Original Article

Structural Health Monitoring
2025, Vol. 24(6) 3849–3867

DOI: 10.1177/14759217241279784
journals.sagepub.com/home/shm

A novel HI construction method based on healthy-state data training for rotating machinery components

Hongliang Song[1], Yi Sun[1], Hongli Gao[1], Liang Guo[1] and Tingting Wu[2]

Abstract

In predictive maintenance for rotating machinery components, constructing health indicators is crucial for improving operational efficiency and extending the equipment lifespan. However, traditional processes of building health indicators face two significant challenges: the high cost of collecting full lifecycle data and the requirement for the operating conditions of the equipment to remain consistent with those under which the indicators were developed. These limitations make health indicators expensive and time-consuming to construct and use. This article proposed a health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model (JPCCED-HI) to overcome these challenges. This innovative method utilizes only healthy state data, combining principal component analysis with the empirical cumulative distribution function to construct a health indicator that ranges from 0 to 1, thereby eliminating the reliance on complete lifecycle data and relaxing the requirements for specific operating conditions. The effectiveness and practical benefits of JPCCED-HI are illustrated through two detailed case studies. In the first case study, we explore the impact of different parameter settings on model performance and evaluate its anti-interference capabilities in high-noise environments based on a gearbox dataset. This study demonstrates the model's robustness in maintaining accurate health assessments despite external noise. The second case study employs a publicly available bearing dataset to compare the JPCCED-HI method against other models trained on health-stage data. This analysis reveals the superior performance of JPCCED-HI in trendability and scale similarity, affirming its effectiveness in various operational conditions. These case studies not only prove the adaptability and robustness of JPCCED-HI but also highlight its potential as a scalable solution for real-time health monitoring and predictive maintenance in industrial applications.

Keywords

Prognostics, health indicator, degradation assessment, PCA, empirical cumulative distribution function

Introduction

Rotating components, such as bearings, gears, and shafts, play a critical role in various mechanical devices, and their failure can lead to inefficiencies in the entire system and downtime. The construction of health indicators (HIs) for rotating machinery components is crucial to ensure the reliable operation of these essential parts. By establishing HIs, the performance status of these components can be monitored in real-time, thereby enabling the early detection of potential issues and anomalies.[1,2] This type of monitoring can significantly reduce unexpected downtime, save costs, and enhance production efficiency. Moreover, the practical construction and application of HIs also help extend the mechanical equipment's service life, optimize maintenance and repair schedules, and improve the overall reliability of operations, which are vital for ensuring the continuity and stability of industrial production.

The construction of HIs based on data-driven methods relies heavily on substantial prior knowledge and a deep understanding of their failure mechanisms.[3] These methods typically require data from the entire lifecycle of the equipment, particularly degradation and failure data, to construct detailed mechanistic or data models. Widodo and Yang[4] introduced principal component analysis (PCA) to fuse the peak value, kurtosis, and entropy into a one-dimensional variable. Mosallam

[1]School of Mechanical Engineering, Southwest Jiaotong University, Chengdu, China
[2]Ziyang Locomotive and Rolling Stock Co. Ltd., CRRC Corporation Limited, Ziyang, China

Corresponding author:
Hongli Gao, School of Mechanical Engineering, Southwest Jiaotong University, No. 111 North Section 1, Chengdu, Sichuan 610031, China.
Email: hongli_gao@home.swjtu.edu.cn

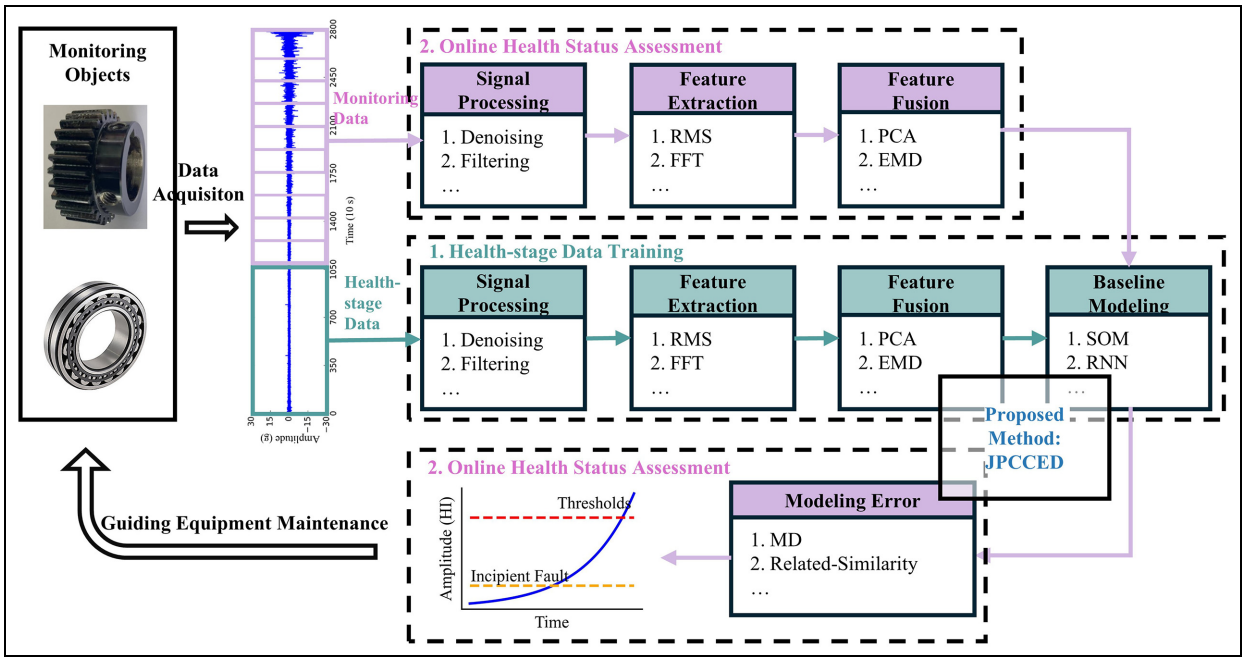

Figure 1. Healthy-state data training flow.

et al.[5] Perform unsupervised feature selection in the feature domain and apply the empirical mode decomposition algorithm to the predictive features to construct HIs. Hong et al.[6] applied the empirical mode decomposition–self-organizing map method to analyze vibration signals and calculate a confidence value on the bearing health state. Zhao et al.[7] used the S transform and Gaussian pyramid to obtain time-frequency representations at multiple scales and employ PCA based on time-frequency domain indicators to form a HI for bearings. In recent years, a significant number of scholars have applied deep neural network techniques in the process of constructing HIs. These advanced computational models have been utilized to enhance the analytical capabilities and accuracy of existing methodologies, including methods such as Convolutional Neural Networks (CNNs),[8,9] Recurrent Neural Networks (RNNs),[10,11] unsupervised Deep Belief Networks,[12,13] and Convolutional Autoencoder Networks.[14,15]

However, these methods typically rely on a critical assumption: there is a high degree of consistency among the monitored objects, meaning they operate under similar operational conditions and exhibit similar vibration amplitudes and spectral signals. Consequently, the proposed models present the following problems:

(1) When there are significant differences in working conditions between the training and testing data, the performance of the models often significantly decreases.
(2) It requires collecting data throughout the entire lifecycle of the monitored objects under different operating conditions, which are expensive.
(3) As the models require complete data for training and deep neural network models are complex, it is not feasible to perform online modeling and training.

To address these issues, in recent years, the method of constructing HIs based on health status data[16] has increasingly attracted researchers' attention. This method utilizes a large amount of monitoring data that are easily obtained under normal operating conditions, thereby avoiding model failure issues when there is a lack of complete lifecycle data or specific operating conditions that are not covered during the training process. The typical process for constructing HIs based on training with healthy-state data is shown in Figure 1.

Measuring the difference between the data in the healthy and tested states is the key idea to constructing HIs based on normal data. In other words, the healthy data are the baseline data for the unidentified data. Such methods can broadly be categorized into those based on similarity measurements and those based on reconstruction error.

Methods based on similarity measurements primarily involve quantifying the similarity or correlation between benchmark data and test data, including distance metrics for quantitative comparison. Guo et al.[10] proposed time-frequency range related-similarity (RS)

features that address the issue of varying statistical feature ranges and unequal contributions. Wang et al.[17] estimated the degradation information by calculating the deviation of the multiple statistics of vibration signals of a bearing from a known healthy state. Schwartz et al.[18] proposed an automatic and unsupervised approach based on the Kernel PCA to enhance the similarity-based Health Index creation. Li et al.[19] calculated the Mahalanobis Distance (MD) depending on the characteristics of vibration signals in the time domain and time-frequency domain and obtained the monotonously increasing HI by using Cumulative Sum of Mahalanobis Distance One (MD1-CUMSUM).

HIs construction based on reconstruction error is achieved through indirect measurement of modeling errors of unidentified monitoring data with a trained model of healthy-state data. Qiu et al.[20] introduced a wavelet filter based method for weak signature enhancement for fault identification and self-organizing map (SOM) based method for performance degradation assessment. Huang et al.[21] used the minimum quantization error (MQE) indicator derived from SOM, which is trained by six vibration features, including a newly designed degradation index for performance degradation assessment. Pei et al.[22] applied gray target decision (GTD) to the health assessment of mechanical systems for the first time, proposing a weighted GTD approach for constructing sensitive health and fault indicators and introducing a Convolutional Long Short-Term Memory (ConvLSTM) framework based on a quadratic transfer function. Yan et al.[23] presented a novel approach called multidomain indicator-based optimized stacked denoising autoencoder to learn more robust and reliable feature representation. Kundu et al.[24] used the correlation coefficient parameter based on the residual vibration signal to monitor and detect the spur gears' pitting progression. Lu et al.[25] obtained the truncated MQE histories by SOM network after feature extraction and adopted the chaos-based parallel multilayer perceptron network and polynomial fitting for residual errors to generate the predicted MQEs and failure times following the truncation times.

The problem with healthy-state data training is that most indicators lack fixed upper and lower limits, which poses significant challenges in determining the health status of rotating equipment.[26] It becomes necessary to introduce fault data to ensure defined boundaries, as it is often challenging to obtain extreme characteristic values before the failure of rotating equipment occurs. Moreover, few methods inherently restrict HIs within a 0–1 range, providing clear information about the health status of rotating equipment to monitoring personnel. Exploring new techniques to extract statistical features within specific ranges is essential to address this issue.

