified by regressing the effect on the cause with a non-linear method and test the regression residual for statistical independence with the cause. If the independence hypothesis is rejected then the underlying causal model is rejected as well. Note that there is no constraint on the selection of a specific type of regression method. Even though this statement significantly simplifies the applicability of ANM for detecting complex causal relationships, however, any randomization in regression via perturbation techniques (e.g., random forest regression) may yield correlated noise that will violate the former assumption.

By considering the ANM from Eq. (10) we jointly determine the conditional distribution $P(E|C)$ which represents the generating mechanism that transforms cause C into effect E. In the causal direction, we further assume that the mechanism is independent of the marginal distribution of the cause $P(C)$, which is entailed by the *independence causal mechanism* principle (Janzing and Schölkopf, 2010). Intuitively, this means that any change on the input of the causal model does not influence whatsoever the causal mechanism $P(E|C)$. On the other hand, trying to predict the cause from the effect, which is referred in Schölkopf et al. (2012) as *anticausal prediction*, the marginal distribution of the effect $P(E)$ and the conditional distribution $P(C|E)$ share information, and they become dependent, which means that knowing the input distribution may help for learning the output conditional distribution $P(C|E)$. Therefore, such anticausal schemes may greatly benefit semi-supervised settings, since only a small amount of labeled data is available for training. It should be noted that no concrete semi-supervised learning (SSL) method, like in Zhou et al. (2005), is utilized on the proposed anticausal framework. Since the assumption for SSL from the anticausal model holds, just few data from the reference (healthy) cycles are used to train the regression models. Besides the former useful attribute, anticausal predictions may be extremely robust with strong generalization, provided that the underlying causal model is known, as rigorously presented in Kilbertus et al. (2018). In PHM applications, we usually obtain a rather small subset at the beginning of the operation life of the system which corresponds to the healthy state of the system and it is also used for training the prediction models. In real-world scenarios, on-line data are evaluated with strong non-stationarities due to distributional shifts that may affect the predictions of the HI. By intertwining such powerful techniques from the causal inference field, we achieve robustness on the predictions of the output that are in the later step used for constructing the HI of the investigated engineered system.

4. HI construction via anticausal learning

Establishing the right HI is a key factor for the overall success of a PHM maintenance strategy, since it strongly affects the prediction of the RUL, and ultimately the health assessment of the system. Besides the monotonic trend of the HI over time that need to be sufficiently fulfilled, one has to consider the robustness to continually changing operating conditions, an ubiquitous challenge in PHM applications (Khan et al., 2021). Especially, in case of cyclic datasets with large non-stationarities in their operational conditions between consecutive cycles, training directly machine learning models with data in healthy state might result in HIs of poor performance. Mainly, this happens since the model has learned the underlying patterns originating from specific operating schemes and any departure from this behavior will be falsely captured in the HI as a deterioration state. The complete procedure of the proposed approach for the estimation of the HI is depicted in Fig. 1. It consists of two main modes: first establishing and training an anticausal model from the selected variables and second the online prediction of the HI from incoming data. In anticausal learning mode, two phases are further unfolded. First, a heuristic method is introduced for ranking the operational and environmental parameters that corresponds to various time-varying operating conditions. The time series of the parameter with the largest weighted complexity metric yields the causal driver with the most of the temporal variability due to exogenous interventions and is selected as the target of the anticausal prediction model. Background knowledge also confirms this outcome. Second, bivariate causal indices are computed with nonlinear Granger causality between the measuring parameters and the causal driver from the first step. The set of the selected sensor parameters yield the definition of the individual structural causal models that jointly formulate the multivariate regression model in the anticausal direction. In the online mode, the HI of incoming data is estimated from the mean absolute prediction error and is evaluated based on its resulted trajectory. Note that in the offline mode, it is assumed that all training data correspond to the initial non-degraded system state, and thus represent the healthy reference model. Finally, we validate our framework with real-world operating conditions that reflect a wide range of flight scenarios.

4.1. Weighted complexity estimate

Usually, complex machinery need to operate throughout its lifetime under different time-varying conditions that may accordingly result to different degradation behavior, but same causal mechanism. For example, machine tools operate under various working conditions that are directly affected by many factors, such as cutting speed, load, vibration, etc. These parameters are recorded via sensor signals and holistically may represent specific operational schemes. However, not all these operational parameters that describe the external conditions have equal influence on the system's state. Hence, we introduce a metric to rank all operational parameters and ultimately select the causal driver that we later integrate into the proposed framework.

In the context of causality, as we already defined in Section 3.2, performing *controlled interventions* is the only way to verify causal relationships. In practice, however, it may be difficult or even infeasible to perform such interventions, especially in safety–critical complex systems such as commercial aircrafts. Under these pure observational settings, and assuming that there are no hidden common causes, we introduce the following metric for selecting the causal driver in the system.

Let us consider an univariate time series $Y = [y_1, y_2, \dots, y_T]$ with a length of T, from which we obtain *complexity estimate* (Batista et al., 2014) for every sliding window of size s_w as follows:

$$CE(Y_{t:t+s_w-1}) = \sqrt{\sum_{k=t}^{t+s_w-1} (y_k - y_{k+1})^2} \tag{11}$$

Intuitively, the complexity estimate expresses the volatility in terms of measuring lengths of stretched time series. For example, within the same time period a time series is more complex than the other, if we stretch them in a straight line and the more complex one would result to a greater length. In our proposed method for PHM, we extend

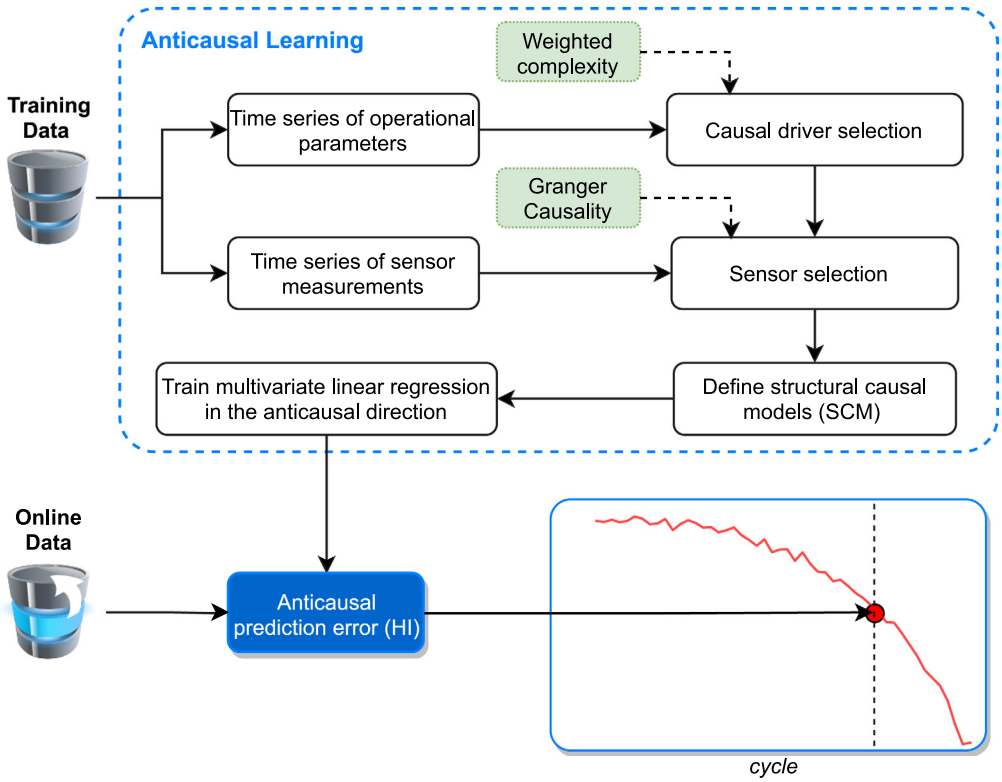


Fig. 1. Schematic representation of the proposed framework that integrates anticausal learning. The reference anticausal model captures the underlying generative process of the causal driver from the selected measuring parameters in the initial healthy state. Any deviation in the online mode of the predicted causal driver from the reference state represents the comprehensive HI from the corresponding cycle.

the existing complexity metric from Eq. (11) and introduce a *weighted* version

$$C\mathcal{E}(Y_{t:t+s_w-1}) = CE(Y_{t:t+s_w-1}) \times \mathbb{1}(\sigma(Y_{t:t+s_w-1}) > 1) \tag{12}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function and $\sigma(\cdot)$ the standard deviation of the sliding window. Note also that the time series is normalized by removing the mean and scaling to unit variance. In this way, only those windows are accounted which contain both large volatility and variability. Among a set of candidate causal drivers, the time series with the largest weighted complexity estimate is selected as the parameter that will be used for establishing the structural causal models, from which the anticausal predictive scheme is formulated. It is worth to point out that we choose to pick a single potential cause to ensure the underlying causal process is disentangled from the presence of other potential causal parents that might influence the anticausal learning process.

