

Gear pitting fault diagnosis using disentangled features from unsupervised deep learning

Yongzhi Qu¹ , Yue Zhang¹, Miao He², David He^{1,2} , Chen Jiao¹ and Zude Zhou¹

Abstract

Effective feature extraction is critical for machinery fault diagnosis and prognosis. The use of time–frequency features for machinery fault diagnosis has prevailed in the last decade. However, more attentions have been drawn to machine learning–based features. While time–frequency domain features can be directly correlated to fault types and fault levels, data-driven features are typically abstract representations. Therefore, classical machine learning approaches require large amount of training data to classify these abstract features for fault diagnosis. This article proposed a fully unsupervised feature extraction method for “meaningful” feature mining, named disentangled tone mining. It is shown that disentangled tone mining can effectively extract the hidden “trend” associated with machinery health state, which can be used directly for online anomaly detection and prediction. Compared with wavelet transform and time domain statistics, disentangled tone mining can better extract fault-related features and reflect the fault degradation process. Shallow methods, such as principal component analysis, multidimensional scaling and single-layer sparse autoencoder, are shown to be inferior in terms of disentangled feature learning for machinery signals. Simulation analysis is also provided to demonstrate and explain the potential mechanism underlying the proposed method.

Keywords

Gear fault, pitting, trending analysis, deep learning, wavelet analysis, sparse autoencoder

Date received: 1 June 2018; accepted: 5 December 2018

Introduction

Rotational machines are extensively used in modern industry. Machinery fault diagnostics can help prevent catastrophic failures in advance and reduce the lifetime maintenance cost. Gear fault diagnostics are one of the most important tasks in research and application. Many different gear fault analysis methods have been researched in literature. Among those, time–frequency analysis has drawn many attentions.^{1,2} However, these methods require strong domain knowledge to interpret and implement. The desire for automated fault feature extraction and diagnostics has driven the development of many machine learning–based methods. Machine learning methods, especially deep neural networks, have been actively applied to both the features and the raw data for fault diagnostic.³ proposed to use deep belief network (DBN) for health state classification. Compared with shallow networks such as support vector machine (SVM) and self-organizing maps (SOMs), DBN achieved the best classification performance for

aircraft engine diagnosis. Jia et al.⁴ utilized deep auto-encoder on frequency spectra and then fed the features into a classification model for fault classification. Lei et al.⁵ proposed to use neural network on raw vibration signal directly for fault diagnostics. In their proposed method, an unsupervised two-layer sparse filtering neural network was first used to directly learn features from mechanical vibration signals. Then, a simple softmax regression model was employed to classify health conditions based on the learned features. Qu et al.⁶ integrated dictionary learning into stacked autoencoder for direct feature learning from vibration signals with

¹School of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan, China

²Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, Chicago, IL, USA

Corresponding author:

Yongzhi Qu, School of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan 430070, Hubei, China.
 Email: quwong@whut.edu.cn

application in gear tooth pitting fault detection. Deep autoencoder was also studied in previous works^{7–9} for gear and bearing fault diagnostics. Denoising autoencoder and corrective autoencoder had also been investigated in fault diagnosis.¹⁰ Lu et al.¹¹ used a stacked denoising autoencoder-based model for health state identification. A Gaussian–Bernoulli deep belief network (GB-DBN) for electronics-rich analog system fault detection was proposed by Liu et al.¹² It was argued that GB-DBN is superior to traditional time-frequency feature extraction method—wavelet transform. Liu utilized restricted Boltzmann machine (RBM) for signal representation to predict remaining useful life (RUL) in the work by Liao et al.¹³ Deutsch and He¹⁴ proposed a method using RBM for bearing RUL prediction. In their work, root mean square (RMS) value was first computed from vibration signals of bearing test. Then, a lagged order of the RMS time series was reconstructed to represent the trend of vibration RMS in different timesteps. This lagged version of RMS served as the input of RBM to learn the degradation features, which were further fed into a regression model for RUL prediction. Although the physical model relying on complex signal processing was replaced by a model learned by RBM and linear regression, their work still falls into the realm of supervised learning. Convolutional neural network (CNN) had also been used for fault diagnosis in several studies.^{15–18} Ince et al.¹⁵ proposed a CNN method for motor fault detection. In their paper, raw time signals of motors were fed to an end-to-end CNN model. Guo et al.¹⁶ proposed a hierarchical adaptive deep convolutional neural network for bearing fault pattern and fault size recognition. Abdeljaber et al.¹⁷ utilized CNN on raw time domain vibration signals for feature extraction. Jing et al.¹⁸ investigated performances of CNN on raw time data, frequency spectrum data, time–frequency data and several hand-crafted features for gearbox fault detection. Zhang et al.¹⁹ combined deep CNN with the wide kernels for feature extraction. An extension of CNN, called convolutional autoencoder, was also tested for signal feature extraction.^{20,21} Cody et al.²¹ proposed a convolutional autoencoder for gear pitting diagnosis using AE signals. [AQ: 1] Deep features were then combined with condition indicators as the final features for classification. Besides, recurrent neural network (RNN) models were also used in fault diagnosis owing to its capability of superior performance on modeling time series. Prognosis of defect using RNN was reported by Malhi et al.²² and Tse and Atherton.²³ Zhao et al.²⁴ presented a long short-term memory (LSTM)-based method for machine health monitoring system in tool wear test. Gated recurrent unit (GRU) network was used for classification in Zhao et al.²⁵ In their paper, multi-sensory input was extracted into local features by tri-domain analysis and the features were fed into GRU model for further classification. Zhao et al.²⁶ gave a comprehensive survey on deep learning-based machine health monitoring. For more

recent applications of deep learning in fault diagnosis and prognosis, refer to the work by Zhao et al.²⁶

Although there are numerous research works using machine learning, almost all of the existing works are based on supervised learning, which needs to feed the features learned from deep neural networks into classification models for diagnostic task.

Early tooth surface wear and gear pitting detection are considered a challenging task due to weak fault signatures. Gear pitting fault detection has been investigated by both model-based and data-driven methods. The model-based methods rely on accurate dynamic models of the system to predict the desired behavior, while the data-driven approaches require historical data to train fault detection models. Rahmounea and Benazzouz²⁷ derived a pitting defect model based on gear mesh stiffness and further established the amplitude modulation function for pitting fault identification. Feki et al.²⁸ modeled the transmission error in the presence of localized gear tooth pitting. Although model-based methods can effectively explain the anomaly in the corresponding signals, it requires strong knowledge of the target system and dynamic modeling expertise. On the other hand, data-driven approaches can be viewed as black boxes, which automatically extract features and classify the potential state of the systems.²⁹ Lee et al.³⁰ proposed a zoomed phase map of continuous wavelet transform to detect gear pitting. The mean frequency of a scalogram was studied in Ozturk et al.³¹ to get features for gear pitting fault detection. In the work by Lewicki et al.,³² condition indicators were extracted from time-averaged vibration data. Spectral kurtosis was applied to extract features for gear pitting fault detection.³³ In the work by Ümütlu et al.,³⁴ statistical parameters of vibration signals in frequency domains were extracted as inputs to artificial neural network for gear pitting fault classification.

