Multiscale Representations Fusion With Joint Multiple Reconstructions Autoencoder for Intelligent Fault Diagnosis

Hui Yu, Kai Wang, *Member, IEEE*, and Yan Li

***Abstract*—Existing intelligent fault diagnosis methods depend mostly on single-scale vibration signals, which not only ignore the latent useful information of other different scales, but also underestimate the complementary benefits across scales. In this work, we show the advantage of learning the combination of multiscale information by aiming to automatically capture complementary and discriminative feature representations from different scales of vibration signals. Specifically, we combine the merit of different activation functions in a joint learning fashion and propose a novel joint multiple reconstructions autoencoder (JMRAE), whose training objective is to jointly optimize multiple reconstruction losses. The JMRAE aims to jointly learn discriminative and robust scale-specific feature representations. In addition, we design a new multiscale representations fusion network (MRFN) model to effectively fuse multiscale feature representations learned concurrently by per-scale JMRAE model for maximizing the discriminative capability of the scale-fused features. The objective of MRFN is to benefit all different scales from each other for improving the classification performance. Extensive experimental results demonstrate the effectiveness of the proposed method on two rolling bearing datasets. The code of the proposed method is available at https://github.com/KWflyer/mrfn.**

***Index Terms*—Autoencoder, features fusion, intelligent fault diagnosis, joint learning, multiscale representations learning.**

I. Introduction

INTELLIGENT fault diagnosis aims at recognizing the fault categories according to the extracted features from the measured vibration signals [1], [2]. This is an extremely challenging classification task because the measured vibration signals may change dramatically in different operational conditions, e.g., working loads, shaft speeds [3]–[7]. Traditional intelligent fault diagnosis methods mainly focus on designing the hand-crafted features, which largely rely on much prior knowledge about signal processing techniques and diagnostic expertise, such as Wavelet Transform [8], Empirical Mode Decomposition (EMD) [9], [10] and Short Time Frequency Transform (STFT) [11]. Recent advances have shown that deep learning architectures such as convolutional neural network (CNN) [12], deep belief network (DBN) [13], stacked denoising autoencoders (SDAE) [14] and recurrent neural network (RNN) [15] have achieved unprecedented success in the field of fault diagnosis with the advantage of automatically learning the high-level features. In particular, autoencoder (AE) has been gradually emerged as a significant group of specialized research for fault diagnosis due to its simplicity and efficiency [16]–[18]. Although these methods have achieved reasonably good performance, there are still some limitations. For example, existing methods based on AE often learn feature representations from input data with only one activation function, which not only ignore the joint benefits and synergistic effects of multiple activation functions. To that end, we propose a novel joint multiple reconstructions autoencoder (JMRAE) network by jointly optimizing multiple reconstruction losses. The mutual complementary benefits of the sigmoid and rectified linear units (RELU) activation functions allow our JMRAE to jointly learn more discriminative and robust feature representations than using a single activation function. Furthermore, existing intelligent fault diagnosis approaches mainly focus on single-scale vibration signals, ignore the latent useful information of other different scales. Therefore, we design a novel multiscale representations fusion network (MRFN), which aims to concurrently learn complementary and discriminative multiscale representations from different scales of vibration signals by all scale-specific JMRAE models, and then obtain the scale-fused features of the vibration signal. Extensive experimental results validate our method achieves more reliable fault diagnosis performance.

Manuscript received August 22, 2018; revised October 11, 2018; accepted October 18, 2018. Date of publication October 29, 2018; date of current version November 8, 2018. This work was supported in part by the National Natural Science Foundation of China under Grants 51435011 and 51505309; and in part by the Sichuan Province Science and Technology Support Program Project under Grant 2018JY0588. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wei Li. *(Corresponding author: Kai Wang.)*

The authors are with the School of Manufacturing Science and Engineering, Sichuan University, Chengdu 610065, China (e-mail: hui_yu_1130@163.com; kai.wang@scu.edu.cn; liyan@scu.edu.cn).