4.2. HI construction with multivariate anticausal regression

Let us first denote $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_N\}$ as the multivariate time series that represent the operating and environmental conditions, $\mathbf{X} = \{X_1, X_2, \dots, X_M\}$ as the measuring parameters and $\{\mathbf{X}, \mathbf{Y}\} \in D_l$, where D_l the domain with index $l \in \{1, 2, \dots, L\}$. A domain,[1] for instance, in the context of PHM, could represent a complete flight route of a commercial aircraft. Note that in this phase all L domains are represented as set of observations from the healthy condition of the system and time series from these domains are concatenated to form the training set. Assuming that these operational time series are independent with each other, since each parameter describes separate temporal interventions on a specific system's component, we compute the weighted complexity estimate introduced in the previous section. Yet, any dependence that might be present can be explained either by background knowledge (e.g., physical laws) or directly by a domain expert so that these variables to be omitted from the selection set. Finally, we select the causal driver Y_{i^*} such that

$$i^* = \underset{i}{\text{argmax}}(C\mathcal{E}_i), \quad i \in \{1, 2, \dots, N\}. \tag{13}$$

Once the causal driver Y_{i^*} is inferred, the next step is sensor selection from which all measuring time series are recorded. Thus, all causal indices $ci_{Y_{i^*} \to X_j}$ from Eq. (7) are calculated for each domain D_l to yield the final set S_{X^*} of cardinality M^* such that

$$S_{X^*} = \bigcup \{X_{j^*} : j^* = \underset{j}{\text{argmax}}(ci_{Y_{i^*} \to X_j})\}_{D_l}. \tag{14}$$

Since all causal dependencies are inferred we define the set of the underlying structural causal models as follows

$$X_j = f_j(Y_{i^*}) + \epsilon_j, \quad \text{for } j = 1, \dots, M^*, \tag{15}$$

where ϵ_j the independent noise terms that need to satisfy the assumption for the ANM, which is presented in the previous section. Graphically, these equations are illustrated in Fig. 2 as a causal graphical model. Further, the opposite direction, which represents the *anticausal prediction*, is shown with red, where we consider the input to be the set of the effect variables and the target the cause variable itself.

Although this setting may seem rather counterintuitive, in machine learning field it is indeed ubiquitous. For example, one can say that in the image classification task of handwritten digits, there is a causal relationship between the class label and the generated image. This might be justified from the causal reasoning that is expressed via a person's will to write down a specific digit that will be later captured in a structured digital form. In this regard, the class label causes the image, however, the predictive task is actually anticausal.

In the later phase of our proposed approach, a multivariate regression model is trained from the concatenated data in all healthy domains

[1] The terms cycle and domain, are interchangeably used throughout this paper.

D_l such that $f_a : S_{X^*} \mapsto Y_{i^*}$ which represents the anticausal prediction. Unlike other methods (Ye and Yu, 2021; Zhai et al., 2021) that use the probabilistic reconstruction error of associational models to quantify the degradation of the system, we simply estimate the mean absolute error from the test predictions for each cycle. The underlying idea of our model is that it does not need to explicitly learn the patterns in each healthy domain. Large shifts, due to system's degradation, in the input distribution $P(S_{X^*})$ can be accurately captured on the output conditional distribution $P(Y_{i^*}|S_{X^*})$ via the invariant causal mechanism, as discussed in the previous section. Hence, generalization on new unseen domains is achieved in a quite efficient and robust way. For online evaluation of a new incoming domain $\mathcal{D}'$ the degradation is computed as follows

$$d_{\mathcal{D}'} = \frac{1}{T}\sum_{k=1}^{T} |\hat{Y}_{i^*}(k) - Y_{i^*}(k)| \tag{16}$$

where T is the length of the time series and $\hat{Y}_{i^*}$ the anticausal prediction of the causal driver. Finally, the health index (HI) of a new cycle can be obtained by normalizing the degradation values into the range $[0, 1]$ by

$$HI = 1 - \frac{max(d_{\mathcal{D}'}^{(j)}) - d_{\mathcal{D}'}^{(j)}}{max(d_{\mathcal{D}'}^{(j)}) - min(d_{\mathcal{D}'}^{(j)})}, \quad j = 1, \dots, \mathcal{L} \tag{17}$$

where $\mathcal{L} > L$ the total number of cycles for the investigated unit.

4.3. Metrics for evaluation of HI

Since the constructed HIs are deployed in safety–critical systems, in which the accuracy of the RUL prediction is strongly dependent on them, they must fulfill several specific requirements. Specifically, at the beginning of the system's faultlessly operation, the HI should start at a maximum threshold value, i.e., HI=1; which outlines the "as-new" state. Based on the physical degradation laws, HI is decreasing over time, usually with a bilinear manner (but not necessarily), up to the onset of the abnormal degradation. Mathematically, the aforementioned behavior of the HI trajectory may be expressed with three evaluation metrics that are presented in Lei et al. (2018).

Fundamentally, a machine's health condition decreases over operating time, as most of the mechanical components in the system, some more and some less, is subjected to dynamical loads that may accelerate the overall degradation rate. Hence, the HI is expected to demonstrate a negative trend with the cycle or domain index. In order to handle non-linearities that occur more often in practice, the trendability is calculated from the Spearman correlation coefficient between the HI and the time index.

$$Tre(X, C) = \left| \frac{K\left(\sum_{k=1}^{K} \tilde{x}_k \tilde{c}_k\right) - \left(\sum_{k=1}^{K} \tilde{x}_k\right)\left(\sum_{k=1}^{K} \tilde{c}_k\right)}{\sqrt{\left[K\sum_{k=1}^{K} \tilde{x}_k^2 - \left(\sum_{k=1}^{K} \tilde{x}_k\right)^2\right]\left[K\sum_{k=1}^{K} \tilde{c}_k^2 - \left(\sum_{k=1}^{K} \tilde{c}_k\right)^2\right]}} \right|, \tag{18}$$

where $\{\tilde{x}\}_{k=1:K}$ and $\{\tilde{c}\}_{k=1:K}$ are the rank sequence of the HI $\{x\}_{k=1:K}$ and cycle $\{c\}_{k=1:K}$, respectively. K denotes the total number of cycles at the 100% of the operational life. Since the correlation coefficient should be negative to represent the degradation trajectory and as close as possible to -1, we put trendability in absolute value to ensure positivity on the final metric.