In sum, most of the above-mentioned methods require either strong domain knowledge or complex signal processing techniques to identify and extract the fault features. It is always desired to have a method which can automatically extract fault-related features without human knowledge. Although many of the above-reviewed machine learning methods devoted to fill this gap, most of the works still view the machine learning methods as basic representation learning technique which prepares inputs for further classification models. In machinery signal learning task, what we really need to learn is those hidden structure closely related to the faults other than features from other sources. A complete or overcomplete representation of the raw data (as learned by most deep learning methods), which also contains noise signals we may not need, can potentially harm the learning task. Therefore, how can we disentangle the interaction of many sources without explicit prior knowledge is more important.³⁵ This work aims to demonstrate that a fully unsupervised learning algorithm can intelligently recognize the

hidden structures in the input space and disentangled them into separate tones. In that sense, the proposed methods are called disentangled tone mining (DTM). This article also evaluates the sensitivity of deep learning methods on incipient pitting.

The main technical contributions of this work are summarized as follows:

1. This work presents a fully unsupervised method named DTM for feature extraction and fault-level trending. Since the proposed method is fully unsupervised, the need of historical training data is completely removed from the online fault detection task. To the best knowledge of the authors, this is the first work that presents a deep learning-based fully unsupervised fault detection approach.
2. It is shown in this work that DTM can extract meaningful fault features that are comparable or superior to traditional signal processing methods such as wavelet decomposition.
3. Detailed comparisons are conducted with shallow signal decomposition methods, such as principal component analysis (PCA), multidimensional scaling (MDS) and single-layer sparse autoencoder (SAE). Wavelet decomposition is also compared with the proposed DTM method. It is shown that DTM outperforms other feature extraction method toward fault feature extraction.
4. Simulation work is performed to interpret and validate the proposed methods.

The rest of the article is organized as follows. Section “The methodology” introduces the idea of the methodology. Section “Experimental setup” presents the experimental work. Section “Results and discussion” gives the results and discussion. Section “Simulation analysis” presents the simulation study on gear fault signals and further validates the proposed methods. Section “Conclusion” concludes the article.

The methodology

The proposed DTM method for sensor signal analysis and fault diagnostics includes three steps:

- Slice the time domain sensor signal into equal length segments and compute the spectrum of each segment.
- Obtain n -dimensional features using deep SAE ($n \ll N$, N is the input dimension) with the spectrum data.
- Use rank estimation-based methods to further reduce the n -dimensional feature space into uncorrelated d -dimensional feature space.

The extracted features can be used for fault-level identification, fault diagnosis, trend prediction, and so on, using unsupervised clustering or regression analysis.

In this article, K -means is used for unsupervised fault-level clustering for the purpose of results validation and comparison.

The motivation of using frequency spectrum instead of the raw time domain signals lies in the fact that Fourier transform can effectively map the temporal correlation over multiple timesteps in the amplitude domain to specific frequency in the frequency domain. This ensures that different source information associated with different frequencies can be represented with independent components. Each frequency component serves one dimension of the inputs of the proposed algorithms. In the second step, a deep neural network is built to learn a low-dimensional embedding for the input frequency vector. Notice that autoencoder does not enforce different dimensions of the learned representations to be orthogonal or fully uncorrelated; we devise a procedure to eliminate strongly correlated dimensions. The idea is to prune dimensions corresponding to small eigenvalues based on QR decomposition. Detailed steps of the proposed procedure are as follows:

1. Conduct QR decomposition on the feature matrix learned from SAE to obtain the upper triangle matrix R .
2. Obtain the eigenvalues of R , which are also the eigenvalues of the feature matrix.
3. Determine the approximate rank of the feature matrix by a threshold over the eigenvalues. For example, one can specify the threshold as the eigenvalue decreases by one order. In this case, we will drop dimensions associated with eigenvalues smaller than 0.1 times of maximum eigenvalue.
4. Select the dimensions associated with the remaining eigenvalues.

One advantage of this proposed method is that number of final dimensions can be automatically determined. Notice that the proposed linearity checking procedure cannot be substituted with PCA because PCA would find the principle components in an affined space other than the original space learned by SAE. Therefore, PCA would not keep some useful features learned by SAE.

Given that the latent features are first disentangled via deep models and a discriminate feature reduction step is followed, we name the overall method DTM.

A brief introduction of autoencoder is given as follows. Autoencoder is an unsupervised learning algorithm, which aims to reconstruct the input through encoding and decoding. It contains two steps: encoding and decoding. By setting a lower number of nodes in the hidden layer compared with the input layer, autoencoder can learn a low-dimensional embedding of features. The structure of an autoencoder is shown Figure 1 with one input layer, one output layer and a fully connected hidden layer. The output layer typically has the same number of nodes as the input layer as it

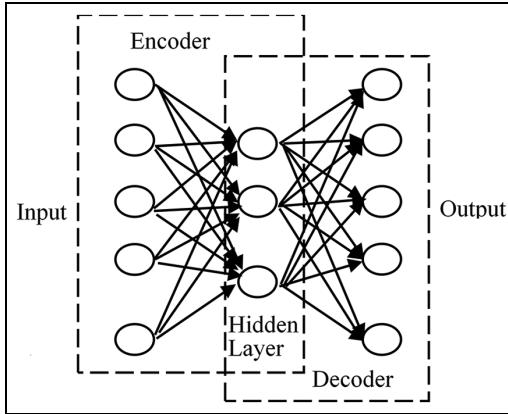


Figure 1. The structure of an autoencoder.

tries to reconstruct the inputs. For an encoder, if input data are represented as x , the output data are \hat{x} and the hidden layer can be represented as h . The encoder and decoder are described in equations (1) and (2)

$$h = \sigma(wx + b) \quad (1)$$

$$\hat{x} = \sigma'(w'h + b') \quad (2)$$

where σ is an activation function, for example, sigmoid function or a rectified linear unit (ReLU); w is the weight matrix and b is the bias vector. In the case of a regenerating autoencoder, w and w' are tied weights.

