Memory Residual Regression Autoencoder for Bearing Fault Detection

Xin Huang, Guangrui Wen, Shuzhi Dong, Haoxuan Zhou, Zihao Lei, Zhifen Zhang, and Xuefeng Chen, *Senior Member, IEEE*

***Abstract*— Anomaly detection is the cornerstone for the health management of rolling element bearings. The unsupervised learning model for anomaly detection driven only by normal data has received increasing attention in recent years. In this article, an innovative deep-learning-based model, namely, memory residual regression autoencoder (MRRAE), is developed to improve the accuracy of anomaly detection in bearing condition monitoring. The memory module and autoregressive estimator are applied to calculate the probability density distribution of the latent memory residual representation. The reconstruction errors and surprisal values of the proposed model are used to detect the abnormal condition of bearing. To verify the superiority of the proposed method in anomaly detection, two sets of run-to-failure experimental data set gathered from the laboratories are studied and analyzed. The result demonstrates that the proposed MRRAE model achieves superior performance compared with several conventional and deep-learning-based anomaly detection methods. Furthermore, the proposed method pays close attention to the special structure of bearing vibration signal and provides a new way for explaining the decision-making processes of deep neural networks.**

***Index Terms*— Anomaly detection, condition monitoring, frequency-domain analysis, machine learning, rolling bearings.**

I. Introduction

As ONE of the most important components in rotating machinery, rolling element bearings are fragile due to their severe working environment, such as high rotating speed, alternating load, and randomly changing conditions [1]–[3]. These harsh working environment may cause micro defects on the component of bearing, which can gradually develop into failures and cause unexpected accidents. Consequently, real-time monitoring of bearing condition, detection, and isolation of early developing faults as well as prediction of fault propagation are very important research topics [4], [5]. As an effective technology to ensure the safety operation of bearings, Prognostic and Health Management (PHM) is beneficial to avoiding major accidents and reducing the maintenance costs [6].

Manuscript received December 11, 2020; revised March 11, 2021; accepted March 22, 2021. Date of publication April 9, 2021; date of current version May 7, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1710002 and in part by the National Natural Science Foundation of China under Grant 51775409. The Associate Editor coordinating the review process was Dr. Qiang Miao. *(Corresponding author: Guangrui Wen.)*

Xin Huang, Shuzhi Dong, Haoxuan Zhou, Zihao Lei, Zhifen Zhang, and Xuefeng Chen are with the School of Mechanical Engineering, Xi'an Jiao-tong University, Xi'an 710049, China.

Guangrui Wen is with the School of Mechanical Engineering, Xi'an Jiao-tong University, Xi'an 710049, China, and also with the Key Laboratory of Education Ministry for Modern Design and Rotor-Bearing System, Xi'an 710049, China (e-mail: grwen@mail.xjtu.edu.cn).

Digital Object Identifier 10.1109/TIM.2021.3072131

Anomaly detection is the cornerstone of bearing health management, which can be used to determine whether the bearing is working at a normal condition [7]. Once the early appearance of the abnormal state is detected, some subsequent works will be conducted based on the result of anomaly detection. For instance, the sampling frequency of vibration signal can be increased to analyze the cause of abnormal (fault diagnosis). Moreover, the cut-off point between normal class and abnormal class determined by the anomaly detection method can be used as the first predicting time (FPT) [8] in the field of remaining useful life prediction (RUL prediction) [9].

With the development of information technology and sensor technology, the quantity of data in the industrial field has been increased greatly, but the quality has not been correspondingly increased. For instance, bearings usually work in healthy state for a long time, and a large number of normal data can be collected, but sufficient abnormal data are often unavailable [10]. Even if occasional faults emerge, the lack of corresponding failure tags will limit the application of these data. Therefore, it is impractical to train the machine learning models with multiple classes' data, which restricts the application of PHM in industrial fields.

Therefore, the unsupervised learning model for anomaly detection driven only by normal data has received increasing attention in recent years, which is essentially a one-class classification task. The main solution to this task is mapping the normal data distribution to high-dimensional hyperplane constructed by rule-based approach or neural network.