Any random fluctuations that reflect on the HI curve may have a great impact on the final evaluation results. To express any variability of the HI, that might originate either from the stochasticity of the degradation process or from the variability of the operating conditions, into a comprehensible value, we compute the robustness of HIs as follows,

$$Rob(X) = \frac{1}{K}\sum_{k=1}^{K} \exp\left(-\left|\frac{x_k - x_k^T}{x_k}\right|\right), \tag{19}$$

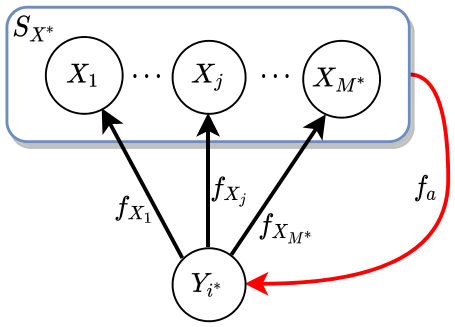


Fig. 2. Structural causal models between the causal driver Y_{i^*} derived by the maximum weighted complexity (potential cause) and the selected set S_{X^*} of the measuring parameters (effects). Red line indicates the opposite direction that maps the set of the effect variables with the cause as the target via a function f_a and represents the anticausal prediction model.

where x_k is the estimated value of the HI at cycle k and x_k^T is the mean trend value of the HI at cycle k that is obtained by smoothing or decomposition techniques. A high value of robustness will yield a smoother HI curve, which ultimately will enable more robust predictions of the RUL.

In real operating conditions, the machinery fails most of the time gradually in terms of the severity-based stages. HIs should be able to capture these degradation transitions in time, which is expressed via the identifiability metric. Similarly, as with trendability, identifiability yields the nonlinear correlation between the HI and the stage sequence.

$$Ide(X, C) = \frac{(m_e - m_l)^2}{\sigma_e^2 + \sigma_l^2}, \tag{20}$$

where m_e, σ_e^2 and m_l, σ_l^2 are the mean and the variance of the early and later degradation class, respectively. High identifiability further indicates that the HI can very well identify the true onset of the degradation, and hence accurately represents its stage sequence. Identifiability can be easily extended to more than two degradation classes (Lei et al., 2018), however we focus on a two-stage degradation pattern. Besides the three former metrics, for the convenience of comparison and validation of the proposed method, we introduce the *Applicability* (***App*** = ***Tre*** + ***Rob*** + ***Ide***) for the final evaluation of the computed HIs.

5. Experimental results

5.1. N-CMAPSS turbofan engine dataset

In the last decade, an extensively large amount of research in the field of PHM has been conducted, evaluated, and assessed with the CMAPSS (Commercial Modular Aero-Propulsion System Simulation) dataset of a large turbofan engine (Saxena et al., 2008b; Lei et al., 2018). A great advantage of the CMAPSS dataset is the incorporation of run-to-failure trajectories, an indispensable attribute for evaluation of data-driven prognostics algorithms. Usually, such real-worlds datasets are proprietary and they are quite seldom shared publicly by their operators. Although in the CMAPSS dataset, a wide range of operating values with the appropriate amount of mixed noise are set to approximate as close as possible the actual flight conditions, still the multifaceted complexity that is met in real data is missing. Further, individual measurement snapshots per flight are employed to the CMAPSS, so that health indices of the components of interest are inferred. This means that no temporal information during each flight is accounted into the simulation model, which seems a rather strong limitation that might lead to dubious conclusions regarding the efficiency of the developed algorithms. Finally, such aggregation on the flight data would not enable to capture any dependencies between

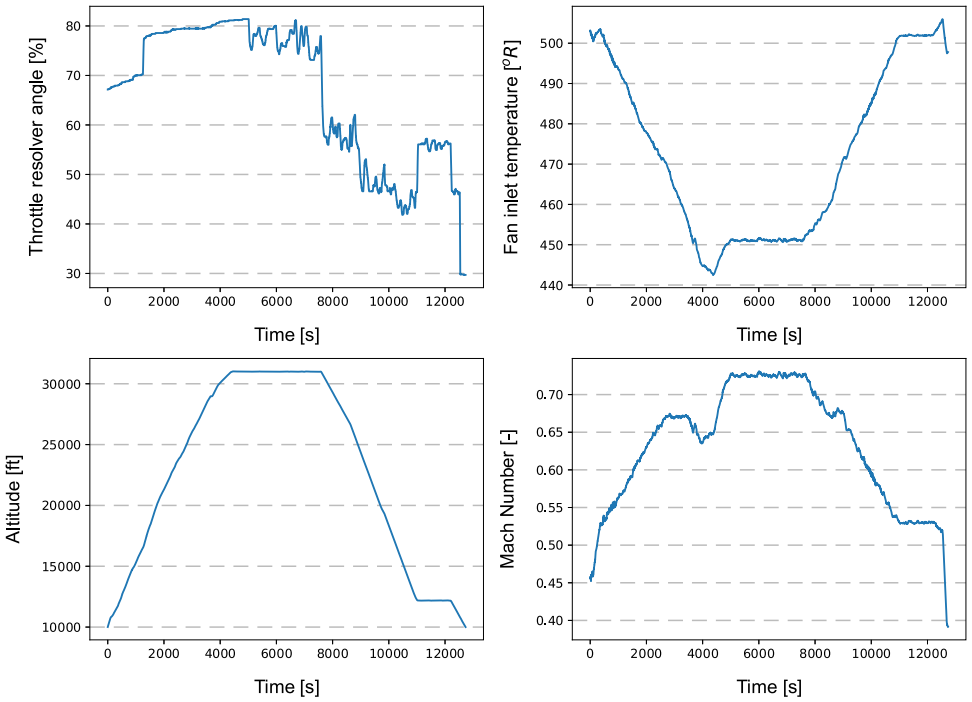


Fig. 3. Real-world time series of throttle-resolver angle (TRA), temperature of inlet fan (T2), altitude (alt) and speed (Mach) for a single unit across an entire flight course cycle (climb–cruise–descend).

the operation history and the abnormal degradation profile, which is a ubiquitous characteristic in every safety–critical system.

Recently, an improvement of the existing CMAPSS dataset is introduced in Arias Chao et al. (2021), also known as N-CMAPSS, that addresses all aforementioned gaps and challenges. First and foremost, real-world data that correspond to various climb, cruise and descend conditions are employed to the CMAPSS model, which realistically cover a wide range of flight routes. These flight conditions are captured by four operational and environmental parameter: flight Mach number, altitude, thrust-resolver angle and total temperature at fan inlet. An exemplary flight route of a given unit (cycle) with its four operational parameters is shown in Fig. 3. Since the entire temporal profile of the most critical operational parameters is observed, it is stochastically associated with the degradation modeling process by controlling the onset of the abnormal degradation. At the initial operation cycles of the engine, the components in various modules of interest are subjected to a normal degradation, mainly due to manufacturing imperfections, however, their state is still classified as healthy. In the next phase a normal continuous degradation is modeled by linearly adjusting the flow and efficiency parameters of five critical engine modules: fan, low pressure compressor (LPC), high pressure compressor (HPC), high pressure turbine (HPT), low pressure turbine (LPT). Finally, at the abnormal degradation, the estimated wear trend on the specific modules follows an exponential damage propagation law, as in the original CMAPSS model (Saxena et al., 2008b), and hence the system is labeled as unhealthy up to the point where the engine fails.

In general, the CMAPSS system model consists of a set of non-linear equations, in which the input is the four scenario-descriptor variables and the unobserved degradation trajectories of flow and efficiency from the affected engine modules. The outputs of the system model are the measuring parameters in various components that are presented in Table 2 with their descriptions, and an additional set of parameters from virtual sensors. Since we propose a purely data-driven solution, we exclude from our analysis the inferred set of the virtual sensors. On closer examination of the CMAPSS equation presented in Arias Chao et al. (2021), the degraded flight condition variables appear on the right-hand side enable a directed generating process of the variables in the left-hand side (i.e., observed and virtual sensors), and not the other way around. Within the paradigm of structural causal models the operational parameters represent the cause variables, the degradation process the causal mechanism, and finally the measuring parameters the effect variables.

Table 2
Physical measuring parameters that are estimated from the CMAPSS model.

