Applied Soft Computing Journal 89 (2020) 106119

Contents lists available at ScienceDirect

Applied Soft Computing Journal

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

Constructing a health indicator for roller bearings by using a stacked auto-encoder with an exponential function to eliminate concussion

Fan Xu [a], Zhelin Huang [a,b,*], Fangfang Yang [a], Dong Wang [c], Kwok Leung Tsui [a]

[a] School of Data Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong 990777, PR China
[b] Department of Statistics, School of Economics, Shenzhen University, Shenzhen 518061, PR China
[c] The State Key Laboratory of Mechanical Systems and Vibration, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 17 March 2019
Received in revised form 23 December 2019
Accepted 17 January 2020
Available online 23 January 2020

Keywords:
Roller bearings
Deep learning
Stacked auto-encoder
Health indicator
Exponent function

ABSTRACT

Most deep-learning models, especially stacked auto-encoders (SAEs), have been used in recent years for the diagnosis of faults in rotating machinery. However, very few studies have reported on health indicator (HI) construction by using SAEs in deep learning. SAEs have a good feature-extraction ability when several hidden layers are used to reconstruct the original input. In this study, we first introduce a method to reduce dependence on prior knowledge that is based on SAEs and enables extraction of the preliminary degradation trend from the bearing's frequency domain directly. Second, to construct the final HI and improve the monotonicity of the indicators, an exponential function is used to eliminate global severe vibration after an SAE has extracted the preliminary degradation trend. To prove the effect of our presented method, some other HI construction models, such as root mean square, kurtosis, approximate entropy, permutations entropy, empirical mode decomposition-singular value decomposition, K-means/K-medoids, and various time–frequency fusion indicators are used for comparison. Moreover, to prove that the exponential-function effect exceeds other severe vibration-eliminating methods, examples of the latter methods such as exponentially weighted moving-average and outlier detection are used for comparative analysis. Finally, the results shows that our proposed model is better than the above-mentioned existing models.

1. Introduction

Roller bearings ensure the reliable operation of a mechanical system and are some of the most commonly used and vulnerable mechanical parts in industry. Reducing the maintenance cost of such equipment is very important because it is known that increased bearing running-time leads to degradation of the bearing's performance [1–3]. To achieve a quantitative assessment of the bearing health status, the concept of a health indicator (HI) is invoked to describe and model the entire degradation procedure from the intact state to a series of different degraded states. Such HI-derived information can assist in predicting the remaining useful life for a bearing.

Many HI construction models have been developed. Shen et al. studied root-mean-square (RMS) analysis to determine the useful degradation characteristics of bearings [4]. Lei et al. also used RMS analysis and constructed an HI to evaluate the abrasion status of bearings [5]. Kosasih et al. used the RMS and kurtosis to detect bearing-status after a low-frequency filter had been applied to the original vibration signal. Tse et al. used various time–frequency indicators with principal component analysis (PCA) to build an HI construct model. Accordingly, time–frequency indicators, including RMS, kurtosis, and variance, were used to calculate the original vibrational signal. PCA was then used to reduce the dimension of the extracted feature vector, and the first principal component was selected as the extracted degradation trend and used to construct HI [6].

Yan et al. used approximate entropy (AE) and permutation entropy (PE) to build an HI and achieved good results [7,8]. Ye et al. also used PE to extract the vibrational signal fault-feature to diagnose faults [9]. Because the vibration signal has nonlinear and random complexity, these time–frequency statistical indicators cannot adaptively decompose the vibrational signal. However, empirical mode decomposition (EMD) can adaptively disintegrate the signal into a series of intrinsic mode functions (IMFs). Rai et al. devised a combined method based on EMD, singular-value decomposition (SVD), and K-medoids to construct an HI for bearing performance-assessment and showed that this model was superior to RMS, kurtosis, and K-means approaches [10]. However, even though these traditional HI construction models yielded good results, they also presented some challenges. First, prior knowledge is needed to understand the collected

[*] Corresponding author.
E-mail addresses: fanxu8@cityu.edu.hk (F. Xu), zhuang44@cityu.edu.hk (Z. Huang).

https://doi.org/10.1016/j.asoc.2020.106119
1568-4946/© 2020 Elsevier B.V. All rights reserved.

data, and abundant experiments are required to construct HI. However, this knowledge is not easy to obtain. Second, these HI models are usually produced for specific problems, so their utility depends on knowledge of the degradation process for a specific, potentially complex engineering problem, which is not always possible. Therefore, a feature-learning approach is needed to enable automatic HI construction without the requirement for prior knowledge.

Unlike traditional HI construction models, an SAE has a strong learning ability that enables it to extract potential features through several hidden layers and is composed of encoders and decoders to reconstruct the input data without any prior knowledge. Therefore, many researchers are gradually paying more attention to this technique and using SAEs to complete classification, monitoring, and forecasting tasks. SAEs have therefore been widely and successfully applied in image and speech recognition [11,12] and especially for fault diagnosis. Feng et al. used an SAE for roller bearing fault-diagnosis [13].

To strengthen the robustness of the SAE, a stacked de-noising auto-encoder (SDAE) [14,15] was proposed and used to de-noise the original vibrational signal and obtain a higher classification accuracy. An SDAE uses the destructed input to reconstruct the input data. Lu et al. used an SDAE for fault diagnosis in rotary machinery components, which revealed useful characteristics and showed that an SDAE can extract robust information via several hidden layers by iterative learning [16]. Xu et al. proposed an unsupervised method on the basis of an SDAE and Gath–Geva (GG) clustering to diagnose the different roller-bearing faults under various working conditions [17].

In addition, Saufi et al. used a method for bearing fault diagnosis based on the stacked sparse auto-encoder (SSAE), whereby fused evolution and resilient backpropagation were used to enhance the robustness capability of SSAE [18]. To extract the robust vibration signal with noise, Shen et al. proposed a method for rotating machinery, including gears and bearings, based on the contractive autoencoder (CAE). In this study, CAE was used to obtain the internal factors and potential robust features via a Jacobian matrix with Frobenius norm to penalize the hidden layer [19]. To help the SSAE obtain more useful potential features, Qi et al. used ensemble empirical mode decomposition (EEMD) and autoregression to preprocess the original vibration signal. The intrinsic mode function components (IMFs) from the EEMD were then chosen as the input of the SSAE for diagnosing faults [20].

However, all these basic SAEs and some improved and mutated SAE models are usually applied in fault diagnosis. These models combined with a variant of an SAE or an SAE itself, necessitates a multistep data preprocessing method. In contrast, only the fast Fourier transformation (FFT)–converted frequency signal is needed to extract useful fault characteristics in [13]. Moreover, very few reports focus on constructing an HI for bearings using SAEs. Therefore, we use an SAE to extract the useful feature information and construct an HI after FFT data preprocessing for roller bearings without prior knowledge being required.

Another problem is the fact that the preliminary extracted HI line clearly exhibits severe vibration during an SAE analysis, such as that displayed in Fig. 1(a). The constructed HI obtained from the SAE in Fig. 1(a) demonstrates that some noise and severe vibration is occurring. In this case, the extracted HI curve shows poor stability and monotonicity. Compared with Fig. 1(a), Fig. 1(d) shows that the extracted degradation trend is stable and smooth, and the monotonicity of the curve is significantly enhanced. The even growth of the curve in Fig. 1(d) is clear when the exponential function is used, corresponding to engineering requirements.

For further refinement, there are two common ways to remove the oscillations and improve monotonicity: (a) an exponential weighted moving average (EWMA) [21,22] and (b) outlier detection [23,24]. For EWMA, a key parameter such as the weighting coefficient must be preset depending on prior knowledge for its computation; the weighting coefficient therefore represents the EWMA of the historical measurement value. The closer the weighting coefficient value is to 1, the greater the weighting of the measurement of the previous value. The result of the extracted degradation when EWMA is used can be seen in Fig. 1(c), which shows obvious instability at 300–600 s (blue dotted elliptical area) instead of a smooth growth curve. Moreover, the EWMA method gives the current point a larger weight value, so if the current point oscillates and deviates greatly, the EWMA cannot obtain a poor smoothing effect.

In contrast, ideal outlier detection enables local outlier areas to be found first, such has O1 and O2 (red elliptical area) in Fig. 1(e). However, these detected areas are formed by some continuous outlier points, and yet these are sometimes absent, such as Fig. 1(b), in which several outlier points within the blue dotted ellipse are discrete and thus cannot produce an outlier area. Simultaneously, distinct from EWMA, the outlier detection model can only eliminate the vibration phenomenon locally, according to the outlier areas.

In this study, to overcome these weaknesses of EWMA and outlier detection, an exponential function is introduced that enables the elimination of shocks and the construction of an HI from a global perspective. Furthermore, the exponential function is monotonic and does not require a parameter to be set before its calculation. The exponential function uses the average from the starting point to the current point (the same weight is given from the starting point to the current point, as shown in Eq. (11)). This statistics not only extract the central tendency but also smooth the curve, thereby eliminating shocks and improving the monotonicity of the extracted HI curve. Overall, this study describes the construction of an HI for roller bearings based on an SAE that incorporates an exponential function.

The main contributions of this study are given below in the following points.

(1) Unlike various traditional HI construction models, this study uses an SAE to extract the preliminary HI of the roller bearings, as few reports have focused on HI construction. In addition, the proposed model reduces the dependence on prior knowledge for use of an SAE because only the frequency domain vibration obtained from FFT preprocessing is regarded as the input of SAE.

(2) To eliminate the global severe vibration and improve the monotonicity of the HI curve in this study, an exponential function is used. To prove that our proposed model is superior to other models, such as RMS, kurtosis, AE, PE, EMD-SVD-K-means/K-medoids, various time–frequency fusion indicators, EWMA, and outlier detection, we compare and analyze both qualitative and quantitative aspects of these other models with our own.

The rest of this paper is organized as follows. Section 2 provides the basic theory of SAE and the method procedures, and evaluation indicators are then used to assess the effects of various methods. In Section 3, the experiment platform and dataset are first detailed, and an SAE with an exponential function is then used to construct an HI and is compared with other methods. Conclusions are shown in Section 4.

2. Review of SAE and procedure of proposed method

2.1. Procedure of model presented

The three steps in the proposed method are: (1) data pretreatment; (2) HI construction; and (3) comparison and evaluation. More details are described as follows and in Fig. 2.

(1) Data pretreatment:

In [13,25], the authors used the amplitude matrix after FFT as the input of SAE for bearing fault diagnosis after FFT operation

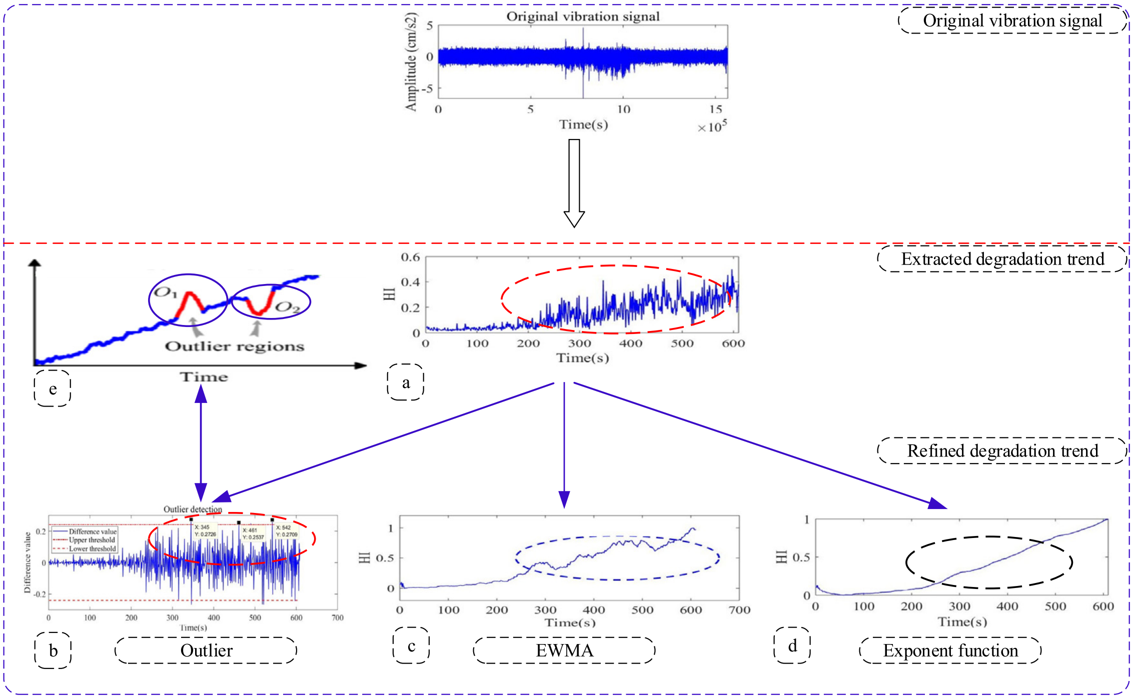

Fig. 1. Use of SAE to extract degradation trend with different data-smoothing methods. (a) Preliminary degradation trend extracted by SAE. (b) Use of outlier detection model to filter some outlier points. (c) Removal of severe vibration by use of EWMA. (d) Removal of the severe vibration by use of an exponential function. (e) Two examples of outlier areas.

because the frequency domain signal contains useful information. Inspired by this idea, we also use the amplitude matrix as the input of SAE to extract HI. The detailed data pretreatment procedure is shown in Fig. 3 and given as follows.

