A Shape-Constrained Neural Data Fusion Network for Health Index Construction and Residual Life Prediction

Zhen Li, *Member, IEEE,* Jianguo Wu, *Member, IEEE,* and Xiaowei Yue, *Member, IEEE*

***Abstract*— With the rapid development of sensor technologies, multisensor signals are now readily available for health condition monitoring and remaining useful life (RUL) prediction. To fully utilize these signals for a better health condition assessment and RUL prediction, health indices are often constructed through various data fusion techniques. Nevertheless, most of the existing methods fuse signals linearly, which may not be sufficient to characterize the health status for RUL prediction. To address this issue and improve the predictability, this article proposes a novel nonlinear data fusion approach, namely, a shape-constrained neural data fusion network for health index construction. Especially, a neural network-based structure is employed, and a novel loss function is formulated by simultaneously considering the monotonicity and curvature of the constructed health index and its variability at the failure time. A tailored adaptive moment estimation algorithm (Adam) is proposed for model parameter estimation. The effectiveness of the proposed method is demonstrated and compared through a case study using the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) data set.**

***Index Terms*— Condition monitoring, health index, neural data fusion network, remaining useful life (RUL) prediction, shape constrained.**

NOMENCLATURE

N	Number of historical units.
$\boldsymbol{T}$	Total life cycles of these N units.
S	Number of sensors.
$\boldsymbol{D}$	Multisensor data.
$\boldsymbol{X}_n$	Multisensor data of size $T_n \times S$ for unit n.
$\boldsymbol{x}_{n,s}$	Sensor data for unit n of sensor s.
$x_{n,s,t}$	Sensor data for unit n and sensor s at observation epoch t.
$h_i(t)$	Health index of unit i.
$H(\cdot)$	Nonlinear data fusion function.
$L(\cdot)$	Loss function.
$\boldsymbol{\theta}$	Model parameters.
$\delta(x)$	Indicator function.
λ_1	Monotonicity penalty coefficient.
λ_2	Convexity penalty coefficient.
$\boldsymbol{g}_n$	Gradient of loss function for sample n.
α	Step size.
$\boldsymbol{\theta}_0$	Initial value for θ to be optimized.
$\boldsymbol{m}_0$	First-moment vector.
$\boldsymbol{v}_0$	Second-moment vector.
k	Time step.
β_1 and β_2	Exponential decay rates for the moment estimates.
y_t	Degradation signal or health index at time step t.
$\phi(\cdot)$	Parametric linear or nonlinear function.
$\boldsymbol{\alpha}$	Vector of the fixed-effect parameters.
$\boldsymbol{\gamma}$	Vector of the random-effect parameters.
$\boldsymbol{z}_t$	Vector of polynomial basis functions.
ε_t	Noise term.
σ^2	Variance of the noise ε_t.
$\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_t$	Prior and posterior means of $\boldsymbol{\gamma}$.
$\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_t$	Prior and posterior variances of $\boldsymbol{\gamma}$.
ρ	Pairwise correlation threshold.
$\hat{R}_n$	Predicted residual life for unit n.
R_n	True remaining useful life (RUL) of unit n.

Manuscript received October 2, 2019; revised April 7, 2020 and July 28, 2020; accepted September 22, 2020. This work was supported by the Natural Science Foundation of China under Grant 51875003 and Grant 71932006. *(Corresponding author: Jianguo Wu.)*

Zhen Li and Jianguo Wu are with the Department of Industrial Engineering and Management, College of Engineering, Peking University, Beijing 100871, China (e-mail: zhen.li@pku.edu.cn; j.wu@pku.edu.cn).

Xiaowei Yue is with the Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: xwy@vt.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.3026644

I. INTRODUCTION

DEGRADATION is often unavoidable and invertible, in nature, in the utilization of machines, tools, equipment, or systems. The unexpected breakdown of machines due to degradation may cause severe consequences, such as the suspension of production, the occurrence of safety hazards, economic losses, and delay in delivery. Therefore, monitoring the health condition and predicting the remaining useful life (RUL) are critically important to prevent unexpected failures and ensure the system/process reliability [1]–[3]. There are generally two types of prognostic methods, physics-based and data-driven approaches [4]. Physics-based methods require a full understanding of the fundamental degradation mechanisms, which may not be realistic due to high system complexity. The data-driven approaches, on the other hand, do not need the knowledge of the failure processes and, thus, have gained wide popularity as data acquisition becomes more and more convenient. With the collected sensor signals in hand, the progression of the degradation process can be well characterized using various fit degradation models, such as stochastic process models and general path models [5], [6].

Most of the existing studies concentrate on the degradation modeling based on a single-sensor signal [7], [8]. The underlying assumption is that the single sensor is sufficient to capture the degradation process. However, as the machines or engineering systems become increasingly more complex, it may not be possible for a single sensor to characterize the health condition accurately [9], [10]. With the rapid development of sensor technologies, multisensor signals are now readily available for health condition monitoring and residual life prediction. Multisensor signals have the advantage of capturing multichannel information on the degradation process from various physical aspects. Consequently, there is a trend of incorporating multiple sensors to monitor the health status of a working unit.

While multisensor systems are widely adopted in health condition monitoring, how to make full use of multisensor signals to better capture the health condition and predict the RUL is very challenging. In general, it is often not possible for every sensor to have the same capability of capturing the degradation process. Some signals may possess stronger relationships than others with the underlying degradation process. As each signal only contains partial information and multisensor signals may be correlated with each other, a properly justified data fusion of multisensors may yield a more precise and robust health condition estimation and RUL prediction.

Based on the implementation strategy, data fusion techniques can be generally classified into three categories [11]: data-level fusion, feature-level fusion, and decision-level fusion. Data-level fusion approaches directly utilize the multisensor data as input in the model building. The feature-level fusion, on the other hand, first extracts various features from the raw data and then combines these features using data-level fusion techniques. In the decision-level fusion techniques, individual models are built based on each sensor data, and then, the outputs of all the individual models are fused to generate the final prediction result. In this article, we focus on the data-level fusion. We assume that the underlying degradation process can be accurately characterized by one unobservable metric, which is referred to as the health index. Once the health index reaches a predefined threshold, the unit or system is considered failed. The objective of this work is to develop a novel data-level fusion technique in the construction of a health index for health condition assessment and RUL prediction. It should be noted that some neural network-based RUL prediction methods [12]–[14] treat the predicted percentage of residual life as a health index or health indicator, which linearly decreases to 0 and is fundamentally different from the one defined earlier.

Health index construction using data fusion techniques has been intensively studied recently [3], [15]. Compared with those directly using multiple signals as model input to predict the residual life, e.g., various neural network-based approaches [13], [16], [17], the health index-based approaches have the advantage of not only providing straightforward visualization of the health condition but also facilitating RUL prediction using various existing univariate prognostics. Liu *et al.* [3] developed a composite health index by linearly combining these sensor data by jointly minimizing the variance of the constructed index at the failure time and maximizing the degradation monotonicity. Later, another linear fusion approach is developed by jointly minimizing the model fitting errors and the variance of constructed health index at the failure time [18]. Song and Liu [19] developed a linear fusion approach by leveraging a quantile regression for fusion coefficient optimization, where the prediction error is directly utilized as the loss function. However, all these data fusion methods for health index construction are linear, which may have inherent limitations due to high nonlinearity between sensor signals and health conditions. To capture the nonlinearity, Song *et al.* [20] integrated the kernel method with a linear fusion approach. Nevertheless, the form of the nonlinearity is constrained by the type of kernel functions. There are also some distance- or deviation-based health index construction methods, which are inherently nonlinear, such as the minimum quantization error (MQE) method [21], [22], the Mahalanobis distance (MD)-based method [23], and the support vector data description (SVDD)-based approach [24]. The basic idea of these methods is to calculate the distance (e.g., the Euclidean distance) or deviation between the degrading states and normal/initial conditions using the multivariate degradation signals and treat the distance as a health index. In these methods, it is often assumed that the normal/initial condition is identical no matter what kind of fault the unit will eventually have. However, due to unit-to-unit heterogeneity, different units may have different initial health conditions. Besides, although the distance can somehow measure the severity of the degradation, it may not be able to accurately capture the true health condition. Therefore, it is hard to set a unified failure threshold for all the units.

To overcome these issues, this article develops a novel shape-constrained neural data fusion network for nonlinear health index construction. Especially, an unsupervised neural network is first proposed to model the nonlinear relationship between the true health index and the multisensor signals. Besides the property of monotonicity and minimum variance at the failure time, we also innovatively incorporate the convexity property into loss function to improve the quality of health index construction. In addition, a new penalty function is proposed to capture the shape constraints in the loss function formulation. A tailored adaptive moment estimation algorithm (Adam) is further proposed to optimize the loss function for parameter estimation. Different from the existing neural network-based methods where the RULs are used as labels in a supervised learning process, the proposed method is unsupervised and is intrinsically a dimension reduction technique for health index construction.