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

Digital Object Identifier 10.1109/LSP.2018.2878356

II. Proposed Method

A. Joint Multiple Reconstructions Autoencoder (JMRAE)

Unlike the traditional AE network which usually uses a single activation function to reconstruct the original input data [19], [20], e.g., tanh, sigmoid, linear, and RELU, this ignores their joint benefits. In this work, we formulate a JMRAE model to jointly optimize multiple reconstruction losses embedded into the objective function by synergistically combining the merit of two different activation functions. The goal of the JMRAE is to capture salient information and jointly learn discriminative and robust feature representations from the input data. The architecture of the proposed JMRAE is depicted in Fig. 1. Specifically, this JMRAE model consists of two branches, namely the sigmoid branch and RELU branch. Each branch shares the same

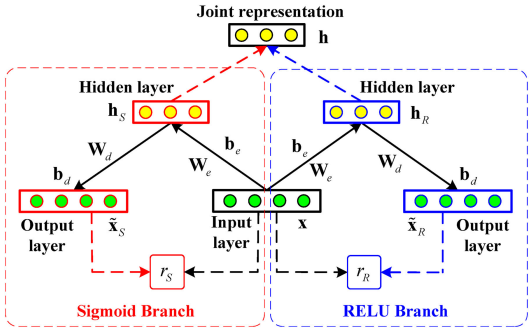

Fig. 1. The architecture of the proposed JMRAE.

architecture with AE network, which consists of an input layer, a hidden layer, and an output layer. The two branches reconstruct the same input vector $\mathbf{x}$ in the output layer using the sigmoid and RELU activation functions, respectively and learn respective feature representations in the hidden layer from the input data concurrently. The two different hidden representations $\mathbf{h}_S, \mathbf{h}_R$ and the corresponding reconstructed vectors $\tilde{\mathbf{x}}_S, \tilde{\mathbf{x}}_R$ can be calculated as follows:

$$\mathbf{h}_S = f_S(\mathbf{W}_e\mathbf{x} + \mathbf{b}_e), \mathbf{h}_R = f_R(\mathbf{W}_e\mathbf{x} + \mathbf{b}_e) \quad (1)$$

$$\tilde{\mathbf{x}}_S = f_S(\mathbf{W}_d\mathbf{h}_S + \mathbf{b}_d), \tilde{\mathbf{x}}_R = f_R(\mathbf{W}_d\mathbf{h}_R + \mathbf{b}_d) \quad (2)$$

where $f_S(z) = 1/[1 + \exp(-z)]$ and $f_R(z) = \max(0, z)$ are the sigmoid and RELU activation functions respectively. Thus, the parameter set of the JMRAE is $\boldsymbol{\theta} = \{\mathbf{W}_e, \mathbf{b}_e, \mathbf{W}_d, \mathbf{b}_d\}$, where $\mathbf{W}_e \in \Re^{m\times n}$ and $\mathbf{W}_d \in \Re^{n\times m}$ are the weight matrixs, and $\mathbf{b}_e \in \Re^m$ and $\mathbf{b}_d \in \Re^n$ are the bias vectors.

Therefore, the two reconstruction losses r_S, r_R at two branches are defined as:

$$r_S = \|\mathbf{x} - \tilde{\mathbf{x}}_S\|_2^2, r_R = \|\mathbf{x} - \tilde{\mathbf{x}}_R\|_2^2 \quad (3)$$