First, for a given a raw bearing vibration signal dataset $X = [X_1, X_2, \ldots, X_i, \ldots, X_N]$ $1 \le \mathrm{i} \le N$ (where $X_i = \left[x_i^1, x_i^2, x_i^j, \ldots x_i^M\right]^T$ $1 \le j \le M$). N is the total number of samples, M is the length of each sample, FFT transforms X from the time domain to frequency domain and get the Fourier coefficient matrix $\overline{X} = [\overline{X_1}, \overline{X_2}, \ldots, \overline{X_i}, \ldots, \overline{X_N}]$ $1 \le \mathrm{i} \le \mathrm{N}$ (where $\overline{X_i} = \left[\overline{x_i^1}, \overline{x_i^2}, \overline{x_i^j}, \ldots \overline{x_i^M}\right]^T$ $1 \le j \le M$ and $\overline{X}$ is a complex matrix).

Second, because the input of the SAE model is positive, the absolute amplitude matrix $\overline{|X|} = [|\overline{X_1}|, |\overline{X_2}|, \ldots, |\overline{X_i}| \ldots, |\overline{X_N}|]$ $1 \le \mathrm{i} \le \mathrm{N}$ (where $\overline{|X|}$ is the amplitude (Module length of complex number in the first step) of the Fourier coefficient matrix $\overline{X}$) is obtained from $\overline{X}$. For each SAE input sample $|\overline{X_i}| = \left[|\overline{x_i^1}|, |\overline{x_i^2}|, |\overline{x_i^j}|, \ldots |\overline{x_i^m}|\right]^T$ $1 \le j \le m, m = M/2$ (We uniformly use $|\overline{X}|$ to represent the SAE input). Noted that $\overline{X}$ is a complex matrix (Note that the complex coefficients obtained by Fourier transform are all conjugate complex numbers), hence the matrix $|\overline{X}|$ is a symmetric matrix. It means that the first half amplitude vector ranged in [1, m] is the same as the rest half amplitude vector ranged in [m + 1, M], here $m = M/2$. Therefore we use the first half amplitude matrix of $|\overline{X}|$ as the input of SAE for HI extraction. In addition, to generalize and unify the statistical distribution of each sample, all (samples $|\overline{X}|$) are normalized in range of [0, 1] before SAE training and testing. The maximum and minimum normalization method is used to normalize the input data $|\overline{X}|$. The details of the normalization steps are given below.

$$|\overline{X_i}| = |\overline{X_i}| - |\overline{X}|_{\min} \Big/ \left(|\overline{X}|_{\max} - |\overline{X}|_{\min}\right) \quad 1 \le \mathrm{i} \le \mathrm{N} \tag{1}$$

where $|\overline{X}|_{\max}$ and $|\overline{X}|_{\min}$ are the maximum and minimum of $|\overline{X}|$.

(2) HI construction:

After data preprocessing, SAE is used to construct the HI.

The SAE input layer connects $|\overline{X}|$ to the first hidden layer. In multiple hidden layers of SAE, the output of the previously hidden layer will be used as the input of the adjacent next hidden layer. Each hidden layer uses autoencoder (AE) to extract potential features h at the hidden layer. The purpose of AE is to use the encoder and decoder to reconstruct the input data $|\overline{X}|$ and extract features by using sigmoid function, the reconstruction error between the input $|\overline{X}|$ and output Z is iteratively reduced through the gradient descent method [26–28]. The detailed information about AE is shown in [29,30]. The backpropagation method is used to iteratively update the network parameters at each hidden layers. The extracted features h (at of) the last hidden layer are used as the input of the output layer for HI construction. Therefore, SAE includes three parts: (a) input layer, (b) several hidden layers are composed by the AE unit, (c) the output layer. The basic structure of SAE is shown in Fig. 4 and the basic structure of AE is shown in Fig. 5.

L represents the total number of hidden layers in Fig. 2. For the input layer, the number of input neural nodes is m because each input sample$|\overline{X_i}|$ is an m*1-dimension vector. For the hidden layer, the author in [17] used multiple hidden layer structures based on triangles (In the selection of the number of hidden layer nodes, the author in [17] adopted the structure of decreasing layer by layer (or in other words, triangular structure, e.g. [100, 50, 20, 10, 5, 1])) to effectively extract bearing fault features and reduce the data dimension. Inspired by the idea, the hidden layer also uses a triangular structure in this paper, that is, the number of neural nodes in the next adjacent hidden layer is half the current hidden layer. For the output layer, because each input sample $|\overline{X_i}|$ corresponds to a HI output, the number of neuron nodes at the output layer is set to 1. Therefore, $HI = [HI_1, HI_2, \ldots, HI_i, \ldots, HI_N]$ $1 \le i \le N$. The input data $|\overline{X}|$ was normalized before SAE training, hence the output range of HI

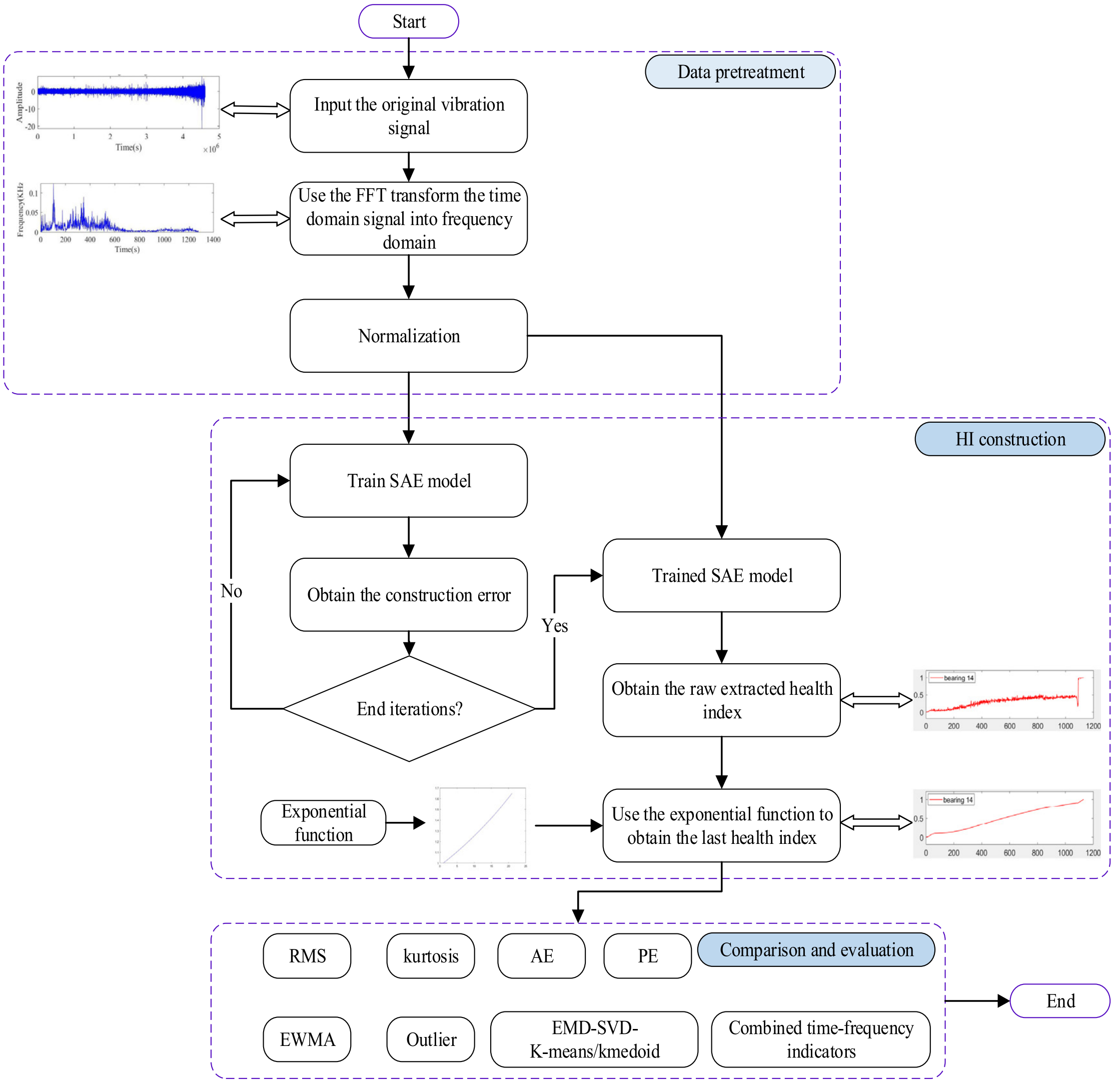
Start
Data pretreatment
Input the original vibration signal
Amplitude
Time(s)
Use the FFT transform the time domain signal into frequency domain
Frequency/KHz
Normalization
HI construction
Train SAE model
Obtain the construction error
End iterations?
No
Yes
Trained SAE model
Obtain the raw extracted health index
bearing 14
Exponential function
Use the exponential function to obtain the last health index
Comparison and evaluation
RMS
kurtosis
AE
PE
EWMA
Outlier
EMD-SVD-K-means/kmedoid
Combined time-frequency indicators
End

Fig. 2. Detailed process step to construct HI and compare it with other models.

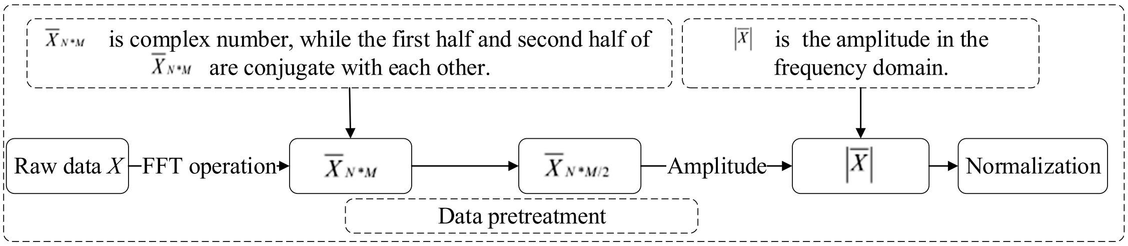
$\overline{X}_{N*M}$ is complex number, while the first half and second half of $\overline{X}_{N*M}$ are conjugate with each other.
$|\overline{X}|$ is the amplitude in the frequency domain.
Raw data X
FFT operation
$\overline{X}_{N*M}$
$\overline{X}_{N*M/2}$
Amplitude
$|\overline{X}|$
Normalization
Data pretreatment

Fig. 3. The detailed calculation procedure of data pretreatment.

should be [0, 1]. Due to the output range of the Sigmoid function is [0, 1]. the AE unit and function at the output layer uses a sigmoid function to extract HI. The AE calculation is given as follows.