By adding a sparse regularization term in the objective function of an autoencoder, it becomes an SAE. The loss equation of SAE can be written as follows

$$L(w, w', b, b'; x) = \frac{1}{2m} \sum_m \|x - \hat{x}\|_2^2 + \lambda \|w\|^2 + \beta \sum_{j=1}^{s_2} KL(\rho || \hat{\rho}) \quad (3)$$

where $KL(\rho || \hat{\rho})$ represents the KL divergence. m is the number of training examples and s_2 is the number of hidden units. ρ is the sparsity parameter. λ and β are two weight parameters. The first term of equation (3) is the reconstruction error, the second term is the weight regularization term and the third term is for sparse penalty.

Deep neural network has proved that it has a better performance for data representation than shallow network in many fields.³⁶ In this article, a deep SAE is constructed by connecting the output layer of the last SAE to the input layer of the next SAE. The deep SAE is optimized in a greedy manner layer by layer.

Experimental setup

Experimental test rig

The pitting fault diagnosis experiments were performed on the test rig shown in Figure 2. It includes two 45 kW Siemens servo motors, where one of the motor acts as

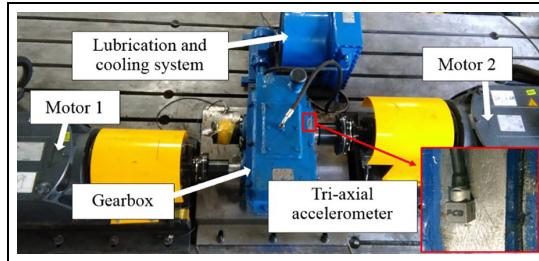


Figure 2. The gearbox test rig.

Table I. The major parameters of gearbox.

Gear parameter	Driving gear	Driven gear
Tooth number	40	72
Module	3 mm	3 mm
Pitch diameter	120 mm	120 mm
Base circle diameter	112.763 mm	202.974 mm
Pressure angle	20°	20°
Tooth width	85 mm	85 mm

the drive, the other one configured as the load motor. The testing gearbox is a single-stage spur gearbox. Major gear parameters are shown in Table 1.

Gear fault simulation and data description

We simulated the pitting fault on the surface of the driven gear. Five consecutive sets of experiments were conducted, including one set of healthy state and four faulty states. For faulty cases, the fault levels were gradually increased, as shown in Figure 3. The healthy gear contains no sign of wear and is not shown here. The pitting faults were manually created using a drill bit without any dismantling of the gearbox except the lube oil cover. Therefore, the gear test rig remains in a consistent condition except the pitting faults.

As can be seen from Figure 3, the pitting faults are created on the tooth surface with more area pitted and depth deepened after each set of experiments. We mark test case #1 as healthy (not shown here). Test case #2 only contains a few scratches on the gear surface. Case #3 and #4 include more and more dents. Pitting fault #5 contains minor pitting on a second gear tooth besides the pitting on the first gear tooth.

The approximate percentages of pitting area are shown in Table 2. The pitting percentages are normalized to the area of one tooth surface. The pitting area is estimated using image processing tools in Adobe Photoshop® software.

The vibration signals are collected with a sampling rate of 20.48 KHz. Multiple speed and load conditions are tested. The data under 1000 RPM/200 Nm are used in this work. Vibration signals from the vertical direction are chosen. For each test condition, five sets of signals are collected.

Table 2. The approximate percentage of pitting area.

Experiment number	1	2	3	4	5
Percentage (%)	0	4.33	12.12	20.47	24.91

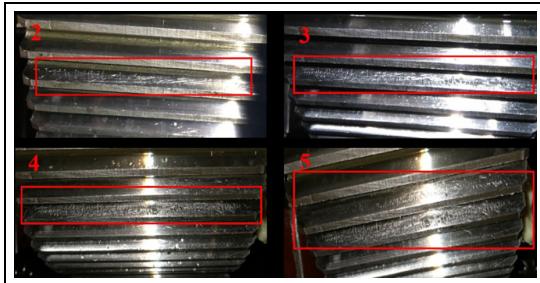
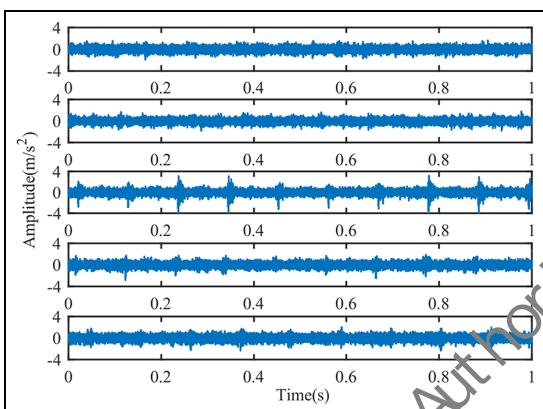
**Figure 3.** Simulated pitting faults.**Figure 4.** The waveforms of the raw vibration signals, from top to bottom: cases 1, 2, 3, 4, and 5.

Figure 4 shows the raw vibration signals for the corresponding five cases. It can be seen that the signals look quite similar except pitting case #3, where more noise and peaks are presented. By evaluating typical

condition indicators of the time domain signals, the fault trends cannot be correctly revealed. As shown in Figure 5, P2P is not increasing consistently with the fault levels. RMS shows a slight upward trend over five tests. However, the trend is inconsistent neither.

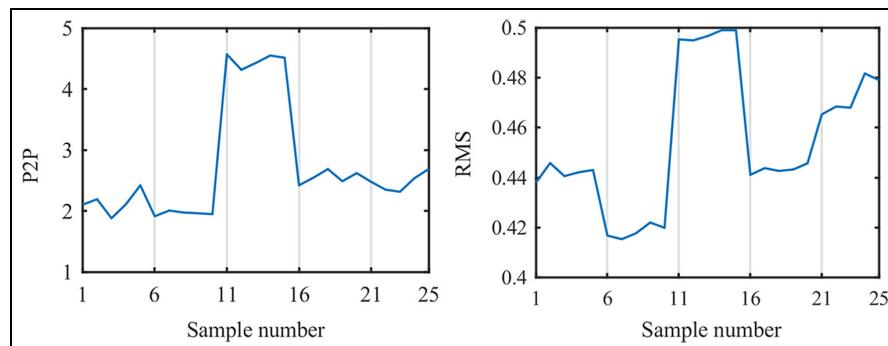
To evaluate the proposed methods, the vibration signals are sliced into equal length segments, which roughly cover one revolution of the driven gear. In total, 280 samples for each test case are generated. Since there are five different test cases (five test conditions), there are a total of 1400 samples. The detailed procedure is presented as follows:

1. As introduced in the section “The methodology,” a fast Fourier transform is first performed on the sliced vibration data. The 1000 frequency points are obtained for each segment, as shown in Figure 6. The spectra data serve as inputs to DTM.
2. The spectra data are normalized to 0 to 1.
3. The structure of the deep SAE is designed as 1000-500-200-50, with three hidden layers. It is found that the results are relatively robust to hyperparameters of the SAE. The hyperparameters were set as follows: the sparsity parameters in SAE are set as $\rho = 0.15$ and $\beta = 3$, weight regularization parameter $\lambda = 0.001$.