The rest of this article is organized as follows. Section II provides an overview of the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) data set as a motivating example and presents a general mathematical formulation toward this problem. Section III describes the detailed methodologies of the shape-constrained neural data fusion approach to construct a health index. Section IV demonstrates the effectiveness of the proposed method through a case study of the NASA C-MAPSS data set for commercial aircraft gas turbine engines [9]. Section V concludes this article and outlines future directions.

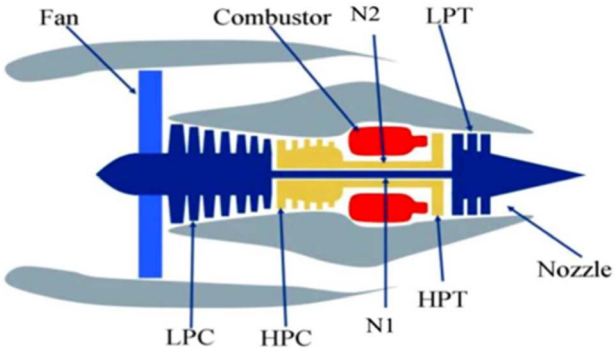

Fig. 1. Simplified engine diagram simulated in C-MAPSS [9].

II. Problem Description

A. Overview of the C-MAPSS Data Set

Before introducing the proposed methodology, we first provide an overview of the motivating example, i.e., C-MAPSS data set, which has been widely used as a benchmark problem in the prognostics and health management (PHM) field. C-MAPSS is a simulation tool developed by the National Aeronautics and Space Administration (NASA) for a realistic simulation of a large commercial turbofan engine, whose health condition is continuously monitored by multiple embedded sensors. Fig. 1 shows a schematic of a commercial aircraft gas turbine engine that was simulated by C-MAPSS [9].

In the C-MAPSS simulation, an engine model of the 90 000-lb thrust class is developed, and simulations are run for operations at various levels of altitude (from 0 to 40 000 ft), Mach number (from 0 to 0.90), and throttle resolver angle (TRA, from 20 to 100). Users can freely adjust these three attributes (aircraft altitude, Mach number, and TRA) to simulate various environmental conditions. C-MAPSS has 14 input variables to simulate various degradation scenarios of the five rotating components of the simulated engine. The outputs consist of 58 different variables in the form of sensor response surfaces and operability margins, of which a total of 21 variables were used for health condition monitoring and prognosis, as shown in Table I.

To model unit-to-unit heterogeneity, the initial degradation levels, the degradation trajectory parameters, and the process noises were set random in the simulation. A hidden health index that is not accessible to users was defined, and once the health index exceeds the failure threshold, the unit is considered failed. In total, there are four data sets with each corresponding to certain failure modes and operating conditions. The FD001 data set has a single failure mode (HPC degradation) and a single operating condition, while the FD002 data set contains six operating conditions, all of which seriously affect the sensor measurements and the degradation process. The FD003 data set has one operating condition but two failure modes (HPC degradation and fan degradation). For the FD004 data set, there are two failure modes and six operating conditions mixed together. As illustrated in the original work for the C-MAPSS data set [9], the degradation trajectories significantly vary for different failure modes and change as a function of operational conditions

TABLE I
Description of the 21 C-MAPSS Outputs

Symbol	Description	Units
T2	Total temperature at fan inlet	°R
T24	Total temperature at LPC outlet	°R
T30	Total temperature at HPC outlet	°R
T50	Total temperature at LPT outlet	°R
P2	Pressure at fan inlet	psia
P15	Total pressure in bypass-duct	psia
P30	Total pressure at HPC outlet	psia
Nf	Physical fan speed	rpm
Nc	Physical core speed	rpm
epr	Engine pressure radio(P50/P2)	–
Ps30	Static pressure at HPC outlet	psia
phi	Ratio of fuel flow to Ps30	pps/psi
NRf	Corrected fan speed	rpm
NRc	Corrected core speed	rpm
BPR	Bypass Ratio	–
farB	Burner fuel-air ratio	–
htBleed	Bleed Enthalpy	–
Nf_dmd	Demanded fan speed	rpm
PCNfR_dmd	Demanded corrected fan speed	rpm
W31	HPT coolant bleed	lbm/s
W32	LPT coolant bleed	lbm/s

(e.g., TRA, altitude, and ambient temperature). Therefore, if there are multiple failure modes or operating conditions, a classification or clustering algorithm could be used first to group the data set, and then, different fusion functions can be applied accordingly [25]–[27]. In this article, we only consider the first data set in the case study, which has been widely used in performance evaluation and comparison. In future studies, our approach will be extended to multiple failure modes and multiple operating conditions.

The considered data set consists of 100 training units and 100 testing units. In the training data set, each unit is run until failed, while, in the testing data set, the run of each unit ends at some random time prior to system failure. A file of the actual remaining lifetime of the 100 testing units is also included. Sensor readings from the 21 outputs are collected at each observation epoch for each unit. The objective of this research is to develop a novel nonlinear data-level fusion approach to combining these multisensor data so that a health index can be constructed to accurately characterize the health condition for better condition monitoring and RUL prediction.

B. Problem Formulation

Let N denote the total number of historical units, e.g., the training units, and $\boldsymbol{T}=(T_1,\ldots,T_N)$ denote the total life cycles of these N units. Suppose that the corresponding multisensor data are represented by $\boldsymbol{D}=\{\boldsymbol{X}_n, n=1,\ldots,N\}$, where $\boldsymbol{X}_n$ is the multisensor data of size $T_n \times S$ for unit n given as

$$\begin{aligned}\boldsymbol{X}_n &= (\boldsymbol{x}_{n,1}, \boldsymbol{x}_{n,2}, \ldots, \boldsymbol{x}_{n,S}) \\ &= \begin{bmatrix} x_{n,1,1} & x_{n,2,1} & \cdots & x_{n,S,1} \\ x_{n,1,2} & x_{n,2,2} & \cdots & x_{n,S,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1,T_n} & x_{n,2,T_n} & \cdots & x_{n,S,T_n} \end{bmatrix}\end{aligned} \tag{1}$$

where $\boldsymbol{x}_{n,s}=(x_{n,s,1},\ldots,x_{n,s,T_n})'$ is the sth sensor data for unit n, $x_{n,s,t}$ is the sensor data for unit n, sensor s is at observation epoch t, and S is the total number of sensors. Without loss of generality, we assume that all the measurements are equally spaced in time with interval 1.

This article aims to find a nonlinear data fusion function $H(\cdot)$ to combine multisensor signals for unit i into a health index $h_i(t)$

$$h_i(t) = H(x_{i,1,t}, x_{i,2,t}, \ldots, x_{i,S,t}). \tag{2}$$

In order to guarantee that the resulting health index has desirable properties for health condition monitoring and RUL prediction, appropriate fusion function and loss function need to be formulated for the nonlinear function learning. Suppose that the formulated loss function is $L(H, \boldsymbol{D}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the model parameters and H is the selected fusion function, and the learning process is then formulated to the following optimization problem:

$$\text{obj} = \min_{\boldsymbol{\theta}} L(H, \boldsymbol{D}, \boldsymbol{\theta}).$$

In Section III, technical details will be provided on the selection of nonlinear functions and appropriate loss function. For the sake of completeness, the degradation and RUL prediction based on a sequential Bayesian updating scheme will also be introduced.

III. Shape-Constrained Neural Data Fusion Network

In this section, we first introduce several key properties that are essential in the construction of the health index and then formulate the loss function by considering these key properties. A tailored Adam algorithm is then proposed for parameter estimation of the neural data fusion model. For completeness, the Bayesian updating scheme is also provided for RUL prediction based on a single degradation signal.

A. Key Properties for a Health Index

To facilitate health assessment and RUL prediction, it is essential to set appropriate properties that the constructed health index should satisfy. The most desirable properties are the monotonicity of the degradation path and the consistency of degradation level at the failure time across all units [3]. The details of these two properties are given as follows.

Property 1: The health index should be monotonic after the onset of the degradation of a monitored unit.

Property 2: The variance of the health index at the failure time for different units should be minimal under the same environmental condition and failure mode.

Property 1 suggests that the health index should have a clearer monotonic trend despite the fact that the raw sensor data may be contaminated by nonmonotonic noisy samples. This property is very natural, as, in most cases, the damage accumulates, and the degradation is irreversibly or monotonically growing in the operation process. Property 2, on the other hand, ensures the health index to be consistent at the failure time across different units under the same failure mode. Although the degradation paths across different units may differ significantly due to unit-to-unit heterogeneity, the degradation values at their failure times should be close to each other. This requirement is consistent with the soft failure assumption that, once the degradation value reaches a predefined threshold, the unit is considered failed. Therefore, it is essential to incorporate the minimal variance property to ensure the prognostic accuracy.