For a given input dataset $|\overline{X}|$, the encoder is used to mapping the input vector $|\overline{X}|$ into a hidden representation $h = [h_1, h_2, h_3, \ldots, h_i, \ldots, h_N]$ (here $h_i = [h_i^1, h_i^2, h_i^3, \ldots, h_i^d]^T$) by

$$h = S\left(W^1 \left|\overline{X}\right| + b^1\right) \ and\ S\left(p\right) = 1/1 + e^{-p} \qquad (2)$$

where S is the sigmoid function, e denotes the exponent function. Each input sample $\left|\overline{X_i}\right|$ and h_i are the m-dimension and d-dimension vector, d is the number of neural nodes at each hidden layer. W^1 is the d^*m-dimension weight matrix. b^1 is the bias vector with d-dimension. For example, for each input sample $\left|\overline{X_i}\right|$ and h_i are the m-dimension and d-dimension vector, d is the number of neural nodes at each hidden layer. W^1 is the d^*m-dimension weight matrix. b^1 is the bias vector with d-dimension. For example, for each input sample $\left|\overline{X_i}\right| = \left[\left|\overline{x_i^1}\right|, \left|\overline{x_i^2}\right|, \left|\overline{x_i^j}\right|, \ldots \left|\overline{x_i^m}\right|\right]^T 1 \leq j \leq m, m = M/2$, according to the matrix calculation rule, the $d * m$ -dimension matrix W^1 and m^*1-dimension matrix $\left|\overline{X_i}\right|$ are multiplied by Eq. (2) to obtain the d^*1-dimension matrix h_i. Hence the dimension of $W^1 \left|\overline{X_i}\right| + b^1$ is d^*1. Hence for each input sample $\left|\overline{X_i}\right|$, S function ($S\left(p\right) = 1/1 + e^{-p}$) uses each element of

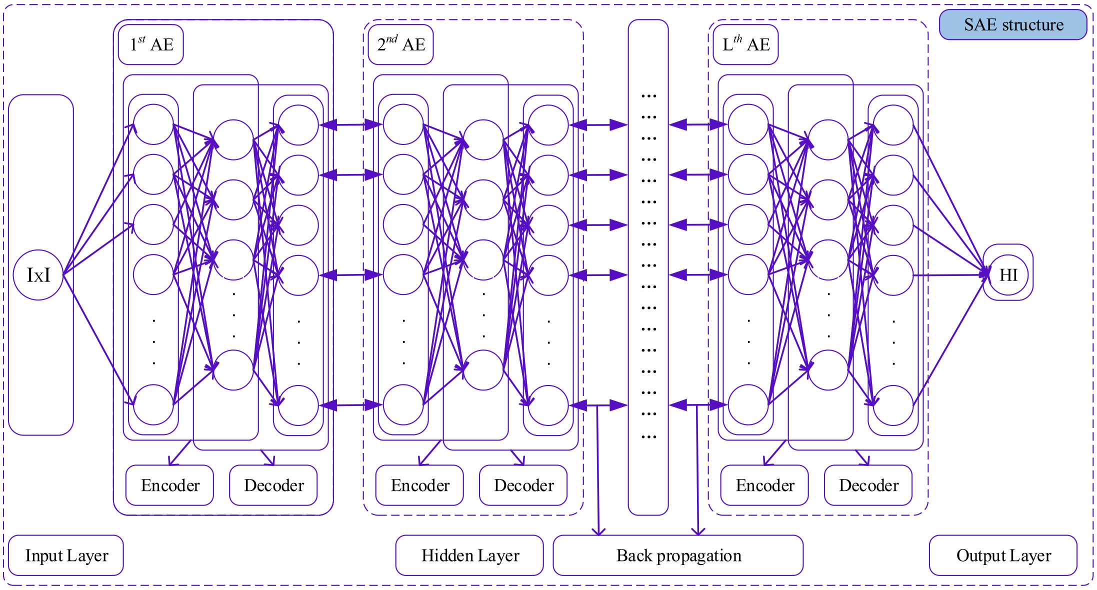

Fig. 4. Structure of SAE.

$W^1\left|\overline{X_i}\right| + b^1$ and get the extracted hidden represent h_i with d*1-dimension. Because there are N input samples $\left|\overline{X_i}\right|$ to calculate h_i. Hence the dimension of h is d*N.

In Fig. 4, decoder is used to mapping the hidden representation h into a reconstruction output matrix $\mathrm{Z} = [Z_1, Z_2, Z_3, \ldots Z_i, \ldots Z_N]$ $1 \le \mathrm{i} \le N$ (Each $\mathrm{Z}_i = \left[z_1, z_2, z_3, \ldots z_j, \ldots z_m\right]$ $1 \le j \le m$) by the following equation. (The calculation process is similar to the encoder, it not be explained here again).

$$Z = S\left(W^2 h + b^2\right) \; and \; S\left(p\right) = 1/1 + e^{-p} \tag{3}$$

where Z is a matrix with m-dimension. W^2 is the m*d-dimension weight matrix. b^2 is the bias vector with m-dimension.

Therefore, the reconstruction error J_l between the input $\overline{|X|}^l$ and the output Z^l at the lth hidden layer is defined as the function below (Eq. (4)):

$$J_l = \left[\frac{1}{2}\sum\left(\left\|\left(\overline{|X|}^l - Z^l\right)\right\|^2\right)\right] \tag{4}$$

In the entire encoding and decoding procedure, J_l thus regarded as the optimization function that is utilized to find the optimization and update the weight matrix $W\left(l\right) = \left[W^1, W^2\right]$ and bias item $b\left(l\right) = \left[b^1, b^2\right]$ according to gradient descent and backward propagation algorithm. Backpropagation starts from the last hidden layer and ends at the first hidden layer. The partial derivatives ρ^l of each parameter at each hidden layer is calculated to obtain the gradient information. ρ^l propagates the gradient information of the former adjacent hidden layer to the previously hidden layers to update the parameters. This process is repeated until the first hidden layer.

Therefore, J_l is used to compute the partial derivative ρ^l of the output Z^l at the lth hidden layer:

$$\begin{aligned}\rho^l &= \frac{\partial J_l}{\partial Z^l} = \frac{\partial\left[\frac{1}{2}\sum\left(\left\|\left(\overline{|X|}^l - Z^l\right)\right\|^2\right)\right]_l}{\partial Z^l} \\ &= -\left(\overline{|X|}^l - Z^l\right)\partial Z^l = -\left(\overline{|X|}^l - Z^l\right)S'\left(Z^l\right)\end{aligned} \tag{5}$$

where $S\left(Z^l\right) = S\left(W_2 h + b_2\right) = S\left(W_2(W_1\overline{|X|}^l + b_1) + b_2\right) = S\left(W\overline{|X|}^l + b\right)$. The term S' is the derivative function of S, therefore, $S'\left(Z^l\right) = S\left(Z^l\right)\left(1 - S\left(Z^l\right)\right)$ is the gradient information of the lth hidden layer.

Similarly, J_l is also used to calculate the partial derivative of the output Z^{l-1} at the $(l - 1)$th hidden layer as follows

$$\begin{aligned}\rho^{l-1} &= \frac{\partial J_l}{\partial Z^{l-1}} = \frac{\partial J_l}{\partial Z^l} \bullet \frac{\partial Z^l}{\partial Z^{l-1}} = \rho^l \bullet \frac{\partial\left[\sum S(W\overline{|X|}^l + b)\right]^l}{\partial Z^{l-1}} \\ &= \rho^l \bullet \frac{\partial\left[\sum S(WZ^{l-1} + b)\right]_l}{\partial Z^{l-1}} \\ &= \rho^l \bullet \frac{\partial\left[\sum S(W_2(W_1\overline{|X|}^{l-1} + b_2) + b_1)\right]_l}{\partial Z^{l-1}} = \rho^l \sum WS'\left(Z^{l-1}\right)\end{aligned} \tag{6}$$

Because the output Z^{l-1} is regarded as the input $\overline{|X|}^l$ at the lth hidden layer, hence $\overline{|X|}^l = Z^{l-1}$. The reconstructed error J_{l-1} is reduced according to the gradient information $S'\left(Z^l\right)$, hence the parameter W and the bias item b are updated by.

$$W = W - \eta\frac{\partial}{\partial W}J_{l-1} \tag{7}$$

$$b = b - \eta\frac{\partial}{\partial b}J_{l-1} \tag{8}$$

where η is the learning rate. $\frac{\partial}{\partial W}J_{l-1}$ and $\frac{\partial}{\partial b}J_{l-1}$ are the partial derivatives of W and b at l—1th hidden layer, respectively, and the calculation by the following equations. (The detailed derivation calculation process is similar to Eq. (6))

$$\frac{\partial}{\partial W}J_{l-1} = Z^{l-1}\rho^l \tag{9}$$

$$\frac{\partial}{\partial b}J_{l-1} = \rho^l \tag{10}$$

To eliminate severe vibration in the extracted HI curve, the exponential function is used to smooth and improve the monotonicity of the extracted HI curve, according to the following equation we denote as the exponent health indicator (EHI) (Eq. (11)):

$$EHI_i = e^{\left(\sum_{j=1}^{i} HI_j/i\right)} \; 1 \le \mathrm{i} \le N \tag{11}$$

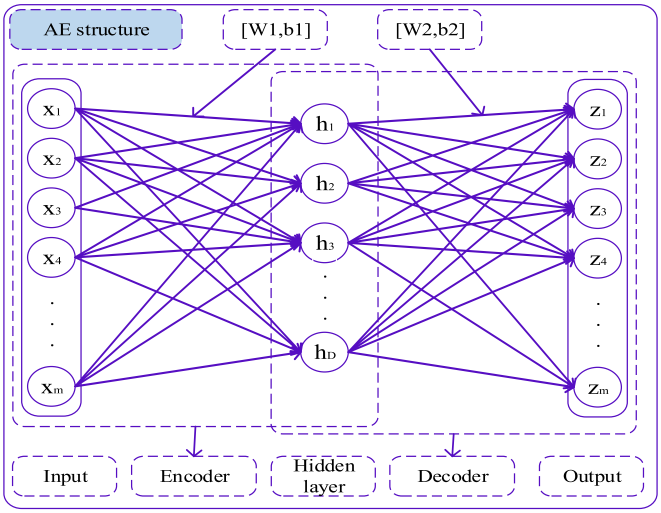

Fig. 5. Network structure of AE.

Table 1
Detailed information of vibrational signal for different conditions.

Operation condition	Bearing no.	Speed (rpm)	Load (N)	Total sample numbers	Length
Condition 1	Bearing 11	1800	4000	2803	2560
	Bearing 12	1800	4000	871	2560
	Bearing 13	1800	4000	1802	2560
	Bearing 14	1800	4000	1139	2560
	Bearing 15	1800	4000	2302	2560
	Bearing 16	1800	4000	2302	2560
	Bearing 17	1800	4000	1502	2560
Condition 2	Bearing 21	1650	4200	911	2560
	Bearing 22	1650	4200	797	2560
	Bearing 23	1650	4200	1202	2560
	Bearing 24	1650	4200	612	2560
	Bearing 25	1650	4200	2002	2560
	Bearing 26	1650	4200	572	2560
	Bearing 27	1650	4200	172	2560
Condition 3	Bearing 31	1500	5000	515	2560
	Bearing 32	1500	5000	1637	2560
	Bearing 33	1500	5000	352	2560

where e denotes the exponent function.

Eq. (11) use the mean value from the first value HI_1 to the current jth value HI_j to calculate the final ith EHI value EHI_i. If the current HI_j value increases or decreases sharply, it means that the HI curve exhibits a significant oscillation phenomenon at HI_j point. The average HI value (EHI_i) from the first value HI_1 to the current jth HI value HI_j is used to replace the HI_j when it increases dramatically, thus it can weaken the sharp increase or decrease of the HI value. Moreover, the exponential function is a monotonically increasing function.

Therefore, using Eq. (11) makes the construction EHI curve smoother and enhances the monotonicity.

(3) Comparison: To verify the superiority of our method, its performance is compared with that of some other HI construction models such as EMD-SVD-K-means/K-medoids [10], RMS [4,24], kurtosis [5,31–33], AE [7], PE [8,9], and various time–frequency fusion indicators [6]. Also, two evaluation indicators—Correlation ($Corr$) and Monotonicity (Mon) [34,35]—are used to evaluate the performance of various models.

(a) $Corr$ represents the degree of linear correlation between the HI value and the sampling time t. As time increases, the ideal HI curve should also gradually increase. If obvious oscillation occurs, the linear correlation with time is weak. Therefore, the closer the $Corr$ absolute value is to 1, the stronger the linear correlation with the sampling time, and vice versa. $Corr$ is calculated as follows (Eq. (12)):

$$Corr(EHI) = \frac{\left|\sum_{i=1}^{N}\left(EHI_i - \overline{EHI}\right)\left(i - \left(\sum_{i=1}^{N} i/N\right)\right)\right|}{\sqrt{\sum_{i=1}^{N}\left(EHI_i - \overline{EHI}\right)^2 \sum_{i=1}^{N}\left(i - \left(\sum_{i=1}^{N} i/N\right)\right)^2}} \quad 1 \leq \mathrm{i} \leq N \tag{12}$$

where EHI_i denotes the ith EHI value, $\overline{EHI}$ denotes the mean value of all the EHI values.

(b) Mon represents the monotonicity of the curve [34,35] and uses the positive and negative numbers of the difference values of the two adjacent EHI values to evaluate the monotonicity of the EHI curve. If the total number of positive values is more than the total number of negative values, this shows that the monotonicity is rising, and vice versa. If the curve displays obvious noise and oscillation, the total number of positive values is close to the total number of negative values, thus the Mon value is close to 0 and the monotonicity of the HI curve is not good. Therefore, the closer the Mon value is to 1, the better the monotonicity of the HI curve. The calculation of Mon is as follows:

$$\begin{aligned} Mon &= \left|\frac{Number\ of\ dF > 0}{N-1} - \frac{Number\ of\ dF < 0}{N-1}\right| \mathrm{dF} \\ &= \frac{EHI_{i+1} - EHI_i}{i} \quad 1 \leq \mathrm{i} \leq N \end{aligned} \tag{13}$$

where dF indicates the difference value of any two adjacent EHI values.