#	Symbol	Description	Units
1	Wf	Fuel flow	pps
2	Nf	Physical fan speed	rpm
3	Nc	Physical core speed	rpm
4	T24	Total temperature at LPC outlet	°R
5	T30	Total temperature at HPC outlet	°R
6	T48	Total temperature at HPT outlet	°R
7	T50	Total temperature at LPT outlet	°R
8	P15	Total pressure in bypass-duct	psia
9	P2	Total pressure at fan inlet	psia
10	P21	Total pressure at fan outlet	psia
11	P24	Total pressure at LPC outlet	psia
12	Ps30	Static pressure at HPC outlet	psia
13	P40	Total pressure at burner outlet	psia
14	P50	Total pressure at LPT outlet	psia

We evaluate the proposed framework with the DS02 dataset from N-CMAPSS dataset, which has already been used for data-driven prognostics in Arias Chao et al. (2022). The dataset in total contains nine units and each unit is consisted of a specific number of cycles ranging from the beginning of the operation up to the point of engine failure. Each cycle comprises different kinds of flights in terms of duration and operational profiles (i.e., Mach speed, altitude). Table 3 summarizes the units from the DS02 dataset with the transition times t_k of the cycle k that designate the onset of the component's degradation, the total number of samples, the end-of-life time cycle t_{EOL} and the types of failure mode. In this work, we aim at finding the best representation of the entire system's health, and hence the investigation of the degradation of each component is out of scope. However, we show that our framework might indeed provide an intuition to domain experts from the sensor selection step that would indicate the possible root cause of the upcoming engine degradation.

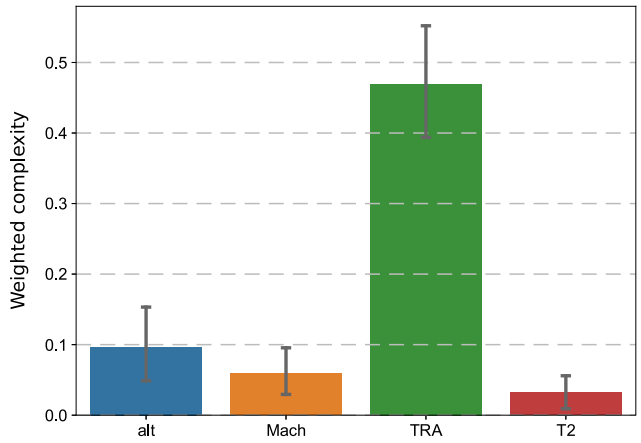


Fig. 4. Mean values of the weighted complexity from the four operational and environmental parameters that characterize the time-varying operating conditions. The bootstrap technique with 100 iterations is performed to calculate 95% confidence intervals.

Table 3
Dataset characteristics of the investigated engine units with their failure modes. Superscripts (f), (e) denote the flow capacity and efficiency of the corresponding sub-components, respectively.

Unit	Time steps	t_k	t_{EOL}	Failure mode
2	853,142	17	75	HPT^e
5	1,033,420	17	89	HPT^e
10	952,711	17	82	HPT^e
16	765,295	16	63	HPT^e+LPT^e+LPT^f
18	890,719	17	71	HPT^e+LPT^e+LPT^f
20	768,160	17	66	HPT^e+LPT^e+LPT^f
11	663,495	19	59	HPT^e+LPT^e+LPT^f
14	156,778	36	76	HPT^e+LPT^e+LPT^f
15	433,470	24	67	HPT^e+LPT^e+LPT^f

5.2. Data preparation for anticausal learning

In real-world settings, each turbofan engine unit manifests a unique degradation footprint due to its multivariate operational profile, which need to be captured by those parameters that encode invariant causal structures. Even though in the N-CMAPSS dataset the operational variables are already established, it is non-trivial to infer which one of those operational variables is the causal driver. Based on the heuristic method that we introduce in Section 4.1, we compute the *weighted complexity* for all four operational variables across all units. In particular, within each unit the first 20% of the operating cycles are considered as healthy and are included for the computation of the weighted complexity. From here onwards, we refer as the training set the first 20% of the operational useful life and as test set the 80% remaining cycles for each investigated unit. The complexity within the training set from each unit is computed using a sliding window approach with a size of 10% of the total cycle length. In the purpose of computation of the complexity the data is normalized with the z-score method by removing the mean and scaling to unit variance.

Fig. 4 presents the averaged weighted complexity of the four operational and environmental parameters. The results clearly rank TRA as the operation parameter that varies most in time with potentially the largest casual footprint on the measuring parameters. Our findings may easily be derived based on engineering principles that are mainly applicable in civil aviation industry. Flight operational parameters, such as cruise speed (Mach) and altitude (alt), are mostly changing gradually due to adjustment of the cabin pressure. Any sudden large fluctuations of those parameters might have an impact for both the crew, the passengers, and the normal operation of the aircraft. Temperature at the fan inlet (T2) is basically influenced by the altitude, provided that no other external conditions are applied. Throttle-resolver angle (TRA) controls the overall thrust of the engine, which subsequently influences the speed of the aircraft, and ultimately the ability of the aircraft to fly.

Table 4
Sensor selection per unit via bivariate Granger causality.

#Unit	Wf	Nf	Nc	T24	T30	T48	T50	P15	P2	P21	P24	Ps30	P40	P50
2	✓		✓	✓		✓	✓	✓		✓		✓	✓	✓
5	✓	✓	✓	✓	✓			✓	✓		✓	✓	✓	
10	✓	✓		✓	✓	✓	✓	✓	✓	✓			✓	✓
16			✓	✓		✓		✓	✓	✓	✓		✓	
18		✓		✓	✓	✓		✓			✓	✓	✓	
20		✓	✓	✓		✓	✓		✓	✓		✓	✓	✓
11			✓	✓	✓	✓			✓	✓			✓	
14	✓		✓	✓	✓		✓			✓				
15	✓		✓	✓				✓	✓			✓		✓

Once the causal driver is inferred, we perform sensor selection on the measuring parameters from the bivariate Granger causality indices introduced in Section 3.1. Each unit is investigated separately since the flight scenarios are from real-world data and generate unique degradation trajectories. In all experiments, we set the lag to a single time step so that to capture the closest in time causal dependencies. Table 4 shows the selected measurements for each investigated unit. The same training set as in the ranking phase is used here as well. Each of these feature set S_{X^*} denotes the effect variables which, based on our hypothesis, are mainly generated from the causal driver, namely TRA. Hence, individual structural causal models are defined for each unit that will be later used for training the multivariate regression in the anticausal direction.

5.3. HI construction

We further perform experiments to compare the effectiveness and robustness of the anticausal approach. In the first set of evaluation methods, two deep learning architectures from Malhotra et al. (2016) and Liu et al. (2020) are used as state-of-the-art, which aim to estimate the reconstruction error of the input for deriving the health index of the system. For the sake of comparison, feature extraction methods are not included, as the sampling frequency of 1 Hz in the N-CMAPSS dataset is relatively low, and hence frequency spectra may be inherently biased. In the second set, both linear and non-linear regression with the anticausal approach are evaluated to highlight and verify the power of the proposed framework. As in the previous steps of our framework, we used as training set the 20% of the operational life of each unit and the rest of the data for evaluation purposes.

Deep learning models. First, a reconstruction-based approach with LSTM layers (LSTM-AE) of the entire time series parameter set (operational and measuring) is considered for comparison purposes. LSTM-AE contains an input layer of 32 LSTM cells in the encoder, and accordingly an output layer of 32 LSTM cells in the decoder. The dimension of the latent space is set to 5, from which the data are reconstructed via the decoder layer. Previous hyperparameters are set in accordance with Liu et al. (2020). Note that no hidden layers are inserted, since few data are available for training, and as such, lower model complexity is preferred (Khan et al., 2021). Nevertheless, we further conducted experiments with an additional hidden layer on the architecture and it showed no improvement on the results from the main setting. This fact verifies indeed the previous assertion regarding model complexity. Since the input of the LSTM-AE needs to be in multiple sliding windows, the length of each window and its stride is set to 10 and 1, respectively. Moreover, as a baseline approach we build an autoencoder (AE) network architecture, which integrates a single-layered encoder, a hidden latent representation, and finally, the single-layered decoder that reconstructs the input data. Considering the limited amount of training data, we set one input and output layer of size 32 neurons for both encoder and decoder, respectively. The latent dimension of the middle layer is set to 8 neurons. For training of both architectures, the

Adam optimizer is used with a learning rate of 0.001 and a batch size of 64 for 50 epochs. Throughout both network architectures, rectified linear units (*ReLu*) are used as activation functions. To further prevent overfitting, early stopping with a patience of 5 epochs is implemented. All experiments are performed on Python 3.8.6 with Tensorflow 2.4.0 software platform under a laptop with Intel-i7 8565U 1.8 GHz CPU, 16 GB RAM.