Classification-based approaches aim to construct representative one-class decision boundaries such as hyperplanes or hyperspheres [11] around the normal distribution to detect abnormal samples [12]. Liu and Gryllias [13] used cyclic spectral indicators to build the feature space, and the Euclidean distance in hypersphere was applied to isolate the healthy and faulty data; the experimental results show that the support vector data description model with cyclic spectral coherence indicators can precisely detect the bearing faults. Zhu *et al.* [14] proposed a novel one-class classifier, namely, rough support vector data description, to assess the performance degradation of bearing, which can solve the over-fitting problem that exists in support vector data description. A robust one-class support vector machine for bearing fault detection is proposed in [15], which designed penalty factors to depress the influence of outliers. The simulation example shows that this method is superior to the general one-class support vector machine, especially when the training data set is corrupted by outliers. Saari *et al.* [16] used one-class support vector machine to detect and identify wind turbine bearing faults with the assistant of fault-specific features extracted from vibration signals. In [17], a novel method, namely, one-class classification-based convex hull, was proposed for bearing fault detection, which calculated the nearest point to the origin from the reduced convex hull of training samples. The experimental results demonstrate that this method performs more efficiently than one-class SVM (OCSVM).

However, the anomaly detection performance of classification-based method can be affected when the data points lay in a high-dimensional space, as modeling the high-dimensional data is notoriously challenging [18], [19].

Reconstruction-based methods are another basic approach currently being adopted in anomaly detection, which usually learn a parametric projection and reconstruction model of normal data [20]. As one of the representative models of deep learning, deep autoencoder (AE) [21] is a powerful tool to process the high-dimensional data in the unsupervised setting. A novel framework integrating adaptive sparse contractive autoencoder algorithm and optimized unsupervised extreme learning machine is presented in [22], which can achieve effective sparse and sensitive feature information extraction to avoid over-fitting. A semisupervised deep sparse autoencoder with local and nonlocal information is proposed in [23] to extract the fault feature. The weighted cross-entropy techniques are applied to improve the generalization performance of this model for labeled data. In [24], a three-layer sparse autoencoder is constructed to extract the features of vibration signal, and the prediction accuracy is better than the results of other comparison algorithms. A novel framework named data-enhanced stacked autoencoders is proposed in [25], in which stacked autoencoders can generate simulated signals to augment the insufficient training data. In the field of anomaly detection, AE is usually trained by minimizing the reconstruction errors on normal data [18]. Thus, the distribution of normal data can be recorded in the structure and parameters of AE. In the testing stage, the reconstruction error of abnormal sample will increase definitely. A new approach integrating AE and on-line sequential extreme learning machine is proposed in [26] for on-line condition monitoring of bearing. The simple design of this method is promising for easy hardware implementation in industrial applications. Chen *et al.* [27] used stacked denoising AEs to detect the abnormal condition of wind turbines based on the reconstruction of condition parameters. Demonstration on real SCADA data shows that this method is effective for anomaly detection and early warning of an actual wind turbine. Xu *et al*. [28] proposed a moving window-based stacked autoencoder with an exponential function, which incorporates a slope local minimum point to extract the degradation trends of roller bearing.

However, limited by the simple structure of AE, the reconstruction error of some abnormal inputs is not changed, especially when the abnormal samples are not much different from the normal samples.

The running state of the bearing is a continuous process. When failures occur in the component of bearing, there is no significant change in the time domain and frequency domain of the vibration signal compared with the normal state. The autoencoder may misjudge an early abnormal state as a normal state, which limits its application in bearing anomaly detection. VAEs are directed probabilistic graphical models whose posterior is approximated by a neural network, forming an AE-like architecture [29], [30]. They use a variational approach for latent representation learning, which results in an additional loss component and a specific estimator for the training algorithm called the Stochastic Gradient Variational Bayes estimator [31]. VAEs have been applied for anomaly detection in many research fields [32]–[34]. However, in the context of anomaly detection, VAE also has the problem of being insensitive to subtle changes.

To mitigate the drawback of AE (VAE), a recent research trend considers using generative adversarial networks (GANs) for anomaly detection [35]–[39]. During the training stage, the optimization algorithm is used to make the discriminator unable to distinguish between normal samples and reconstructed samples. During the testing stage, the reconstruction error and the output features of the discriminator are used as the indicator for anomaly detection. Wu *et al.* [10] proposed a fault-attention generative probabilistic adversarial AE (FGPAA) and constructed the fault-attention abnormal state indictor to achieve high accuracy in machine anomaly detection. A novel anomaly detection approach based on GANs is proposed in [40] to distinguish abnormal samples from normal samples. However, the simultaneous training of generator and discriminator models in GANs is inherently unstable. A lot of trial and error is required regarding the network structure and training strategy.