3. Experimental setup and validation

3.1. Experimental data and data pretreatment

The experimental data come from the PROGNOSTIC device [36], and the structure of this experimental platform is shown in Fig. 6 [36]. At the early stage of the experiment, the roller bearings have no fault. When the bearing speed and the radial load are added, the accelerated life-test can be performed to determine the overall degradation progress of the bearing over several hours. Two sensors in the experimental platform are used to collect the horizontal and vertical vibration information of the bearing. In this study, the vibration value in the horizontal direction is used. The dataset is an accelerated degradation test of the bearing under various operating conditions. The measured data include a total of three working conditions for three bearings, as follows: for condition 1, the load is 4000 N and the speed is 1800 rpm; for condition 2, the load is 4200 N and the speed is 1650 rpm; for condition 3, the load is 5000 N and the speed is 1500 rpm. Seven bearings (11–17 and 21–27) are used to collect experimental data for conditions 1 and 2, respectively, and for condition 3, three bearings (31–33) are used to collect data. The vibrational signal of the bearing in the experimental platform collects one data sample at a frequency of 25.6 Hz every 10 s. The recording time of each sample is 0.1 s, and each data-sample has a length of 2560. Other detailed related information of the data is given in Table 1.

In Table 1, "sample numbers" are the total number of samples taken for each bearing. The original time-domain vibration signal wave figures for bearing 24 and bearing 25 and its FTT decomposition result are shown in Fig. 6. Here, only one sample from each of bearing 24 and bearing 25 is used to show the FTT result. For bearing 24, we use the no. 25 sample (612 samples in total; randomly selected because samples from the same bearing are

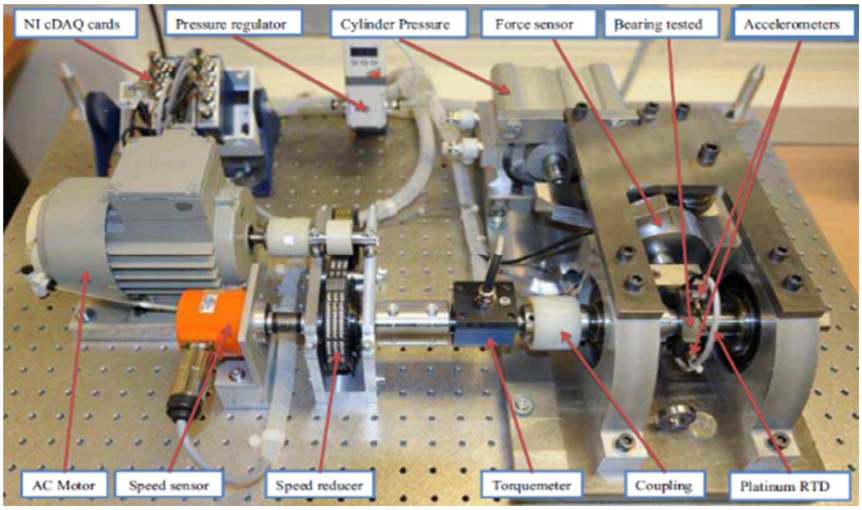

Fig. 6. Overview of prognostic experimental device.

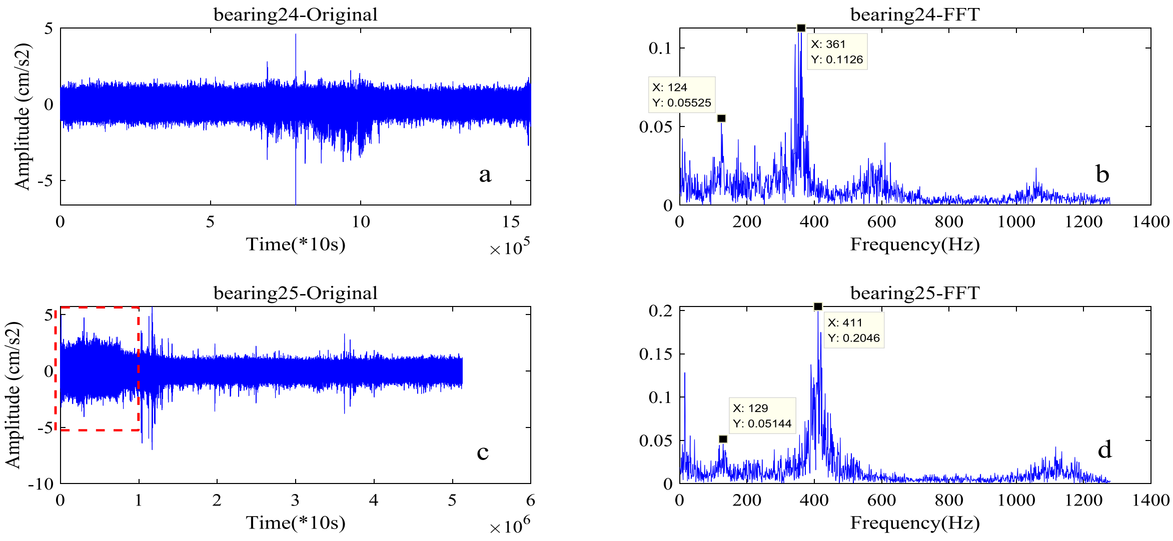

Fig. 7. Original time–frequency domain signal wave for bearing 24 and bearing 25, with (a) original vibration signal for bearing 24; (b) FFT result for bearing 24; (c) original vibration signal for bearing 25; (d) FFT result bearing 25; '10 s' denotes sampling interval.

quite similar) to subject to FFT transformation, thus the corresponding sampling time is between 240 s to 250 s. For bearing 25, we use the no. 125 sample, thus the corresponding sampling time is between 1240 s until 1250 s.

Fig. 7 shows that the amplitude of the vibration of bearing 24 is not changed at the early stage, whereas the amplitude of the vibration of bearing 25 decreases significantly in the early stage (red-dotted area) and then reaches a stable state. This result demonstrates the difficulty of judging the relevant information of the bearing with the naked eye and prior knowledge from the time-domain signal.

We therefore use FFT to transfer the time-domain vibration signal to the frequency domain. As shown in Fig. 7(b) and (d), the bearing 24 and bearing 25 frequency bands are focused mainly on 0–500 Hz, particularly 124 and 129 Hz because they are close to 128 Hz. Given that 128 is ∼ four times 25, 361 is ∼14 times 25.6, and 411 is ∼16 times 25.6, this indicates that the frequency domain signal contains useful information. Therefore, FFT is used in the first step at the data-preprocessing stage. After FFT operation, all datasets are normalized into [0, 1] for HI construction using an SAE with an exponential function.

3.2. HI construction using SAE with exponential function

In this section, SAE is first used to extract a preliminary HI, and an exponential function is used to improve the smoothness and monotonicity. Before training and testing, some key parameters should be preset, such as the learning rate and the network structure at the hidden layer in the SAE. These are detailed below.

(a) If the learning rate is too slow, the convergence rate will be slow. If the learning rate is too fast, it will hinder convergence. We take a bearing dataset as an example to finish the performance analysis of the SAE network under different learning rates by observing the convergence state of the reconstruction error.

(b) For the SAE network structure, the authors in reference [17] adopted the triangular structure at the hidden layer in the SAE because it enabled the extraction of useful fault-features and directly and effectively reduced the number of dimensions to two or three. Thus, we also use this approach, meaning that the number of neural nodes at the next hidden layer is half that of the previously hidden layer. The coefficient matrix for each sample is symmetrical after the FFT operation, and each row of the coefficient matrix denotes one sample. Thus, the first half of the coefficient matrix for each sample is regarded as the

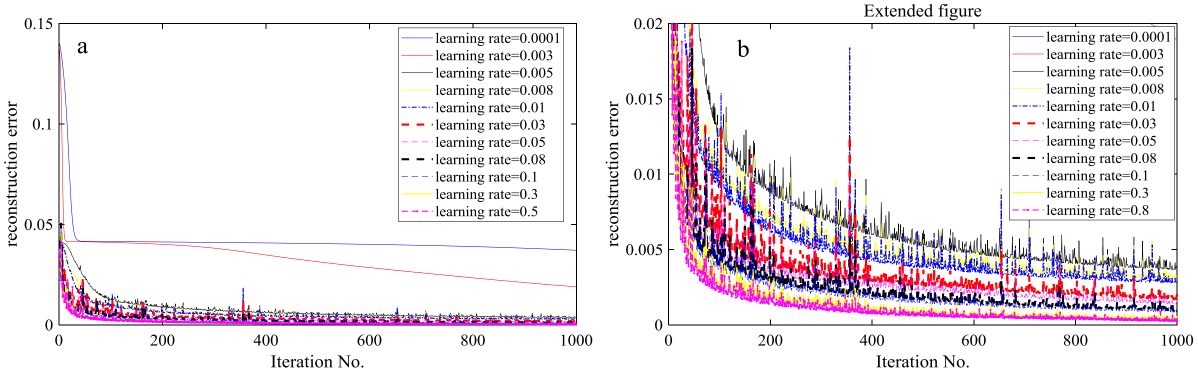

Fig. 8. Result of reconstruction error at various learning rates, where (b) is extended version of figure (a).

SAE input. Because one sample has an HI value, the neural node number at the output layer is thus set as 1. For SAE training and testing, the dataset in Table 1 is divided into training data and test data.

Here, we take bearing 24 as an example to construct an HI by using an SAE and examine the performance of the SAE network at various learning rates. Condition 2 has seven bearings (21–27), so if bearing 24 is selected as the testing data, the remaining six bearing datasets (for 21, 22, 23, 25, 26, and 27) are selected as the training data. Because the length of each sample for bearing 24 is 2560, the first half of the coefficient matrix after FFT operation is used as the SAE input and its input size is set to 1280, and the neural nodes of triangular structure at the hidden layer are set as [640, 320, 160, 80, 40, 20, and 10], meaning that seven hidden layers (seven AE units) are used in this study.

The actual target output value is the degradation percentage of the *i*th sample at time *t* [35]. For instance, supposing that the full time of bearing 24 is 5,120,000 s, then the degradation percentage is 0.6 when the *i*th sample is collected for 3,072,000 s and the output size is set as 1. The maximum iteration number is temporarily set as 1000, the learning rate is set as [0.0001, 0.003, 0.0005, 0.01, 0.03, 0.05, 0.1, 0.5, and 0.5], and the result of the reconstruction error under different learning rates is shown in Fig. 8.

(1) From Fig. 8(a), it can be seen that the reconstruction error values when the learning rate is 0.001 and 0.003 are higher than those at other learning rates. This is because the convergence rate is slow when the learning rate is too small. Therefore, we do not consider 0.001 and 0.003 in this study.

(2) Fig. 8(b) shows that all reconstruction error curves converge faster as the learning rate increases. For example, the reconstruction error curve sharply decreases in the first 200 cycles and slowly decreases from 200 cycles with learning rates of 0.1, 0.3, and 0.8. With learning rates of 0.0001, 0.003, and 0.005, the corresponding convergence speeds of the reconstruction error curve are slower than 0.1, 0.3, and 0.8, respectively. In contrast, with learning rates of 0.1, 0.3, and 0.8, the reconstruction error curve sharply decreases in the first 400 cycles but slowly decreases after 400 cycles. However, all reconstruction error curves in Fig. 8(b) slowly decrease starting from 600 cycles, which indicates that the reconstruction error decreases slowly because the number of cycles does not increase sharply. In addition, the reconstruction errors are similar after 600 cycles when the learning rate is greater than 0.005. With respect to the convergence speed and the reconstruction error value, the learning rate is 0.01 and the maximum iteration number at each hidden layer is chosen as 1000.

The SAE with an exponential function used to smooth the HI curve for bearing 24 is shown in Fig. 8. *Corr* and *Mon* index results are also shown in Table 2. In Fig. 9(a), SAE-HI represents the initial HI extracted directly by SAE, whereas SAE-EHI represents the extracted EHI after use of the exponential function defined in Eq. (11). The SAE-HI curve significantly increases from the 179th data point (HI = 0.03864) to the 180th data point (HI = 0.2306), but the SAE-EHI curve does not. The average value in Eq. (11) is used in SAE-EHI to calculate the EHI value (0.06424). Thus, the exponential HI average value that is calculated by the first 180 HI values can significantly weaken the sharp increase or decrease of the HI curve.