For concurrently optimizing the learning of per-branch discriminative features, a joint learning scheme that both sigmoid and RELU branches not only share the same network parameters but also are synergistically correlated is considered. To that end, the JMRAE model is designed to perform joint multiple reconstructions on the shared network parameters by endowing each branch with a separate reconstruction loss. Mathematically, a novel objective function of the JMRAE model is computed by the linear combination of reconstruction losses of two branches. In order to prevent overfitting, the weight regularization term J_{weight} is added, and the final objective function of JMRAE can be defined as follows:

$$\begin{aligned}
&\min_{\boldsymbol{\theta}} J_{\text{JMRAE}} \\
&= \frac{1}{2N}\sum_{i=1}^{N}\left[\eta r_S^{(i)} + (1-\eta)\, r_R^{(i)}\right] + \lambda J_{\text{weight}} \\
&= \frac{1}{2N}\sum_{i=1}^{N}\left[\eta\left\|\mathbf{x}^{(i)} - \tilde{\mathbf{x}}_S^{(i)}\right\|_2^2 + (1-\eta)\left\|\mathbf{x}^{(i)} - \tilde{\mathbf{x}}_R^{(i)}\right\|_2^2\right] \\
&\quad + \frac{\lambda}{2}\left(\|\mathbf{W}_e\|_{\text{F}}^2 + \|\mathbf{W}_d\|_{\text{F}}^2\right) \quad (4)
\end{aligned}$$

where N denotes the total number of training samples, $0 < \eta < 1$ is a trade-off parameter to control the balance between the two reconstruction loss terms, and λ is a positive regularization parameter.

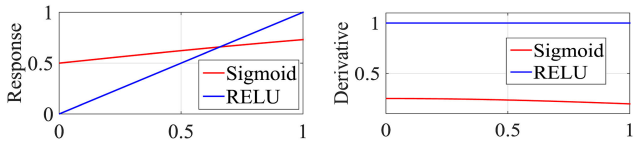

Fig. 2. The activation responses (a) and derivatives (b) of the sigmoid and RELU activation functions concerning an input in the range of [0, 1] .

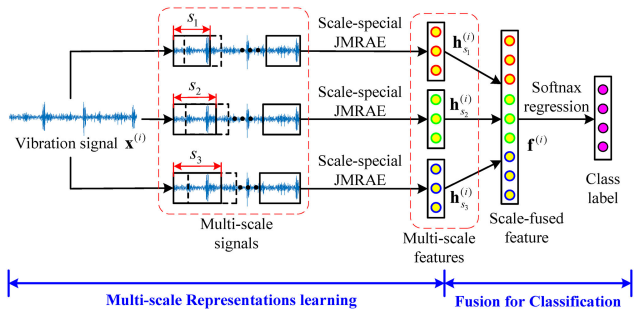

Fig. 3. Overall architecture of the proposed MRFN for fault classification.

The adaptive L-BFGS optimization algorithm with back propagation is applied to train our JMRAE model, and we can obtain the joint representation $\mathbf{h}$ of the input $\mathbf{x}$ by the combination of per-branch hidden representations.

$$\begin{aligned}
\mathbf{h} &= \eta\mathbf{h}_S + (1-\eta)\,\mathbf{h}_R \\
&= \eta f_S(\mathbf{W}_e\mathbf{x} + \mathbf{b}_e) + (1-\eta)\, f_R(\mathbf{W}_e\mathbf{x} + \mathbf{b}_e) \quad (5)
\end{aligned}$$

Our JMRAE with different activation functions can learn different characteristic of the input data. Intuitively, Fig. 2 shows that the RELU function has a wider range of activation responses and yields much larger derivatives than the sigmoid function in the input range of [0, 1] , which indicates that the RELU function is more sensitive to small perturbations in the input, and thus this function is prone to explore detailed and local features of the input. In contrast, the sigmoid function with the more concentrated distribution of activation responses and lower derivatives is more insensitive and invariant to small input changes and can ignore these detailed changes, which is more suitable for capturing relative global and robust information of the input.

In this work, through the joint optimizing both reconstruction losses of RELU and sigmoid branches in (4) and the synergistic integration of the two activation functions in the feature learning in (5), the proposed JMRAE can not only optimize the joint learning of local and global discriminative features, but also preserve both local saliency and global robustness in the learned hidden representations concurrently, therefore maximizing the discriminative power of the learned representations from vibration signals.