On the other hand, it is crucial to understand the reasons behind the decision made by neural networks. In recent years, attention mechanism has been widely studied and applied in the field of prognostic and health management. Li *et al*. [41] provided a new perspective in understanding the hidden mechanism of intelligent fault diagnosis by introducing attention mechanism into the deep neural network. A new attention module considering the characteristics of rolling bearing faults is designed in [42] to enhance fault-related features and to ignore irrelevant features. Multiattention 1-D convolutional neural network is further proposed to diagnose wheelset bearing faults, which outperforms eight state-of-the-art networks. Wang *et al.* [43] proposed a new deep framework named multiscale convolutional attention network, in which self-attention modules are constructed to fuse the input multisensor data. The experimental results demonstrate that this framework achieve superior performance in fusing multisensor information and improving RUL prediction accuracy [43]. An attention-based deep learning framework for machine RUL prediction is proposed in [44], which can assign larger weights to more important features and time steps automatically.

Qin *et al.* [45] proposed gated dual attention unit to estimate the RUL of the rolling bearing, which achieved higher prediction accuracy and convergence speed than several conventional RUL prediction methods. An attention-based LSTM network is proposed in [46] for machine RUL prediction, in which the attention mechanism can solve the problem of information loss in the long-distance signal transmission of LSTM effectively. Liu *et al.* [47] proposed a novel feature-attention-based RUL prediction approach. The feature-attention mechanism gives greater weights to more important features dynamically which make the prediction model pay more attention to key inputs [47]. However, the current anomaly detection methods of bearing condition monitoring only give the final results. Few writers have been drawn on any systematic research to interpret the decision-making processes of these models. Moreover, there has been little consideration about the special structure and characteristic of bearing vibration signal in anomaly detection. When an abnormal sample is detected, the sensitive frequency band needs to be further analyzed in which the fault symptoms are enhanced by structural resonances.

To address these problems, an innovative deep-learning-based model, namely, memory residual regression AE (MRRAE), is proposed in this article. The 1-D convolutional layer is constructed as the main component of the encoder and decoder part, the memory module is used to calculate a sparse approximation of the latent representation, and the difference between the input and the output of the memory module is input into the autoregressive estimator. The reconstruction errors and surprisal values of the proposed model are used to detect the abnormal condition of bearing, which achieves superior performance than the comparison methods. The contribution of this letter can be summarized as follows.

1) An MRRAE is proposed in this article, and memory module and parametric density estimator are integrated to pinpoint bearing failure at an early stage. Compared with the conventional and deep-learning-based anomaly detection methods, the proposed model shows a better recognition performance.
2) The sensitivity and interpretability of the features in latent space are considered simultaneously via parametric density estimator. In this way, the decision-making processes of the proposed method can be monitored and analyzed.
3) The proposed method takes the characteristics of convolution and the distribution of bearing vibration signals in the frequency domain into account. The abnormal part in the frequency domain can be localized with the assistance of probability density distribution.

The rest of this article is organized as follows. The technical preliminaries are introduced in Section II. The architecture and detailed procedures of the proposed model are presented in Section III. The proposed method is verified on two run-to-failure experimental data sets in Section IV, where the common evaluation metrics is used to describe the performance of different methods quantitatively. Finally, conclusions are drawn in Section V.

II. Technical Preliminaries

A. Autoencoder

AE is one of the classic unsupervised machine learning algorithms. The aim of AE is to generate target values as close as possible to its original input by training the network through gradient backpropagation. The number of units in the hidden layer is less than that in the input and output layers. Therefore, the input samples of AE are compressed into a low-dimensional latent space which is also recognized as information bottleneck.