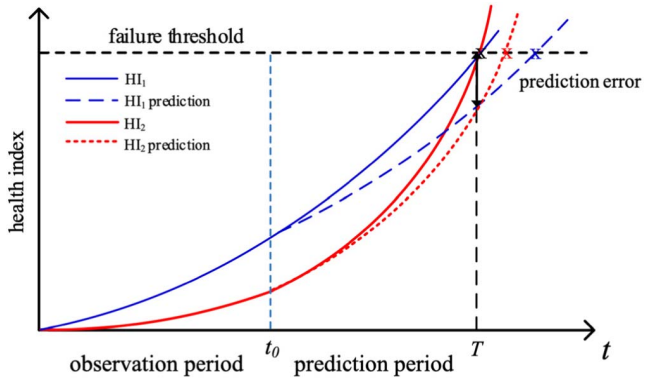

Fig. 2. Influence of the curvature property on the prediction accuracy.

To further improve the quality of the constructed health index, here, we propose another property regarding the degradation rate, which is given in Property 3, as follows.

Property 3: The health index should be convex when monotonically increasing or concave when monotonically decreasing.

While Property 1 guarantees that the overall degradation trend should be monotonic with the operation time, i.e., increasing or decreasing, Property 3 further puts a constraint on the shape of the health index in terms of degradation rate. To be specific, this property indicates that the health index should increase or decrease faster and faster, or equivalently, the degradation rate is monotonically increasing with operating time. This phenomenon of accelerated degradation occurs in a wide variety of engineering systems and degradation processes, such as bearings [28]–[31], batteries [32], [33], gyros [34], and fatigue crack growth in aircraft structures and other mechanical components [34], [35]. Taking this property into consideration could enrich the prior knowledge of the degradation path and, thus, improve the quality of the health index. Another rationale of including this property is that it could further increase the monotonicity of the health index when the in-service unit approaches the end of its lifetime, which, as a result, can increase the prognostic accuracy.

As illustrated in Fig. 2, two health indices, i.e., HI_1 and HI_2, with different curvatures are developed for the same unit. According to the definition of soft failure, the working unit fails at time T as the health index reaches the failure threshold. HI_2 has a more convex degradation trend than HI_1, i.e., the derivative of the degradation rate of HI_2 is larger than that of HI_1. The two solid curves represent the constructed health indices, while the dashed curves are the predictions of the health indices at the prediction time t_0. To make a fair

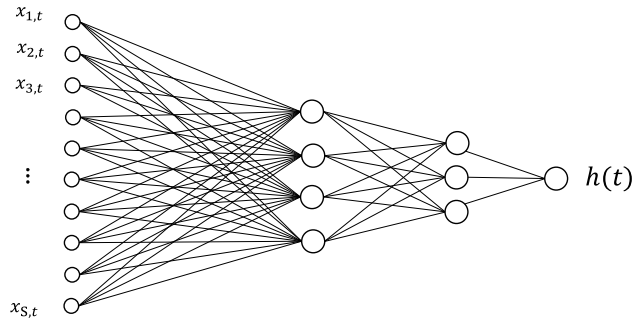

Fig. 3. Illustration of the neural data fusion network.

comparison, we assume that the errors of the predicted health indices are the same at the failure time T (marked with a double-sided arrow). It is clear that the RUL prediction using HI_2 tends to be more accurate than using HI_1. It is worth noting that, in Fig. 2, the predicted degradation level is below the failure threshold, which results in an overestimated RUL. On the other hand, if the predicted degradation level at T is above the failure threshold, the estimated failure time will be earlier than T, and HI_2 will still be more accurate than HI_1 in RUL prediction.

Without loss of generality, we assume that the health index is monotonically increasing, under which the Property 3 is simply a convexity shape constraint on the health index. As all units are assumed to fail under the same failure mode and operating condition, this article aims to develop a composite health index that exhibits a consistent pattern for all units in the degradation process, which can be achieved by jointly maximizing the monotonicity, convexity property, and minimizing the variance of the health index at the failure time in the health index construction. Most of the current methods for health index development are based on linear fusion approaches that have inherent limitations. Thus, this article proposes an innovative neural data fusion model with the consideration of the three aforementioned key properties for health index construction.

B. Neural Data Fusion Model

In this article, the underlying health status for a unit at time t is characterized by the health index $H(x_{1,t}, x_{2,t}, \ldots, x_{S,t})$. Due to its excellent capability of approximating any nonlinear functions, the artificial neural network is proposed to model the fusion function $H(\cdot)$. Unlike the traditional neural network where the output is just a single value for each instance, the proposed neural data fusion model generates a health index curve or profile for each unit.

Fig. 3 shows an illustrative example of the neural data fusion model. The hyperbolic tangent function $\tanh(\cdot)$ is selected as the activation function among the input layer and the hidden layers due to its advantages over the sigmoid function in overcoming the vanishing gradient problem. Between the last hidden layer and the output layer, the identity activation function is used.

In the loss function formulation, since the true health index is not available, we propose to apply the unsupervised learning approach, i.e., jointly maximizing the monotonicity and convexity properties and minimizing the variance at the failure time. To minimize the variance, we could use the following loss function based on the variance definition:

$$\sum_{n=1}^{N}[h_n(T_n) - \overline{h}]^2 \tag{3}$$

where $\overline{h} = (1/N)\sum_{n=1}^{N} h_n(T_n)$. The abovementioned loss function actually does not involve the bias term or the intercept of the last hidden layer; therefore, we can arbitrarily select a bias term. Note that changing the bias term is equivalent to shifting the whole health index upward or downward, which will not influence the RUL prediction. In this article, we specifically set the failure threshold to 1 to determine the bias term, which leads to the following loss function:

$$\sum_{n=1}^{N}[h_n(T_n) - 1]^2. \tag{4}$$

It can be easily shown that

$$\sum_{n=1}^{N}[h_n(T_n) - 1]^2 = \sum_{n=1}^{N}[h_n(T_n) - \overline{h}]^2 + N(1-\overline{h})^2. \tag{5}$$

Therefore, minimizing (4) is actually equivalent to minimizing (3) and then adjusting or shifting the bias parameter in the last hidden layer so that $\overline{h} = 1$. To impose the monotonicity and convexity constraint, penalty terms are typically introduced. The most commonly used penalty term for the monotonicity constraint is [3], [36]

$$\lambda_1 \sum_{n=1}^{N}\sum_{t=2}^{T_n} \max(0, h_n(t-1) - h_n(t)) \tag{6}$$

or equivalently

$$\lambda_1 \sum_{n=1}^{N}\sum_{t=2}^{T_n} [h_n(t-1) - h_n(t)]_+ \tag{7}$$

where $[x]_+ = \max(0, x)$, and λ_1 is a tuning parameter. In this penalty function, the penalty linearly increases with the difference $h(t-1) - h(t)$ when it violates the monotonicity. Alternatively, we could use the exponential penalty $\max(0, \exp[h(t-1) - h(t)] - 1)$ or the hyperbolic tangent penalty $\max(0, \tanh[h(t-1) - h(t)])$. The three types of penalty as a function of the monotonicity violation amount x are shown in Fig. 4.

It is clear that, for the exponential penalty, the penalty exponentially increases with the amount of monotonicity violation, while, for the hyperbolic tangent penalty, the penalty gradually goes saturated. Intuitively, the exponential penalty is more likely to eliminate a large decrease in the health index than the other two functions. Indeed, in the case study, we found that the exponential penalty outperforms the other two in terms of prognostic accuracy. It is worth noting that, to impose strict monotonicity constraint, $[h_n(t-1) - h_n(t) + c]_+$ can be used instead in (7), where c is a positive constant.

For the convexity constraint, a penalty term could be proposed in the same way. Especially, if the exponential penalty

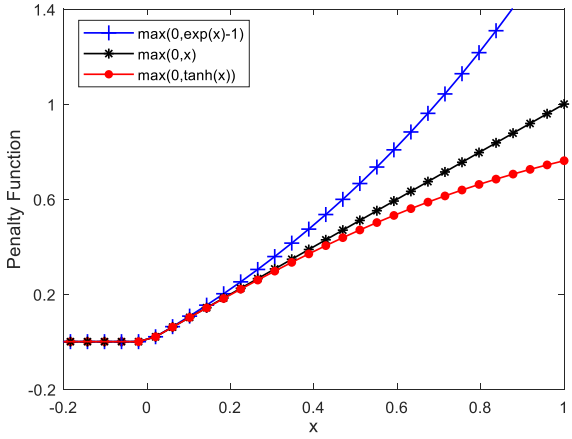

Fig. 4. Three penalty functions.

function is utilized, the penalty term can be expressed as

$$\lambda_2 \sum_{n=1}^{N} \sum_{t=3}^{T_n} [\exp((h_n(t-1) - h_n(t-2)) - (h_n(t) - h_n(t-1))) - 1]_+ \tag{8}$$

where λ_2 is a tuning parameter.