B. Multiscale Representations Fusion Network (MRFN)

The overall framework of the proposed MRFN is depicted in Fig. 3. This MRFN model consists of three scale-specific subnetworks, including the following two stages:

1) Multiscale Representations Learning: The three subnetworks can concurrently learn the discriminative feature representations with three scale-specific JMRAE models from the per-scale signal. Specifically, we train the scale-specific JMRAE model on the per-scale signal, which aims to capture the

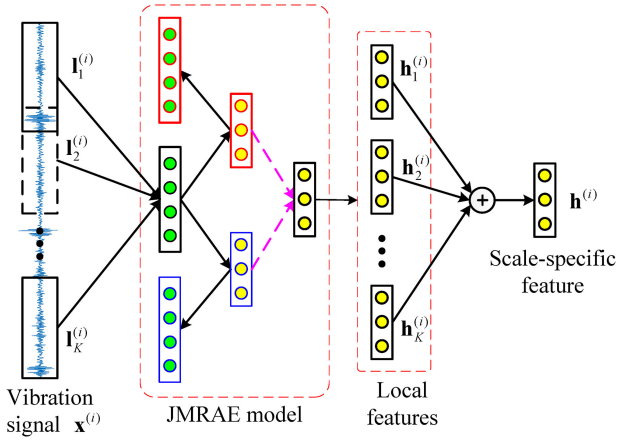

Fig. 4. Overview of the scale-specific representation learning.

intrinsic useful information at each scale and learn discriminative scale-specific representations. The overall framework of the scale-specific representation learning is depicted in Fig. 4. Given a set of N training vibration signals $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$, we alternately slice every vibration signal $\mathbf{x}^{(i)}$ into K local fragments $\{\mathbf{l}_j^{(i)}\}_{j=1}^{K}$ with a fixed size of overlapping windows. All fragments on whole training vibration signals compose the new training set, which are used to train the scale-specific JMRAE model. Once the JMRAE model is learned, we can obtain the feature of each fragment $\mathbf{l}_j^{(i)}$

$$\mathbf{h}_j^{(i)} = \eta f_S(\mathbf{W}_e \mathbf{l}_j^{(i)} + \mathbf{b}_e) + (1-\eta)\, f_R\left(\mathbf{W}_e \mathbf{l}_j^{(i)} + \mathbf{b}_e\right) \quad (6)$$

Then, the scale-specific feature $\mathbf{h}^{(i)}$ of each vibration signal $\mathbf{x}^{(i)}$ is obtained by averaging K local features $\{\mathbf{h}_j^{(i)}\}_{j=1}^{K}$.

$$\mathbf{h}^{(i)} = \frac{1}{K}\sum_{j=1}^{K} \mathbf{h}_j^{(i)} \quad (7)$$

This is motivated by the average pooling of CNN. This intuition for the averaging operation aims to eliminate the shift-variance of vibration signals [21], [22] whilst discovering and capturing more discriminative feature representations.

2) Multiscale Representations Fusion: We then apply a feature-level fusion technique with the aggregation operation to combine the benefits of different scale feature representations. Specifically, the multiscale feature representations learned by all scale-specific JMRAE models are concatenated into a vector as the scale-fused feature, which cumulatively contains salient and complementary information of different scales. Finally, the fused feature is input into the softmax regression [23], [24] for fault classification. The scale-fused feature of the vibration signal $\mathbf{x}^{(i)}$ is given by:

$$\mathbf{f}^{(i)} = \left[\mathbf{h}_{s_1}^{(i)}; \mathbf{h}_{s_2}^{(i)}; \mathbf{h}_{s_3}^{(i)}\right] \quad (8)$$

where $\mathbf{h}_{s_t}^{(i)} (t = 1, 2, 3)$ denotes the scale-specific feature of the scale s_t of the vibration signal $\mathbf{x}^{(i)}$.

III. Experiments and Results

To investigate the performance of the proposed method, we conduct the experiments on rolling element bearing dataset provided by the Case Western Reserve University (CWRU) [25] and Machinery Failure Prevention Technology (MFPT) [26].