The AE network is composed of two basic components: the encode and the decoder. As presented in the following equation, an encoder is a neural network f_e that compresses the input signal $\boldsymbol{X}$ into a low-dimensional representation $\boldsymbol{Z}$, where θ_e represents the parameters of encoder:

$$\boldsymbol{Z} = f_e(\boldsymbol{X}; \theta_e). \tag{1}$$

The decoder f_d is a neural network (usually the same network structure as the encoder) which maps the latent space representation $\boldsymbol{Z}$ to the reconstruction sample $\hat{\boldsymbol{X}}$ of the same dimension as the original input $\boldsymbol{X}$

$$\hat{\boldsymbol{X}} = f_d(\boldsymbol{Z}; \theta_d). \tag{2}$$

The AE can be trained by minimizing the reconstruction error, which measures the mean squared error between the input sample $\boldsymbol{X}$ and the reconstruction sample $\hat{\boldsymbol{X}}$

$$\mathcal{L}(\theta_e, \theta_d) = \left\| \boldsymbol{X} - \hat{\boldsymbol{X}} \right\|_2^2. \tag{3}$$

B. Memory-Augmented AE

To improve the robustness of AE in anomaly detection, machine learning strategies with a memory component are applied in the proposed memory-augmented deep autoencoder (MemAE). The latent space representation is decomposed into the product of coefficients and memory atoms, which is determined by normal data in the training stage. Thus, the reconstruction of the decoder tends to be close to the normal distribution which can make the reconstruction errors on abnormal data samples strengthened [18]. This advantage of MemAE offers a strong guarantee for anomaly detection.

The MemAE model includes the following three parts: encoder, memory module, and decoder. The encoder compresses the input samples into latent space representation $\boldsymbol{Z}$ which performs as a query to retrieve relevant atoms in the memory $\boldsymbol{M}$ [18]. The decoder is used to reconstruct the samples from the latent space representation $\hat{\boldsymbol{Z}}$, which is the sparse approximation of the latent representation.

The memory module is defined as a matrix $\boldsymbol{M} \in \mathbb{R}^{N \times C}$ which includes N real-valued vectors of dimension C. The row vector $\boldsymbol{m}_i$ denotes the ith row of $\boldsymbol{M}$, and each of them is recognized as a memory atom. The memory module is a recording component which aims at finding a sparse approximation $\hat{\boldsymbol{Z}}$ of the latent representation $\boldsymbol{Z}$ in the form of a linear combination of memory atoms $\boldsymbol{m}_i$. The principle of this attention-based addressing strategy is shown in (4), where the attention weight $\boldsymbol{A}$ is a row vector with nonnegative entries that sum to one and a_i denotes the ith entry of $\boldsymbol{A}$

$$\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{M} = \sum_{N}^{i=1} a_i \boldsymbol{m}_i. \tag{4}$$

The cosine similarity is used to evaluate the similarity between the latent space presentation $\boldsymbol{Z}$ and each memory atom $\boldsymbol{m}_i$ in the memory module

$$d(z, \boldsymbol{m}_i) = \frac{z\boldsymbol{m}_i^\top}{\|z\|\|\boldsymbol{m}_i\|}. \tag{5}$$

Thus, each attention weight a_i in (4) is obtained via a softmax operation on similarity measurement

$$a_i = \frac{\exp(d(z, \boldsymbol{m}_i))}{\sum_N^{j=1} \exp\big(d\big(z, \boldsymbol{m}_j\big)\big)}. \tag{6}$$

The hard shrinkage operation is applied to promote the sparsity of $\boldsymbol{A}$, which encourages fewer but more relevant memory atoms used to approximate the latent representation

$$\hat{a}_i = s(a_i; \lambda) = \begin{cases} a_i, & a_i > \lambda \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

$\hat{a}_i$ denotes the ith entry of the memory addressing weight vector $\hat{\boldsymbol{A}}$ after shrinkage and λ denotes the shrinkage threshold [18]. After the shrinkage operation, the sparse approximation $\hat{\boldsymbol{Z}}$ can be obtained by $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{A}}\boldsymbol{M}$.

Similar to the dictionary update strategy of sparse dictionary learning, given the normal data, the memory contents are trained to be close to the normal distribution. In the test stage, the learned memory will be fixed, and the reconstruction is calculated from a few selected memory atoms, which can make the reconstruction errors on abnormal data samples strengthened [18]. In this way, the performance of MemAE in anomaly detection can be improved.

C. Latent Space Autoregression

A new generative unsupervised model is proposed in [20] to detect the anomaly by minimizing the reconstruction errors and surprisal values of normal samples. A parametric density estimator is designed to calculate the surprisal values of latent representation via an autoregressive procedure [20].