As in Liu et al. (2020), the health index HI_r at both cases is calculated from the root mean square error of the original multivariate input data $\mathbf{X}$ and the reconstructed input $\hat{\mathbf{X}}$.

$$HI_r = RMS(\hat{\mathbf{X}} - \mathbf{X}) \tag{21}$$

In both aforementioned methods, both operational and measuring parameters are concatenated and used for training the networks across all units. Note that no feature selection, like in the proposed approach, is applied, since all parameters are included. Hence, the dimension of the input space (i.e., $X \in \mathbb{R}^{T\times 18}$) is constant, where T is the size of the concatenated time series of the reference (healthy) cycles and 18 of the input size, namely 14 measuring and 4 operational parameters.

Anticausal regression. To demonstrate the low model complexity of our approach we employ three linear regression methods with anticausal learning: ordinary least-squares (LR^+), support vector machine with a linear kernel ($LSVR^+$) and least-squares with l2 regularization (RR^+). Moreover, a multi-layer perceptron (MLP) is selected as a non-linear regression with wide applicability so that to investigate the behavior of the proposed approach in settings with higher model complexity. We employed a rather simple MLP architecture that comprises two layers with 64 neurons each that are activated by *ReLu* functions. Except least-squares regression, all other methods require adjustment of their hyperparameter (e.g., λ_2 for ridge regression and C regularization parameter for the linear support vector machine). For that purpose, we employ a grid search scheme with a fivefold cross validation for obtaining the optimal set of the above hyperparameters. The search procedure for RR^+ yielded a value of $\lambda_2 = 0.01$ and for $LSVR^+$ a regularization parameter of $C = 40$. Note that the dimension of the input space for anticausal learning varies, as it depends on the sensor selection step for each unit, which is displayed in Table 4. For example, in Unit 2 the dimension of the input space is $X \in \mathbb{R}^{T\times 10}$ that shows that the number of the measuring parameters is reduced to 10.

5.4. Results and discussion

In this section, the constructed HIs are extrapolated on test cycles for all units to compare their performance based on the evaluation metrics that are presented in Section 4.3. In particular, Table 5 summarizes each of these metrics that show the effectiveness of the proposed method. Note that no smoothing is applied on the HI values, which renders the evaluation process more challenging, and at the same time highlights the robustness of the proposed approach. Interestingly, based on the overall applicability metric, all three linear regression methods that are used with the proposed anticausal learning outperform both state-of-the-art deep learning methods, as well as the non-linear regression. In contrast to other methods with increased model complexity, linear regression disentangles the individual structural causal models in the anticausal direction by avoiding any deterioration of the causal power via its inherent additivity. MLP^+ outperforms both autoencoder architectures, still it has worse performance than the linear regression, as deep neural networks have been found to not strongly generalize for anticausal learning schemes (Kilbertus et al., 2018). Two regression methods that employ l2-regularization ($LSVR^+$, RR^+) deliver similar results, and indeed, better that the least-squares method, highlighting the positive impact of the regularization on the final performance.

Comparing the trendability values shown from different models (see Table 5), all three linear methods (LR^+, $LSVR^+$, RR^+) consistently have the best results. This indicates how well the monotonic trend of the engine's degradation is captured by the anticausal approach, even with the linear models. On the other hand, AE-based methods exhibit lower trendability values, since they captured associational relationships that may collapse under changeable operating conditions. Traditionally, these methods require huge amounts of data to perform well. Training such architectures with fewer data may lead to large bias, which in our case is also reflected to the identifiability that is dropped almost to the half of the one computed from the linear models.

Table 5
Resulted metrics for all HIs values averaged across all units. Best results for each category metric is highlighted in bold.

HI method	Trendability	Robustness	Identifiability	Applicability
LSTM-AE	0.6739	0.9461	0.4720	2.0921
AE	0.7494	0.9506	0.4131	2.1131
LR^+	0.9120	0.9714	0.8497	2.7330
$LSVR^+$	**0.9325**	**0.9745**	**0.8798**	**2.7868**
RR^+	0.9311	0.9735	0.8716	2.7763
MLP^+	0.7097	0.9500	0.6684	2.3281

To further highlight the generalization performance of the proposed approach, we obtain the aggregated values from the true individual degradation parameters used in the C-MAPSS simulation model. In each engine unit, and for a specific operating cycle, flow and efficiency of various critical sub-components is subjected to a degradation that is captured by health degradation parameters $\theta(k)$, where k is the index cycle. Finally, the minimum value is computed by the aggregation of all non-zero degradation parameters and is denoted as:

$$HI_\theta(k) = min([\theta_c^f(k), \theta_c^e(k)]) \tag{22}$$

where $\theta_c^f(k), \theta_c^e(k)$ is the health degradation of flow and efficiency from a corresponding sub-component c respectively.

Fig. 5 illustrates the predicted HI values with the proposed approach and the true aggregated values. Note that all values are min–max normalized so that to be in the range of 0 and 1. It is evident that the approximation of the underlying degradation is very well captured from the proposed approach. Especially, in unit 16 the predictions in the test cycles are almost overlapping with the true values. It is worth noticing that in unit 15 with a multi-modal failure mode, some outliers in the HIs are present, however the overall degradation trajectory is not distorted since the minimum predicted HI corresponds to the actual failure cycle of the engine. Such implication is probably a result of the causal sufficiency assumption, since it might affect the definition of the structural causal models by selecting a single causal driver. Furthermore, any deterioration in the HI predictions might originate from the linearity of the anticausal model, however the overall degradation trend is well captured without any significant deviations.

Fig. 6 clearly illustrates that both least-squares and linear SVM with l2-regularization outperform the state-of-the-art with very small averaged RMSE values. In almost all units, both methods show a consistent behavior, while least-squares regression performs similarly well with an exception in unit 2, which is due to the aforementioned implications. Nevertheless, across all units linear regression is better than both deep learning methods that shows the power of the anticausal scheme combined with low complexity models. On the other hand, computational models with higher complexity, such as MLP, performs worse in some of the units. This deterioration in performance of the MLP is mainly caused due to limitations of this family of models in anticausal learning (Kilbertus et al., 2018). Finally, lack of training data from few cycles is an additional factor that largely influences the performance of the former non-linear models.

As this article primarily focuses on the development of the HI construction methodology for degradation monitoring, the prognosis of the RUL was not deeply investigated. However, we used the proposed HI to further evaluate its prognostic performance to highlight the effectiveness, and hence, the utility in practical applications.