In practical applications, the total lifetime often varies significantly across different units. To reduce the penalty unbalance caused by lifetime difference, we propose to average the penalty in (6)–(8) by the number of terms for each unit. Suppose that the exponential penalty function is adopted. Let $d_{n,t} = h_n(t-1) - h_n(t)$, and then, the overall loss function considering the three key properties can be formulated as follows:

$$\begin{aligned} L(H, D, \theta) = & \sum_{n=1}^{N} [h_n(T_n) - 1]^2 \\ & + \lambda_1 \sum_{n=1}^{N} \sum_{t=2}^{T_n} \frac{1}{T_n - 1} [\exp(d_{n,t}) - 1]_+ \\ & + \lambda_2 \sum_{n=1}^{N} \sum_{t=3}^{T_n} \frac{1}{T_n - 2} [\exp(d_{n,t} - d_{n,t-1}) - 1]_+. \end{aligned} \tag{9}$$

C. Model Fitting Through Adaptive Moment Estimation Algorithm

Stochastic gradient descent (SGD) with a backpropagation algorithm is a generic and the most popular optimization method for minimizing the loss function of artificial neural networks. In each iteration, the SGD only uses a random subset (minibatch) of the training samples to evaluate the gradient, which can significantly reduce the computational cost. Adam [37] is an extension of SGD that has been widely adopted recently for model training in deep learning applications. It is an algorithm for the first-order gradient-based optimization of stochastic objective functions based on adaptive estimates of lower order moments. As it has many advantages over the traditional SGD approach, e.g., straightforward to implement, computationally efficient, and having little memory requirements, the Adam algorithm is adopted in our work to minimize the loss function.

The key step in SGD or Adam is to compute the gradient at each iteration. However, the loss function in (9) is not differentiable everywhere. To apply Adam, we can use the subgradient instead at those nondifferentiable points. Let $\delta(x)$ be an indicator function, which is defined as

$$\delta(x) = \begin{cases} 0, & x \le 0 \\ 1, & x > 0. \end{cases}$$

Then, the gradient or one subgradient of $L_n(H, \boldsymbol{D}, \boldsymbol{\theta})$ for sample n can be derived as

$$\begin{aligned} \boldsymbol{g}_n = & 2[h_n(T_n) - 1] \frac{\partial h_n(T_n)}{\partial \boldsymbol{\theta}} + \frac{\lambda_1}{T_n - 1} \sum_{t=2}^{T_n} \delta(\exp(d_{n,t}) - 1) \\ & \times \exp(d_{n,t}) \frac{\partial d_{n,t}}{\partial \boldsymbol{\theta}} + \frac{\lambda_2}{T_n - 2} \sum_{t=3}^{T_n} \delta(\exp(d_{n,t} - d_{n,t-1}) - 1) \\ & \times \exp(d_{n,t} - d_{n,t-1}) \frac{\partial (d_{n,t} - d_{n,t-1})}{\partial \boldsymbol{\theta}}. \end{aligned} \tag{10}$$

In the iteration process, the gradient or subgradient $\boldsymbol{g}_n$ can be efficiently calculated through the backpropagation algorithm.

The detailed Adam algorithm for the model fitting is provided in Algorithm 1, where q_k^2 represents the elementwise square. Following the suggestions in [37], we set the hyperparameters as $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ in the training process.

Algorithm 1: Adam for Neural Data Fusion Model Fitting

Initialize:
α : step size;$\boldsymbol{\theta}_0$: the initial value for $\boldsymbol{\theta}$ to be optimized;
$k \leftarrow 0$ (time step);
$\boldsymbol{m}_0, \boldsymbol{v}_0 \leftarrow \boldsymbol{0}$ (the 1st and 2nd moment vector);
$\beta_1, \beta_2 \in [0, 1)$: exponential decay rates for the two moment estimates.
While $\boldsymbol{\theta}_k$ *not converged* **Do**
  $k \leftarrow k + 1$
  Calculate the gradient or sub-gradient $\boldsymbol{q}_k$ based on (10)
  $\boldsymbol{m}_k \leftarrow \beta_1 \cdot \boldsymbol{m}_{k-1} + (1 - \beta_1) \cdot \boldsymbol{q}_k$ (update biased 1st moment estimate)
  $\boldsymbol{v}_k \leftarrow \beta_2 \cdot \boldsymbol{v}_{k-1} + (1 - \beta_2) \cdot \boldsymbol{q}_k^2$ (update biased 2nd raw moment estimate)
  $\hat{\boldsymbol{m}}_k \leftarrow \frac{\boldsymbol{m}_k}{1 - \beta_1^k}$ (compute bias-corrected 1st moment estimate)
  $\hat{\boldsymbol{v}}_k \leftarrow \frac{\boldsymbol{v}_k}{1 - \beta_2^k}$ (compute bias-corrected 2nd raw moment estimate)
  $\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_{k-1} - \alpha \cdot \hat{m}_k / (\sqrt{\hat{v}_k + \epsilon})$ (update parameters)
End
Return $\boldsymbol{\theta}_k$

D. Bayesian Linear Modeling and RUL Prediction

For the sake of completeness, we provide a brief review of the Bayesian inference-based RUL prediction approach.

The Bayesian modeling and online updating approach has been widely used for RUL prediction based on a single degradation signal or a health index [5], [30]. It leverages both the historical data in the form of prior distributions and *in situ* measured degradation signals of an in-service unit for residual life prediction. Therefore, it is capable of capturing both the population degradation trajectory and unit heterogeneity. There are often two stages: the off-line modeling of the historical data using mixed-effect model and the online Bayesian model updating of an individual unit for RUL prediction.

In the off-line modeling state, the mixed-effects model can be generally formulated as

$$y_t = \phi(\boldsymbol{\alpha}, \boldsymbol{\gamma}, t) + \varepsilon_t \tag{11}$$

where y_t is the degradation signal or health index at time step t, $\phi(\cdot)$ is a parametric linear or nonlinear function, $\boldsymbol{\alpha}$ is a vector of the fixed-effect parameters, $\boldsymbol{\gamma}$ is a vector of the random-effect parameters, and ε_t is a noise term following i.i.d. normal distributions, e.g., $\varepsilon_t \sim N(0, \sigma^2)$. A linear model, particularly a polynomial model, is often specified in $\phi(\cdot)$ for its simplicity and flexibility. The linear model with a polynomial form can be expressed as

$$y_t = \boldsymbol{z}_t \boldsymbol{\gamma} + \varepsilon_t \tag{12}$$

where $\boldsymbol{z}_t = (1, t, t^2, \ldots, t^q)$ is a $(q+1)$-D vector of polynomial basis functions, and $\boldsymbol{\gamma}$ is the regression parameter of dimension $q+1$ following a multivariate normal distribution. To make the model more flexible to capture the noise variance heterogeneity, σ^2 can also be assumed random. To facilitate the online Bayesian inference, a joint distribution is often specified [30], [38], i.e., $\sigma^2 \sim \text{IG}(a_1, a_2)$, $\boldsymbol{\gamma}|\sigma^2 \sim N(\boldsymbol{\mu}_0, \sigma^2 \boldsymbol{\Sigma}_0)$, where IG denotes the inverse Gamma (IG) distribution. The hyperparameters $(a_1, a_2, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ can be estimated through the maximum likelihood estimation method.