This study introduces an innovative method based on healthy-state data training called the Joint Principal Component Cumulative Empirical Distribution Model (JPCCED-HI). This model effectively mitigates the failure of HI assessments caused by varying operating conditions under healthy-state data training while ensuring that the HIs are theoretically confined within the 0–1 range. The core strategy of JPCCED-HI includes initially applying PCA to extract the main components from the feature space, thereby eliminating collinearity among features. Subsequently, utilizing the joint cumulative empirical distribution function exploits these principal components to construct HIs with fixed upper and lower limits and distinct trend characteristics in an unsupervised environment. The primary contributions of this research include the following:

(1) By using the healthy-state data of the monitored object as a baseline for analysis, the process not only simplifies data handling but also significantly enhances the model's adaptability and robustness under different operational conditions, effectively overcoming the issue of traditional methods failing under varying conditions and providing a more reliable and practical tool for mechanical equipment health monitoring.
(2) The development of the JPCCED-HI model uniquely ensures precise delimitation of HIs within a 0–1 range without additional fault data. This method uses PCA and the joint cumulative empirical distribution technique to effectively extract critical features from healthy-state data for health monitoring, ensuring stable model performance without explicit fault data.
(3) The model's key parameters were analyzed for their impact on results, its noise resistance was tested, and it was compared with several typical healthy-state data training models, validating its superior performance regarding trendability and scale similarity.

The subsequent organization of this article is as follows: The "Basic methodology" section introduces the theoretical foundations of manual feature extraction and PCA. The "Proposed methods" section describes the proposed methods' motivations, processes, and key parameters. "Case study 1" presents the application of this method to gearbox datasets, including the results of HI construction, analysis of the impact of different parameter values on the model, and its noise interference capabilities. "Case study 2" discusses the experimental results on public datasets and comparisons with several typical healthy-state data training models. Finally, the "Conclusion" section summarizes the findings of this article.

Basic methodology

Manual feature extraction

Manual feature extraction is a crucial technique in constructing HIs for equipment. This method relies on the expertise and knowledge of domain experts to identify and extract features from raw data that are closely related to the performance and health status of the equipment. By analyzing these parameters of the time series data, experts can define indicators that reflect the equipment's health condition. Although this method may be time-consuming for feature extraction and heavily dependent on expert knowledge, it still demonstrates unique effectiveness and accuracy in handling complex and highly individualized needs. Feature extraction of vibration signals has been extensively studied in the field of equipment HI construction, and many features have shown good trends and correlation with health stages.[27,28] A significant advantage of manually extracted features over those extracted by neural networks is that they do not require data training, facilitating the online completion of the entire HI construction process.

The features used in this study are listed in the attached Tables 1 and 2, where T, f, and F denote time domain signals, frequency components, and magnitude of the corresponding frequency components, respectively.

Principal component analysis

PCA[29–31] is a fundamental technique in multivariate statistics and data analysis, aimed at reducing the complexity of high-dimensional data while retaining as much variability in the data as possible. This statistical method transforms the original variables into a new set of variables. These new variables are linear combinations of the original ones, arranged in descending order of their ability to explain the variance in the data.

At its core, PCA involves converting many potentially correlated variables into a smaller number of uncorrelated variables, known as principal components. The first principal component accounts for the highest possible variance, thus explaining the maximum variability in the data. Each subsequent component, under the constraint of being orthogonal to the preceding ones, also aims to account for the maximum possible variance. The mathematical steps include the following:

(1) *Standardization*: For a feature set $\boldsymbol{F}$ consisting of multiple features, the range of features is standardized using the following equation:

$$f_i' = \frac{(f_i - \mu)}{\sigma} \tag{1}$$

where f_i is the value of the feature variable, μ is the mean, and σ is the standard deviation.

(2) *Covariance matrix computation*: The next step is to calculate the covariance matrix of the variables. For standardized feature values, the covariance matrix is the matrix of correlation coefficients:

$$\boldsymbol{C} = \frac{1}{n-1}\boldsymbol{F}'^{T}\boldsymbol{F}' \tag{2}$$

Table 1. Time domain feature list.

Time domain features			
Mean (Z_1)	$\frac{1}{N}\sum_{i=1}^{N} T_i$	Max	$\max(T)$
Mean absolute deviation	$\frac{1}{N}\sum_{i=1}^{N} \lvert T_i - Z_1 \rvert$	Peak to peak	$\lvert \max(T) - \min(T) \rvert$
Root mean square	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} T_i^2}$	Kurtosis	$\frac{N\sum_{i=1}^{N}(T_i - Z_1)^4}{\left(\sum_{i=1}^{N}(T_i - Z_1)^2\right)^2} - 3$
Skewness	$\frac{1}{N}\sum_{i=1}^{N} \frac{(T_i - Z_1)^4}{\left(\frac{1}{N-1}\sum_{i=1}^{N}(T_i - Z_1)^2\right)^2}$	Absolute energy	$\sum_{i=1}^{N} \lvert T_i \rvert^2$
Average power	$\frac{\sum_{i=1}^{N} T_i^2}{t_N - t_1}$	Mean difference	$\frac{1}{N-1}\sum_{i=1}^{N-1}(T_{i+1} - T_i)$
Mean absolute diff	$\frac{1}{N-1}\sum_{i=1}^{N-1} \lvert T_{i+1} - T_i \rvert$	Median	Median(T)
Absolute energy	$\sum_{i=1}^{N} \lvert T_i \rvert^2$		

Table 2. Frequency domain feature list.

Frequency domain features			
Maximum frequency	$\max(F)$	Median frequency	$\mathrm{Median}(F)$
Mean frequency (Z_2)	$\frac{1}{N}\sum_{i=1}^{N} F_i$	Spectral centroid (Z_3)	$\sum_{i=1}^{N} f_i \left(\frac{F_i}{\sum_{i=1}^{N} F_i}\right)$
Spectral distance	$\sum_{i=1}^{N}\left(\frac{i\cdot\sum_{j=1}^{N} F_j}{N-1} - \sum_{j=1}^{i} F_j\right)$	Spectral entropy	$-\sum_{k=1}^{N}\left(\frac{F_i}{Z_2 N}\log_2\frac{F_i}{Z_2 N}\right)$
Spectral kurtosis	$(f_i - Z_3)^4 \cdot \frac{F_i}{\sum_{i=1}^{N} F_i}$	Spectral decrease	$\frac{\sum_{i=1}^{N-1}\frac{F_{i+1}-F_1}{i}}{\sum_{i=1}^{N-1} F_{i+1}}$

where $\boldsymbol{F}'$ is the matrix of standardized features, n is the number of features.

(3) *Eigenvalue decomposition*: Decompose the covariance matrix to obtain its eigenvalues and the corresponding eigenvectors. Solve for the eigenvalues λ_i and the corresponding eigenvectors v_i of the covariance matrix $\boldsymbol{C}$,

$$\boldsymbol{C}v_i = \lambda_i v_i \tag{3}$$

where λ_i is the eigenvalues and v_i is the corresponding eigenvectors of the covariance matrix $\boldsymbol{C}$.

(4) *Principal component selection*: Based on the percentage of variance (information) contributed by the components, select the m principal components with the largest proportions to form a vector matrix, which we refer to as the feature vector $\boldsymbol{V}$.

(5) *Feature projection*: Transform the feature matrix $\boldsymbol{F}'$ into the new feature space:

$$\boldsymbol{P} = \boldsymbol{F}'\boldsymbol{V} \tag{4}$$

where $\boldsymbol{P}$ is the projection of the features in the principal component space.

The effectiveness of PCA as a dimensionality reduction tool is widely recognized, especially when the data are of high dimensionality. PCA is often used to visualize high-dimensional features, reduce noise in the original signal, or simplify machine learning inputs. The principal components extracted by PCA are orthogonal to each other. This means that each principal component is statistically linearly independent of the others. In the new coordinate system of the data, this orthogonality ensures that there will be no linear correlation between the features and helps to avoid the problem of multicollinearity. This property of PCA is essential for the method proposed in this article.

Proposed methods

Motivation and high-level idea

To establish a concrete and clear health metric for equipment with defined upper and lower bounds, ideally situated between 0 and 1, adopting a probability distribution-based approach has become an intuitive and effective strategy. Mainly, when the distribution of a specific feature is more pronounced in the tail compared to the distribution associated with a healthy state, this often indicates that the equipment in question is less healthy. More critically, the degree of health deterioration can be quantified by calculating probability values, providing a mathematical basis and support for equipment maintenance and early warning.

In this section, we will detail how to construct an HI based on PCA and the concept of the joint cumulative empirical distribution function, referred to hereafter as JPCCED-HI. Initially, we utilize the PCA method to extract key features from a large set of equipment status data, which capture the main variabilities in the data. This approach not only simplifies the data structure and eliminates feature collinearity but also highlights those variables that most significantly impact the equipment's health status. Subsequently, based on these critical features, a joint cumulative empirical distribution function is constructed. This distribution reflects the probability of transitioning the equipment from a completely healthy state to a potential failure state. Establishing this distribution relies on the cumulative probability changes of the observed feature values throughout the equipment's lifecycle, forming a continuous indicator from 0 to 1 that intuitively represents the health level of the equipment.

The advantage of the JPCCED-HI method is its provision of a statistically quantified assessment tool for equipment health status. It not only indicates

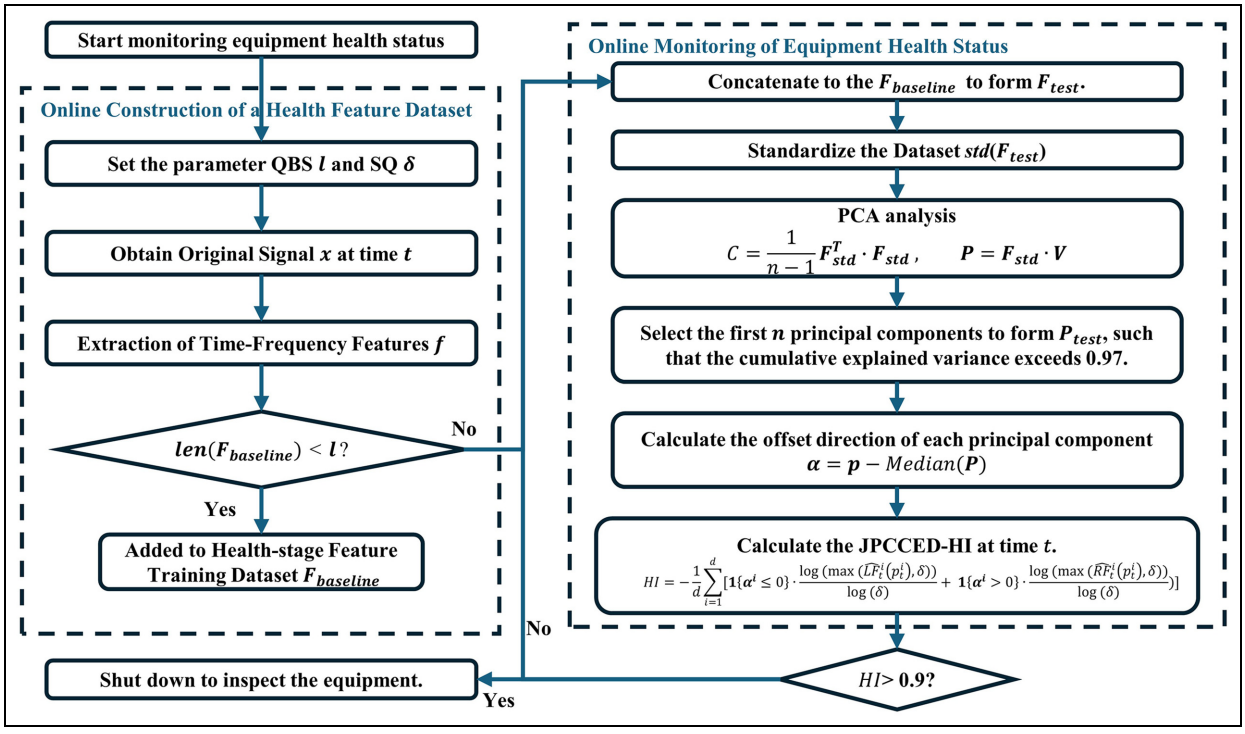

Figure 2. Equipment health monitoring process using the JPCCED-HI.
JPCCED-HI, health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model.

changes in the health of the equipment but also supports maintenance decisions with statistical data. This section will further explore the steps involved in constructing JPCCED-HI and the advantages and potential challenges of this method in practical applications.