Table 2
Result of *Corr* and *Mon* for bearing 24.

Model	*Corr*	*Mon*
SAE-HI	0.8325	0.0049
SAE-EHI	0.9670	0.6059

It can be seen with the naked eye that the SAE-EHI curve is smoother than the SAE-HI curve. The different values of SAE-HI and SAE-EHI in Eq. (10) are also shown in Fig. 9(b) and (c). In Fig. 9(c), few difference values of SAE-EHI in the red dashed area are less than zero compared with the number of these in SAE-HI. However, Fig. 9(b) shows that the number of difference values of SAE-HI that are more than 0 ($dF > 0$) is similar to when the number is lower than 0 ($dF < 0$). Thus, the *Mon* absolute value (0.0049) of SAE-HI in Table 2. In contrast, the *Mon* value of SAE-EHI in Table 2 (0.6059) is significantly higher than that of SAE-HI (0.0049) because the SAE-HI curve in Fig. 9(a) has significant oscillation. Table 2 shows that the *Corr* absolute value (0.8325) of SAE-HI is also smaller than the SAE-EHI value (0.9670). These results show that the exponential function operates effectively to eliminate oscillation and improve the monotonicity of the curve.

3.3. Comparison with other models

To verify the performance of our proposed method, we compared some other construct HI models, such as EMD-SVD-K-means/K-medoids [7], RMS [4,6,23], kurtosis [5,6,24,25,31], AE [7], PE [8,9], and various time–frequency indicators fusion models in reference [6]. We again take bearing 24 as an example, which means that it is the testing dataset and that the remaining condition 2 datasets are the training data. Therefore, each bearing 24 original vibrational signal sample is used to calculate the HI value by a different model (see reference [6]).

(1) For RMS, kurtosis, AE, and PE, we take RMS for an example. For a given bearing-24 dataset with an original vibration signals dataset $X = [X_1, X_2, \ldots, X_i, \ldots, X_N]$ $1 \leq \mathrm{i} \leq \mathrm{N}$, the HI calculation by using RMS is given as follows (Eq. (11)):

$$\mathrm{HI}_{i-RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} X_i^2} \tag{14}$$

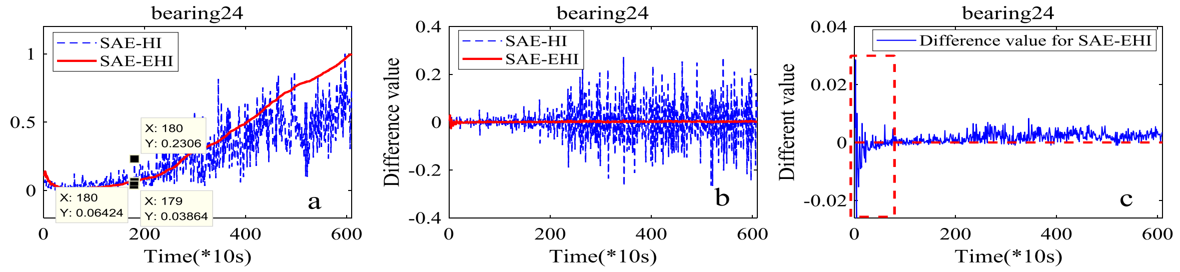

Fig. 9. Extracted HI and EHI and its corresponding difference value obtained by using SAE. (a) SAE-HI and SAE-EHI. (b) Difference value for SAE-HI and SAE-EHI. (c) Difference value of SAE-EHI in (b).

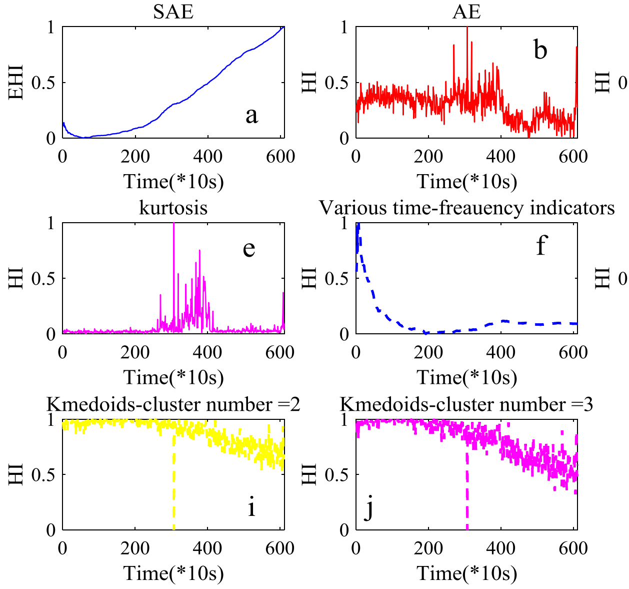

Fig. 10. Extracted HI via various models, where "2" and "3" denote the cluster number, which is set as 2 and 3 in K-means and K-medoids.

The HI construction for kurtosis, AE, and PE, is similar to RMS. Each sample corresponds to an HI value, and some parameters in AE and PE should be preset before HI construction and calculation, as described below.

(a) AE: Two key parameters of AE should be set before computation: similar tolerance and embedded dimension. The larger the embedded dimension, the more useful the information included in AE, although this unavoidably increases the computation load. The embedded dimension is often fixed at 2, as suggested in reference [7]. A similar tolerance is usually set at 0.1—0.25 *SD, where SD represents the standard deviation. We use 0.15* SD in this study.

(b) PE: The PE model has two important parameters: embedded dimension (*mm*) and time delay. In reference [37], the authors advise that the *mm* should be in the range of 3–7. If *mm* < 3, PE cannot accurately detect the tiny change in the signal; if *mm* > 8, the corresponding calculation efficiency is poor because the reconstruction of phase space will homogenize the vibration signals. Therefore, the embedded dimension *mm* is set at 5. Time-delay has a minimal influence on PE [8,9,37], and thus the time delay is set as 1.

(2) For EMD-SVD-K-means/K-medoids, EMD is first used to decompose each of bearing 24's original vibrational signal samples into IMFs. Second, the first two IMFs are selected to compute the single value via SVD. Finally, two single values (SV1-SV2) are taken as the input of K-means/K-medoids to construct HI. The detailed calculation steps are given in reference [10]. In addition, the cluster number in K-means/K-medoids is selected as 2 ($k = 2$) or 3 ($k = 3$), because bearings usually have two or three degenerate forms: normal and abnormal, or normal, slightly abraded, and severely abraded. Finally, the HI is calculated by the distance between each sample and each cluster center-point. The detailed procedure for this is given in reference [10].

(3) For the various time–frequency indicators fusion model, 18 time–frequency indicators in reference [6] are used to construct an HI by using bearing 24's original vibrational signal sample. Then, PCA is used to reduce the dimension of the 18 time–frequency indicators, and the first PC (PC1) is selected as the extracted HI.

The extracted HI from these models and our proposed method are shown in Fig. 10, and the *Mon* and *Corr* results are shown in Table 3.

(1) In Fig. 10, noise and oscillation are obvious in (b–d) and (g–j), but not in (a) and (f). These noises and oscillations seriously affect and easily submerge the smoothness of the HI curve, particularly in (e).

(2) In Fig. 10(g–j), all HI curves are very similar to each other under different cluster numbers, following use of EMD-SVD-K-means/K-medoids. Moreover, the *Mon* and *Corr* results of EMD-SVD-K-means/K-medoids in Table 3 are similar, and the *Mon* value in K-means (2) is the same as that in K-medoids (2). This result indicates that the preset cluster number in the presented model in reference [10] does not have a significant effect on the HI construct. However, these curves have a naked-eye straight line at the early stages in Fig. 10(g–j). No up-and-down trend is seen in EMD-SVD-K-means/K-medoids. All these HI curves, except that from our proposed method, sharply increase and

Table 3
Result of *Corr* and *Mon* for bearing 24 by using various models, where "2" and "3" denote cluster number setting in K-means and K-medoids.

Model	RMS	kurtosis	AE	PE	K-means (2)	K-means (3)	K-medoids (2)	K-medoids (3)	Ref. [6]	SAE-EHI
Corr	−0.6503	0.1454	0.5167	0.9162	−0.7717	0.8431	−0.7715	−0.8422	−0.4219	0.9670
Mon	0.0115	0.0115	0.0016	0.0049	0.0213	0.0115	0.0213	0.0049	0.0507	0.6059

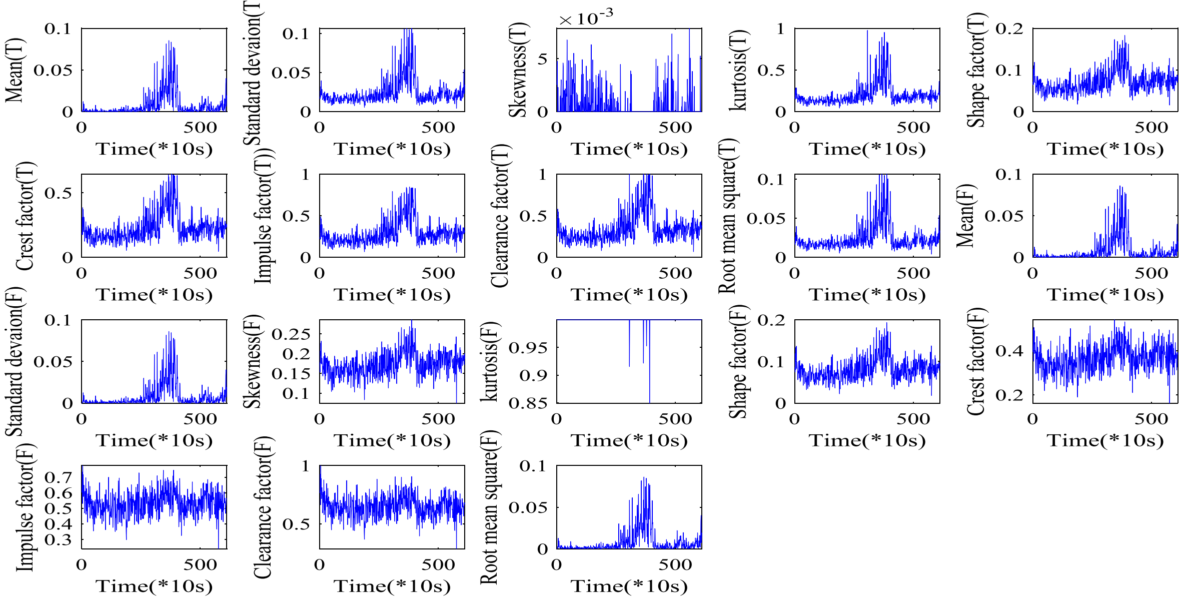

Fig. 11. Eighteen time–frequency indicators in reference [6] are calculated by using bearing 24's original vibrational signal, where 'T' denotes time-domain, and 'F' denotes frequency-domain (where FFT is used to transfer time-domain signal into frequency domain).

decrease, rather than smoothly increasing and decreasing as per Fig. 10(a) and (c).

(3) In Fig. (a) and Fig. 10(c), the two HI curves are smooth. However, oscillation is obvious in Fig. 10(c). Below, bearing 24 is used to illustrate why this is the case. The 18 time–frequency indicators that are calculated from the original signal of bearing 24 in reference [6] are shown in Fig. 11.

(a) In Fig. 11, 'T' represents the time domain; 'F' represents the frequency domain, and all curve shapes are very similar except for skewness (T), skewness (F), kurtosis (F), shape factor (F), crest factor (F), and impulse factor (F). PCA is used to obtain the first principle component (PC). PCA uses the 18 indicators and covariance matrix to calculate the Eigen matrix in PCA and finally calculates the contribution rate of each indicator via the Eigen matrix. The contribution rate is the ratio of the eigenvalue of each PC in the Eigen matrix. If the eigenvalue of the PC is larger, the cumulative contribution rate is larger, and the amount of information corresponding to the original data is larger.

(b) Here, we take the RMS(T) indicator as an example. Fig. 11(a) shows the covariance matrix result between the RMS (T) and other indicators. The RMS(T) and kurtosis(F) (the 13th point in Fig. 11(a)) have the largest absolute value of covariance, and Fig. 12 also shows that the RMS (T) and kurtosis (F) differ significantly. If the covariance value is negative, the two indicators are mutually exclusive. The larger the covariance absolute value, the stronger the repulsive effect between the two indicators. When using PCA calculations, the relationship between the RMS (T) and other indicators, like skewness (T), skewness (F), kurtosis (F), shape factor (F), crest factor (F), impulse factor (F), and clearance factor, is mutually exclusive.