In the online stage, the individual model is updated for RUL prediction. Define

$$\boldsymbol{Z}_{1,t} = \begin{bmatrix} 1 & 1 & \cdots & 1^q \\ 1 & 2 & \cdots & 2^q \\ \cdots & \cdots & \cdots & \cdots \\ 1 & t & \cdots & t^q \end{bmatrix}.$$

Suppose that the real-time measured degradation data up to time t are $y_{1:t} = (y_1, y_2, \ldots, y_t)$, and then, it can be proven that $\sigma^2|y_{1:t}$ follows an IG distribution and $(\boldsymbol{\gamma} \mid \sigma^2, y_{1:t})$ follows a multivariate normal distribution, which is given as follows:

$$\begin{aligned} \sigma^2|y_{1:t} &\sim \text{IG}\left(a_1 + \frac{t}{2}, a_2 + \frac{H_t}{2}\right) \\ (\boldsymbol{\gamma} \mid \sigma^2, y_{1:t}) &\sim N(\boldsymbol{\mu}_t, \sigma^2 \boldsymbol{\Sigma}_t) \end{aligned} \tag{13}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_t &= (\boldsymbol{Z}_{1,t}^T \boldsymbol{Z}_{1,t} + \boldsymbol{\Sigma}_0^{-1})^{-1} \\ \boldsymbol{M}_t &= \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \boldsymbol{Z}_{1,t} y_{1:t} \\ \boldsymbol{\mu}_t &= \boldsymbol{\Sigma}_t \boldsymbol{M}_t \\ H_t &= \|y_{1:t}\|^2 + \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{M}_t^T \boldsymbol{\Sigma}_t \boldsymbol{M}_t. \end{aligned}$$

The proof can be found in [30]. Based on (13), we can further prove that the predictive distribution of future observations $y_{t+1:t+L}$ given the observations up to the current time follows a multivariate t-distribution:

$$\begin{aligned} y_{t+1:t+L}|y_{1:t} \sim \text{MT}\Bigg(&2a_1 + t, \boldsymbol{Z}_{t+1,t+L} \boldsymbol{\mu}_t, \frac{2a_2 + H_t}{2a_1 + t} \\ &\left(\boldsymbol{I} + \boldsymbol{Z}_{t+1,t+L} \boldsymbol{\Sigma}_t \boldsymbol{Z}_{t+1,t+L}^T\right)\Bigg) \end{aligned} \tag{14}$$

where MT denotes a multivariate t-distribution with the three parameters representing the degree of freedom, the mean, and the shape matrix, respectively. The proof is provided in the Appendix. With this predictive distribution, we can efficiently predict the evolution of the degradation path and the time the degradation signal reaches the failure threshold.

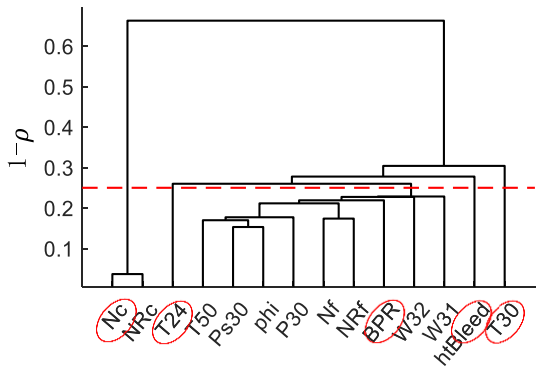

Fig. 5. Dendrogram of the hierarchical clustering for variable selection.

IV. Case Study

In this section, numeric experiments are conducted on the benchmarking C-MAPSS data set to evaluate the performance of the proposed data fusion methodology. The detailed introduction of the C-MAPSS data set is provided in Section II-A.

A. Variable Selection and Data Preprocessing

Before applying the proposed method, data preprocessing and variable selection are first conducted. Among the 21 outputs, only 14 outputs have obvious increasing or decreasing degradation trajectories. The others are almost unchanged throughout the whole life cycle. Therefore, only these 14 degradation signals are included for further selection. In the correlation analysis, it is found that some signals are highly correlated, with correlation coefficient ρ reaching up to 0.96. To reduce the information redundancy and, at the same time diversify, the degradation signals to capture various aspects of the degradation process, we set a pairwise correlation threshold, e.g., 0.75 in this work, above which only one degradation signal in each cluster is kept. Especially, we propose to use a hierarchical clustering algorithm [39] for variable selection, where $1 - \rho$ is used as the distance or dissimilarity measure.

Fig. 5 shows the clustering dendrogram. In total, five clusters are identified with the correlation threshold set to 0.75. Within each cluster, a degradation signal is arbitrarily selected. Here, we select Nc, T24, BPR, htBleed, and T30 for health index construction. Fig. 6 shows a representative set of

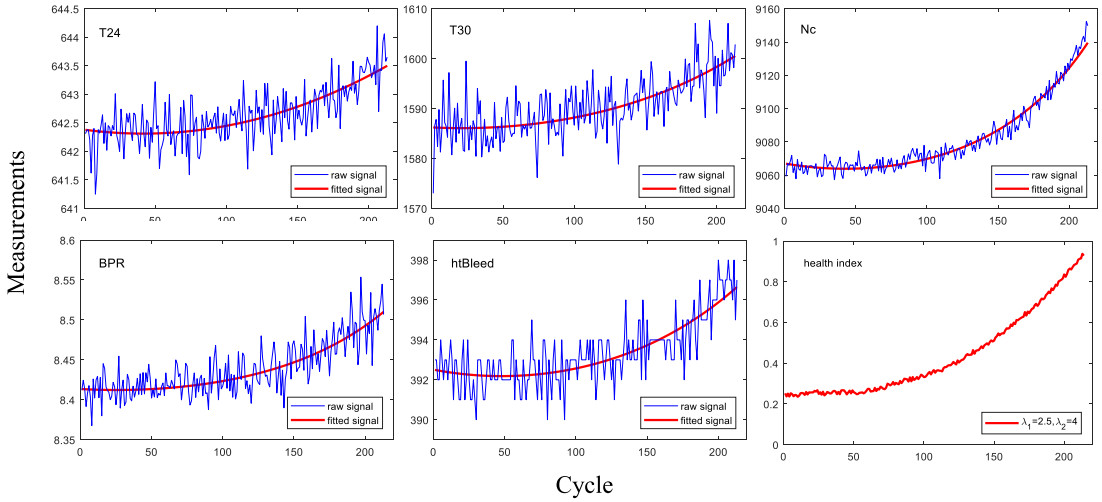

Fig. 6. Representative raw measurements with fit curves using exponential functions and a constructed health index.

selected degradation signals, and all of them demonstrate an exponential growth for degradation. The corresponding R-squares of model fitting using exponential functions are 0.7071, 0.6336, 0.9603, 0.7987, and 0.6637, respectively, for the five selected signals. A constructed HI is also provided as an illustration. Following the work of Liu *et al.* [3], we use log-transformation to preprocess the data and apply a linear model with quadratic form, i.e., $z_t = (1, t, t^2)$, to model the degradation path. To facilitate model training, the transformed data are normalized to the interval [0, 1].

B. Model Training and Illustration

In the model training process, we need to select appropriate network structures and penalty coefficients λ_1 and λ_2. The K-fold cross-validation method is typically applied to fulfill this goal. Here, we use the commonly adopted fivefold cross validation on the 100 training units. Since the eventual goal is to achieve an accurate estimation of the residual life, we used the average prediction error at multiple time epochs throughout the whole life cycle as the validation error. Note that a more natural way is to use the averaged prediction error across all the observation epochs as the validation error to achieve the optimal overall performance. However, this method is very computationally expensive. Here, we selected three time epochs, 1/4, 2/4, and 3/4, of the entire life cycle to reduce the computational cost. The mean value of the absolute percentage error is used as the performance metric, which is defined as

$$\text{err} = \frac{1}{N}\sum_{n=1}^{N}\frac{|\hat{R}_n - R_n|}{T_n}$$

where $\hat{R}_n$ is the predicted residual life at the time of prediction, R_n is the true residual life, T_n is the total life from the beginning to the failure time, and N is the total number of units for prediction. In the prediction, the Bayesian updating approach described in Section III-D is used to predict each of the selected degradation signals, and the time when the health index exceeds 1 is treated as the predicted failure time.

TABLE II
SELECTED MODEL STRUCTURES AND PENALTY COEFFICIENTS

Penalty type	λ_1	λ_2	Model structure
tanh	2.1	3.8	5-5-3-1
exp	2.7	4.2	5-5-3-1
linear	4.3	6.4	5-4-2-1

In the early stage of the lifetime, the predicted degradation path may be temporary decreasing due to small degradation rate and measurement noise, which may cause huge errors in predicting future degradation levels. To avoid such a situation, the coefficient of the quadratic term is constrained to be positive, i.e., the posterior distribution is truncated at 0.

The model is trained on a Windows server with two Tesla V100 GPUs. The initial learning rate is set to 0.01. As the training is inherently stochastic, the optimization for each model structure is repeated 100 times with randomly generated initial weights and biases, and the model parameters with minimum fitting loss are selected as the trained parameters.

The select network structure and penalty coefficients for the three penalty functions are shown in Table II.