JPCCED-HI construction method

The equipment health monitoring process using the JPCCED-HI is shown in Figure 2. First, for each moment t, apply the PCA algorithm to the manually extracted feature set $\boldsymbol{F}_t$, selecting principal components that account for a cumulative variance explanation rate of 97%. This process has two main objectives: to eliminate the linear correlations among features, which are crucial for subsequent calculations of the joint cumulative empirical distribution, and the second is to reduce data noise and redundancy, thereby enhancing analytical efficiency. Through the cumulative variance explanation rate, we can determine how many principal components need to be retained to effectively represent the essential characteristics of the data while ensuring that the data maintain its core structure and properties after simplification.

Let each moment t be represented by a one-dimensional vector composed of d principal components denoted by $\boldsymbol{p}$, where the i-th principal component is represented as $p^{(i)}$. Specifically, samples collected from the start of the device operation up to a designated moment are defined as the baseline, and the principal components are obtained from the healthy sample principal component set $\boldsymbol{P}$. The quantity of baseline samples (QBS) in $\boldsymbol{P}$ is l, with the dimension also being d. Concatenating $\boldsymbol{p}$ with $\boldsymbol{P}$ forms a matrix, referred to as $\boldsymbol{P}_{\text{test}}$. Let F denote the joint cumulative distribution function (CDF) across all d dimensions.

For the same monitoring subject, it is assumed that each $\boldsymbol{p}$ obtained from sampling originates from the same distribution; thus, each principal component within $\boldsymbol{P}$ also comes from the same distribution. Then, according to the definition of the joint CDF F, for any $\boldsymbol{p} \in \mathbb{R}^d$,

$$F(\boldsymbol{p}) = \mathbb{P}\left(\boldsymbol{P}^{(1)} \leqslant p^{(1)},\ \boldsymbol{P}^{(2)} \leqslant p^{(2)},\ \ldots,\ \boldsymbol{P}^{(d)} \leqslant p^{(d)}\right) \tag{5}$$

The equation is abbreviated as

$$F(\boldsymbol{p}) = \mathbb{P}(\boldsymbol{P} \leqslant \boldsymbol{p}) \tag{6}$$

Therefore, the probability obtained by $F(\boldsymbol{p})$ is a measure of how much $\boldsymbol{p}$ deviates to the "left-tail" from a healthy state: the smaller $F(\boldsymbol{p})$ is, the smaller the likelihood that a sample taken from $\boldsymbol{p}$, with each dimension having the same distribution, will satisfy $\boldsymbol{P} \leqslant \boldsymbol{p}$. This indicates a deviation from the healthy state of the

equipment at time t. Similarly, $1 - F(\boldsymbol{p})$ measures the deviation of $\boldsymbol{p}$ from a healthy state from the right-tail. Therefore, if $F(\boldsymbol{p})$ or $1 - F(\boldsymbol{p})$ is small, it indicates that the state of the equipment at time t is unhealthy.

However, since we cannot ascertain the true joint CDF, it must be estimated and simplified for computational purposes. We can obtain a set of components with greater independence by using PCA to remove linear correlations among features, which lead to more accurate HI assessments. It should be noted, though, that while the components derived from PCA eliminate linear correlations, they do not guarantee complete independence. Nevertheless, for the sake of computational convenience, we assume that each principal component is independent, allowing the joint CDF to be effectively factorized:

$$F(\boldsymbol{p}) = \prod_{i=1}^{d} F^{(i)}\left(p^{(i)}\right) \tag{7}$$

where $F^{(i)}$ represents the univariate CDF of the i-th principal component. for any $p^{(i)} \in \mathbb{R}$,

$$F^{(i)}\left(p^{(i)}\right) = \mathbb{P}\left(\boldsymbol{P}^{(i)} \leqslant p^{(i)}\right) \tag{8}$$

A univariate CDF can be precisely estimated using the empirical cumulative distribution function (ECDF).

$$\hat{L}F^{(i)}\left(p^{(i)}\right) := \frac{1}{l}\sum_{t=1}^{l} \mathbf{1}\left\{\boldsymbol{P}^{(i)} \leqslant p^{(i)}\right\} \tag{9}$$

where $\mathbf{1}$ is the indicator function that enumerates as 1, otherwise 0.

Theoretically, we can also use the "right-tail" ECDF to assess the health status of the monitored object.

$$\hat{R}F^{(i)}\left(p^{(i)}\right) := \frac{1}{l}\sum_{t=1}^{l} \mathbf{1}\left\{\boldsymbol{P}^{(i)} \geqslant p^{(i)}\right\} \tag{10}$$

Therefore, based on the assumption that each principal component is independent, we can quickly calculate the joint left-tail and joint right-tail ECDF for all principal components.

$$\hat{L}F(\boldsymbol{p}) = \prod_{i=1}^{d} \hat{L}F^{(i)}\left(p^{(i)}\right) \tag{11}$$

$$\hat{R}F(\boldsymbol{p}) = \prod_{i=1}^{d} \hat{R}F^{(i)}\left(p^{(i)}\right) \tag{12}$$

Based on this, we can establish JPCCED-HI. Considering that some parameters with significant trends might lead to a very low ECDF for the corresponding principal component, it is necessary to set a scale parameter δ. If the value of a principal component at a specific moment is below this scale parameter, it is set to δ. We operate in the negative logarithmic probability space to optimize performance and simplify calculations:

$$\text{LEFT\_HI} = -\frac{1}{d}\sum_{i=1}^{d} \frac{\log\left(\max\left(\hat{L}F^{(i)}\left(p^{(i)}\right), \delta\right)\right)}{\log(\delta)} \tag{13}$$

$$\text{RIGHT\_HI} = -\frac{1}{d}\sum_{i=1}^{d} \frac{\log\left(\max\left(\hat{R}F^{(i)}\left(p^{(i)}\right), \delta\right)\right)}{\log(\delta)} \tag{14}$$

First, we obtained the left-tail and right-tail ECDF scores for each principal component at time t. To address the issue of exceedingly small values caused by a principal component $\boldsymbol{p}$ being at the extreme edge of the distribution at time t, we take the maximum value between $\hat{L}F^{(i)}\left(p^{(i)}\right)$ and δ to lessen this component's impact on the HI, while also ensuring that the HI remains compressed between 0 and 1. This method proposes to prevent the ECDF from taking values that are too low and exclude outliers in specific principal components.

However, we found that using either the left-tail or right-tail uniformly for each principal component does not yield good results because we cannot determine in advance which direction each component will shift as the equipment's health deteriorates. Additionally, this shift might not consistently move in a fixed direction as the equipment degrades. For instance, there may be cases where a component shifts toward the left in the mid-term and then rapidly shifts toward the right in the later stages. Therefore, we recommend comparing each principal component of $\boldsymbol{p}$ at time t with the median of that principal component in the healthy state $\boldsymbol{p}$, denoted as Median($\boldsymbol{P}$).

$$\alpha = \boldsymbol{p} - \text{Median}(\boldsymbol{P}) \tag{15}$$

Based on the sign of $\alpha^{(i)}$, decide whether to use the "left-tail" or "right-tail" to assess that principal component, and finally, take the average of the scores of all principal components:

$$\begin{aligned}\text{HI} = -\frac{1}{d}\sum_{i=1}^{d}\Bigg[&\mathbf{1}\left\{\alpha^{(i)} \leqslant 0\right\} \cdot \frac{\log\left(\max\left(\hat{L}F^{(i)}\left(p^{(i)}\right), \delta\right)\right)}{\log(\delta)} \\ &+ \mathbf{1}\left\{\alpha^{(i)} > 0\right\} \cdot \frac{\log\left(\max\left(\hat{R}F^{(i)}\left(p^{(i)}\right), \delta\right)\right)}{\log(\delta)}\Bigg]\end{aligned} \tag{16}$$

Since we operate in the negative logarithmic probability space, lower probabilities correspond to higher negative logarithmic probabilities. Therefore, the closer HI is to 1, the poorer the health of the equipment, and the closer it is to 0, the better the health of the equipment.

Key parameters in JPCCED-HI

The method primarily involves two parameters: the Quantity of Baseline Samples (QBS), denoted by l, and the Scale Parameter (SP), denoted by δ.

QBS l determines how much data we use to build a baseline, which is a relatively experience-based parameter because the average lifespans of different subjects under different operating conditions can vary significantly. For example, two milling cutters of the same material and model might only last a few hours when machining workpiece A, which is very hard, but could last hundreds of hours when machining workpiece B is softer and less tough. If l is set too high, the system might not have collected enough healthy data before signs of damage appear on the milling cutter used on workpiece A. Conversely, if l is set too low, the milling cutter used on workpiece B might still be in its initial break-in phase, and data on its most stable operating state might not have been collected. Therefore, we recommend that l should represent about 0.2–0.3 of the average lifespan based on the understanding of the lifespan of the object being monitored. In case 1, we will demonstrate the impact of different QBS on the construction results of the HI.

The SP δ mainly serves two functions:

(1) To prevent the ECDF from taking too low values, which also help exclude outliers in certain principal components.
(2) To ensure that the HI remains compressed between 0 and 1.