A. Data Preparation

The CWRU dataset was a series of vibration signals collected under four different shaft speeds of the motor (1730, 1750, 1772 and 1797 rpm). The dataset contains 10 classes of health conditions (1 normal, 3 inner race faults, 3 outer race faults and 3 roller faults). There are 100 samples for each health condition under one speed, and each sample contains 1200 data points. Therefore, the dataset totally contains 4000$(100 \times 4 \times 10)$ samples. The MFPT dataset contains 3 classes: normal, outer race fault, and inner race fault, but the vibration signals were collected under 7 different loads. In our experiment, there are 840 samples for each class, and each sample contains 1200 data points. Therefore, the dataset totally contains 2520 (840×3) samples. Note that the same health condition under different loads and speeds is treated as one class.

B. Experimental Setup

For model training, all weights are randomly initialized (including the weights of softmax regression), and all biases are set to zeros at the beginning of training. All parameters of three scale-specific JMRAE models are concurrently optimized at different scales in a parallel fashion, and thus obtaining the scale-fused features about the training dataset, which are directly used to train the softmax regression. In the test procedure, the trained softmax regression model is used to predict the corresponding class labels with the extracted scale-fused features on the test dataset. In our experiments, the optimal trade-off parameters η and λ of the proposed JMRAE model were selected by grid search in the range $[0.1, 0.9]$ with a step size of 0.1 and $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ respectively, and then optimal η and λ are respectively fixed as 0.6 and 10^{-4} in all the remaining experiments. For simplicity, the number of units in the hidden layer of JMRAE is set to the same as the input layer according to the size of each scale. The regular parameter of the softmax regression is empirically fixed to 10^{-5}. The input data fed to the proposed JMRAE are first pre-processed by normalizing each input to the scale of [0, 1] . For each dataset, we randomly select 10% of samples per class for training, and the remaining samples are used for test. To avoid bias, independently random partitions are conducted for 20 times. The average classification accuracy and standard deviation are reported as the final classification performance.

C. Results and Analysis

We first evaluated the classification performance of the proposed MRFN model under different scales (corresponding to different lengths of the fragment) and the fusion of different scales on the CWRU dataset. Specifically, for 5 single-scale settings: $s_1 = 50$, $s_2 = 75$, $s_3 = 100$, $s_4 = 125$, and $s_5 = 150$, it could be observed that the MRFN with the single-scale representation learning (black bars as shown in Fig. 5) such as s_1, s_2 and s_3 performs relatively better (above 98.5%) than with the single scale s_4 and s_5 (below 98%). While, as for 10 two-scale fusions, it is shown that the average classification accuracy of the MFRN with any of the two-scale fusion (green bars, i.e., s_{12}) achieves a significant improvement than the MFRN with each of corresponding single scale (i.e., s_1 and s_2). Moreover,

TABLE I
COMPARISON OF CLASSIFICATION PERFORMANCE WITH DIFFERENT ACTIVATION FUNCTION UNDER DIFFERENT SCALES FUSION

Datasets	Methods	s_1	s_2	s_3	s_{12}	s_{13}	s_{23}	s_{123}
CWRU	Sigmoid	69.98±1.60	69.73±1.56	67.04±1.51	75.40±1.45	74.59±1.46	72.31±1.36	75.59±1.47
	RELU	33.04±1.08	32.15±0.94	32.98±0.98	31.63±0.76	33.08±1.03	32.08±0.91	32.24±0.95
	JMRAE-wd	**92.53±1.76**	**90.24±1.93**	**86.07±1.96**	**94.95±1.27**	**93.65±1.75**	**91.14±1.74**	**95.21±1.59**
	JMRAE	**98.73±0.44**	**98.88±0.45**	**98.82±0.45**	**99.64±0.15**	**99.57±0.19**	**99.50±0.26**	**99.73±0.13**
MFPT	Sigmoid	77.65±2.50	74.07±1.60	72.26±1.73	79.12±2.05	77.68±1.96	75.45±1.26	78.47±1.65
	RELU	80.12±1.00	79.24±0.75	78.27±0.90	78.94±1.13	78.55±0.88	77.39±1.15	77.71±1.40
	JMRAE-wd	**91.61±2.24**	**91.97±2.00**	**91.06±1.60**	**94.58±1.49**	**94.83±1.36**	**93.77±1.35**	**95.14±1.41**
	JMRAE	**94.01±1.21**	**95.19±0.97**	**93.50±1.41**	**96.19±0.89**	**96.04±1.21**	**96.21±0.92**	**96.74±0.77**