Therefore, the PC curve that is chosen as the final HI curve generated by the fusion of indicators (such as RMS (T) and kurtosis (F)) have a mutually exclusive effect. The RMS (T) curve in Fig. 10 has a slight downward trend in the early stage and then begins to rise gradually, whereas the final PC curve has a significant downward trend in the early stage and has a slight upward trend because of the mutual exclusion that exists in the 18 indicators (such as RMS (T) and kurtosis (F)). Therefore, the first PC (PC1) has a slight upward trend after a sharp decline, as seen in Fig. 12(c). Note that the final HI curve is PC1 because the eigenvalue of PC1 is higher than PC2 in Fig. 12(b). Compared with PC2-HI in Fig. 12(d), the PC1-HI curve in Fig. 12(c) oscillates more obviously (as seen in the dotted red area), which indicates that the mutual exclusion is more obvious. Violent oscillation occurs, and the monotonicity of the HI curve is weakened.

3.4. Comparison with EWMA and outlier detection

To prove that the capability of the exponential function exceeds that of EWMA and outlier detection, *Corr* and *Mon* are also used in this section. After the SAE has extracted the preliminary degradation trend, EWMA and outlier detection are considered to smooth the extracted curves. The corresponding result is shown in Fig. 13.

(1) EWMA: a key parameter such as the weighting coefficient parameter λ must be preset depending on manual experience for its computation [38–47]. The HI construction of EWMA is $HI_i\,(EWMA) = \lambda HI_{i-1} + (1-\lambda) HI_i \;\; 1 \leq i \leq N$. λ represents the EWMA of the historical measurement value. The closer the weighting coefficient value is to 1, the greater the weighting of the measurement of the previous value [38–47]. Therefore, we set the weighting coefficient parameter $\lambda = [0.01, 0.03, 0.05, 0.08, 0.1, 0.2, 0.3, 0.4, 0.5]$ to compare with SAE-EHI. Fig. 13(a) shows a severe vibration in SAE-EWMA compared with SAE-EHI. The HI curve (SAE-EWMA) has obvious oscillations around the 300th data point in the red rectangular area; it does not always rise gradually but rises first and then falls. In addition, the parameter λ has a poor effect on curve smoothing. All SAE-EWMA curves

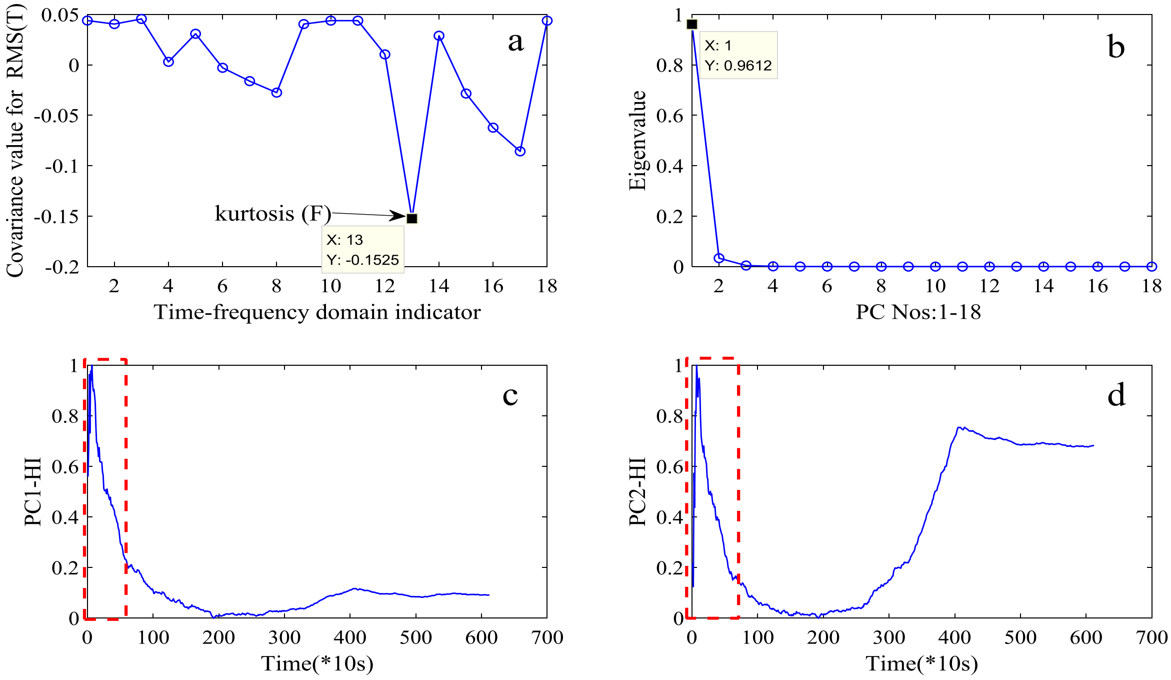

Fig. 12. Results generated by PCA. (a) Covariance result for RMS(T) and other indicators. (b) Eigenvalue for each PC in PCA. (c) First PC. (d) Second PC.

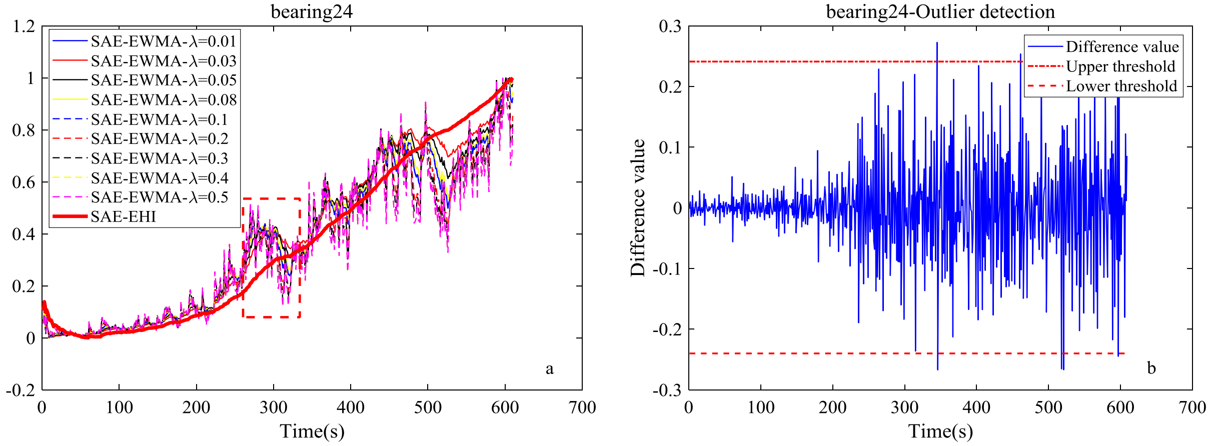

Fig. 13. Using EWMA and outlier detection to eliminate shocks for SAE-HI of bearing 24: (a) EWMA; (b) Outlier detection.

are similar to each other under different parameters λ, hence we choose $\lambda = 0.05$ in this paper. Moreover, EWMA needs to rely on manual experience to set the parameter λ. For SAE-EHI, Eq. (11) use the mean value from the first value HI_1 to the current jth value HI_j to calculate the final ith EHI value EHI_i. For example, the 300th EHI value is calculated by the mean value which is from the first value HI_1 to the 300th value HI_{300}. Thus, the exponential average value that is calculated by the first 300 HI values can significantly weaken the sharp increase or decrease of the HI curve, With the exception of the SAE-EHI curve at approximately the 300th data point in the red rectangular area, the SAE-EHI curve rises gradually.

(2) Eq. (11) shows that as the number increases, the weight coefficient of the current HI value becomes smaller and smaller when EHI is calculated, especially eliminate concussion effect is good when the curve has a significant vibration phenomenon. For example, the SAE-HI curve in Fig. 9(a) significantly increases from the 179th data point (HI = 0.03864) to the 180th data point (HI = 0.2306), but the SAE-EHI curve does not. Therefore, the weight coefficient of the 180th HI value of EWMA is 0.95. However, in the proposed method, the same weight is given from the starting point to the current point, as shown in Eq. (11), hence the weight coefficient for the 180th HI value is 1/180. This can significantly weaken the oscillation phenomenon. 0.95 is much larger than 1/180. Therefore, the smoothing and eliminate concussion effect using the exponential function Eq. (11) is better than EWMA, as shown in Fig. 13(a).

(3) Outlier detection: Outlier detection is also used to find the outlier area in the first step. The upper threshold in Fig. 13(b) is $\mu + 3\sigma$, and the lower threshold is $\mu - 3\sigma$. Here, μ is the mean value of the difference value of SAE-HI in Fig. 9(a), and σ denotes the variance value of the difference value of SAE-HI in Fig. 9(a). From Fig. 13(b), few discrete and discontinuous points exceed the upper and lower threshold lines, and these cannot generate the outlier area. Therefore, outlier detection cannot deal with the curves that have oscillations.

(4) The corresponding *Corr* absolute values for EWMA and our proposed model are 0.9670 and 0.9708, respectively, and those for *Mon* are 0. 6059 and 0.0378, respectively. Where the *Mon* value is higher than the EWMA, SAE-EHI is used. Although the

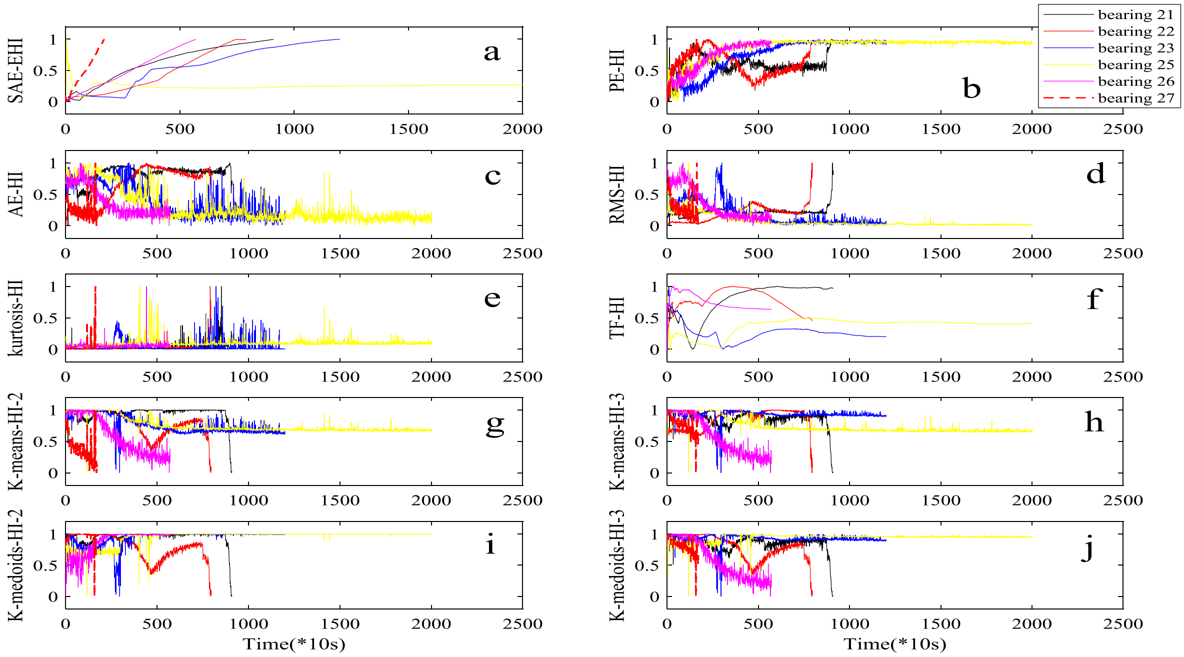

Fig. 14. HI extracted via various models for Condition 2, except bearing 24. "TF-HI" denotes 18 time–frequency indicators fusion model in Ref [6]. "2" and "3" denote cluster number, set as 2 and 3 in K-means and K-medoids.