Since we operate in the negative logarithmic probability space, a minimal ECDF value for a principal component can easily result in a disproportionately large value. Because HI is derived from the mean, such large values can overshadow the results of other principal components. We know that $0 \leqslant \hat{LF}^{(i)}\left(p^{(i)}\right) \leqslant 1$ and $0 \leqslant \delta \leqslant 1$; therefore,

$$\log(\delta) \leqslant \log\left(\max\left(\hat{LF}^{(i)}\left(p^{(i)}\right), \delta\right)\right) \leqslant 0 \tag{17}$$

$$0 \leqslant \frac{\log\left(\max\left(\hat{LF}^{(i)}\left(p^{(i)}\right), \delta\right)\right)}{\log(\delta)} \leqslant 1 \tag{18}$$

Given that the "right-tail" also satisfies the equation, it follows that

$$0 \leqslant \mathrm{HI} \leqslant 1 \tag{19}$$

This article analyzes two cases to validate the proposed method's effectiveness. The first case involves a complete lifecycle dataset of gearboxes collected in 2023 by our team at Southwest Jiaotong University. In Case 1, we analyze and demonstrate the impact of different QBS and SP on the construction results of HI. The second case utilized a publicly available bearing experimental dataset from the PRONOSTIA setup, which is considered a benchmark dataset for HI construction and lifespan prediction and is widely used by researchers.

Case study 1: HI construction of gearbox datasets

Dataset description

In this experiment, vibration signals across the entire lifecycle of gears operating at a constant speed were collected. The gearbox setup involved a two-stage parallel shaft configuration with rolling drum bearings. Two CMSONE-TES001T accelerometers were mounted at the end cap to acquire vibration signals from the x and y axes of the gearbox's input shaft. Vibration signals from the intermediate shaft's x, y, and z axes were also collected. Additionally, a photoelectric sensor was employed to continuously capture the rotational speed pulse signal from the gear's input shaft at a sampling frequency of 12,800 Hz and a sample duration of 5.12 s. The layout of the experimental setup is depicted in Figure 3. The positional relationship of the gear set is shown in Figure 4. Specific parameters of the test gear type are detailed in Table 3. Data from the x-axis sensor on the load end was used for analysis in this study. Furthermore, each set of 8192 data points was segmented into a sample, with corresponding feature values extracted as indicated in Tables 1 and 2.

We conducted a series of constant-speed tests on gears during the experiment to analyze vibration signals under various load conditions. We meticulously set multiple conditions to collect data throughout the entire lifecycle of the gears. Particularly toward the end of their service life, we observed various fault states, detailed in Table 4.

Process of constructing JPCCED-HI

In this case, vibration signals from the X-axis of the intermediate shaft were used because we observed that the X-axis's vibration amplitude varied significantly throughout the entire lifecycle. We extracted manual features from the gearbox vibration signals, and Figure 5 illustrates some of these features throughout the entire lifecycle of the G1 gearbox. Certain features exhibit a noticeable correlation with the gearbox's degradation process. This suggests that these characteristics could be used to monitor the gearbox's condition and predict its future performance.

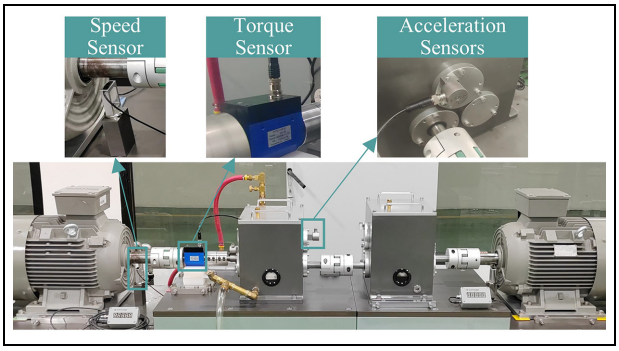

Figure 3. Gearbox test platform.

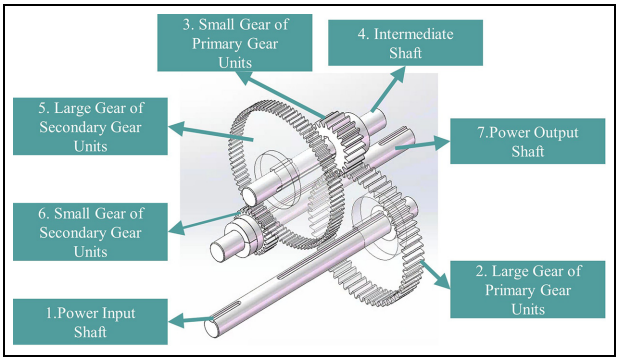

Figure 4. The positional relationship of the gear set.

Table 3. Gear parameter information for secondary drive gearbox test platform.

Parameters	Primary gear	Secondary gear
Number of small gear teeth	29	36
Number of large gear teeth	95	90
Small gear tooth width/mm	15	15
Large gear tooth width/mm	15	15
Pressure angle/°	20	20
Modulus/mm	1.5	1.5

By performing PCA and feature extraction on G1 sample data, we have plotted the situation of the principal components. It is observed that the first eight principal components have already explained 97% of the total variance in Figure 6. Therefore, we use the space constructed by these first eight principal components to map the original features, achieving feature dimensionality reduction. Further, we plotted the joint distribution of the first three principal components in Figure 7. We found some correlation between them, and these principal components do not conform to the standard normal distribution. Therefore, we adopted a method based on the joint CDF, which does not require the feature set to follow a normal distribution or any other specific distribution model.

The samples were evenly divided into eight groups according to the sequence of sampling times. Box plots for the first three principal components were generated as depicted in Figure 8 to observe distribution changes. The first box comprises principal component statistics for samples 1–2937, the second box for samples 2938–5874, and so forth. Principal Component 1 (PC1) exhibits a strong trend, with the mean of each statistical stage progressively shifting upward. PC2 and PC3 show weaker trends, with noticeable shifts in distribution only in the last two statistical stages. We can quickly identify which principal components significantly influence the changes in JPCCED-HI. For example, if the data from stages 1 to 5 are considered baseline data, the change in PC1 during the sixth and seventh statistical stages relative to stages 1–5 is significantly more significant than that of PC2 and PC3. According to equations (9) and (16), it is easy to see that during this stage, the changes in JPCCED-HI will be mainly caused by PC1. In stages 8 and 9, since PC1 has already significantly exceeded the baseline range, its continued increase will not lead to significant changes in JPCCED-HI due to the constraints of the scale parameter. The deviations of PC2 and PC3 relative to the baseline mainly occur in stages 8 and 9, so the changes in PC2 and PC3 during these stages will significantly impact the value of JPCCED-HI. It should be emphasized that we have only analyzed the first three principal components as examples to explain the changes in JPCCED-HI to a certain extent. However, the actual JPCCED-HI value will result from the combined effects of all selected principal components.

Additionally, the last three box sizes for principal components visibly increase, indicating more significant fluctuations in principal component values. This suggests that JPCCED-HI may undergo relatively

Table 4. Detailed information of gearbox full lifecycle test dataset under different torque load.

Gear no.	Torque load (%)	Speed (rpm)	Failure modes	Sample size
G1	50	40	Pitting gear at load end	26,434
G2	60	40	Broken pinion gear at load end	21,988
G3	70	40	Broken pinion gear at load end	8192

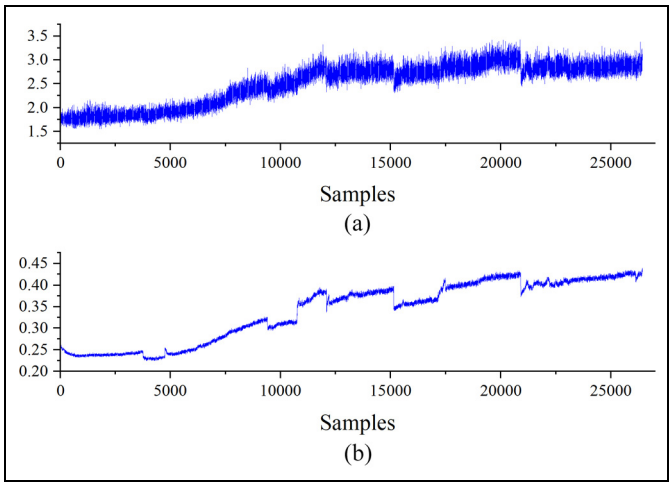

Figure 5. Peak and root mean square (RMS) of vibration signals throughout the gearbox lifecycle: (a) peak and (b) RMS.
JPCCED-HI, health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model.

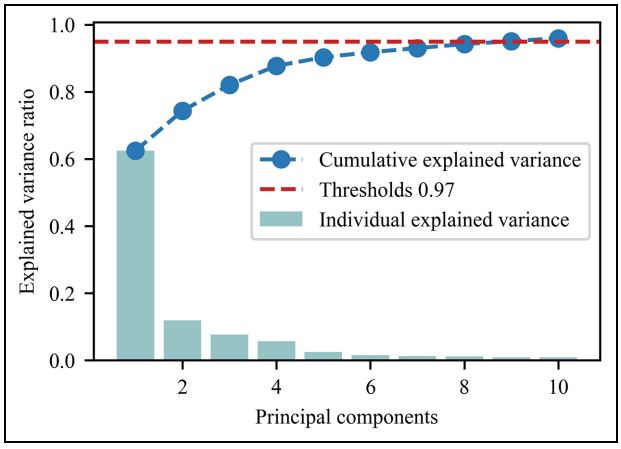

Figure 6. Principal component explained variance ratio of G1.
JPCCED-HI, health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model.

larger changes during this phase, which is advantageous for capturing stages of accelerated equipment degradation. PC1 and PC2 shift toward higher values, whereas PC3 shifts toward lower values, underscoring the importance of using both "left-tail" and "right-tail" ECDF.

During the construction of the JPCCED-HI, the parameter settings were as follows: Scale at 0.02, QBS at 5000, and the number of PCA components at 8. This configuration means that the first 5000 features of vibration data were used as the health samples. The first eight principal components were utilized to build the HIs. Median smoothing was applied to these HIs. As shown in Figure 9, the results indicate that all HIs are between 0 and 1, demonstrating good trending characteristics.

Key parameter analysis

To more clearly demonstrate the effectiveness of the method proposed in this article, we conducted a detailed analysis, sequentially comparing the specific impacts of different parameters QBS and SP on the analysis results. First, we adjusted the QBS parameter to observe its influence on the calculation and trend analysis of HIs. Subsequently, we changed the SP parameter to analyze how this scale parameter affects the construction results of the model's HIs. Through this approach, we could thoroughly illustrate the specific effects of each parameter setting on the final results, thereby validating the reliability and applicability of the method proposed in this article. This comparative analysis enhances the understanding of the method and provides data support and a theoretical basis for future applications and optimizations.

To set the QBS parameter, we must have an initial understanding of the overall lifespan of the monitoring device. It is recommended that the first 20%–40% of the estimated lifespan of the target monitoring device be defined as healthy data, serving as a baseline for the model. Subsequent data collected will be compared against this benchmark using the algorithm proposed in this study to calculate HIs. In this experiment, the QBS values were set at 1000, 3000, 5000, and 7000 to observe the changes in the HIs of three gearboxes. Figure 10 shows that due to different QBS settings, the starting points for monitoring the three gearboxes vary. Although the quantity of different training samples may affect the assessment of HIs by the model proposed in this article, all models exhibit similar trends.