Considering the RUL estimation from the constructed HI, a HI-based RUL prediction approach as in Yang et al. (2016) is adopted that

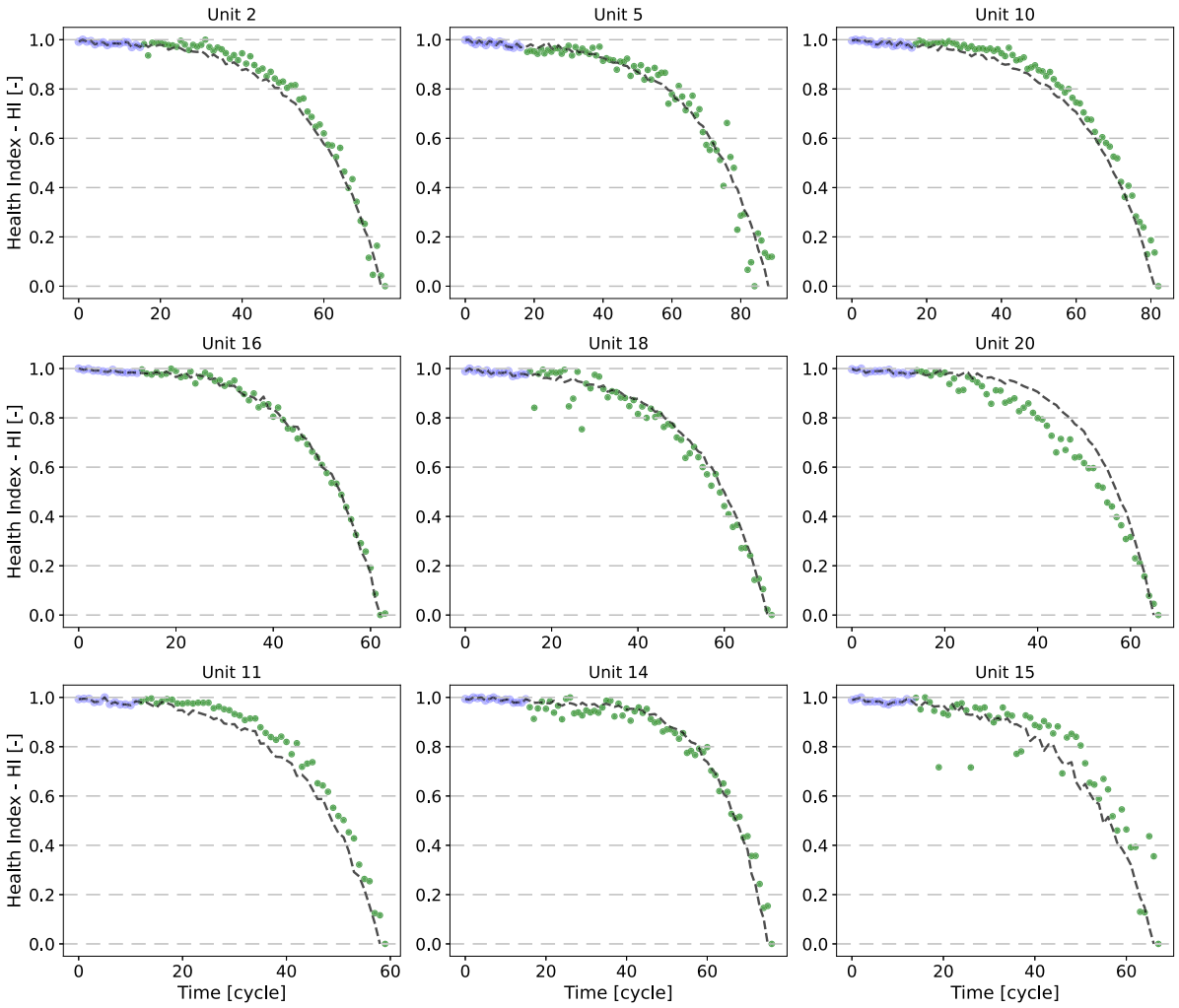


Fig. 5. Predicted HI values via anticausal learning with linear support vector regression (LSVR$^+$). Black dashed line represents the ground truth of the aggregated health degradation values for each unit. Blue points depict the cycles that are used for training and green points depict the test cycles for extrapolating HIs.

takes account units with different lifetimes. Since the method needs to aggregate the predictions in an ensemble manner, first, the resulted HIs from each training unit (2, 5, 10, 16, 18 and 20) and the corresponding RUL are used to build a model for each of these units. For obtaining the final RUL prediction on each cycle within a testing unit, single predictions from multiple models are averaged (ensemble). According to Arias Chao et al. (2021), units 11, 14 and 15 are selected for testing as their operational flight conditions are significantly different than those of the training units. Due to the limited number of instances (cycles) within each unit, k-nearest neighbors (KNN) is selected as the non-linear regression method for direct mapping HI → RUL. We used the default parameter from the package scikit-learn for the number of neighbors to obtain unbiased final results of the prognostic metrics.

In this study two common metrics (Saxena et al., 2008a) are used to evaluate the system health prognostics of the proposed method: root mean square error (e) and PHM score (s). The equations of calculating both metrics are expressed as follows

$$s = \begin{cases} \sum_{i=1}^{N} \exp(-\Delta r_i/13) - 1, & \text{if } d_i < 0 \\ \sum_{i=1}^{N} \exp(\Delta r_i/10) - 1, & \text{if } d_i \geq 0 \end{cases} \tag{23}$$

$$e = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \Delta r_i^2} \tag{24}$$

where Δr_i denotes the error between the predicted and the true RUL of the ith testing cycle, respectively. Due to the asymmetry of the prediction metric s the overestimation of the RUL (degradation level is lower than failure threshold) is penalized more than the underestimation (degradation level is higher than failure threshold). Therefore, lower values of s indicate that the proposed method can successfully address the overestimation issue that may have more severe implications in practice. Table 6 presents the prognostic results from the RUL that are derived by the HI-based ensemble approach. Interestingly, HIs that are constructed with anticausal learning yield better results in both metrics. It is worth noting that the anticausal HI with the linear regression achieves the lowest error values, which demonstrates its high performance with very low model complexity. The findings from the prognostics evaluation are completely consistent with the previous results (presented in Table 5) that show all linear anticausal-based HIs have better performance in terms of monotonicity and robustness than the rest of the compared methods. These results show a promising aspect of the proposed approach regarding its utility for prognostics in further studies.

6. Conclusion

In this work, a novel VHI is developed for degradation monitoring of safety–critical engineered systems that operate under time-varying conditions. By incorporating limited domain knowledge with a two-phase heuristics approach, a causal driver and a set of measuring parameters

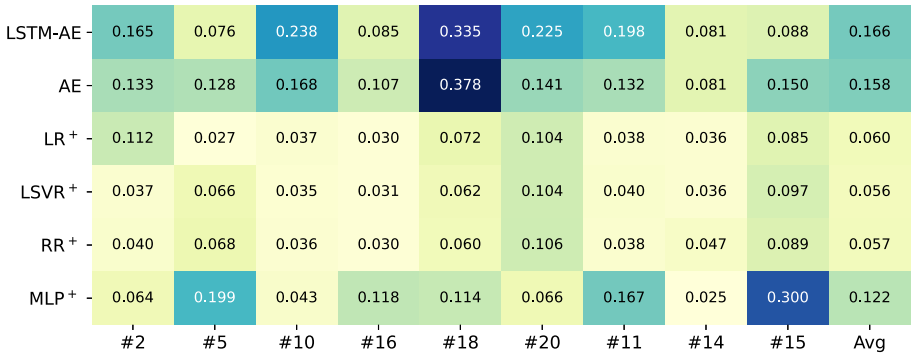


Fig. 6. Matrix with RMSE values of predicted HI values across all investigated units. Brighter colors indicate smaller RMSE values, while darker colors larger RMSE values. Averaged values are displayed in the first column on the right-hand side.

Table 6
Prognostic performance of all HIs across testing units 11, 14 and 15. The first column for each unit presents the root mean square e of the RUL in cycles, while the second one presents the PHM score s. Smaller values of both metrics indicate better performance.

HI method	Unit 11		Unit 14		Unit 15	
	e [cycles]	s [–]	e [cycles]	s [–]	e [cycles]	s [–]
LSTM-AE	14.24	161	17.76	314	14.70	232
AE	13.84	149	20.35	404	10.38	107
LR+	**11.49**	**130**	**10.91**	**114**	**8.92**	**82**
LSVR+	12.27	148	11.24	124	9.36	86
RR+	12.50	154	12.88	151	8.95	83
MLP+	11.53	132	11.15	127	18.25	249

are selected. In the next phase, structural causal models are built that entail the invariance properties of the causal mechanism. Anticausal learning from only few degradation-free cycles is then performed via models with low computational complexity and high interpretability. Finally, robust HIs are extrapolated on noisy test cycles for capturing holistic degradation trends.