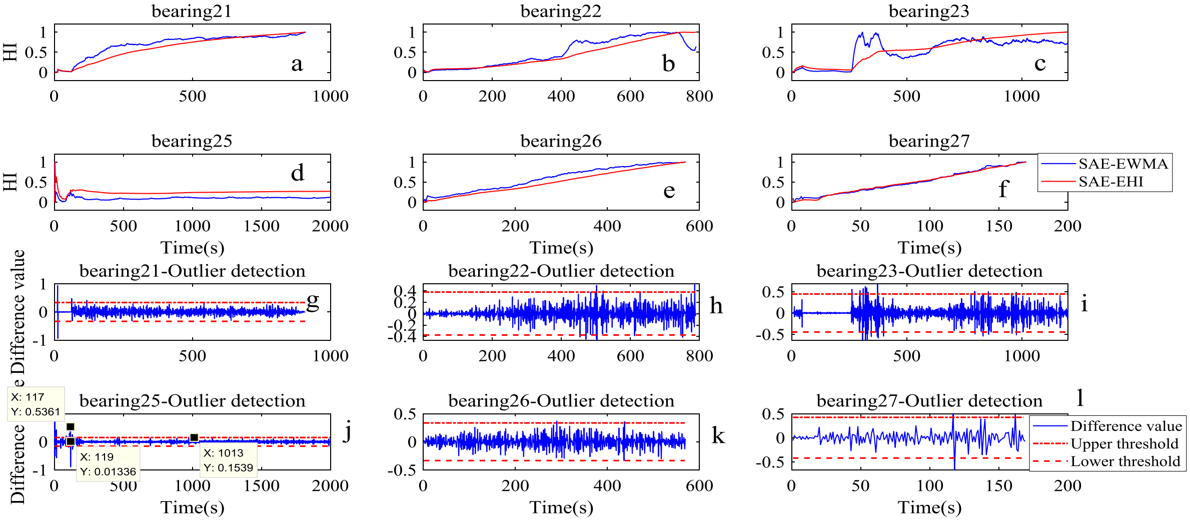

Fig. 15. Use of EWMA to eliminate shocks, and outlier detection for checking outlier point of Condition 2 (except bearing 24). (a–f) EWMA (g–l) Outlier detection.

Corr absolute value is slightly lower than the EWMA value, the *Mon* value is significantly higher than the EWMA value. These results indicate that the exponential function can effectively eliminate global shocks.

3.5. Cross validation

To further prove that our proposed method is superior to other models, the results for seven bearings in condition 2 are randomly divided into seven groups for cross validation. Analogous to before, this means that the six folds generated by six bearings are training data, and the seventh fold is the testing data; that is, if bearing 21 is selected as the testing data, the other six bearings are the training data.

We use the same network structure and same parameters of bearing 24's experiment for cross-validation. Different parameters in different models are also excepet SAE the same, as in the above-mentioned configuration. Finally, we also used *Mon* and *Corr* to compare the performance of the proposed method with other models from an overall angle. Because the result of bearing 24 was already presented in the previous sections, we show the results of six bearings but not those of bearing 24 in this section. The extracted HI obtained by using different modes is shown in Fig. 14, and Table 4 shows the *Mon* and *Corr* result.

(1) In Fig. 14, noise and oscillation are obvious in (a–j), but not in (a) and (f). These noises and oscillations seriously affect and easily submerge the growth and decrease of the HI curve, particularly the kurtosis in Fig. (e).

(2) Fig. 14(g–j) have a straight line in the early stages. There is no up-and-down trend in EMD-SVD-K-means/K-medoids. All these models, except our proposed method, increase and decrease sharply, rather than smoothly and gradually increasing as in Fig. 14(a). For example, take bearing 22: the EHI curve (red line) in (a) is gradually increasing without noise, but other HI curves

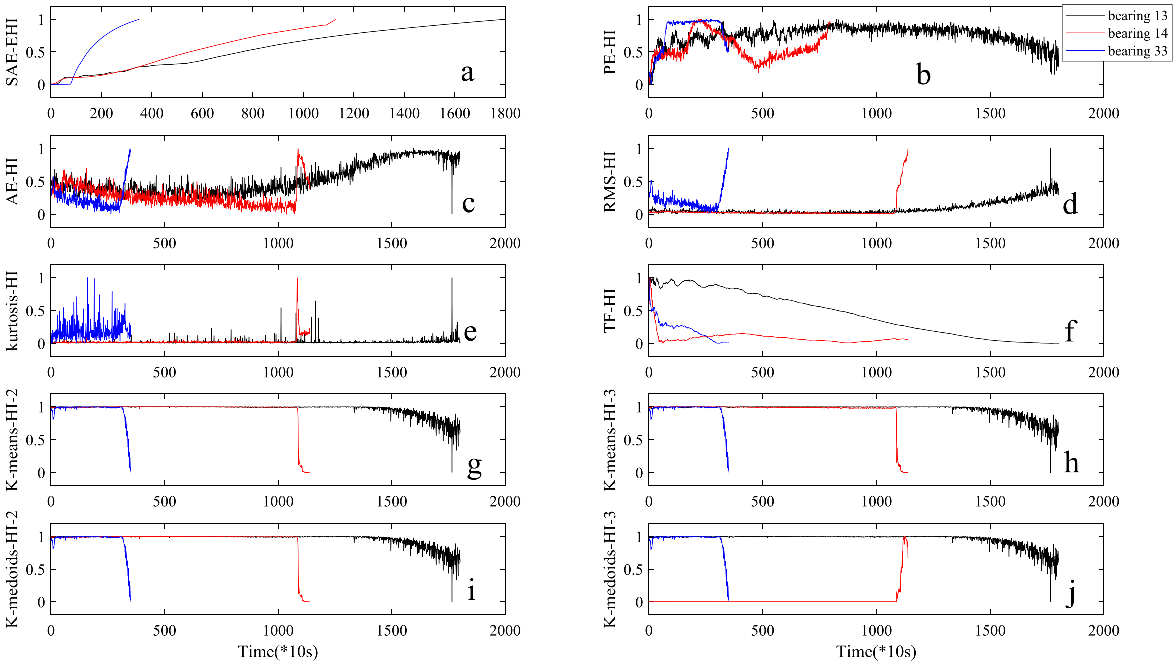

Fig. 16. Extracted HI from different models for bearing 13, bearing 14 and bearing 33. TF-HI denotes 18 time–frequency indicators fusion model in Ref [6], and "2" and "3" denote that cluster number is set as 2 and 3 in K-means and K-medoids.

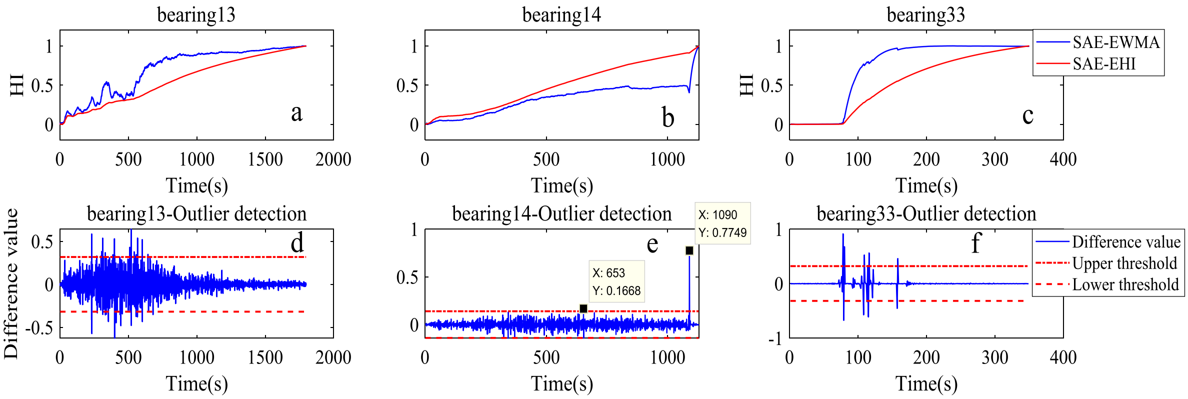

Fig. 17. Use of EWMA and Outlier detection to eliminate shocks for bearings 13, 14, and 33. (a–c) EWMA, (d–f) Outlier detection.

extracted from other models display the "decrease and increase" phenomenon, such as Fig. 14(b), (g) and (i).

(3) For bearing 25, all HI curves in Fig. 14 obtained from different models do not always increase; the red area in Fig. 7(c) shows that the bearing 25 vibrational signal has obvious instability at the early stage. For example, the signal amplitude decreases first and then increases, and the final amplitude decreases again before reaching a stable state. Therefore, the *Corr* absolute and *Mon* values for bearing 25 are significantly lower than those for other bearings, on the whole.

(4) Table 4 shows that all the *Corr* and *Mon* values of the proposed method are higher than those from other models, except those from bearing 25 in the last row of the table. The highest *Corr* absolute and *Mon* values are 0.999 and 0.894. These results demonstrate that the performance of the SAE with an exponential function is better than that of other models.

We also compare the proposed method with EWMA and outlier detection. The EWMA and outlier detection result for Condition 2 (except bearing 24) is shown in Fig. 15. *Corr* and *Mon* values are also shown in Table 5.

(1) EWMA: All EHI curves except that for bearing 25 are smoothly and continuously increasing in Fig. 15(a–f). However, there are significant severe vibrations for different bearings; in Fig. 15(c), the HI curve (SAE-EWMA) shows obvious oscillation from 300 to 500 s. It does not always rise gradually, but rises first and then falls.

(2) From Fig. 15(g–l), it can been see that few discrete and discontinuous points exceed the upper and lower threshold lines, and these points cannot generate the outlier area. Therefore, outlier detection cannot deal with curves that have oscillations.

(3) The *Corr* absolute and *Mon* values of SAE-EHI, except for bearing 25, are higher than those from EWMA. These results indicate that our method also efficiently eliminates global shocks.

Finally, we also selected bearings 13, 14, and 33 as the testing dataset to prove that our method is superior to other models. If bearing 13 is selected as the testing dataset, the other six bearings (11–12 and 14–17) are regarded as the training datasets. The results of this analysis are shown in Figs. 16 and 17, and the *Corr* and *Mon* values are also shown in Table 6.

(1) EWMA: All EHI curves are smoothly, continuously and gradually increasing in Figs. 16(a) and 17(a–c), and significant severe vibrations are seen for different bearings, with noise from other models, such as Figs. 16(b–c) and 17(a).

Table 4
Result of *Corr* and *Mon* using various models for bearing 2 (but not bearing 24).

Model	*Corr*						*Mon*					
Bearing No.	21	22	23	25	26	27	21	22	23	25	26	27
RMS	0.395	0.7450	−0.751	−0.705	−0.878	−0.577	0.013	0.042	0.039	0.023	0.008	0.029
kurtosis	0.276	0.3271	0.051	−8.5e−04	0.1150	0.2108	0.015	0.025	0.002	0.007	0.043	0.017
AE	0.506	0.8254	−0.799	−0.749	−0.847	−0.324	0.015	0.002	8.3e−04	0.026	0.015	0.005
PE	0.169	−0.0605	0.870	0.5991	0.9236	0.8802	0.017	0.025	0.005	0.004	0.026	0.017
K-means (2)	−0.030	−0.6752	−0.752	−0.765	−0.939	−0.693	0.006	0.012	0.007	0.012	0.054	0.005
K-means (3)	−0.335	0.5870	−0.006	−0.769	−0.948	−0.795	0.004	0.002	0.012	0.009	0.001	0.005
K-medoids (2)	−0.031	−0.6735	0.533	0.623	0.7686	0.0267	0.004	0.010	0.007	0.017	0.043	0.040
K-medoids (3)	−0.346	−0.6750	−0.041	0.372	−0.944	−0.776	0.011	0.012	0.010	0.016	0.001	0.005
Ref. [6]	0.7950	−0.1709	−0.254	−0.421	0.541	−0.147	0.283	0.160	0.100	0.020	0.197	0.052
SAE-EHI	0.9667	0.9787	0.970	0.146	0.999	0.998	0.870	0.797	0.514	0.071	0.894	0.810

Table 5
Result of *Corr* and *Mon* by using EWMA and SAE-EHI for Condition 2 (except bearing 24).

Model	*Corr*						*Mon*					
Bearing No.	21	22	23	25	26	27	21	22	23	25	26	27
EWMA	0.8604	0.9322	0.7526	0.1953	0.9896	0.9940	0.1485	0.1381	0.1643	0.1156	0.2091	0.4556
SAE-EHI	0.9667	0.9787	0.9704	0.1469	0.9993	0.9980	0.8702	0.7972	0.5146	0.0715	0.8946	0.8107

Table 6
The result of *Corr* and *Mon* from using various models for bearings 13, 14, and bearing 33, where 2 and 3 in K-means and K-medoids denote that the cluster number is 2 and 3.