The probability distribution $p(\boldsymbol{Z})$ of latent representation $\boldsymbol{Z}$ can be factorized as a joint distribution of conditional probability density (CPD) $p(z_i|\boldsymbol{Z}_{<i})$

$$p(\boldsymbol{Z}) = \prod_{i=1}^{d} p(z_i|\boldsymbol{Z}_{<i}) \tag{8}$$

where the symbol $<$ implies an order over random variables.

The probabilistic model $h(\boldsymbol{Z}; \theta_h)$ is designed to estimate the probability distribution $p(\boldsymbol{Z})$ via the autoregressive procedure, in which each output depends on previous observations [48], [49].

To achieve this aim, the masked fully connection is constructed based on the classic fully connected layer. Given the input $\boldsymbol{h} \in \mathbb{R}^{d\times ci}$(assuming $ci = 1$ at the input layer), the output feature map $\boldsymbol{o} \in \mathbb{R}^{d\times co}$ of the masked fully connection

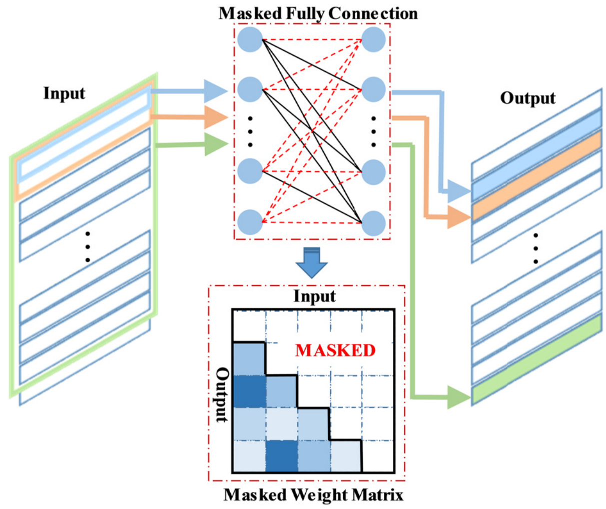

Fig. 1. Schematic of masked fully connection.

is obtained by multiplying the input with a masked weight matrix, which has the same calculation way with the fully connected layer.

The masked weight matrix $\boldsymbol{M}$ is computed by setting the corresponding element of the weight matrix $\boldsymbol{W}$ to 0

$$m_{i,j}^{k,l} = \begin{cases} \omega_{i,j}^{k,l}, & \text{if } i < j \\ \begin{cases} \omega_{i,j}^{k,l}, & \text{if type} = \text{B} \\ 0, & \text{if type} = \text{A}, \end{cases} & \text{if } i = j \\ 0, & \text{if } i > j. \end{cases} \tag{9}$$

Type A forces a strict dependence on previous elements (and is used only as the first estimator layer), whereas type B masks only succeeding elements [20].

As shown in Fig. 1, the red dotted line in the neural network diagram and the right triangular portion of the weight matrix indicate mask operation, which makes the output of MFC in different positions only connected to the specific unmasked neurons. Similar to the concept of the receptive fields in convolutional neural networks, each output of MFC depends on previous observations through this order-stepped mask strategy.

The autoregressive estimator is constructed by stacking multiple MFCs, where the output of the last autoregressive layer provides an approximate estimate of the CPD $p(z_i|\boldsymbol{Z}_{<i})$.

Lowering the surprisal values of the autoregressive estimator for a normal configuration is the same as maximizing the probability of latent representations, which is defined as the negative log-density of the latent space distribution [20]

$$\mathcal{L}(\theta_h) = -\frac{1}{K}\sum_{k=1}^{K} \log(h(z; \theta_h)) \tag{10}$$

where K is the channel number of latent representation and θ_h is the parameters of autoregressive estimator.