Results and discussion

Discussion of results

First, it is necessary to examine the spectra in Figure 6. They are plotted in the same order as raw time domain data. It can be seen that the major frequency tones are

**Figure 5.** P2P and RMS value of the raw signals. There are five samples in each case, plotted in the order of cases #1 to #5.

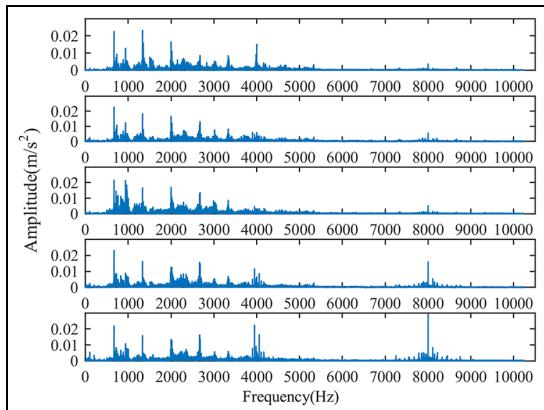


Figure 6. The corresponding frequency spectra of the five tests.

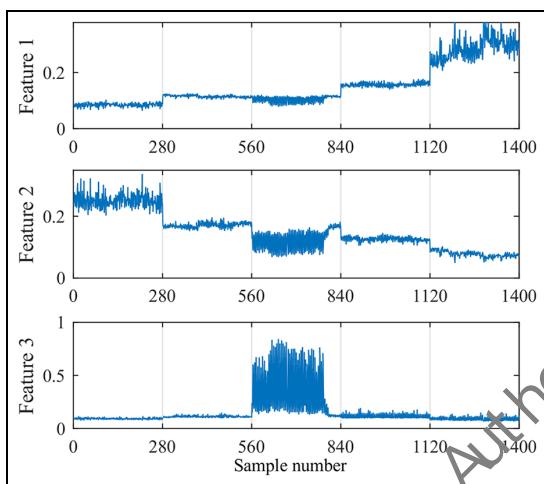


Figure 7. Features obtained from the proposed DTM method.

relatively unchanged. There is little variation with regard to the meshing frequency (666.7 Hz). However, the distribution of the frequency components shows slight variation over different conditions. Some high-frequency components around 8000 Hz show a developing trend in pitting conditions #4 and #5, whereas such trends are not presented in conditions #2 and #3. Overall, the fault level and pitting trend cannot be explicitly observed and quantified from frequency spectra directly.

In Figure 7, three “disentangled” features are captured using the proposed DTM method. In each subplot, the x-axis indicates the sample numbers, where each test condition includes 280 samples. The first 280 samples are healthy data, and data points 281–540 refer to test condition #2 (the first fault level), and so on. It can be seen that five test conditions can be roughly separated by Feature 1 and Feature 2. It is more important to see that Feature 1 and Feature 2 display a clear monotonic trend with conditions #1, #2, #4 and #5,

although test condition 3 contains more fluctuations. It is found that signals from condition #3 contain multiple high-level sidebands around meshing frequency and harmonics, while signals from other conditions do not possess such characteristics. Although the fault feature seems to be overdrawn by excessive noise for condition #3, it is worth to mention that this abnormal feature is “smartly” recognized and represented in Feature 3. It can be seen that the values of Feature 3 remain unchanged for test conditions #1, #2, #4 and #5, while it clearly distinguishes condition #3. Since Feature 3 is mainly caused by the abnormal noise associated with condition #3 and it shows no trend for other test cases, it is concluded that Feature 1 and Feature 2 are more closely related to the level of pitting fault and can be used as potential condition indicators to track the hidden state of the geared systems.

As a fully unsupervised algorithm, it is important to understand how DTM recognized the hidden trends in the spectra data. While a precise mathematical explanation is absented for many of the deep learning algorithms so far, an analogous comparison can be made between the actual fault trend and the extracted features. First of all, since the gearbox remains the same except the pitting levels, the hidden trend in the spectra is believed to be mainly caused by the fault levels. This hidden trend is typically weak and cannot be directly extracted by traditional time-frequency method. However, the proposed DTM method can disentangle uncorrelated noise and recover the discriminant features. Second, the changes in the faulty signals can be represented by certain frequency distribution patterns in the spectrum. One notable explanation is that the amplitude of low-frequency components sees an overall decreasing trend, while that of the high-frequency components see an increasing trend, as shown in Figure 6. The distribution pattern represents the fault level of the gearbox, which can be disentangled and learned.

Comparison with shallow feature extraction methods

In order to validate the performance of the proposed method, two shallow methods are tested for comparison: PCA and MDS. Both methods are unsupervised methods in a shallow manner. A (1000-3) mapping structure is chosen to test their performances.

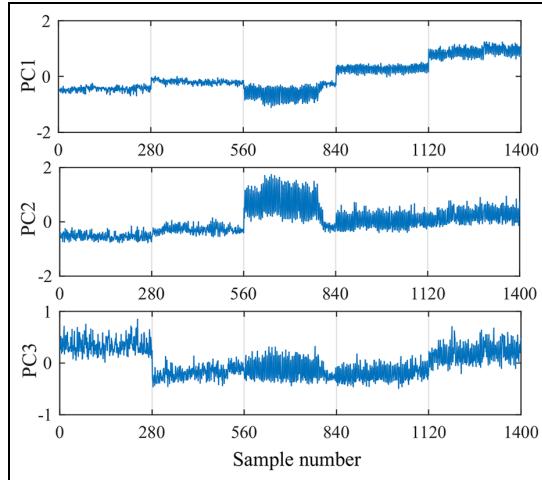
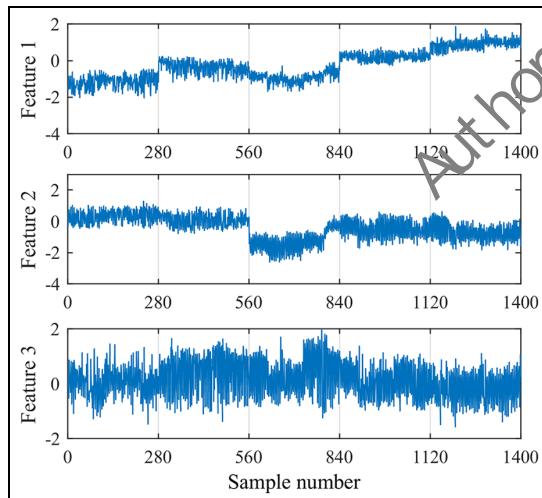
Figure 8 shows the results of PCA. PC1–PC3 are the first three principal components of the data. They represent 17.1%, 13.6% and 4.1% of the sum of the eigenvalues. Compared with the proposed DTM method, three features represent 44.0%, 30.1% and 22.9% of the sum of eigenvalues. It shows that deep SAE generated a more effective and compact representation. Although PC1 of PCA also shows a rising trend that can reflect the fault level, the noise level is much higher than that of DTM features. Besides, we can see from PC2 and PC3 that PCA cannot effectively disentangle the noise feature from the fault features. The results from deep SAE-based methods are improved by the

Table 3. K-means results on DTM, PCA and MDS.