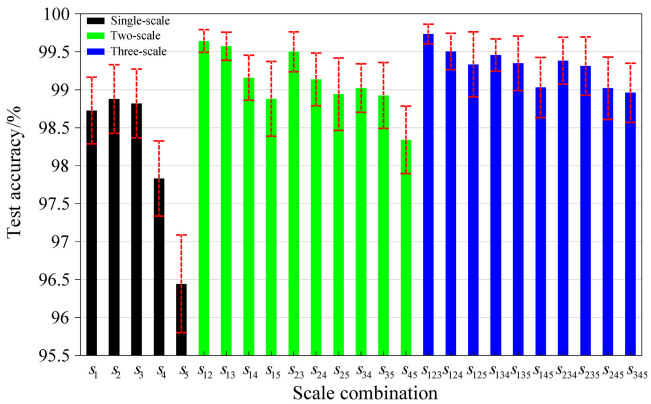

Fig. 5. Classification performance of the proposed MRFN on the CWRU dataset. Experiments are performed by considering the different scales and fusion of different scales. For example, s_{12} means the fusion of the scale s_1 and s_2. And, s_{123} means the fusion of the scale s_1, s_2 and s_3.

it is easily observed that with more scales are fused, such as three-scale fusion, the proposed MRFN can achieve better and more reliable classification performance. We should also note that the proposed MRFN with the three-scale fusion s_{123} obtains the highest average classification accuracy and the lowest standard deviation, followed by the MRFN with the two-scale fusion s_{12}. This means that the MRFN with the fusion of scale s_1, s_2 and s_3 can discover and capture more complementary and meaningful information of vibration signals than the fusion of scale s_4 and s_5. Therefore, based on results above, we only choose the single scale s_1, s_2 and s_3 as well as their fusions for performance analysis of our proposed MRFN in the remaining experiments.

To better understand the advantage of multiscale representations fusion in the proposed MRFN based on JMRAE, we compare the classification accuracy of the proposed JMRAE model with different activation functions by fusing different scale representations on the CWRU and MFPT datasets, respectively. The comparison results are reported in Table I. It is easily observed that the JMRAE by jointly incorporating different activation functions consistently exhibits a more remarkable advantage than only using the sigmoid or RELU activation function over all different scale fusions. Besides, JMRAE consistently achieves better performance than JMRAE-wd (remove the term J_{weight}) over all different scale fusions, which is mainly due to the fact that the weight regularization term prevents overfitting and further improves the generation ability of our JMRAE model. In addition, we can find that our MRFN model with three-scale fusion obtains highest test accuracy ($99.73 \pm 0.13\%$ and $96.74 \pm 0.77\%$ on the CWRU and MFPT datasets respectively), which significantly outperforms the MRFN with single-scale or two-scale fusion. This results further prove that the proposed MRFN can learn complementary and more discriminative representations due to the incorporation of multiscale information of vibration signals.