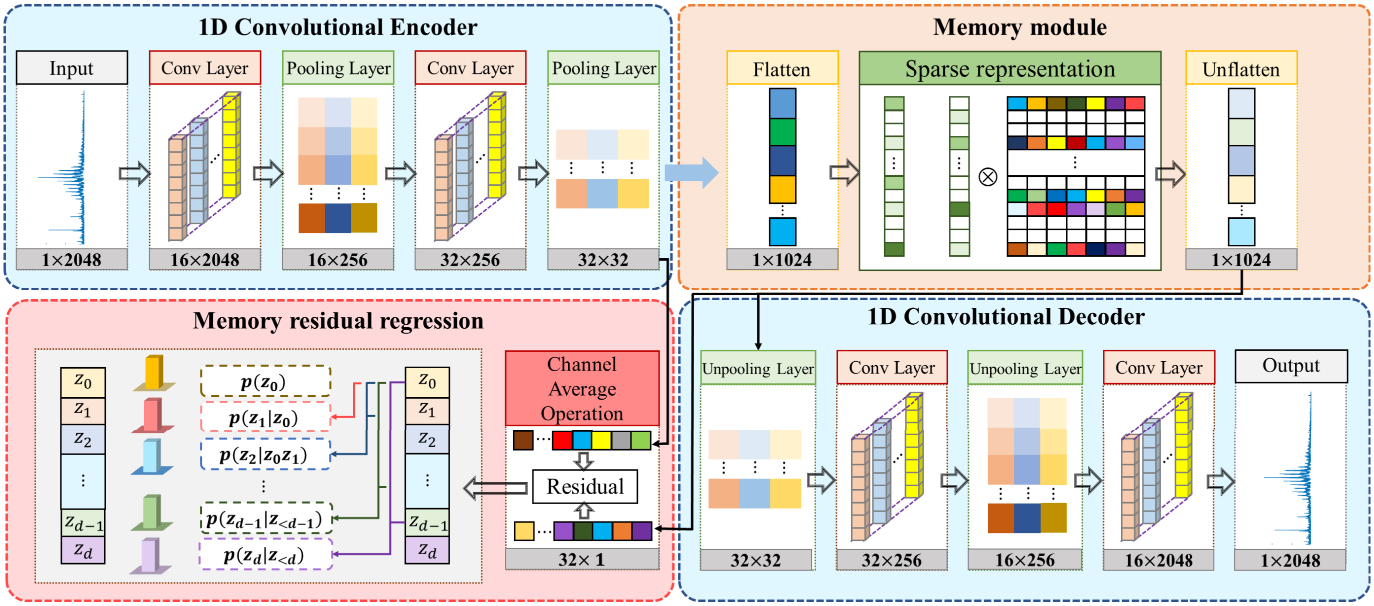

Fig. 2. Architecture of the MRRAE.

III. Methods

A. Architecture of Proposed Model

As presented in Fig. 2, the proposed MRRAE model consists of three parts—the convolutional AE, the memory module, and the latent residual autoregression estimator. First, the input signal is compressed into a low-dimensional representation in the same way as the traditional AE-based anomaly detection method. The memory module is then used to calculate a sparse approximation of the latent representation, which promotes the robustness of the proposed method in anomaly detection. After that, the residual of the memory module is input into the autoregressive estimator, which can compute the surprisal values of different samples. The manifold distribution of the normal samples is recorded in the parameters of the proposed model. The variation in the anomaly indicator implies the appearance of anomalies out of the normal distribution.

B. 1-D Convolutional AE

The 1-D convolutional layer is constructed as the main component of the encoder and decoder part, which can preserve the integrity and the physical meaning of the vibration signals. On the other hand, forward propagation and backpropagation in the 1-D convolutional layer require simple array operations, which means this method is well-suited for real-time and low-cost applications. As shown in Fig. 2, in the convolutional encoder network, the input data pass through two convolutional layers with 7×1 kernels. Sixteen individual 1-D kernels are used in the first layer to compress the input samples into feature maps. Similarly, the feature maps of the second convolutional layer are obtained via 32 1-D kernels. In the proposed model, the rectified linear units (ReLUs) activation function is applied in the convolutional layers. A maximum pooling layer with filters of size 8×1 is constructed after each convolutional layer to reduce the dimensions of feature maps. The padding value of each convolutional layer is set as 3, so the bottleneck has size 32×32 where the first 32 and the second 32 denote the channel numbers and the dimensions of the latent representation, respectively.

The decoder mirrors this architecture with transposed convolutional layers. The feature maps of each convolutional layer are up-sampled by the unpooling layers which can compute a partial inverse of the pooling layers. To recover the vibration signal more accurately, the indices of the maximal values in the encoder pooling layer are recorded and picked as the input of the unpooling layers.