Method	DTM	PCA	MDS	SAE (w/o#3)	PCA (w/o#3)	MDS (w/o#3)
Accuracy (%)	87.93	69.14	57.86	99.11	63.57	54.29

DTM: disentangled tone mining; PCA: principal component analysis; MDS: multidimensional scaling.

W/o #3 refers to without #3 test data.

**Figure 8.** The first three principle components of PCA.**Figure 9.** The first three features using MDS.

deep nonlinear architecture since a linear autoencoder would be mathematically equivalent to PCA. Figure 9 shows the results of MDS. Similar to PCA, MDS can also get a general trend of fault level. However, it can hardly separate five test cases.

To further quantify the performance of the proposed method, we employ unsupervised clustering technique on extracted features. The results of K -means are shown in Table 3. We exclude the Feature 3 since it is

non-discriminative for most test cases and is believed to be mainly noise related. We use the standard unsupervised evaluation metric—accuracy for the evaluation and comparison. The accuracy is defined as follows³⁷

$$ACC = \max_m \frac{\sum_{i=1}^n 1\{l_i = m(c_i)\}}{n} \quad (4)$$

where l_i is the ground truth label, c_i is the cluster assignment produced by the K -means and m ranges over all possible one-to-one mappings between clusters and labels.

It is needed to mention that the number of clusters is assumed to be known here and set as an input to K -means. In practice, the number of clusters can be learned by the clustering algorithm itself. The accuracy of DTM is 87.93%. Accuracy of PCA is 69.14% and MDS is 57.86%. MDS shows large variances over multiple runs, and the given results show the averaged results over five runs. On the contrary, the results obtained by DTM are quite stable. Notice that test condition #3 is highly corrupted with external noise; if test condition #3 were removed, the proposed DTM method achieves 99.11% accuracy toward fault-level clustering.

Comparison with WKT

One of the state-of-the-art features obtained by signal processing method is WKT,³⁸ which was extracted from discrete wavelet transform (DWT) and kurtosis. [AQ: 2] WKT contains two parts: wavelet coefficients and kurtosis. The first part consists of the energy of the detail coefficient from wavelet decomposition, where each energy feature was defined as the sum of squares of the detail coefficients of each decomposition level. The second part is the kurtosis of samples in time domain. In this article, Db4 wavelet is chosen since it gives the best overall results over other wavelets. It is also found that features from high-level decomposition are mostly noise. Therefore, a four-level wavelet decomposition is performed. The corresponding features extracted from the experimental data are shown in Figure 10(a). Four sets of detail coefficients are obtained, denoted as Feature 1, Feature 2, Feature 3, and Feature 4, as shown in Figure 10(a).

It can be seen that Feature 1 in Figure 10(a) also shows an approximate upward trend. However, the first three test cases remain at the same level without clear separation. It is interesting that Feature 4 from wavelet decomposition has a similar trend as the Feature 3 in

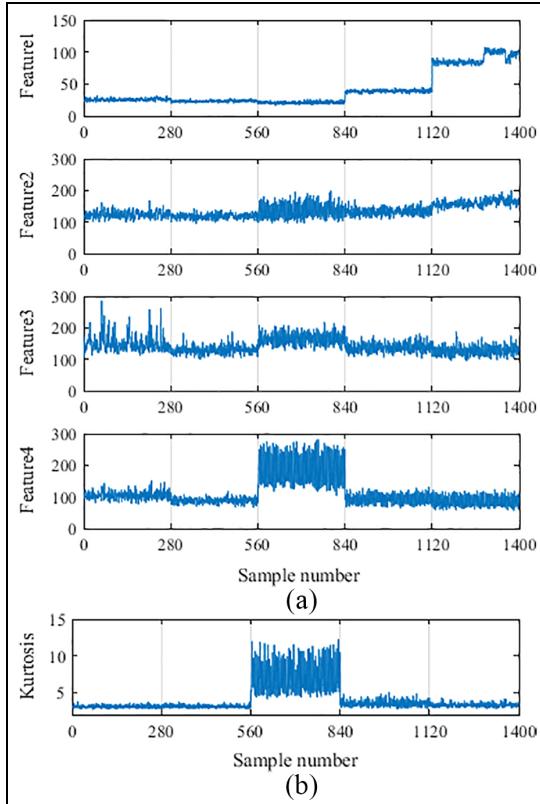


Figure 10. The features of WKT: (a) wavelet features and (b) kurtosis features.

Figure 7 obtained by the proposed DTM method. Using wavelet decomposition, Feature 1–Feature 4 actually reflect the energy level in different frequency bands, where Feature 1 reflects the highest frequency band, Feature 2 the next highest, and so on. More specifically, in this test, the detail coefficient in level 1 reflects the distribution of the high-frequency components in the range of 5120–10,240 Hz. Furthermore, Feature 1 in Figure 10(a) reflects the energy level of signals from 5120 to 10,240 Hz. Although wavelet transform can be regarded as a feature mapping method, each feature still reflects partial information in local frequency ranges. It cannot extract the overall signal patterns in a combinatorial manner. However, the features from DTM do not subject to such limitation. In DTM, the overall distribution trend in the whole frequency bands can be combined and represented with a single latent variable.