We explore the impact of different scale factors on the HI curves. Under constant conditions, we compared the changes in the HI curves when the scale

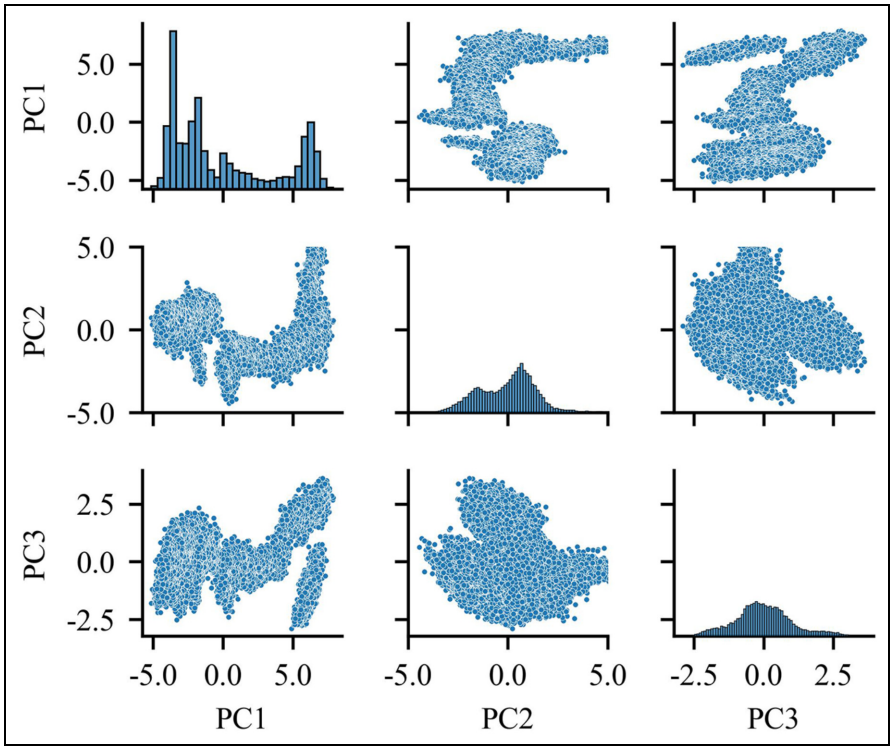

Figure 7. Distribution of the first three principal components and scatter plot matrix of G1.
JPCCED-HI, health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model.

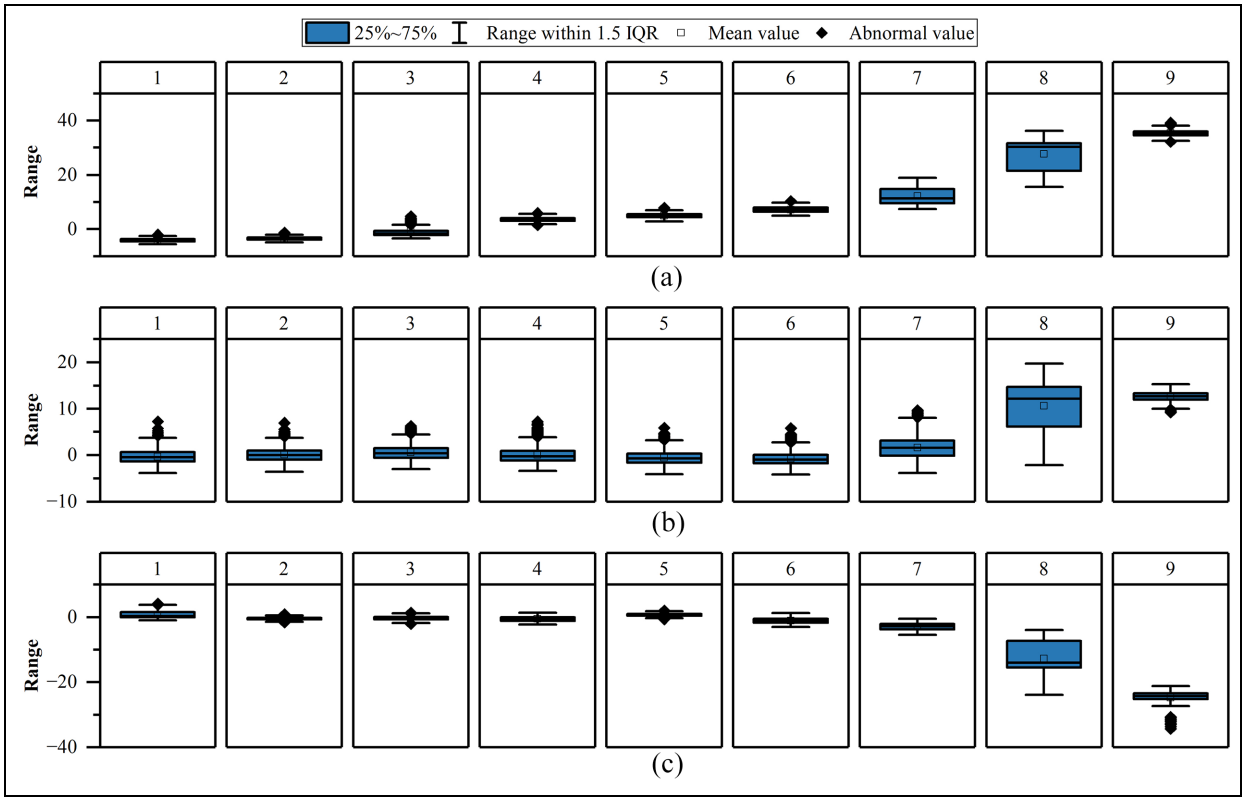

Figure 8. Distribution changes of the first three principal components of G1 throughout its lifecycle: (a) PC1, (b) PC2, and (c) PC3.
JPCCED-HI, health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model.

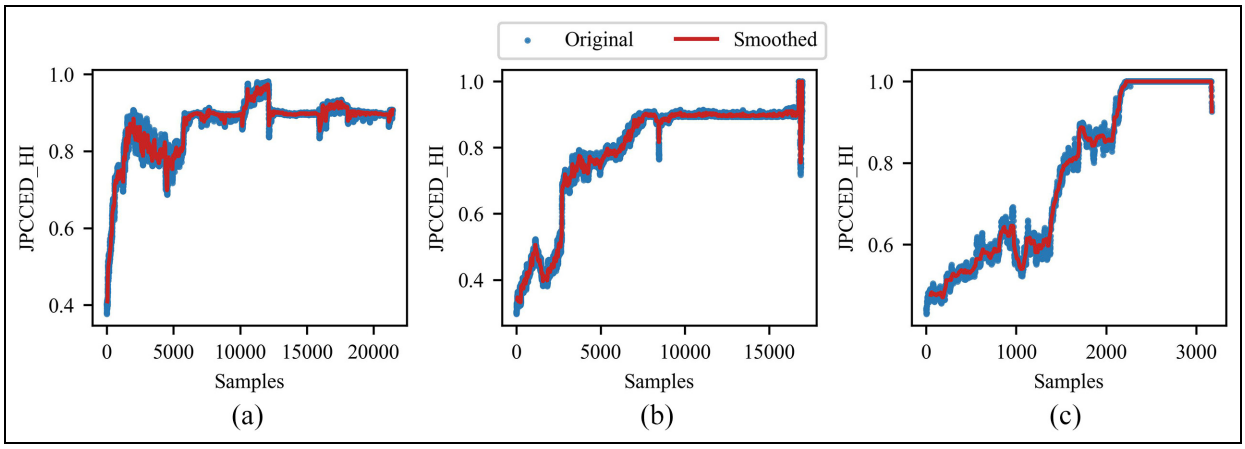

Figure 9. JPCCED-HI construction results for three gearboxes: (a) G1, (b) G2, and (c) G3.
JPCCED-HI, health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model.

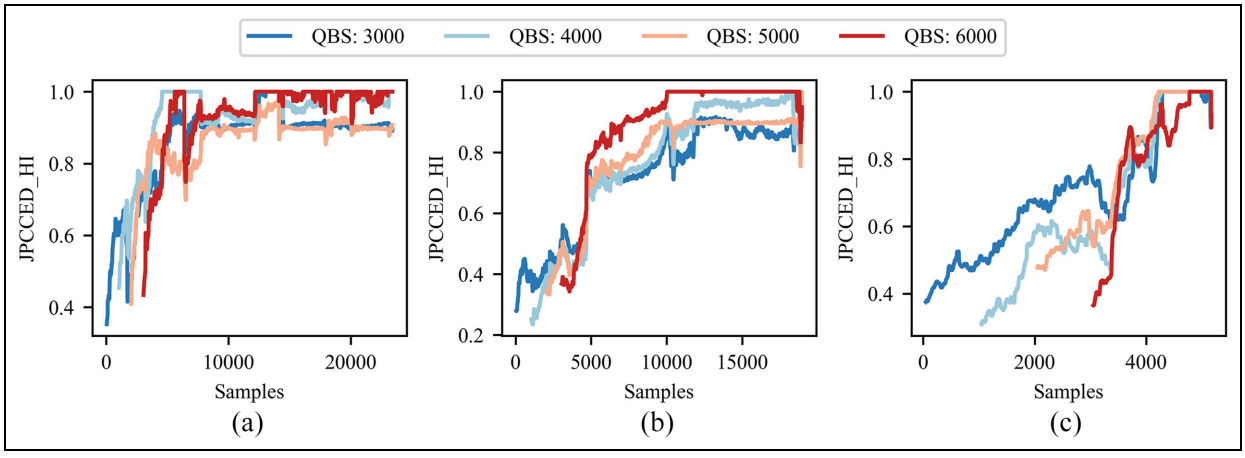

Figure 10. JPCCED-HI construction results for three gearboxes under different QBS parameter conditions: (a) G1, (b) G2, and (c) G3.
JPCCED-HI, health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model; QBS, quantity of baseline samples.

factor was set at 0.001, 0.01, 0.1, and 0.2. Figure 11 clearly shows that the smaller the scale factor, the lower the HI values at corresponding times. However, since the range of HIs is between 0 and 1, a too-large scale factor might compress the HI values within a narrow range, thereby obscuring the precise depiction of the equipment's health deterioration process. Based on this finding, we recommend that the scale factor chosen for practical applications should be less than 0.01 to track and reflect the health status of the equipment more accurately.

We did not attempt larger QBS and SP values. As the total sample size for G3 is 8191, if the QBS value continues to increase, G3 will already have faults before entering the HI evaluation stage. When the SP value is 0.2, it shows a higher value at the initial stage of HI monitoring, which is not conducive to monitoring the health status of the equipment.