TABLE II
COMPARISON OF CLASSIFICATION PERFORMANCE BETWEEN THE PROPOSED METHOD AND STATE-OF-ART METHODS

Methods	No. of health condition	No. of shaft speeds	No. of training samples/%	Test accuracy/%
[21]	10	4	10	99.66±0.19
[21]	10	4	1	85.10
[27]	10	4	50	91.33
[28]	10	1	75	88.90
[29]	4	4	40	95.80
[30]	11	4	40	97.91±0.09
MRFN	**10**	**4**	**10**	**99.73±0.13**
MRFN	**10**	**4**	**1**	**98.49±1.10**

To further demonstrate the effectiveness and superiority of the proposed MRFN, we compare the test accuracy of our MRFN with the originally reported best diagnosis accuracies of several state-art-the-art methods in the literature on the CWRU dataset. The comparison results are summarized in Table II. We can find that our MRFN achieves a more remarkable classification performance of $99.73 \pm 0.13\%$ over other compared methods with 10% of training samples, where four different shaft speeds are considered. Especially when the percentage of training samples decreases to 1%, our proposed MRFN still achieves the higher classification accuracy ($98.49 \pm 1.10\%$), which significantly outperforms Ref. [21] (85.10%) with an improvement of about 13%. This result demonstrates that the proposed MRFN achieves higher classification accuracy for intelligent fault diagnosis.

IV. CONCLUSION

In this letter, we presented a novel MRFN model based on the proposed JMRAE for intelligent fault diagnosis. In contrast to existing single-scale approaches, the proposed MRFN is capability of learning the complementary combination of multiscale information by effectively fusing multiscale representations. Moreover, we constructed a novel JMRAE model by jointly optimizing multiple reconstruction losses in the objective function, which aims to jointly learn more discriminative and robust scale-specific feature representations at different scales. Experimental results have validated the effectiveness and superiority of our method for fault diagnosis.

REFERENCES

[1] F. Jia, Y. Lei, J. Lin, X. Zhou, and 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.*, vol. 72/73, pp. 303–315, May 2016.

[2] L. Song, P. Chen, H. Wang, and M. Kato, "Intelligent condition diagnosis method for rotating machinery based on probability density and discriminant analyses," *IEEE Signal Process. Lett.*, vol. 23, no. 8, pp. 1111–1115, Aug. 2016.

[3] F. Chaari, W. Bartelmus, R. Zimroz, T. Fakhfakh, and M. Haddar, "Gearbox vibration signal amplitude and frequency modulation," *Shock Vib.*, vol. 19, no. 4, pp. 635–652, 2012.

[4] F. Chaari, W. Bartelmus, R. Zimroz, T. Fakhfakh, and M. Haddar, "Effect of load shape in cyclic load variation on dynamic behavior of spur gear system," *Key Eng. Mater.*, vol. 518, pp. 119–126, 2012.

[5] W. Bartelmus and R. Zimroz, "A new feature for monitoring the condition of gearboxes in non-stationary operating conditions," *Mech. Syst. Signal Process.*, vol. 23, no. 5, pp. 1528–1534, 2009.

[6] P. Boskoski and D. Juricic, "Fault detection of mechanical drives under variable operating conditions based on wavelet packet renyi entropy signatures," *Mech. Syst. Signal Process.*, vol. 31, pp. 369–381, 2012.

[7] J. Urbanek, T. Barszcz, R. Zimroz, and J. Antoni, "Application of averaged instantaneous power spectrum for diagnostics of machinery operating under non-stationary operational conditions," *Measurement*, vol. 45, no. 7, pp. 1782–1791, 2012.

[8] D. You, X. Gao, and S. Katayama, "WPD-PCA-based laser welding process monitoring and defects diagnosis by using FNN and SVM," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 628–638, Jan. 2015.

[9] J. Dybała and R. Zimroz, "Rolling bearing diagnosing method based on empirical mode decomposition of machine vibration signal," *Appl. Acoust.*, vol. 77, pp. 195–203, 2014.

[10] R. Ricci and P. Pennacchi, "Diagnostics of gear faults based on EMD and automatic selection of intrinsic mode functions," *Mech. Syst. Signal Process.*, vol. 25, no. 3, pp. 821–838, Apr. 2011.