The performance and efficiency of the proposed framework is evaluated on the N-CMAPSS turbofan engine dataset, which contains real-world operational time series data. Multifaceted degradation trajectories of flow and efficiency in specific components was generated by the CMAPSS simulation model. For comparison and validation purposes, reconstruction-based methods with increased computational complexity were employed for extracting HI values. Anticausal HIs with linear methods exhibited outstanding monotonic behavior, and low prediction error values that proved its consistency on the ground truth degradation trajectories. In addition, the constructed HIs were further evaluated on RUL prediction and it showed that anticausal learning with linear regression even outperformed other deep learning methods, thus highlighting the utility of the proposed approach for practical applications in PHM.

Still, several issues remain open that need further investigation. First and foremost, more than one causal driver may be included for analysis to address the hidden confounder issue. Hence, more complex causal structures could be learned that might accordingly deliver more accurate predictions. Second, shape quality metrics of the HIs could be integrated in existing causal structure learning algorithms for building targeted optimization schemes. Future work will further address how robust RUL predictions with physics-based approaches can be obtained by integrating the proposed health indicator. Towards that direction, model parameter uncertainty on the RUL predictions will be also in-depth investigated.

CRediT authorship contribution statement

Georgios Koutroulis: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Belgin Mutlu:** Project administration, Funding acquisition, Writing – review & editing. **Roman Kern:** Conceptualization, Resources, Writing – review & editing, Supervision.

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.

Acknowledgments

This work has been supported by the FFG, Contract No. 881844: "Pro^{2}Future is funded within the Austrian COMET Program Competence Centers for Excellent Technologies under the auspices of the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology, the Austrian Federal Ministry for Digital and Economic Affairs and of the Provinces of Upper Austria and Styria. COMET is managed by the Austrian Research Promotion Agency FFG".

References

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

Ancona, N., Marinazzo, D., Stramaglia, S., 2004. Radial basis function approach to nonlinear granger causality of time series. Phys. Rev. E 70 (5), 056221. http://dx.doi.org/10.1103/PhysRevE.70.056221.

Arias Chao, M., Kulkarni, C., Goebel, K., Fink, O., 2021. Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. Data 6 (1), 5. http://dx.doi.org/10.3390/data6010005.

Arias Chao, M., Kulkarni, C., Goebel, K., Fink, O., 2022. Fusing physics-based and deep learning models for prognostics. Reliab. Eng. Syst. Saf. 217, 107961. http://dx.doi.org/10.1016/j.ress.2021.107961.

Baptista, M.L., Goebel, K., Henriques, E.M., 2022. Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. Artificial Intelligence 306, 103667. http://dx.doi.org/10.1016/j.artint.2022.103667.

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

Batista, G.E., Keogh, E.J., Tataw, O.M., De Souza, V.M., 2014. CID: An efficient complexity-invariant distance for time series. Data Min. Knowl. Discov. 28 (3), 634–669. http://dx.doi.org/10.1007/s10618-013-0312-3.

Benkedjouh, T., Medjaher, K., Zerhouni, N., Rechak, S., 2013. Remaining useful life estimation based on nonlinear feature reduction and support vector regression. Eng. Appl. Artif. Intell. 26 (7), 1751–1760. http://dx.doi.org/10.1016/j.engappai.2013.02.006.

Bühlmann, P., 2020. Rejoinder: Invariance, causality and robustness. Statist. Sci. 35 (3), 434–436. http://dx.doi.org/10.1214/20-STS797.

Chen, L., Xu, G., Zhang, S., Yan, W., Wu, Q., 2020. Health indicator construction of machinery based on end-to-end trainable convolution recurrent neural networks. J. Manuf. Syst. 54, 1–11. http://dx.doi.org/10.1016/j.jmsy.2019.11.008.

Edinburgh, T., Eglen, S.J., Ercole, A., 2021. Causality indices for bivariate time series data: A comparative review of performance. Chaos 31 (8), 083111. http://dx.doi.org/10.1063/5.0053519.

Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.-J., Ducoffe, M., 2020. Potential, challenges and future directions for deep learning in prognostics and health management applications. Eng. Appl. Artif. Intell. 92, 103678. http://dx.doi.org/10.1016/j.engappai.2020.103678.

Fu, S., Zhong, S., Lin, L., Zhao, M., 2021. A novel time-series memory auto-encoder with sequentially updated reconstructions for remaining useful life prediction. IEEE Trans. Neural Netw. Learn. Syst. http://dx.doi.org/10.1109/TNNLS.2021.3084249.

Geweke, J.F., 1984. Measures of conditional linear dependence and feedback between time series. J. Amer. Statist. Assoc. 79 (388), 907–915. http://dx.doi.org/10.2307/2288723.

Goebel, K., Daigle, M.J., Saxena, A., Roychoudhury, I., Sankararaman, S., Celaya, J.R., 2017. Prognostics: The science of making predictions.

Granger, C., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37 (3), 424–438. http://dx.doi.org/10.2307/1912791.

Groenenboom, J., 2018. The changing MRO landscape: IATA maintenance cost conference, athens, Greece. URL https://www.iata.org/contentassets/81005748740046de878439e6c54f2355/d1-1100-1130-the-changing-mro-landscape_icf.pdf.

Gugulothu, N., Tv, V., Malhotra, P., Vig, L., Agarwal, P., Shroff, G., 2017. Predicting remaining useful life using time series embeddings based on recurrent neural networks. 9,

Guo, L., Lei, Y., Li, N., Yan, T., Li, N., 2018. Machinery health indicator construction based on convolutional neural networks considering trend burr. Neurocomputing 292, 142–150. http://dx.doi.org/10.1016/j.neucom.2018.02.083.

Guo, J., Li, Z., Li, M., 2019. A review on prognostics methods for engineering systems. IEEE Trans. Reliab. 69 (3), 1110–1129. http://dx.doi.org/10.1109/TR.2019.2957965.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hong, Y., Liu, Y., Wang, S., 2009. Granger causality in risk and detection of extreme risk spillover between financial markets. J. Econometrics 150 (2), 271–287. http://dx.doi.org/10.1016/j.jeconom.2008.12.013.

Hoyer, P.O., Janzing, D., Mooij, J.M., Peters, J., Schölkopf, B., et al., 2008. Nonlinear causal discovery with additive noise models. In: NIPS, Vol. 21. Citeseer pp. 689–696.

Hu, C., Smith, W.A., Randall, R.B., Peng, Z., 2016. Development of a gear vibration indicator and its application in gear wear monitoring. Mech. Syst. Signal Process. 76, 319–336. http://dx.doi.org/10.1016/j.ymssp.2016.01.018.

Hu, C., Youn, B.D., Wang, P., Yoon, J.T., 2012. Ensemble of data-driven prognostic algorithms for robust prediction of remaining useful life. Reliab. Eng. Syst. Saf. 103, 120–135. http://dx.doi.org/10.1016/j.ress.2012.03.008.

Janzing, D., Schölkopf, B., 2010. Causal inference using the algorithmic Markov condition. IEEE Trans. Inform. Theory 56 (10), 5168–5194. http://dx.doi.org/10.1109/TIT.2010.2060095.

Javed, K., Gouriveau, R., Zerhouni, N., Nectoux, P., 2014. Enabling health monitoring approach based on vibration data for accurate prognostics. IEEE Trans. Ind. Electron. 62 (1), 647–656. http://dx.doi.org/10.1109/TIE.2014.2327917.

Karasu, S., Altan, A., 2022. Crude oil time series prediction model based on LSTM network with chaotic henry gas solubility optimization. Energy 242, 122964. http://dx.doi.org/10.1016/j.energy.2021.122964.