Online monitoring capability

According to the formulas in the JPCCED-HI construction method, it is evident that JPCCED-HI has $\mathcal{O}(nd)$ time and space complexity, and we break down the analysis in steps. In the estimation steps, computing ECDF for all d dimensions using QBS l samples leads to $\mathcal{O}(ld)$ time and space complexity. In the aggregation steps, tail probability calculation and aggregation also lead to $\mathcal{O}(ld)$ time and space complexity.

We measured the average runtime of the algorithm on our device, which runs Windows 11 and is equipped with a 13th Gen Intel(R) Core(TM) i5-13600K

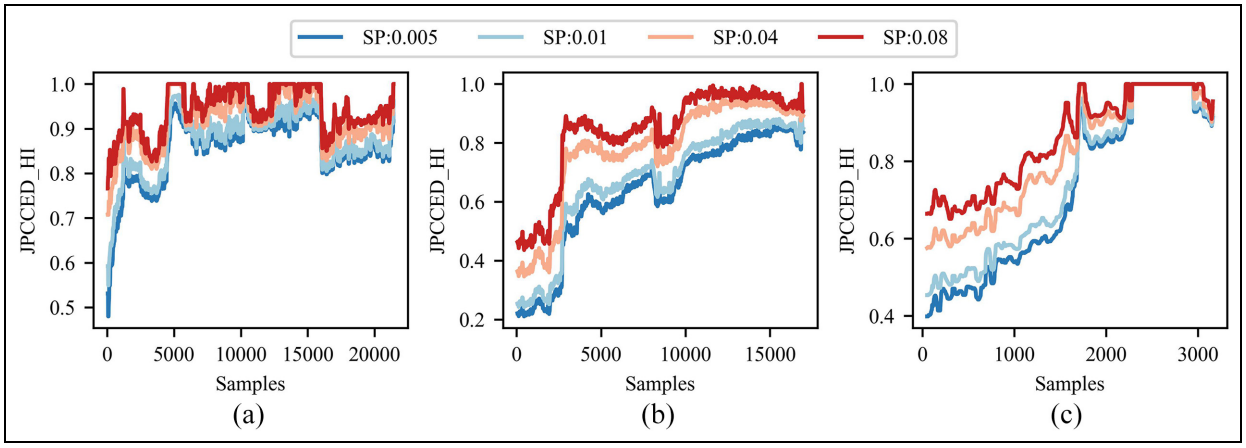

Figure 11. JPCCED-HI construction results for three gearboxes under different SP conditions: (a) G1, (b) G2, and (c) G3.
JPCCED-HI, health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model.

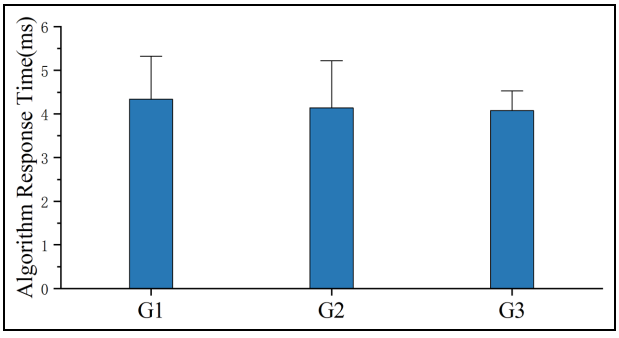

Figure 12. JPCCED-HI algorithm response time.
JPCCED-HI, health indicator developed based based on the Joint Principal Component Cumulative Empirical Distribution Model.

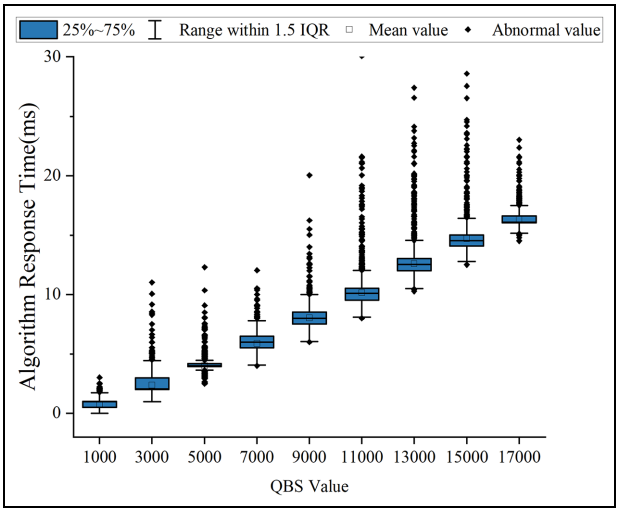

Figure 13. Relationships between the inference time of JPCCED-HI and QBS parameters.
JPCCED-HI, health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model; QBS, quantity of baseline samples.

3.50 GHz processor and 32 GB of RAM. This measurement only includes the time from inputting the original data features into the algorithm to obtaining the JPCCED-HI results. It does not include the time taken for feature extraction. We believe different sensitive features can be tested based on the specific task requirements. Therefore, the feature extraction time was not considered in the runtime calculation. As shown in Figure 12, when QBS is set to 5000, the average runtime for online evaluation of HIs G1, G2, and G3 is 4.25, 4.25, and 4.21 ms, respectively, with an overall average runtime of 4.24 ms. This demonstrates that the method meets the requirements for online evaluation.

We evaluated the impact of different QBS values on the online performance of HIs. It can be seen in Figure 13 that as the QBS value increases, the average response time also shows a linear increase. When the QBS is below 11,000, the response time remains within 10 ms. Therefore, we must pay attention to the QBS value and avoid enormous QBS values to ensure a fast response time. If the device has a long lifespan, it means that during the baseline data collection phase, increasing the data sampling interval can be used to reduce the QBS value.

Anti-interference capability

To comprehensively assess the anti-interference capability of the algorithm developed in this study, we deliberately introduced white noise of varying magnitudes into the original data for testing. The signal-to-noise ratio (SNR) was initially set at 30 dB and reduced by 1 unit in each experiment until reaching a minimum level of 5 dB. We used the method proposed in this article to extract HIs from vibration signals with SNRs

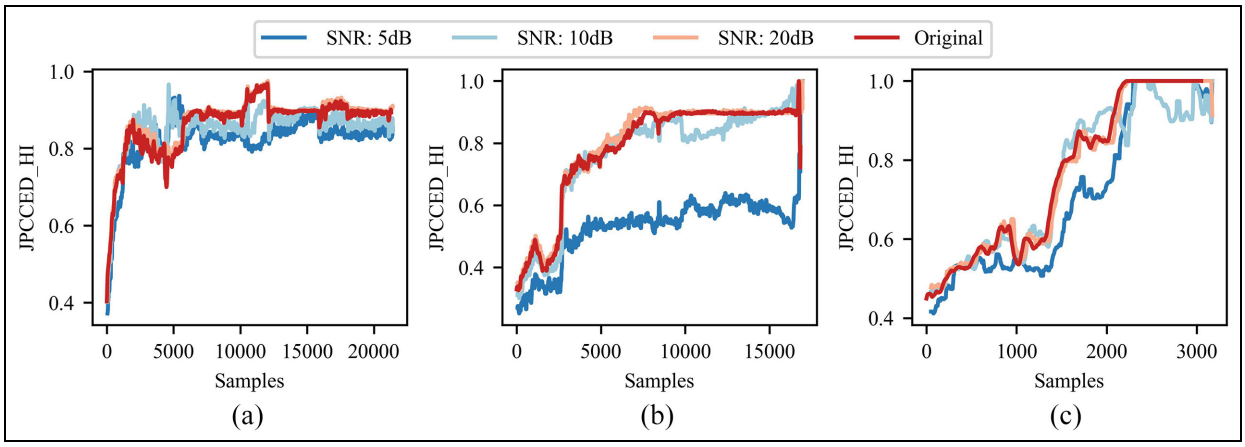

Figure 14. JPCCED-HI construction results for three gearboxes under various SNR conditions: (a) G1, (b) G2, and (c) G3.
JPCCED-HI, health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model; SNR, signal-to-noise ratio.

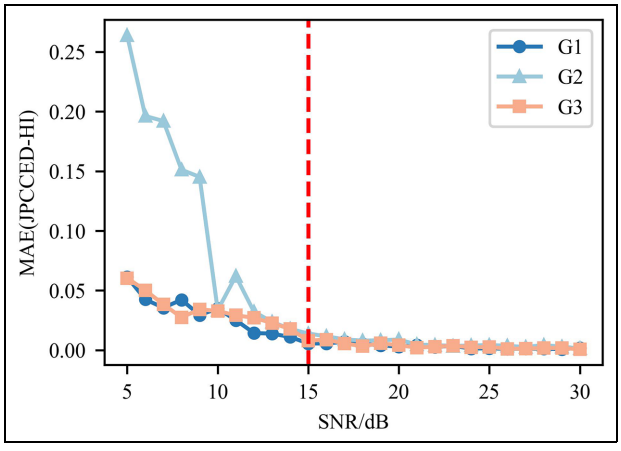

Figure 15. MAE values of JPCCED-HI under 5–30 SNR conditions compared to the original un-noised signal.
JPCCED-HI, health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model; MAE, mean absolute error; SNR, signal-to-noise ratio.

of 5, 10, 20, and 30 dB, with the corresponding results displayed in Figure 14. These charts demonstrate that even under extreme noise conditions, the algorithm still maintains good results in constructing HIs, showcasing its exceptional noise resistance. The stability and accuracy of the system under various SNR conditions confirm our algorithm's advanced design and practicality, especially in dealing with common noise disturbances in real-world scenarios.

Figure 15 illustrates the changes in mean absolute error between the HIs extracted by the algorithm and those constructed from the original signal within a SNR range of 5–30 dB after adding white noise. Figure 15 shows that when the SNR is above 15 dB, the average deviation of the HIs obtained by the algorithm remains below 0.05. This indicates that even in the presence of some noise interference, the extracted HIs are very close to those obtained from the original signal under noise-free conditions. This result highlights the superior performance of the algorithm in resisting noise interference. Moreover, even at lower SNRs, such as within the 5–15 dB range, the deviation in HIs remains acceptable, further demonstrating the algorithm's ability to maintain indicator accuracy under different noise environments. These findings are of significant practical importance for handling noise interference in signals in real applications and provide solid data support for the future application of this technology in more challenging environments.