In Figure 10(b), the condition indicator of kurtosis was shown. It can be seen that kurtosis can only reflect that the third case is abnormal from other four cases due to noise contamination. The fault trend is totally missing. Moreover, we also get similar results with RMS, crest factor (CF) and skewness. It is concluded that (1) traditional time domain condition indicators cannot effectively represent the weak fault features overdriven by background noise and therefore are not

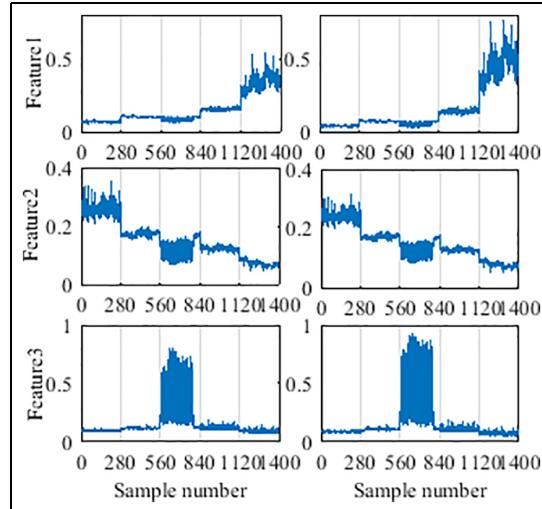


Figure 11. Testing results of another two structures of SAE (left: 1000-500-200-100, right: 1000-500-200-20).

effective in the early fault detection of machines and (2) time domain condition indicators are easily affected by external noise.

For the sake of comparison, 68.86% accuracy was achieved when using K-means to cluster the above WKT features with five dimensions. The accuracy is close to that of PCA and MDS.

Robustness evaluation of the DTM method

Usually, the structure of the deep autoencoder may affect the overall results. We further show that the DTM algorithm is relatively robust with different number of layers and different numbers of hidden nodes. We further show that the algorithms are relatively robust with different numbers of hidden nodes. In terms of varying number of output nodes, two additional structures are tested here: 1000-500-200-100 and 1000-500-200-20. The tested results display similar trends as shown in Figure 11.

The results are also similar when only two hidden layers are employed. Therefore, the proposed DTM method is relatively insensitive to network structures. However, if only one hidden layer is used, for example, with 1000-50 network setting, the latent trends cannot be effectively learned. Following the proposed DTM procedure, seven uncorrelated features were generated instead of three, as shown in Figure 12. Despite that similar trends like those from deep structure can also be roughly extracted, some mixed features were also generated. Moreover, these features cannot effectively differentiate the fault level. Using the obtained seven features for K-means clustering, the accuracy is only 66.57%, which is significantly lower than that of the deep models. This further demonstrates that deep structure outperforms shallow structure in learning latent features and obtaining disentangled representations.

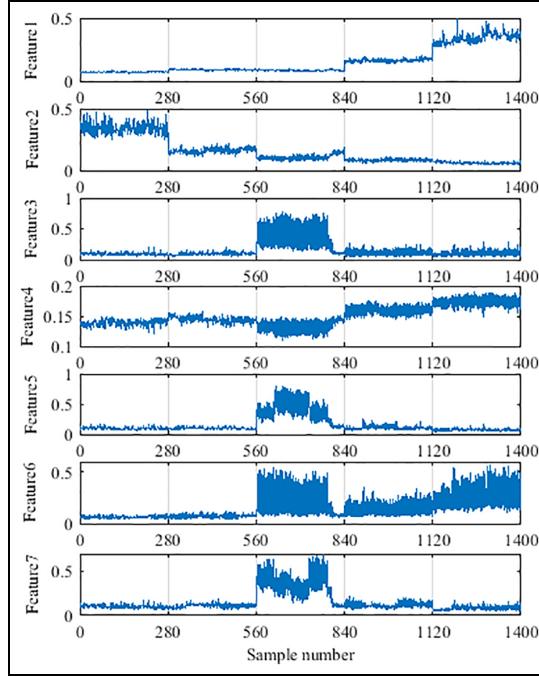


Figure 12. Features obtained with one hidden layer (1000-50).

In addition, it is noted that the threshold of the rank estimation is set as 0.1 when performing the linear correlation check. We tested the thresholds in different orders, such as 0.01 and 0.001, and the results stay the same.

Simulation analysis

To better understand and visualize the effect of proposed DTM method, we conduct a further analysis on simulated gear signals. The procedure adopted here is to change some frequency components gradually as fault injection to see how the proposed method reacts to the changes. A simple model of gear vibration signal is used, as described below.³⁹

The healthy condition is modeled as follows

$$x(t) = \sum_{m=0}^M X_m \cos(2\pi m f_z t + \varphi_m) \quad (5)$$

where f_z is the mesh frequency and X_m is the amplitude of the m th meshing harmonic.

When a localized fault is presented in the gearbox, the vibration signal model becomes

$$a_m(t) = \sum_{n=0}^N A_{m,n} \cos(2\pi n f_n t + a_{m,n}) \quad (6)$$

$$x'(t) = \sum_{m=0}^M X_m (1 + a_m(t)) \cos(2\pi m f_z t + \varphi_m) \quad (7)$$

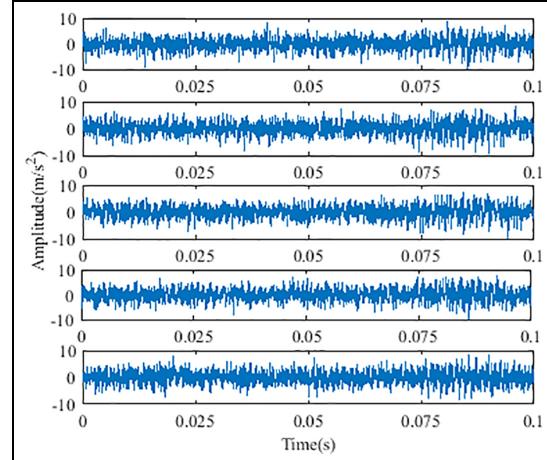


Figure 13. Simulation signals in the time domain. From top to bottom: groups 1–5.

where $a_m(t)$ represents the amplitude modulation function and f_n is the rotational frequency. The speed and the sampling rate are kept consistent with the experimental test. This leads to $f_z = 666.7$ Hz and $f_n = 9.26$ Hz. In the corresponding modulation signals, sidebands will be incorporated due to the modulation effect. To simplify the process of simulation, all the phases and the amplitudes of $a_m(t)$ were set randomly. The amplitude of up to n th-order mesh frequency ($n = 1, 2, \dots, 12$) was changed.

In the simulation, we decrease first to third orders of mesh frequencies gradually with a step size of 0.1, where the initial amplitudes are set to 0.8, 0.6 and 0.6, respectively. Concurrently, the top 3 orders of mesh harmonics (10th–12th harmonics) are increased with a step size of 0.1, where their initial amplitudes were set as 0.3, 0.4 and 0.5. Five groups of simulation signals are generated, corresponding to five different levels of fault. In order to test whether the proposed method is robust to noise interference, we added 6 dB white noise to the signal. Figures 13 and 14 show the simulated signals in the time domain and their corresponding spectra, respectively.