To illustrate the influence of shape constraint on the constructed health index, we use the exponential penalty function and selected model structure with different penalty coefficient combinations. We randomly pick a unit (#41 in the testing data set) to show the influence of penalty weight on the constructed health indices. The vertical and horizontal dashed lines represent the true failure time and the failure threshold. The dashed lines on the curves are the predicted health indices. Besides, increasing the penalty weight of the monotonicity constraint in Fig. 7(a) could effectively reduce the noise and obtain a smoothed and monotonic health index. Besides, increasing the convexity penalty weight can further increase the degradation acceleration and the range of health index from the initial state to the failure time, as indicated in Fig. 7(b). It is worth noting that the range of the health index is a very

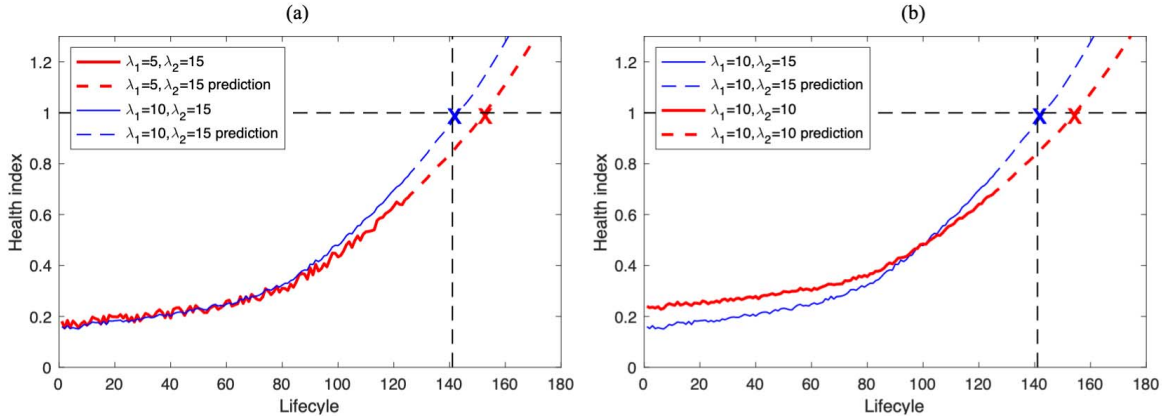

Fig. 7. Influence of the monotonicity and convexity shape constraint on the constructed health index. (a) Increasing the monotonicity penalty weight from $\lambda_1 = 5$ to $\lambda_1 = 10$. (b) Increasing the convexity penalty weight from $\lambda_2 = 10$ to $\lambda_2 = 15$.

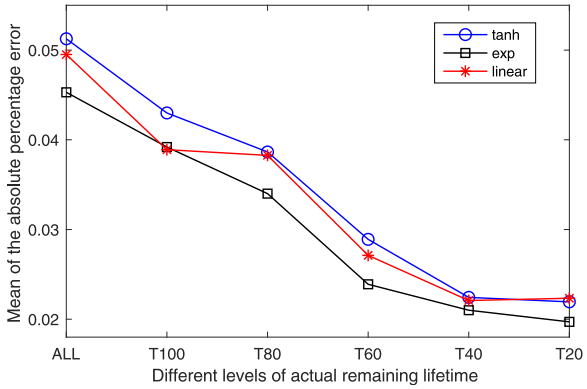

Fig. 8. Comparison of the three penalty functions on the prediction errors of the proposed method.

important characteristic to evaluate the quality of constructed health index [40]. It possesses the advantage of having a low signal-to-noise ratio and providing more accurate residual life prediction. With the convexity constraint posed in the model training, we are able to obtain a health index with not only more accurate shape, e.g., convexity, but also a large signal range, which is a by-product for better prognosis.

C. Performance Evaluation and Comparison

The testing data set with 100 units is used to evaluate and compare the performance of the proposed method with other methods. First, we evaluate and compare the performance of the three penalty functions, i.e., exp, tanh, and linear.

Fig. 8 shows the mean absolute percentage errors of the RUL prediction at different levels of the actual RUL for the proposed method using different penalty functions. The level label "ALL" represents all the 100 testing units, while "T-N" denotes those testing units with residual life not larger than N cycles. From Fig. 8, we can see that, for all three penalty functions, the prediction accuracy increases as the actual RUL decreases, which is expected as more observations are obtained to infer the degradation path, and the prediction length, i.e., the number of cycles required to predict the failure, is reduced. Comparing the three penalty functions, the exponential function outperforms the other two at almost all levels of actual RUL. The hyperbolic tangent penalty function generally has the lowest performance. Note that, in this application, the noise of the constructed health indices is not significant, i.e., $h(t-1)-h(t)$ is small, which reduces the effects of penalty function types.

The proposed method is further compared with the state-of-the-art data fusion approaches, HI-linear [3], HI-kernel [20], and HI-quantile [19]. For the HI-linear method, the health index is constructed by linearly combining these sensor data. The HI-kernel method, on the other hand, utilizes the kernel method to fuse multisensor signals nonlinearly. The HI-quantile method is still a linear fusion approach. However, it leverages the quantile regression in the optimization of the linear fusion coefficient. To study the effects of the convexity shape constraint, the method without the convexity constraint, i.e., ignoring the third term in (9), is also included in the comparison. The network structure is set to be the same as the one with both monotonicity and convexity considered. The optimal penalty coefficient is $\lambda_1 = 4.7$. These two methods are denoted by $\mathrm{HI}_{\mathrm{convex}}$ and $\mathrm{HI}_{\mathrm{no\ convex}}$, respectively. Since the exponential penalty function has the best performance, we adopt it in the comparison. We ignore the comparison of the proposed method with existing neural network-based approaches in RUL prediction, which utilizes the percentage of life occurred or alternatively the residual life percentage as model output. These methods only provide the estimated residual life that linearly decreases to zero (e.g., $\mathrm{RUL}(t) = T - t$), and they are not able to provide direct visualization of the degradation path of the health condition. Therefore, following the convention of existing health index construction work [3], [18], [20], we ignore the comparison although the proposed method has comparable performance in RUL prediction.

Fig. 9 shows the comparison results in terms of the mean of the absolute percentage error. As expected, the prediction error decreases for all methods when the units approach failure. The proposed method outperforms all the other methods at almost all degradation levels, especially in the early degradation stages. At the degradation level T20, all the methods have

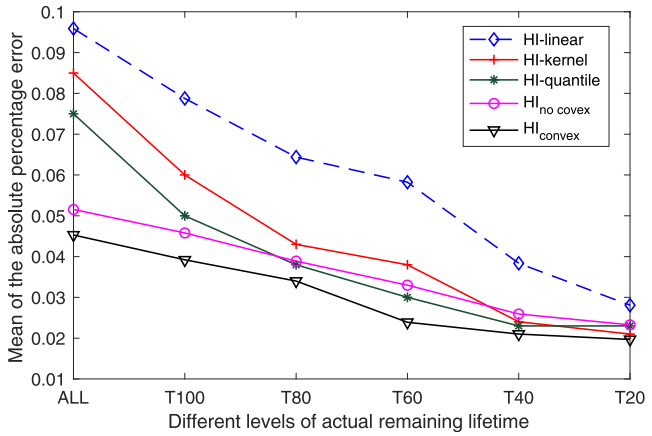

Fig. 9. Comparison of the proposed method with other approaches.

TABLE III
VARIANCES OF NORMALIZED SIGNALS AND HEALTH INDEX

Signal	Nc	T24	BPR	htBleed	T30	HI
Var	0.0075	0.0091	0.0680	0.0048	0.0076	0.0002

comparable prediction accuracy. The HI-linear method has the highest prediction error. Due to the usage of indirect supervised learning, the linear method HI-quantile has significantly lower prediction errors than HI-linear. However, due to the limitation of linear fusion, it still has lower prediction accuracy than the proposed nonlinear method. $\mathrm{HI}_{\text{no convex}}$ has much higher prediction accuracy than HI-kernel at early prediction stages, e.g., ALL, T100, and T60, while its performance is comparable with (or slightly lower than) HI-kernel at high degradation levels (T40 and T20), which demonstrates the advantages of using neural network data fusion approach than using kernel method. Comparing $\mathrm{HI}_{\text{no convex}}$ and $\mathrm{HI}_{\text{convex}}$, we can clearly see that the usage of the convexity shape constraint can effectively improve the prediction accuracy.

To justify the necessity of fusing multiple sensor signals in RUL prediction, we also have to make sure that the proposed data fusion method has a better performance than using any single-sensor signal. Since the HI-linear, HI-quantile, and HI-kernel methods have already demonstrated their superiority over using a single signal for prognosis [3], [19], [20], there is no need to provide that comparison. Instead, we compare the variances of selected sensor signals (normalized) at the failure time and the constructed health index, as shown in Table III. It is clear that the variance of the health index is much lower than that of each sensor signal at the failure time, which also justifies the necessity of data fusion methods.