Case study 2: HI construction of bearing dataset

Dataset description

In Case 2, we selected the vibration signal dataset from the "IEEE PHM 2012 Prognostic Challenge." The dataset is derived from an experimental system named PRONOSTIA, as illustrated in Figure 16. The bearing model tested is NSK 6804DD. This dataset includes a series of run-to-failure experiments under three different load-speed conditions. Table 5 details the conditions of the bearings under each scenario, including load and speed. The sampling frequency for the vibration signals is 25.6 kHz, with signals sampled every 10 s for a duration of 0.1 s. Testing of all bearings is immediately halted if the vibration signal amplitude exceeds the critical threshold of 20 g. During testing, each bearing undergoes natural degradation to avoid premature failures. Due to the diversity among different bearings and the randomness of the degradation processes, the failure modes of different bearings may vary slightly. For more detailed information on the test

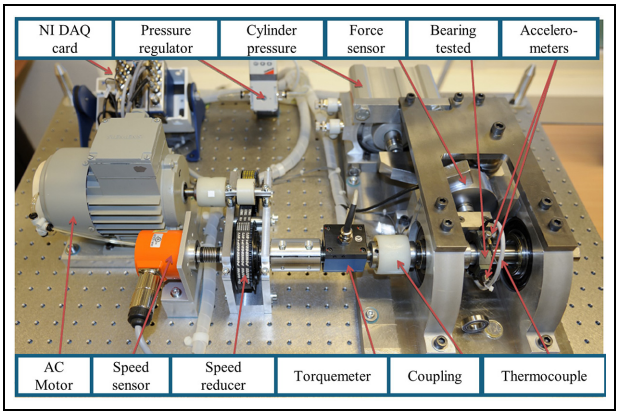

Figure 16. PHM2012 bearing experimental platform.

Table 5. Detailed information of bearings full lifecycle test dataset under different working conditions.

Bearing no.	Speed (r/min)	Load (N)	Sample size
1_1	1800	4000	2803
1_2	1800	4000	871
1_3	1800	4000	1802
1_4	1800	4000	1139
1_5	1800	4000	2302
1_6	1800	4000	2302
1_7	1800	4000	1502
2_1	1650	4200	911
2_2	1650	4200	797
2_3	1650	4200	1202
2_4	1650	4200	612
2_5	1650	4200	2002
2_6	1650	4200	572
2_7	1650	4200	172
3_1	1500	5000	515
3_2	1500	5000	1637
3_3	1500	5000	352

platform and experiments, please refer to Nectoux et al.[32] As the methodology of this study requires sufficient QBS to establish a baseline for each bearing, several bearings with sample sizes below 1500 were excluded. Consequently, six bearings, specifically bearing numbers 1_1, 1_3, 1_7, 2_3, 2_5, and 3_2 were selected for this experiment. These six bearings cover all three operational conditions used in the experiments, enabling the verification of whether the method can address the construction of HIs under different operating conditions.

Comparison model

This article presents a health-stage data training model for constructing HI that monitors the health status of equipment without the need for labeled data or fault data. To comprehensively evaluate and validate the superiority of our model, we compared it with five different types of healthy-state data training HI methods. These comparative methods include those based on statistical features, distance-based methods, similarity-based methods, and methods based on reconstruction error. Each method was tested on the same dataset to ensure consistency and fairness in the evaluation.

- RMS-HI, a HI that evaluates the overall condition of machinery by calculating the root mean square of vibration signals, thereby reflecting the equipment's energy levels and identifying potential issues.
- SOM-HI, utilizes SOMs, a type of artificial neural network, to visualize and cluster machine data, allowing for the detection of anomalous patterns that indicate equipment degradation.
- MD-HI, measures the statistical distance between a normal operational point and the current operational state of the machine.
- AE-HI, employs autoencoders (AE) to model and reconstruct operational data, effectively highlighting discrepancies that could signify machine faults.
- RS-HI, compares the related-similarity of current machine signals to a set of baseline signals to determine relative changes over time.

In this article, the JPCCED-HI method introduces two key parameters: SP is set at 0.01, and QBS is set at 900. To control variables, all methods used in this study utilize manually extracted features mentioned earlier in the text to construct HIs, ensuring consistency in experimental conditions. Additionally, all models employ the same median sliding average to smooth the data and reduce random fluctuations, aiding in a more accurate performance assessment of the models.

It should be emphasized that this article does not compare the proposed HI construction method with traditional methods that use full lifecycle data for HI construction. This decision is based on two primary considerations: first, methods using full lifecycle data (typically supervised HI construction methods fall into this category) are generally not suitable for the health status monitoring needs addressed in this article under no-failure data conditions, as they require a pre-labeled full lifecycle dataset for model training, which is quite demanding for practical industrial applications. Second, the effectiveness of supervised methods largely depends on the quality and quantity of the training samples, making a direct comparison with our proposed method potentially unfair. Therefore, this article explores and validates the potential and practical value

of the health-stage data training model in equipment health monitoring.

To quantitatively assess the performance of the proposed method, the following two metrics and a composite result of these metrics are used for evaluation.

Trendability: It is designed to measure whether HI trends are the same with the run-to-failure degradation as follows:

$$\mathrm{Tred}(H,\ T) = \frac{\left|\sum_{k=1}^{K}(h_k - \bar{H})(t_k - \bar{T})\right|}{\sqrt{\sum_{k=1}^{K}(h_k - \bar{H})^2 \sum_{k=1}^{K}(t_k - \bar{T})^2}} \tag{20}$$

where $\bar{H} = \frac{1}{K}\sum_{k=1}^{K} h_k$ and $\bar{T} = \frac{1}{K}\sum_{k=1}^{K} t_k$. K is the number of HI and h_k is the k-th HI at time t_k.

Scale similarity: This metric assesses the scale similarity of HIs, representing the similarity of HIs range among different cutting tools. Its formula is given by,

$$\mathrm{Scales}(H) = \frac{1}{N}\sum_{j=1}^{N}\left(1 - \frac{\left|h^j_{\max} - h_{\max}\right| + \left|h^j_{\min} - h_{\min}\right|}{\mathrm{len}(H) + \mathrm{len}(H^j)}\right) \tag{21}$$

where N is the number of bearings in the experiment. $h_{\max}$ and $h_{\min}$ denote the maximum and minimum values of H aimed at the selected bearings, respectively. $h^j_{\max}$ and $h^j_{\min}$ indicate those values aimed at the j-th bearings left. $\mathrm{len}(H)$ and $\mathrm{len}(H^j)$ mean the range of H aimed at the selected bearings and the j-th bearings left respectively.

The two metrics estimate different properties of a HI. To comprehensively assess the suitability of a HI, a composite metric (Cme) is proposed as follows:

$$\mathrm{Cme} = \frac{\mathrm{Tred} + \mathrm{Scales}}{2} \tag{22}$$

Comparison result

The HI results of the six models are shown in Figure 17. All HI methods correlate with the lifespan of the bearings, reaching relatively high values at the end-of-life stages, which reflect a poorer health state of the equipment. However, there are significant variations in the scale of HIs for different bearings across RMS-HI, SOM-HI, MD-HI, and AE-HI, especially toward the end of life. These differences in magnitude among various bearings make it challenging to determine a consistent failure threshold. On the other hand, the similarity-based indicator, through its algorithmic design, effectively ensures that its values remain between 0 and 1. However, this method is sensitive to even minor changes in features. It can be observed that the RS-HI health indicator tends to show unstable fluctuations in the early stages of equipment monitoring. For instance, Bearing1_7 showed a peak value close to 1.0 at the beginning of the monitoring, while Bearing3_2 reached an initial peak of 0.8.

In the bar graph of Figure 18, different bearings serve as the horizontal axis, with trendability, scale similarity, and the composite indicator represented by different colors of the bars. It is evident that, compared to other control groups, the method proposed in this article achieves the highest composite indicator scores across multiple bearings. The average values of all metrics for six bearings were calculated and are displayed in Table 6. The results similarly indicate that JPCCED-HI is superior compared to other HIs. Moreover, this method effectively ensures that the constructed HIs consistently remain within the 0–1 range, significantly outperforming other indicators.

This ability to maintain HIs within the 0–1 range not only simplifies the setting of failure thresholds but also enhances the reliability and practicality of the health monitoring system. The consistency and predictability of the indicator values facilitate a smooth and accurate transition from experimental data to real-world applications. By comparing the HIs across different bearings, this method also shows high consistency and comparability across devices, which are particularly important for the health management of multi-bearing systems.

In summary, the HI construction method based on health-stage data training proposed in this article not only outperforms other methods in composite evaluation but also exhibits significant practical advantages in maintaining consistency of indicators and achieving health monitoring across devices, providing a new and effective tool for the field of equipment health monitoring.

Conclusion

This study proposes the JPCCED-HI method for constructing HIs using healthy-state data. By leveraging PCA and the concept of the joint ECDF, it establishes robust HIs constrained within the 0–1 range. This method does not require the preconstruction of a full lifecycle dataset for training, nor does it restrict the operational conditions of the monitored objects, as it builds baselines entirely based on these objects' current health state data. It theoretically ensures that HIs stay between 0 and 1, addressing common challenges in existing HI construction methods. Comprehensive analyses in two case studies—one on gearbox datasets and the other on publicly available bearing datasets—demonstrate the effectiveness of the JPCCED-HI. The

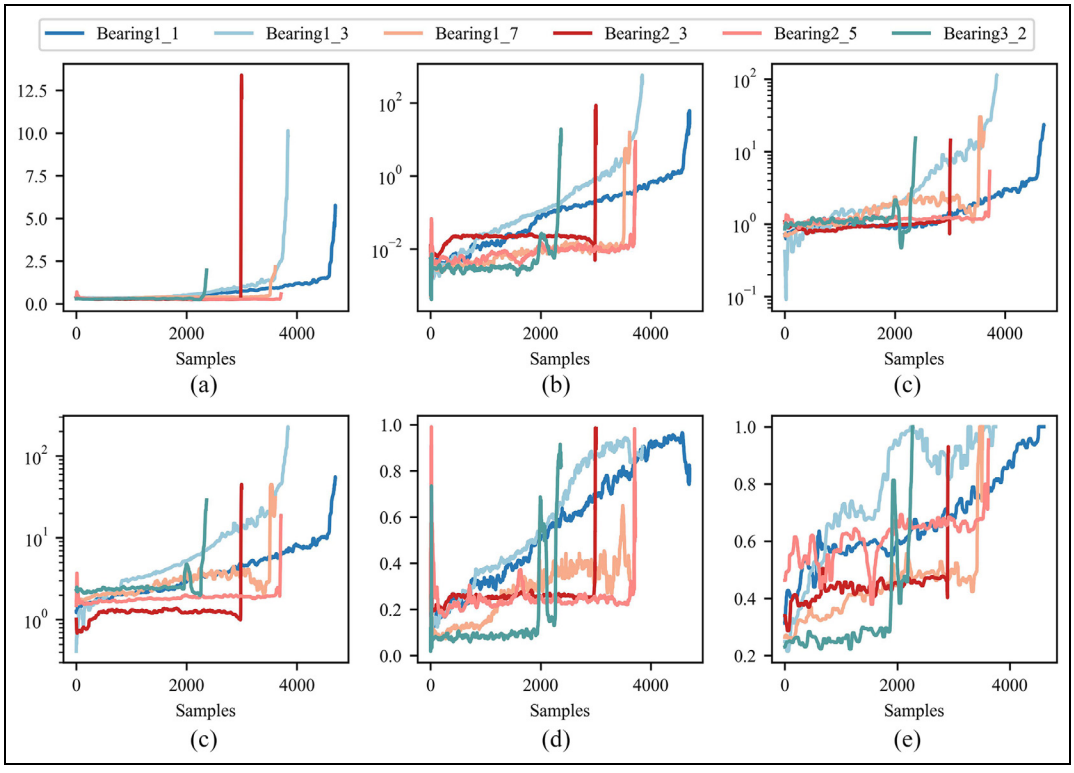

Figure 17. HI results constructed for six test bearings in the PHM2012 dataset by six different healthy state data training models: (a) RMS-HI, (b) AE-HI, (c) MD-HI, (d) SOM-HI, (e) RS-HI, and (f) JPCCED-HI.
HI, health indicator; JPCCED-HI, health indicator developed based on the Joint Principal Component Cumulative Empirical Distribution Model; MD, Mahalanobis Distance; RMS, root mean square; RS, related-similarity; SOM, self-organizing map.