Karasu, S., Altan, A., Bekiros, S., Ahmad, W., 2020. A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series. Energy 212, 118750. http://dx.doi.org/10.1016/j.energy.2020.118750.

Khan, S., Tsutsumi, S., Yairi, T., Nakasuka, S., 2021. Robustness of AI-based prognostic and systems health management. Annu. Rev. Control 51, 130–152. http://dx.doi.org/10.1016/j.arcontrol.2021.04.001.

Kilbertus, N., Parascandolo, G., Schölkopf, B., 2018. Generalization in anti-causal learning. In: NeurIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning. authors are listed in alphabetical order.

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., Lin, J., 2018. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. Mech. Syst. Signal Process. 104, 799–834. http://dx.doi.org/10.1016/j.ymssp.2017.11.016.

Li, N., Gebraeel, N., Lei, Y., Bian, L., Si, X., 2019. Remaining useful life prediction of machinery under time-varying operating conditions based on a two-factor state-space model. Reliab. Eng. Syst. Saf. 186, 88–100. http://dx.doi.org/10.1016/j.ress.2019.02.017.

Liu, C., Sun, J., Liu, H., Lei, S., Hu, X., 2020. Complex engineered system health indexes extraction using low frequency raw time-series data based on deep learning methods. Measurement 161, 107890. http://dx.doi.org/10.1016/j.measurement.2020.107890.

Lozano, A.C., Li, H., Niculescu-Mizil, A., Liu, Y., Perlich, C., Hosking, J., Abe, N., 2009. Spatial-temporal causal modeling for climate change attribution. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 587–596. http://dx.doi.org/10.1145/1557019.1557086.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Lungarella, M., Ishiguro, K., Kuniyoshi, Y., Otsu, N., 2007. Methods for quantifying the causal structure of bivariate time series. Int. J. Bifurcation Chaos 17 (03), 903–921. http://dx.doi.org/10.1142/S0218127407017628.

Luo, B., Wang, H., Liu, H., Li, B., Peng, F., 2018. Early fault detection of machine tools based on deep learning and dynamic identification. IEEE Trans. Ind. Electron. 66 (1), 509–518. http://dx.doi.org/10.1109/TIE.2018.2807414.

Malhotra, P., TV, V., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., Shroff, G., 2016. Multi-sensor prognostics using an unsupervised health index based on LSTM encoder-decoder. arXiv preprint arXiv:1608.06154.

Malhotra, P., TV, V., Vig, L., Agarwal, P., Shroff, G., 2017. TimeNet: PRe-trained deep recurrent neural network for time series classification. arXiv preprint arXiv:1706.08838.

Marcus, G., 2018. Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631.

Marinazzo, D., Pellicoro, M., Stramaglia, S., 2008. Kernel method for nonlinear granger causality. Phys. Rev. Lett. 100 (14), 144103. http://dx.doi.org/10.1103/PhysRevLett.100.144103.

Medjaher, K., Zerhouni, N., Baklouti, J., 2013. Data-driven prognostics based on health indicator construction: Application to PRONOSTIA's data. In: 2013 European Control Conference (ECC). IEEE, pp. 1451–1456. http://dx.doi.org/10.23919/ECC.2013.6669223.

Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., Schölkopf, B., 2016. Distinguishing cause from effect using observational data: methods and benchmarks. J. Mach. Learn. Res. 17 (1), 1103–1204.

Nguyen, K.T., Medjaher, K., 2021. An automated health indicator construction methodology for prognostics based on multi-criteria optimization. ISA Trans. 113, 81–96. http://dx.doi.org/10.1016/j.isatra.2020.03.017.

Pearl, J., 2009. Causality. Cambridge University Press.

Peters, J., Janzing, D., Schölkopf, B., 2017. Elements of Causal Inference: Foundations and Learning Algorithms. The MIT Press.

Qin, Y., Wang, D., Zhao, X., Ma, H., Jia, L., Zhang, Y., 2016. Performance degradation assessment of train rolling bearings based on SVM and segmented vote method. In: 2016 Prognostics and System Health Management Conference (PHM-Chengdu). IEEE, pp. 1–6. http://dx.doi.org/10.1109/PHM.2016.7819891.

Qiu, H., Liu, Y., Subrahmanya, N.A., Li, W., 2012. Granger causality for time-series anomaly detection. In: 2012 IEEE 12th International Conference on Data Mining. IEEE, pp. 1074–1079. http://dx.doi.org/10.1109/ICDM.2012.73.

Rodrigues, L.R., Yoneyama, T., Nascimento, C.L., 2012. How aircraft operators can benefit from PHM techniques. In: 2012 IEEE Aerospace Conference. IEEE, pp. 1–8. http://dx.doi.org/10.1109/AERO.2012.6187376.

Roebroeck, A., Formisano, E., Goebel, R., 2005. Mapping directed influence over the brain using granger causality and fMRI. Neuroimage 25 (1), 230–242. http://dx.doi.org/10.1016/j.neuroimage.2004.11.017.

Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., Schwabacher, M., 2008a. Metrics for evaluating performance of prognostic techniques. In: 2008 International Conference on Prognostics and Health Management. IEEE, pp. 1–17. http://dx.doi.org/10.1109/PHM.2008.4711436.

Saxena, A., Goebel, K., Simon, D., Eklund, N., 2008b. Damage propagation modeling for aircraft engine run-to-failure simulation. In: 2008 International Conference on Prognostics and Health Management. IEEE, pp. 1–9. http://dx.doi.org/10.1109/PHM.2008.4711414.

Schölkopf, B., 2019. Causality for machine learning. arXiv preprint arXiv:1911.10500.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., Mooij, J., 2012. On causal and anticausal learning. In: Proceedings of the 29th International Coference on International Conference on Machine Learning. In: ICML'12, Omni Press, Madison, WI, USA, pp. 459–466.

Shafiee, M., 2015. Maintenance strategy selection problem: an MCDM overview. J. Qual. Maint. Eng. http://dx.doi.org/10.1108/JQME-09-2013-0063.

Thoppil, N.M., Vasu, V., Rao, C., 2021. Deep learning algorithms for machinery health prognostics using time-series data: A review. J. Vib. Eng. Technol. 1–23.

Wang, T., Yu, J., Siegel, D., Lee, J., 2008. A similarity-based prognostics approach for remaining useful life estimation of engineered systems. In: 2008 International Conference on Prognostics and Health Management. IEEE, pp. 1–6. http://dx.doi.org/10.1109/PHM.2008.4711421.

Yang, F., Habibullah, M.S., Zhang, T., Xu, Z., Lim, P., Nadarajan, S., 2016. Health index-based prognostics for remaining useful life predictions in electrical machines. IEEE Trans. Ind. Electron. 63 (4), 2633–2644. http://dx.doi.org/10.1109/TIE.2016.2515054.

Ye, Z., Yu, J., 2021. Health condition monitoring of machines based on long short-term memory convolutional autoencoder. Appl. Soft Comput. 107, 107379. http://dx.doi.org/10.1016/j.asoc.2021.107379.

Yu, W., Kim, I.Y., Mechefske, C., 2019. Remaining useful life estimation using a bidirectional recurrent neural network based autoencoder scheme. Mech. Syst. Signal Process. 129, 764–780. http://dx.doi.org/10.1016/j.ymssp.2019.05.005.

Zhai, S., Gehring, B., Reinhart, G., 2021. Enabling predictive maintenance integrated production scheduling by operation-specific health prognostics with generative deep learning. J. Manuf. Syst. http://dx.doi.org/10.1016/j.jmsy.2021.02.006.

Zhou, Z.-H., Li, M., et al., 2005. Semi-supervised regression with co-training. In: IJCAI, Vol. 5. pp. 908–913.

Zhu, H., 2021. Real-time prognostics of engineered systems under time varying external conditions based on the COX PHM and VARX hybrid approach. Sensors 21 (5), 1712. http://dx.doi.org/10.3390/s21051712.