V. CONCLUSION

With the rapid development in sensor and information technology, deploying multiple sensors for condition monitoring prevails in the modern PHM field. Health indices have gained their wide popularity recently due to their advantage of providing direct visualization of the health condition and facilitating RUL prediction using well-developed univariate prognostic methods. However, the majority of data fusion methods restrict themselves to linear combination, which is incapable of characterizing the underlying nonlinear relationship. To address this issue, this article proposed a shape-constrained neural data fusion approach for health index construction and RUL prediction. In this methodology, an unsupervised learning neural network with shape constraints is proposed to model the potentially nonlinear relationship between the true health index and the observed multisensor signals. A novel loss function is formulated by jointly considering the variance of the health index at the failure time across different units, the monotonicity, and the convexity of the constructed health index. A tailored Adam is developed for model parameter estimation. In the RUL prediction, a Bayesian linear modeling and updating framework is used to model sensor data to capture the population degradation trend and the individual degradation path. The effectiveness of the proposed methodology is demonstrated and validated using the C-MAPSS data set. The results indicate that the proposed neural data fusion model can not only construct a composite health index reasonably but also has much higher prediction accuracy than the existing approaches.

There are still several problems worthy of further investigation. First, the current method is unsupervised or somehow semisupervised in that the variance of the health index at the failure time, instead of the prediction error, is used in the loss function. Incorporating the prediction error in the loss function may further improve the quality and prediction accuracy of the constructed health index. Second, this article focuses on the degradation with only one failure mode and one operational mode. Extensions to multiple failure and operational modes are desirable in practical applications. Last but not least, more information about the health index could be incorporated to obtain a more accurate health index. For example, many degradation processes demonstrate two obvious stages, with the first stage representing the nondefective phase with almost no degradation while the second stage denoting the defective phase with a monotonic degradation trend. The introduction of change points could potentially improve the quality of health indices.

APPENDIX
DERIVATION OF (14)

Based on (13)

$$(\boldsymbol{\gamma} \mid \sigma^2, y_{1:t}) \sim N(\boldsymbol{\mu}_t, \sigma^2 \boldsymbol{\Sigma}_t)$$
$$(\boldsymbol{Z}_{t+j}\boldsymbol{\gamma} \mid \sigma^2, y_{1:t}) \sim N(\boldsymbol{Z}_{t+j}\boldsymbol{\mu}_t, \sigma^2 \boldsymbol{Z}_{t+j}\boldsymbol{\Sigma}_t \boldsymbol{Z}_{t+j}^T)$$
$$(\sigma^2 \mid y_{1:t}) \sim \mathrm{IG}\left(a_1 + \frac{t}{2}, a_2 + \frac{H_t}{2}\right)$$

Since $y_{t+1:t+L} = \boldsymbol{Z}_{t+1,t+L}\boldsymbol{\gamma} + \sigma\varepsilon_{t+1:t+L}$, then $(y_{t+1:t+L}|\sigma^2, y_{1:t}) \sim N(\boldsymbol{Z}_{t+1,t+L}\boldsymbol{\mu}_t, \sigma^2(I + \boldsymbol{Z}_{t+1,t+L}\boldsymbol{\Sigma}_t\boldsymbol{Z}_{t+1,t+L}^T))$.

Let $\boldsymbol{\mu}_* = \boldsymbol{Z}_{t+1,t+L}\boldsymbol{\mu}_t$, $\boldsymbol{\Sigma}_* = \boldsymbol{I} + \boldsymbol{Z}_{t+1,t+L}\boldsymbol{\Sigma}_t\boldsymbol{Z}_{t+1,t+L}^{\mathrm{T}}$

$$p(y_{t+1:t+L} \mid y_{1:t}) = \int p\left(y_{t+1:t+L} \mid \sigma^2, y_{1:t}\right) p\left(\sigma^2 \mid y_{1:t}\right) d\sigma^2$$
$$\propto \int \left(\sigma^2\right)^{-\frac{L}{2}} |\boldsymbol{\Sigma}_*|^{-\frac{1}{2}}$$

$$\times \exp\left[-\frac{(y_{t+1:t+L}-\boldsymbol{\mu}_*)\boldsymbol{\Sigma}_*^{-1}(y_{t+1:t+L}-\boldsymbol{\mu}_*)^{\mathrm{T}}}{2\sigma^2}\right]$$
$$\times \left(\sigma^2\right)^{-a_1-\frac{t}{2}-1}\exp\left[-\frac{2a_2+H_t}{2\sigma^2}\right]d\sigma^2$$
$$\propto \int \left(\sigma^2\right)^{-a_1-\frac{t+L}{2}-1}$$
$$\times \exp\left(-\frac{(y_{t+1:t+L}-\boldsymbol{\mu}_*)\boldsymbol{\Sigma}_*^{-1}(y_{t+1:t+L}-\boldsymbol{\mu}_*)^{\mathrm{T}}+(2a_2+H_t)}{2\sigma^2}\right)$$
$$\times d\sigma^2$$
$$\propto \frac{\Gamma\left(a_1+\frac{t+L}{2}\right)}{\left[\frac{(y_{t+1:t+L}-\boldsymbol{\mu}_*)\boldsymbol{\Sigma}_*^{-1}(y_{t+1:t+L}-\boldsymbol{\mu}_*)^{\mathrm{T}}+(2a_2+H_t)}{2}\right]^{a_1+\frac{t+L}{2}}}$$
$$\propto \left[1+\frac{1}{\upsilon}\frac{(y_{t+1:t+L}-\boldsymbol{\mu}_*)\boldsymbol{\Sigma}_*^{-1}(y_{t+1:t+L}-\boldsymbol{\mu}_*)^{\mathrm{T}}\upsilon}{2a_2+H_t}\right]^{-\frac{L+\upsilon}{2}}$$

where $\upsilon = 2a_1 + t$. Therefore, the future observation vector $y_{t+1:t+L}$ given the observations up to the current time follows a multivariate t-distribution:

$$y_{t+1:t+L}|y_{1:t} \sim \mathrm{MT}\Bigg(2a_1+t, \boldsymbol{Z}_{t+1,t+L}\boldsymbol{\mu}_t, \frac{2a_2+H_t}{2a_1+t} \times \left(\boldsymbol{I}+\boldsymbol{Z}_{t+1,t+L}\boldsymbol{\Sigma}_t\boldsymbol{Z}_{t+1,t+L}^T\right)\Bigg).$$

REFERENCES

[1] A. K. S. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mech. Syst. Signal Process.*, vol. 20, no. 7, pp. 1483–1510, Oct. 2006.

[2] S. J. Bae, W. Kuo, and P. H. Kvam, "Degradation models and implied lifetime distributions," *Rel. Eng. Syst. Saf.*, vol. 92, no. 5, pp. 601–608, May 2007.

[3] K. Liu, N. Z. Gebraeel, and J. Shi, "A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis," *IEEE Trans. Autom. Sci. Eng.*, vol. 10, no. 3, pp. 652–664, Jul. 2013.

[4] S. Simani, C. Fantuzzi, and R. J. Patton, *Model-Based Fault Diagnosis in Dynamic Systems Using Identification Techniques*. Berlin, Germany: Springer, 2003.

[5] N. Gebraeel, "Sensory-updated residual life distributions for components with exponential degradation patterns," *IEEE Trans. Autom. Sci. Eng.*, vol. 3, no. 4, pp. 382–393, Oct. 2006.

[6] X.-S. Si, W. Wang, M.-Y. Chen, C.-H. Hu, and D.-H. Zhou, "A degradation path-dependent approach for remaining useful life estimation with an exact and closed-form solution," *Eur. J. Oper. Res.*, vol. 226, no. 1, pp. 53–66, Apr. 2013.

[7] W. Nelson, *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis*. New York, NY, USA: Wiley, 2008.

[8] J. B. Keats, "Statistical methods for reliability data," *J. Qual. Technol.*, vol. 31, no. 4, pp. 466–468, 1999.

[9] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in *Proc. Int. Conf. Prognostics Health Manage.*, Oct. 2008, pp. 1–8.

[10] T. Brotherton, P. Grabill, D. Wroblewski, R. Friend, B. Sotomayer, and J. Berry, "A testbed for data fusion for engine diagnostics and prognostics," in *Proc. IEEE Proc. Aerosp. Conf.*, Mar. 2002, p. 6.

[11] I. I. Martin Liggins, D. Hall, and J. Llinas, *Handbook of Multisensor Data Fusion: Theory and Practice*, 2nd ed. Norwood, MA, USA: Artech House, 2008.

[12] L. Guo, N. Li, F. Jia, Y. Lei, and J. Lin, "A recurrent neural network based health indicator for remaining useful life prediction of bearings," *Neurocomputing*, vol. 240, pp. 98–109, May 2017.