By adding excessive white noise, the actual order of the five groups of signals cannot be identified by simply comparing the amplitudes of certain frequency components. In Figure 15, an example is given to show the comparison of the amplitudes of the 2nd meshing harmonic and the 10th meshing harmonic. It can be seen from Figure 15(a) that the absolute amplitude of the second harmonic has been contaminated by the injected noise. Similar phenomenon can also be seen from Figure 15(b) for the 10th harmonic.

Next, the proposed DTM method is applied to the spectra of the simulated signals. The structure of deep SAE is chosen as 1000-500-200-50, which is the same for experimental data. We also keep the same setting for hyperparameters in SAE. With regard to the

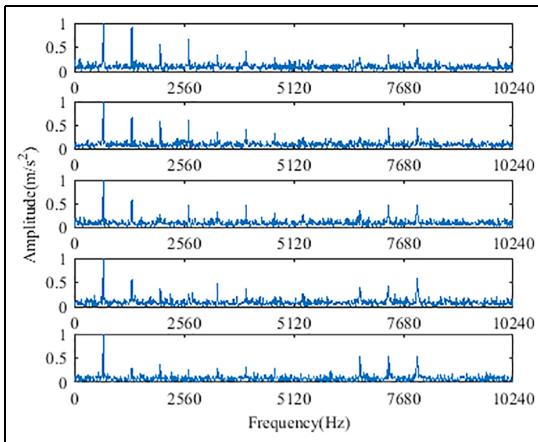


Figure 14. The corresponding spectrum of the simulation signals.

threshold in rank estimation, we also set as 0.1. Since the simulation signals contain much less-frequency components, only one feature is kept by the DTM method. The results are shown in Figure 16.

From Figure 15, one can hardly determine the correct order of the test cases due to the mask of the noise. However, from Figure 16, it can be seen that the proposed DTM can correctly reveal the change of frequency amplitudes in the spectra. It is immune to the interference of noise. It is worth to mention that the decreasing or increasing trend does not mean the increase or decrease in some local frequencies in the spectra, but rather should be viewed as an evolving trend of the distribution.

The same results are obtained when the rank estimation threshold is set within 10^{-1} to 10^{-9} . However, if the rank estimation error is set to 10^{-10} or smaller, two features will be extracted from the simulation signals,

as shown in Figure 17. Further look indicates that they are actually close to linearly correlated with correlation coefficient around -1 . Therefore, the features are redundant if the threshold is set too low. In practice, a threshold of 10^{-1} can be used as a rule of thumb.

The simulation proves the capability of the proposed method to extract the hidden trend overwhelmed in the signal by external noise. The DTM method can track the evolution of the distribution of the spectra. It explains why DTM method can disentangle and extract fault-related trends. For machinery signals that contain a monotonic degradation pattern, the proposed unsupervised deep model can automatically extract the trend and reveal the fault deterioration status.

Conclusion

It is important to intelligently monitor gear health condition online. In this article, an unsupervised feature extraction method called DTM was proposed to learn fault level directly from the frequency spectra of the data. The test data were collected on an industrial gearbox with different levels of seeded pitting faults. The experimental results showed that the proposed method can effectively disentangle the mixed features. The extracted features can clearly indicate the evolution trend of the spectra. Compared with WKT, PCA and MDS, the proposed DTM method can extract better discriminative features with a lower degree of mutual interference. The proposed method can be used online for real-time fault diagnosis and prognosis. The extracted features can then be used with clustering algorithm for fault-type and fault-level diagnostics. It may also be fed into prognostic algorithms for RUL prediction.

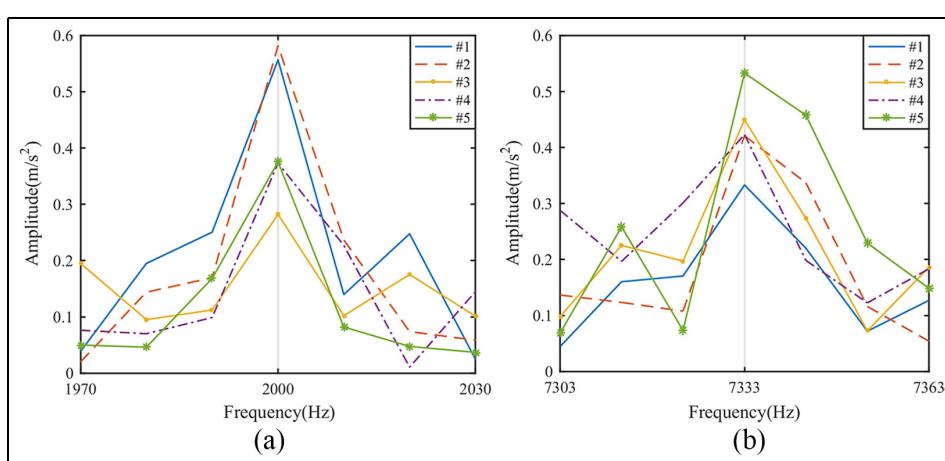


Figure 15. Details of two frequencies: (a) 2nd-order meshing frequency and (b) 10th-order meshing frequency.

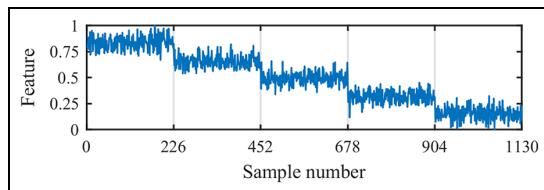


Figure 16. The obtained latent trend from the simulation signals.

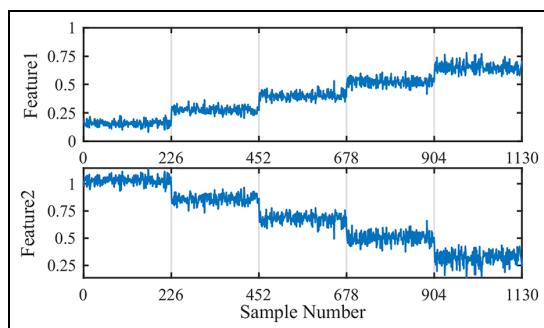


Figure 17. The extracted features from simulation signals with rank estimation error of 10^{-10} .

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

This research is partially supported by Natural Science Foundation of China under project no. 51305353.