Model	*Corr*			*Mon*		
Bearing No.	13	14	33	13	14	33
RMS	0.7090	0.2915	0.1731	0.0083	0.0369	0.0085
Kurtosis	0.1671	0.2902	0.1586	5.552e−04	0.0088	0.0256
AE	0.7889	−0.3240	0.1392	0.0017	0.0105	0.0541
PE	0.0541	0.2575	0.5196	0.0094	0.0070	0.0427
K-means (2)	−0.5563	−0.3543	−0.3975	0.0039	0.0228	0.0256
K-means (3)	−0.5636	−0.3866	−0.3940	0.0150	0.0176	0.0256
K-medoids (3)	−0.5566	−0.3546	−0.3984	0.0050	0.0228	0.0199
K-medoids(3)	−0.5602	0.3054	−0.3927	0.0050	0.0193	0.0427
Ref. [6]	−0.9914	−0.4153	−0.9230	0.4692	0.0264	0.0826
EWMA	0.9112	0.9363	0.8366	0.0817	0.2064	0.3983
SAE-EHI	0.9939	0.9960	0.9759	0.8410	0.9150	0.8453

(2) It can be seen from Fig. 17(d–f) that only a few discrete and discontinuous points exceed the upper and lower threshold lines, such as the 653th and 1090th points in Fig. 17(e). These discrete and discontinuous points cannot generate the outlier area, and therefore outlier detection cannot deal with curves that have oscillations.

(3) Table 4 shows that all *Corr* and *Mon* values of SAE-EHI are higher than those in other models, as can be seen in the last row of the table. These results demonstrate that the proposed method is superior to the other models.

4. Conclusions

To reduce the dependence of prior knowledge on constructing HI, we propose a model based on an SAE with an exponential function. First, a frequency domain signal is selected as the direct input SAE to extract the preliminary HI. The exponential function is used to eliminate the shock of the extracted degradation trend. Because the method uses the mean value, it is calculated by first using current HI values to weaken the oscillating phenomenon. Then, the exponential function is used to improve the HI curve monotonicity, which is possible because the exponential function is a monotonically increasing function.

Unlike the EWMA and outlier detection model, this method can eliminate oscillating phenomena without requiring parameter pre-configuration. It is somewhat limited in terms of outlier detection, in that it is necessary to confirm that the outlier area consists of consecutive outlier points to be able to smooth the curve from a local angle. However, only the mean value is needed to eliminate oscillating phenomena from a global perspective.

In addition, the cross-validation method is used to further verify the effectiveness of the proposed method and to compare it with other HI construction models. We also use the other bearing datasets with different conditions to prove the effectiveness of the proposed method. The resulting data confirm that proposed method is superior to the other HI construction models.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.asoc.2020.106119.

CRediT authorship contribution statement

Fan Xu: Writing - original draft, Writing - review & editing. **Zhelin Huang:** Formal analysis. **Fangfang Yang:** Data curation, Validation, Visualization. **Dong Wang:** Conceptualization, Data curation. **Kwok Leung Tsui:** Funding acquisition, Supervision.

Acknowledgments

This research work was partly supported by the General Research Fund (Project No. CityU 11206417), the Research Grants Council Theme-based Research Scheme under Project T32-101/15-R, the grant (RIF) from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. R-5020-18). The Natural Science Foundation of China under Grant No. 51975355.

References

[1] D. Wang, K.L. Tsui, Q. Miao, Prognostics and health management: A review of vibration based bearing and gear health indicators, IEEE Access 6 (2018) 665–676.

[2] A. Heng, A.C.C. Tan, J. Mathew, N. Montgomery, D. Banjevic, A.K.S. Jardine, Intelligent condition-based prediction of machinery reliability, Mech. Syst. Signal Process. 23 (5) (2009) 1600–1614.

[3] Z.J. He, H.R. Cao, Y.Y. Zi, Developments and thoughts on operational reliability assessment of mechanical equipment, J. Mech. Eng. 50 (2) (2014) 171–186.

[4] X.F. Chen, Z.H. Shen, Z.J. He, C. Sun, Z.W. Liu, Remaining life predictions of rolling bearing based on relative features and multivariable support vector machine, Proc. Inst. Mech. Eng. C 49 (2) (2013) 183–189.

[5] Y.G. Lei, N.P. Li, J. Lin, A new method based on stochastic process models for machine remaining useful life prediction, IEEE Trans. Instrum. Meas. 65 (12) (2016) 2671–2684.

[6] P.W. Tse, D. Wang, J. Vib. Control 23 (12) (2017) 1925–1937.

[7] R.Q. Yan, R.X. Gao, Approximate entropy as a diagnostic tool for machine health monitoring, Mech. Syst. Signal Process. 21 (2) (2007) 824–839.

[8] R.Q. Yan, Y.B. Liu, R.X. Gao, Permutation entropy: A nonlinear statistical measure for status characterization of rotary machines, Mech. Syst. Signal Process. 29 (2012) 474–484.

[9] Y. Tian, Z.L. Wang, C. Lu, Self-adaptive bearing fault diagnosis based on permutation entropy and manifold-based dynamic time warping, Mech. Syst. Signal Process. 114 (1) (2019) 658–673.

[10] A. Rai, S.H. Upadhyay, Bearing performance degradation assessment based on a combination of empirical mode decomposition and K-medoids clustering, Mech. Syst. Signal Process. 93 (2017) 16–29.

[11] S. Sun, B.B. Zhang, L. Xie, Y. Zhang, An unsupervised deep domain adaptation approach for robust speech recognition, Neurocomputing 257 (2017) 79–87.

[12] C. Affonso, A.L.D. Rossi, F.H.A. Vieira, A.C.P.L.F. Carvalho, Deep learning for biological image classification, Expert Syst. Appl. 85 (2017) 114–122.

[13] J. Feng, Y.G. Lei, J. Lin, X. Zhou, N. Lu, Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data, Mech. Syst. Signal Process. 72–73 (2016) 303–315.

[14] P. Vincent, H. Larochelle, Y.S. Bengio, P.A. Manzagol, Extracting and composing robust features with denoising autoencoders, in: Proceeding ICML '08 Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 1096–1103.

[15] P. Vincent, H. Larochelle, I. Lajoie, Y.S. Bengio, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.

[16] C. Lu, Z.Y. Wang, W.L. Qin, J. Ma, Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification, Signal Process. 130 (2017) 377–388.

[17] F. Xu, W.T.P. Tse, Y.L. Tse, Roller bearing fault diagnosis using stacked denoising autoencoder in deep learning and Gath–Geva clustering algorithm without principal component analysis and data label, Appl. Soft Comput. 73 (2018) 898–913.

[18] S.R. Saufi, Z.A.B. Ahmad, M.S. Leong, M.H. Lim, Differential evolution optimization for resilient stacked sparse autoencoder and its applications on bearing fault diagnosis, Meas. Sci. Technol. 29 (2018) 125002.

[19] C.Q. Shen, Y.M. Qi, J. Wang, G.G. Cai, Z.K. Zhu, An automatic and robust features learning method for rotating machinery fault diagnosis based on contractive autoencoder, Eng. Appl. Artif. Intell. 76 (2018) 170–184.

[20] Y.M. Qi, C.Q. Shen, D. Wang, J.J. Shi, X.X. Jiang, Z.K. Zhu, Stacked sparse autoencoder-based deep network for fault diagnosis of rotating machinery, IEEE Access 5 (2017) 15066–15079.

[21] X.Q. Li, H.K. Jiang, X. Xiong, H.D. Shao, Rolling bearing health prognosis using a modified health index based hierarchical gated recurrent unit network, Mech. Mach. Theory 133 (2019) 229–249.

[22] M. Mansouri, A. Al-Khazraji, M. Hajji, M.F. Harkat, H. Nounou, M. Nounou, Wavelet optimized EWMA for fault detection and application to photovoltaic systems, Sol. Energy 167 (2018) 125–136.

[23] V. Chandola, A. Banerjee, V. Kumar, Outlier detection : A survey, ACM Comput. Surv. (2007) 1–83.

[24] M. Gupta, J. Gao, C.C. Aggarwal, J.W. Han, Outlier detection for temporal data: A survey, IEEE Trans. Knowl. Data Eng. 26 (9) (2014) 2250–2267.

[25] Y.R. Wang, Q. Jin, G.D. Sun, C.F. Sun, Planetary gearbox fault feature learning using conditional variational neural networks under noise environment, Knowl.-Based Syst. 163 (2019) 438–449.

[26] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, in: Cognitive Modeling, The MIT Press, 2002, pp. 213–224.

[27] C.A.P. Sarath, S. Lauly, H. Larochelle, M. Khapra, An autoencoder approach to learning bilingual word representations, Adv. Neural Inf. Process. Syst. 3 (2014) 1853–1861.

[28] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[29] H.D. Shao, H.K. Jiang, H.W. Zhao, F. Wang, A novel deep autoencoder feature learning method for rotating machinery fault diagnosis, Mech. Syst. Signal Process. 95 (2017) 187–204.

[30] W.J. Sun, S.Y. Shao, R. Zhao b, R.Q. Yan, X.W. Zhang, X.F. Chen, A sparse auto-encoder-based deep neural network approach for induction motor faults classification, Measurement 89 (2016) 171–178.

[31] B.Y. Kosasih, W. Caesarendra, K. Tieu, A. Widodo, C.A.S. Moodie, A.K. Tieu, Degradation trend estimation and prognosis of large low speed slewing bearing lifetime, Appl. Mech. Mater. 493 (2014) 343–348.

[32] T. Williams, X. Ribadeneira, S. Billington, T. Kurfess, Rolling element bearing diagnostics in run-to-failure lifetime testing, Mech. Syst. Signal Process. 15 (5) (2001) 979–993.

[33] Y.G. Lei, N.P. Li, J. Lin, A new method based on stochastic process models for machine remaining useful life prediction, IEEE Trans. Instrum. Meas. 65 (12) (2016) 2671–2684.

[34] K. Javad, R. Gouriveau, N. Zerhouni, P. Nectoux, Enabling health monitoring approach based on vibration data for accurate prognostics, IEEE Trans. Ind. Electron. 62 (1) (2015) 647–656.

[35] L. Guo, N.P. Li, J. Feng, Y.G. Lei, J. Lin, A recurrent neural network based health indicator for remaining useful life prediction of bearings, Neurocomputing 240 (31) (2017) 98–109.

[36] P. Nectoux, R. Gouriveau, K. Medjaher, E. Ramasso, B. Chebel-Morello. N. Zerhouni, C. Varnier, PRONOSTIA: An experimental platform for bearings accelerated life test, in: IEEE International Conference on Prognostics and Health Management, PHM'12. Denver, Colorado, United States, 2012, pp. 1–8. ffhal-00719503f.

[37] C. Bandt, B. Pompe, Permutation entropy: a natural complexity measure for time series, Phys. Rev. Lett. 88 (2002) 174102–1–174102–4.

[38] Y.L. Tse, M.E. Cholette, P.W. Tse, A multi-sensor approach to remaining useful life estimation for a slurry pump, Measurement 139 (2019) 140–151.

[39] L. Ren, J. Cui, Y.Q. Sun, X.J. Cheng, Multi-bearing remaining useful life collaborative prediction: A deep learning approach, J. Manuf. Syst. 43 (2) (2017) 248–256.

[40] R.L. Tang, Q. Li, Zhou. J.X., S.Y. Zhang, et al., Suppression strategy of short-term and long-term environmental disturbances for maritime photovoltaic system, Appl. Energy (2019) http://dx.doi.org/10.1016/j.apenergy.2019.114183.

[41] Z. Wu, Q. Li, W. Wu, M.B. Zhao, Crowdsourcing model for energy efficiency retrofit and mixed-integer equilibrium analysis, IEEE Trans. Ind. Inform. (2019) http://dx.doi.org/10.1109/TII.2019.2944627.

[42] J.G. Lai, X.Q. Lu, X.H. Yu, A. Monti, Cluster-oriented distributed cooperative control for multiple ac microgrids, IEEE Trans. Ind. Inform. 15 (11) (2019) 2903–5918.

[43] J.G. Lai, X.Q. Lu, A. Monti, G.P. LIu, Stochastic distributed pinning control for co-multi-inverter networks with a virtual leader, IEEE Trans. Circuits Systems-II (2019) http://dx.doi.org/10.1109/TCSII.2019.2950764.

[44] R.L. Tang, Z. Wu, X. Li, Optimal operation of photovoltaic/battery/diesel/cold-ironing hybrid energy system for maritime application, Energy 163 (2018) 697–714.

[45] Z.M. Kong, S.S. Yang, S.S. Wang, D. Wang, H.Z. Lajos, Robust beamforming and jamming for enhancing the physical layer security of full duplex radios, IEEE Trans. Inf. Forensics Secur. 14 (12) (2019) 3151–3159.

[46] F.F. Yang, D. Wang, Y. Zhao, K.L. Tsui, S.J. Bae, A study of the relationship between coulombic efficiency and capacity degradation of commercial lithium-ion batteries, Energy 145 (2018) 486–495.

[47] F.F. Yang, X.B. Song, G.Z. Dong, K.L. Tsui, A coulombic efficiency-based model for prognostics and health estimation of lithium-ion batteries, Energy 171 (2019) 1173–1182.