[11] R. B. Randall, "Applications of spectral kurtosis in machine diagnostics and prognostics," *Key Eng. Mater.*, vol. 293, pp. 21–32, 2005.

[12] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, "Real-time motor fault detection by 1-D convolutional neural networks," *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7067–7075, Nov. 2016.

[13] M. Gan, C. Wang, and C. Zhu, "Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling elements bearings," *Mech. Syst. Signal Process.*, vol. 72, pp. 92–104, 2016.

[14] C. Lu, Z.-Y. Wang, W.-L. Qin, and J. Ma, "Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification," *Signal Process.*, vol. 130, pp. 377–388, Jan. 2017.

[15] T. de Bruin, K. Verbert, and R. Babuska, "Railway track circuit fault diagnosis using recurrent neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 523–533, Mar. 2017.

[16] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 63, no. 3, pp. 2296–2305, Mar. 2017.

[17] W. Sun, S. Shaoa, R. Zhaob, R. Yana, X. Zhangc, and X. Chen, "A sparse auto-encoder-based deep neural network approach for induction motor faults classification," *Measurement*, vol. 89, pp. 171–178, Jul. 2016.

[18] H. Shao, H. Jiang, H. Zhao, and F. Wang, "A novel deep autoencoder feature learning method for rotating machinery fault diagnosis," *Mech. Syst. Signal Process.*, vol. 95, pp. 187–204, Oct. 2017.

[19] P. Baldi, "Autoencoders, unsupervised learning, deep architectures," in *Proc. Int. Conf. Unsupervised Transfer Learn.*, 2012, vol. 27, pp. 37–50. [Online]. Available: http://jmlr.org/proceedings/papers/v27/baldi12a/baldi12a.pdf

[20] A. Ali and F. Yangyu, "Automatic modulation classification using deep learning based on sparse autoencoders with nonnegativity constraints," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1626–1630, Nov. 2017.

[21] Y. Lei, F. Jia, J. Lin, S. Xing, and S. X. Ding, "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data," *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3137–3147, May 2016.

[22] F. Jia, Y. Lei, L. Guo, J. Lin, and S. Xing, "A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines," *Neurocomputing*, vol. 272, pp. 619–628, Jan. 2018.

[23] A. Ng and J. Ngiam, "UFLDL tutorial, softmax regression," Apr. 2013. [Online]. Available: http://ufldl.stanford.edu/wiki/index.php/Softmax_Regression

[24] C. B. Do and A. Y. Ng, "Transfer learning for text classification," in *Advances in Neural Information Processing System*. Cambridge, MA, USA: MIT Press, 2006, pp. 299–306.

[25] K. A. Loparo, "Case western reserve university bearing data center," 2013. [Online]. Available: http://csegroups.case.edu/bearingdatacenter/pages/download-data-file. Accessed on Mar. 2017.

[26] E. Bechhoefer, "Condition based maintenance fault database for testing of diagnostic and prognostics algorithms," Aug. 2018. [Online]. Available: https://mfpt.org/fault-data-sets/

[27] Y. Lei, Z. He, Y. Zi, and Q. Hu, "Fault diagnosis of rotating machinery based on multiple ANFIS combination with GAs," *Mech. Syst. Signal Process.*, vol. 21, no. 5, pp. 2280–2294, Jul. 2007.

[28] W. Du, J. Tao, Y. Li, and C. Liu, "Wavelet leaders multifractal features based fault diagnosis of rotating mechanism," *Mech. Syst. Signal Process.*, vol. 43, no. 1, pp. 57–75, Feb. 2014.

[29] W. Li, S. Zhang, and G. He, "Semisupervised distance-preserving self-organizing map for machine-defect detection and classification," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 5, pp. 869–879, May 2013.

[30] X. Zhang, Y. Liang, and J. Zhou, "A novel bearing fault diagnosis model integrated permutation entropy, ensemble empirical mode decomposition and optimized SVM," *Measurement*, vol. 69, pp. 164–179, Jun. 2015.