ORCID iDs

Yongzhi Qu <https://orcid.org/0000-0002-5314-023X>
David He <https://orcid.org/0000-0002-5703-6616>

References

- Ricci R and Pennacchi P. Diagnostics of gear faults based on EMD and automatic selection of intrinsic mode functions. *Mech Syst Signal Pr* 2011; 25(3): 821–838.
- Wu TY, Chen JC and Wang CC. Characterization of gear faults in variable rotating speed using Hilbert-Huang Transform and instantaneous dimensionless frequency normalization. *Mech Syst Signal Pr* 2012; 30: 103–122.
- Tamil Selvan P and Wang P. Failure diagnosis using deep belief learning based health state classification. *Reliab Eng Syst Safe* 2013; 115: 124–135.
- Jia F, Lei YG, Lin J, et al. Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mech Syst Signal Pr* 2016; 72–73: 303–315.
- Lei YG, Jia F, Lin J, et al. An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. *IEEE T Ind Electron* 2016; 63: 3137–3147.
- Qu Y, He M, Deutsch J, et al. Detection of pitting in gears using a deep sparse autoencoder. *Appl Sci* 2017; 7(5): 515–529.
- Shao H, Jiang H, Zhao H, et al. A novel deep autoencoder feature learning method for rotating machinery fault diagnosis. *Mech Syst Signal Pr* 2017; 95: 187–204.
- Tao S, Zhang T, Yang J, et al. Bearing fault diagnosis method based on stacked autoencoder and softmax regression. In: *Proceedings of the control conference (CCC)*, Hangzhou, China, 28–30 July 2015, pp.6331–6335. New York: IEEE.
- Chen R, Chen S, He M, et al. Rolling bearing fault severity identification using deep sparse auto-encoder network with noise added sample expansion. *Proc IMechE Part O: J Risk and Reliability* 2017; 231(6): 666–679.
- Shao H, Jiang H, Wang F, et al. An enhancement deep feature fusion method for rotating machinery fault diagnosis. *Knowl Based Syst* 2017; 119: 200–220.
- Lu C, Wang ZY, Qin WL, et al. Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification. *Signal Pr* 2017; 130: 377–388.
- Liu Z, Jia Z, Vong C, et al. Capturing high-discriminative fault features for electronics-rich analog system via deep learning. *IEEE T Ind Inform* 2017; 13(3): 1213–1226.
- Liao L, Jin W and Pavel R. Enhanced restricted Boltzmann machine with prognosability regularization for prognostics and health assessment. *IEEE T Ind Electron* 2016; 63(11): 7076–7083.
- Deutsch J and He D. Using deep learning-based approach to predict remaining useful life of rotating components. *IEEE T Syst Man Cyb* 2018; 48(1): 11–20.
- Ince T, Kiranyaz S, Eren L, et al. Real-time motor fault detection by 1-D convolutional neural networks. *IEEE T Ind Electron* 2016; 63(11): 7067–7075.
- Guo X, Chen L and Shen C. Hierarchical adaptive deep convolutional neural network and its application to bearing fault diagnosis. *Measurement* 2016; 93: 490–502.
- Abdeljaber O, Avci O, Kiranyaz S, et al. Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. *J Sound Vib* 2017; 388: 154–170.
- Jing L, Zhao M, Li P, et al. A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox. *Measurement* 2017; 111: 1–10.
- Zhang W, Peng G, Li C, et al. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors* 2017; 17(2): 425.
- Cabrera D, Sancho F, Li C, et al. Automatic feature extraction of time-series applied to fault severity assessment of helical gearbox in stationary and non-stationary speed operation. *Appl Soft Comput* 2017; 58: 53–64.
- Cody T, Adams S and Beling PA. Unsupervised deep learning for gear health monitoring. In: *Proceedings of the conference of the prognostics & health management*, 2018. [AQ: 3]
- Malhi A, Yan R and Gao RX. Prognosis of defect propagation based on recurrent neural networks. *IEEE T Instrum Meas* 2011; 60(3): 703–711.

23. Tse P and Atherton D. Prediction of machine deterioration using vibration based fault trends and recurrent neural networks. *J Vib Acoust* 1999; 121(3): 355–362.
24. Zhao R, Wang J, Yan R, et al. Machine health monitoring with LSTM networks. In: *Proceedings of the IEEE international conference on sensing technology*, Nanjing, China, 11–13 November 2016. New York: IEEE.
25. Zhao R, Wang D, Yan R, et al. Machine health monitoring using local feature-based gated recurrent unit networks. *IEEE T Ind Electron* 2017; 65(2): 1539–1548.
26. Zhao R, Yan R, Chen Z, et al. Deep learning and its applications to machine health monitoring. *Mech Syst Signal Pr* 2019; 115: 213–237.
27. Rahmounnea C and Benazzouz D. Early detection of pitting failure in gears using a spectral kurtosis analysis. *Mech Ind* 2016; 13: 245–254.
28. Feki N, Cavoret J, Ville F, et al. Gear tooth pitting modelling and detection based on transmission error measurements. *Eur J Comput Mech* 2013; 22: 106–119.
29. Liu J and Wang G. A multi-step predictor with a variable input pattern for system state forecasting. *Mech Syst Signal Pr* 2009; 23: 1586–1599.
30. Lee SK, Shim JS and Cho BO. Damage detection of a gear with initial pitting using the zoomed phase map of continuous wavelet transform. *Key Eng Mater* 2006; 306–308: 223–228.
31. Ozturk H, Yesilyurt I and Sabuncu M. Detection and advancement monitoring of distributed pitting failure in gears. *J Non Destruct Eval* 2010; 29: 63–73.
32. Lewicki DG, Dempsey PJ, Heath GF, et al. Gear fault detection effectiveness as applied to tooth surface pitting fatigue damage. In: *Proceedings of the American Helicopter Society 65th annual forum*, Grapevine, TX, 27–29 May 2009 [AQ: 4].
33. Persin G, Viintin J and Juric D. Gear pitting detection based on spectral kurtosis and adaptive denoising filtering. In: *Proceedings of the 11th international conference on condition monitoring and machinery failure prevention technologies (CM 2014/MFPT)*, Manchester, 10–12 June 2014 [AQ: 5].
34. Ümütlu R, Rafet C, Hizarci B, et al. Pitting detection in a worm gearbox using artificial neural networks. In: *Proceedings of the INTER-NOISE 2016—45th international congress and exposition on noise control engineering: towards a quieter future*, Hamburg, 21–24 August 2016 [AQ: 6].
35. Bengio Y, Courville A and Vincent P. Representation learning: a review and new perspectives. *IEEE T Pattern Anal Mach Intell* 2013; 35(8): 1798–1828.
36. Hinton G. To recognize shapes, first learn to generate images. *Prog Brain Res* 2007; 165: 535–547.
37. Yang Y, Xu D, Nie F, et al. Image clustering using local discriminant models and global integration. *IEEE T Image Pr* 2010; 19(10): 2761–2773.
38. Vasan AS, Long B and Pecht M. Diagnostics and prognostics method for analog electronic circuits. *IEEE T Ind Electron* 2013; 60(11): 5277–5291.
39. McFadden PD. A revised model for the extraction of periodic waveforms by time domain averaging. *Mech Syst Signal Pr* 1987; 1(1): 83–95.