[13] J. Zhang, P. Wang, R. Yan, and R. X. Gao, "Long short-term memory for machine remaining life prediction," *J. Manuf. Syst.*, vol. 48, pp. 78–86, Jul. 2018.

[14] L. Chen, G. Xu, S. Zhang, W. Yan, and Q. Wu, "Health indicator construction of machinery based on end-to-end trainable convolution recurrent neural networks," *J. Manuf. Syst.*, vol. 54, pp. 1–11, Jan. 2020.

[15] J. Franklin, "The elements of statistical learning: Data mining, inference and prediction," *Publications Amer. Stat. Assoc.*, vol. 99, no. 466, p. 567, 2010.

[16] Y. Gao, Y. Wen, and J. Wu, "A neural network-based joint prognostic model for data fusion and remaining useful life prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 11, 2020, doi: 10.1109/TNNLS.2020.2977132.

[17] X. Li, Q. Ding, and J.-Q. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks," *Rel. Eng. Syst. Saf.*, vol. 172, pp. 1–11, Apr. 2018.

[18] K. Liu and S. Huang, "Integration of data fusion methodology and degradation modeling process to improve prognostics," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 1, pp. 344–354, Jan. 2016.

[19] C. Song and K. Liu, "Statistical degradation modeling and prognostics of multiple sensor signals via data fusion: A composite health index approach," *IISE Trans.*, vol. 50, no. 10, pp. 853–867, Oct. 2018.

[20] C. Song, K. Liu, and X. Zhang, "Integration of data-level fusion model and kernel methods for degradation modeling and prognostic analysis," *IEEE Trans. Rel.*, vol. 67, no. 2, pp. 640–650, Jun. 2018.

[21] H. Qiu, J. Lee, J. Lin, and G. Yu, "Robust performance degradation assessment methods for enhanced rolling element bearing prognostics," *Adv. Eng. Informat.*, vol. 17, nos. 3–4, pp. 127–140, Jul. 2003.

[22] J.-B. Yu and S. Wang, "Using minimum quantization error chart for the monitoring of process states in multivariate manufacturing processes," *Comput. Ind. Eng.*, vol. 57, no. 4, pp. 1300–1312, Nov. 2009.

[23] J. Zhang, N. Jiang, H. Li, and N. Li, "Online health assessment of wind turbine based on operational condition recognition," *Trans. Inst. Meas. Control*, vol. 41, no. 10, pp. 2970–2981, Jun. 2019.

[24] T. Benkedjouh, K. Medjaher, N. Zerhouni, and S. Rechak, "Fault prognostic of bearings by using support vector data description," in *Proc. IEEE Conf. Prognostics Health Manage.*, Jun. 2012, pp. 1–7.

[25] T. Wang, J. Yu, D. Siegel, and J. Lee, "A similarity-based prognostics approach for remaining useful life estimation of engineered systems," in *Proc. Int. Conf. Prognostics Health Manage.*, Oct. 2008, pp. 1–6, doi: 10.1109/PHM.2008.4711421.

[26] A. Chehade, S. Bonk, and K. Liu, "Sensory-based failure threshold estimation for remaining useful life prediction," *IEEE Trans. Rel.*, vol. 66, no. 3, pp. 939–949, Sep. 2017.

[27] A. Chehade, C. Song, K. Liu, A. Saxena, and X. Zhang, "A data-level fusion approach for degradation modeling and prognostic analysis under multiple failure modes," *J. Qual. Technol.*, vol. 50, no. 2, pp. 150–165, Apr. 2018, doi: 10.1080/00224065.2018.1436829.

[28] Y. Wen, J. Wu, D. Das, and T.-L. Tseng, "Degradation modeling and RUL prediction using Wiener process subject to multiple change points and unit heterogeneity," *Rel. Eng. Syst. Saf.*, vol. 176, pp. 113–124, Aug. 2018.

[29] Y. Wen, J. Wu, and Y. Yuan, "Multiple-phase modeling of degradation signal for condition monitoring and remaining useful life prediction," *IEEE Trans. Rel.*, vol. 66, no. 3, pp. 924–938, Sep. 2017.

[30] Y. Wen, J. Wu, Q. Zhou, and T.-L. Tseng, "Multiple-change-point modeling and exact Bayesian inference of degradation signal for prognostic improvement," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 2, pp. 613–628, Apr. 2019.

[31] N. Z. Gebraeel and M. A. Lawley, "A neural network degradation model for computing and updating residual life distributions," *IEEE Trans. Autom. Sci. Eng.*, vol. 5, no. 1, pp. 154–163, Jan. 2008.

[32] Q. Zhou, J. Son, S. Zhou, X. Mao, and M. Salman, "Remaining useful life prediction of individual units subject to hard failure," *IIE Trans.*, vol. 46, no. 10, pp. 1017–1030, Oct. 2014.

[33] K. L. Tsui, N. Chen, Q. Zhou, Y. Hai, and W. Wang, "Prognostics and health management: A review on data driven approaches," *Math. Problems Eng.*, vol. 2015, pp. 1–17, May 2015.

[34] X.-S. Si, W. Wang, C.-H. Hu, D.-H. Zhou, and M. G. Pecht, "Remaining useful life estimation based on a nonlinear diffusion degradation process," *IEEE Trans. Rel.*, vol. 61, no. 1, pp. 50–67, Mar. 2012.

[35] W. F. Wu and C. C. Ni, "A study of stochastic fatigue crack growth modeling through experimental data," *Probabilistic Eng. Mech.*, vol. 18, no. 2, pp. 107–118, Apr. 2003.

[36] Y. Hong, H. Liao, E. Yashchin, and F. Tsung, "Editor's notes on special issue on 'reliability and maintenance modeling with big data,'" *J. Qual. Technol.*, vol. 50, no. 2, pp. 133–134, 2018.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[38] J. Wu, Y. Chen, and S. Zhou, "Online detection of steady-state operation using a multiple-change-point model and exact Bayesian inference," *IIE Trans.*, vol. 48, no. 7, pp. 599–613, Jul. 2016.
[39] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *Comput. J.*, vol. 26, no. 4, pp. 354–359, Nov. 1983.
[40] K. Liu, A. Chehade, and C. Song, "Optimize the signal quality of the composite health index via data fusion for degradation modeling and prognostic analysis," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 3, pp. 1504–1514, Jul. 2017.

Zhen Li (Member, IEEE) received the B.S. degree in computer science from Southwest University, Chongqing, China, in 2017. She is currently pursuing the Ph.D. degree in industrial engineering and management with Peking University, Beijing, China.

Her research interests are focused on data mining, prognosis, and health management through advanced data analytics, quality, and reliability engineering.

Ms. Li is also a member of Institute for Operations Research and the Management Sciences (INFORMS).

Jianguo Wu (Member, IEEE) received the B.S. degree in mechanical engineering from Tsinghua University, Beijing, China, in 2009, the M.S. degree in mechanical engineering from Purdue University, West Lafayette, IN, USA, in 2011, and the M.S. degree in statistics and the Ph.D. degree in industrial and systems engineering from University of Wisconsin–Madison, Madison, WI, USA, in 2014 and 2015, respectively.

He was an Assistant Professor with the Department of Industrial, Manufacturing and Systems Engineering (IMSE), The University of Texas at El Paso (UTEP), El Paso, TX, USA, from 2015 to 2017. He is currently an Assistant Professor with the Department of Industrial Engineering and Management, Peking University, Beijing. His research interests are focused on data-driven modeling, monitoring, and analysis of advanced manufacturing processes and complex systems for quality control and reliability improvement.

Dr. Wu is also a member of Institute for Operations Research and the Management Sciences (INFORMS), Institute of Industrial and Systems Engineers (IISE), and Society of Manufacturing Engineers (SME). He was a recipient of the STARS Award from the University of Texas Systems and the Outstanding Young Scholar Award from China.

Xiaowei Yue (Member, IEEE) received the B.S. degree in mechanical engineering from the Beijing Institute of Technology, Beijing, China, in 2011, the M.S. degree in power engineering and engineering thermophysics from the Tsinghua University, Beijing, in 2013, and the M.S. degree in statistics and Ph.D. degree in industrial engineering with a minor in machine learning from the Georgia Institute of Technology, Atlanta, GA, USA, in 2016 and 2018, respectively.

He is currently an Assistant Professor with the Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, USA. His research interest are focused on engineering-driven data analytics for advanced manufacturing.

Dr. Yue is also a member of American Society of Mechanical Engineers (ASME), Institute for Operations Research and the Management Sciences (INFORMS), and Institute of Industrial and Systems Engineers (IISE). He was a recipient of the IEEE Transactions on Automation Science and Engineering Best Paper Award, the IISE Pritsker Doctoral Dissertation Award, the FTC Early Career Award, and several other best paper awards.