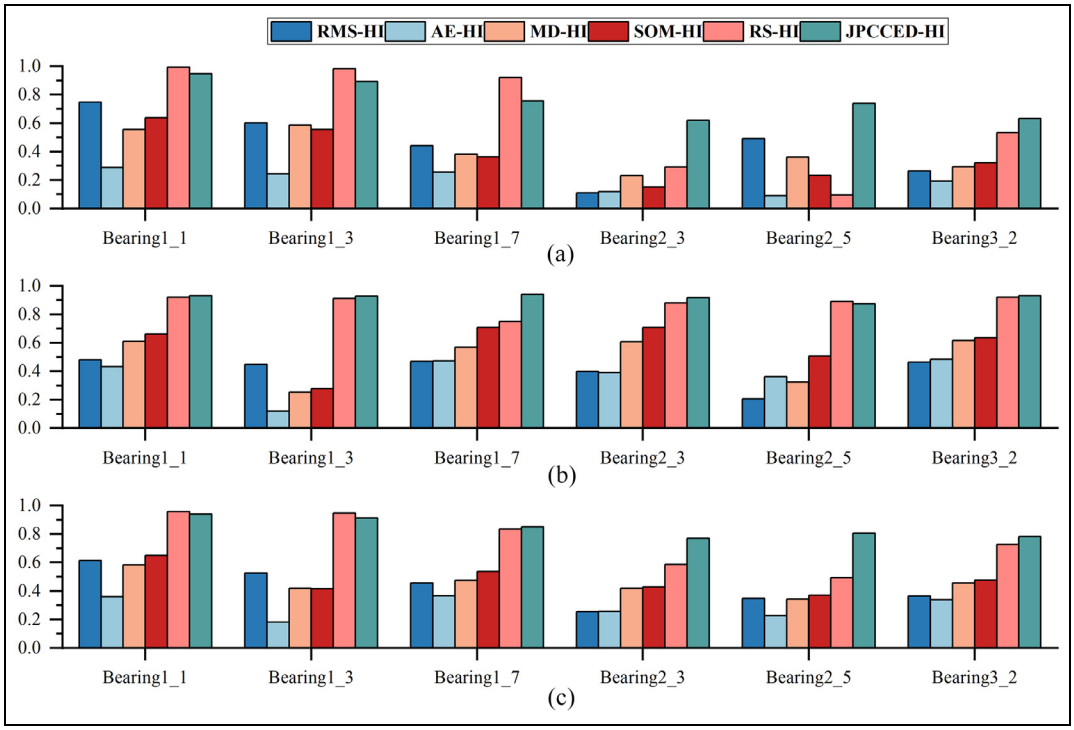

Figure 18. Assessment results for HIs trendability, scale similarity, and condition monitoring on six bearings using six different healthy-state data training models: (a) trendability, (b) scale similarity, and (c) Cme.
Cme, composite metric; HI, health indicator.

Table 6. Mean value of each metric in six bearings.

Model	Trendability	Scale similarity	Cme
RMS-HI	0.442	0.411	0.427
AE-HI	0.199	0.378	0.288
MD-HI	0.400	0.497	0.449
SOM-HI	0.378	0.582	0.480
RS-HI	0.635	0.880	0.758
JPCCED-HI	0.765	0.920	0.842

robust model maintains accurate HI construction even in high-noise conditions, highlighting its practical advantages in real applications. Additionally, the scalability of JPCCED-HI enables its application across various types of machinery and operational environments, demonstrating its extensive industrial potential.

Based on these findings, future research could explore integrating JPCCED-HI with real-time data acquisition systems for dynamic health monitoring and introducing adaptive parameter setting methods to simplify algorithm application.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Hongliang Song https://orcid.org/0009-0008-7884-2343
Liang Guo https://orcid.org/0000-0001-5338-4958

References

1. Kundu P, Darpe AK and Kulkarni MS. A review on diagnostic and prognostic approaches for gears. *Struct Health Monit* 2021; 20(5): 2853–2893.
2. Polverino L, Abbate R, Manco P, et al. Machine learning for prognostics and health management of industrial mechanical systems and equipment: a systematic literature review. *Int J Eng Bus Manage* 2023; 15: 18479790231186848.
3. Wang D, Tsui KL and Miao Q. Prognostics and health management: a review of vibration based bearing and gear health indicators. *IEEE Access* 2018; 6: 665–676.
4. Widodo A and Yang BS. Application of relevance vector machine and survival probability to machine degradation assessment. *Expert Syst Appl* 2011; 38(3): 2592–2599. DOI:10.1016/j.eswa.2010.08.049.
5. Mosallam A, Medjaher K and Zerhouni N. Time series trending for condition assessment and prognostics. *J Manuf Technol Manage* 2014; 25(4): 550–567.
6. Hong S, Zhou Z, Zio E, et al. An adaptive method for health trend prediction of rotating bearings. *Digital Signal Process* 2014; 35: 117–123.
7. Zhao M, Tang B and Tan Q. Bearing remaining useful life estimation based on time–frequency representation and supervised dimensionality reduction. *Measurement* 2016; 86: 41–55.
8. Chen L, Xu G, Zhang S, et al. Health indicator construction of machinery based on end-to-end trainable convolution recurrent neural networks. *J Manuf Syst* 2020; 54: 1–11.
9. Guo L, Lei Y, Li N, et al. Machinery health indicator construction based on convolutional neural networks considering trend burr. *Neurocomputing* 2018; 292: 142–150.
10. Guo L, Li N, Jia F, et al. A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing* 2017; 240: 98–109.
11. Zhang B, Zhang S and Li W. Bearing performance degradation assessment using long short-term memory recurrent network. *Comput Ind* 2019; 106: 14–29.
12. Xu F, Shu X, Li X, et al. Health indicator construction for roller bearing based on an unsupervised deep belief network with a novel sigmoid zero local minimum point model. *Struct Health Monit* 2021; 20(4): 2110–2123.
13. Yu K, Lin TR and Tan J. A bearing fault and severity diagnostic technique using adaptive deep belief networks and Dempster–Shafer theory. *Struct Health Monit* 2020; 19(1): 240–261.
14. Chen Z, Guo R, Lin Z, et al. A data-driven health monitoring method using multiobjective optimization and stacked autoencoder based health indicator. *IEEE Trans Ind Inf* 2021; 17(9): 6379–6389.
15. Guo L, Yu Y, Duan A, et al. An unsupervised feature learning based health indicator construction method for performance assessment of machines. *Mech Syst Signal Process* 2022; 167: 108573.
16. Zhou H, Huang X, Wen G, et al. Construction of health indicators for condition monitoring of rotating machinery: a review of the research. *Expert Syst Appl* 2022; 203: 117297.
17. Wang Y, Peng Y, Zi Y, et al. A two-stage data-driven-based prognostic approach for bearing degradation problem. *IEEE Trans Ind Inf* 2016; 12(3): 924–932.
18. Schwartz S, Montero Jiménez JJ, Vingerhoeds R, et al. An unsupervised approach for health index building and for similarity-based remaining useful life estimation. *Comput Ind* 2022; 141: 103716.
19. Li Q, Yan C, Chen G, et al. Remaining Useful Life prediction of rolling bearings based on risk assessment and degradation state coefficient. *ISA Trans* 2022; 129: 413–428.
20. Qiu H, Lee J, Lin J, et al. Robust performance degradation assessment methods for enhanced rolling element bearing prognostics. *Adv Eng Inf* 2003; 17(3–4): 127–140.

21. Huang R, Xi L, Li X, et al. Residual life predictions for ball bearings based on self-organizing map and back propagation neural network methods. *Mech Syst Signal Process* 2007; 21(1): 193–207.
22. Pei X, Li X and Gao L. A novel machinery RUL prediction method based on exponential model and cross-domain health indicator considering first-to-end prediction time. *Mech Syst Signal Process* 2024; 209: 111122.
23. Yan X, Liu Y and Jia M. Health condition identification for rolling bearing using a multi-domain indicator-based optimized stacked denoising autoencoder. *Struct Health Monit* 2020; 19(5): 1602–1626.
24. Kundu P, Darpe AK and Kulkarni MS. An ensemble decision tree methodology for remaining useful life prediction of spur gears under natural pitting progression. *Struct Health Monit* 2020; 19(3): 854–872.
25. Lu C, Tao L and Fan H. An intelligent approach to machine component health prognostics by utilizing only truncated histories. *Mech Syst Signal Process* 2014; 42(1–2): 300–313.
26. Wang D and Tsui KL. Theoretical investigation of the upper and lower bounds of a generalized dimensionless bearing health indicator. *Mech Syst Signal Process* 2018; 98: 890–901.
27. Xu J, Zhou L, Zhao W, et al. Zero-shot learning for compound fault diagnosis of bearings. *Expert Syst Appl* 2022; 190: 116197.
28. Stack J, Habetler T and Harley R. Fault Classification and fault signature production for rolling element bearings in electric machines. *IEEE Trans Ind Appl* 2004; 40(3): 735–739.
29. Zhao H, Zheng J, Xu J, et al. Fault diagnosis method based on principal component analysis and broad learning system. *IEEE Access* 2019; 7: 99263–99272.
30. Dong Y and Qin SJ. A novel dynamic PCA algorithm for dynamic data modeling and process monitoring. *J Process Control* 2018; 67: 1–11.
31. Jiang Q, Yan X and Huang B. Performance-driven distributed PCA process monitoring based on fault-relevant variable selection and Bayesian inference. *IEEE Trans Ind Electron* 2016; 63(1): 377–386.
32. Nectoux P, Gouriveau R, Medjaher K, et al. PRONOSTIA: an experimental platform for bearings accelerated degradation tests. In *IEEE international conference on prognostics and health management, PHM'12*, volume sur CD ROM, Denver, CO, USA, pp. 1–8. IEEE Catalog Number: CPF12PHM-CDR.