C. Memory Module

To promote the robustness of the proposed method in anomaly detection, the memory module described in Section II is equipped between the encoder and the decoder.

The size of the memory block is set as 100×1024, where 100 and 1024 are the atom number and the dimensions of the memory representation, respectively. The atoms of the memory block are initialized randomly, which will be determined at the training stage. The latent representation matrix with size 32×32 is converted to a vector by flatten operation. The approximation of the latent representation vector was obtained by the memory addressing strategy described in (4) and (5). The approximation vector is converted back into 32×32 matrix and then as input to the decoder, which can generate target values as close as possible to the original input vibration signal. The shrinkage threshold λ in (7) is an important hyperparameter when the sparse representation is calculated at the memory module. The shrinkage threshold λ is set as 0.002 in all experiments of this article.

To avoid the well reconstruction of weakly abnormal signals, the entropy of the addressing weight is minimized at the training stage [18]

$$\Phi\left(\hat{A}\right) = \sum_{m=1}^{M} -\hat{a}_m \cdot \log(\hat{a}_m). \tag{11}$$

Another major advantage of the memory module in the proposed method is that the error between the input and the output of the memory block can be analyzed further, which could generate an additional sensitive indicator to detect the abnormal samples.

D. Memory Residual Regression

As shown in Fig. 2, 32 convolutional kernels are used to transform the feature into latent space in the second convolutional layer of the encoder, and the element in the latent feature vector of each channel represents a nonlinear projection from different region of the frequency band. Thus, the multikernel convolution operation indicates different views of energy distribution in the spectrum.

A concentration vector is obtained by averaging the multichannel latent representations. This averaging strategy can highlight the dominant components between multiple channels and eliminate the influence of random variables introduced by the convolution, which has been applied widely in signal processing [50]–[52] and machine learning [53]. Moreover, the same average operation is conducted on the approximation of the latent representation determined by the memory module.

From the perspective of the reconstruction-based anomaly detection methods, the reconstruction error of the input and the output is applied as the anomaly indicator. Inspired by this idea, the difference between the input and the output of the memory module is considered and studied in this letter. The difference $\boldsymbol{V}_{\text{res}}$ between the concentration vector $\boldsymbol{V}_e$ and the memory approximation concentration vector $\boldsymbol{V}_d$ is input into the parametric density estimator which learns the probability distribution underlying latent representations through an autoregressive procedure

$$\boldsymbol{V}_{\text{res}} = \boldsymbol{V}_e - \boldsymbol{V}_d. \tag{12}$$

Given such modules, the anomaly samples can be detected by monitoring the surprisal values of the latent memory residual representation $\boldsymbol{V}_{\text{res}}$ in the test stage. This operation is conducive to understanding the decision-making processes of the proposed framework. The difference between the input and the output of the memory module in different channels can be monitored and analyzed with the help of the residual part. The same operation on any other layer may weaken the interpretability.

In the proposed architecture, the layer number of the autoregressive estimator is set as one, which means there is only one masked fully connection included in the estimator module. The output channel number of the estimator is set as 50, and the input channel number of the estimator is set as 32, the same as the dimension of the latent representation.

Besides, the element in the latent feature vector $\boldsymbol{V}_{\text{res}}$ represents a nonlinear mapping from different region of the frequency band. Therefore, the probability density distribution $h(\boldsymbol{V}_{\text{res}}; \theta_h)$ of the latent memory residual representation can be used to indicate the energy variations in the spectrum by inverse mapping. In this way, the abnormal part of the spectrum can be detected and localized. Furthermore, the abnormal spectrum band can be applied for envelope demodulation and bearing fault diagnosis. In the past years, envelop analysis has been considered as the benchmark technique for fault diagnosis of rolling element bearings. The vibration signal is processed by a band-pass filter with a selected frequency band in which the fault symptoms are enhanced by structural resonances. The procedures of the envelope spectrum calculation according to the sensitive band in this letter are as follows.

1) The frequency band corresponding to the sensitive channel is detected through back mapping operation.
2) (Band-pass filter with the sensitive frequency band is constructed to process the vibration signal.
3) The filtered signal is demodulated through Hilbert transform which is widely used in signal demodulation.
4) Envelope spectrum is finally calculated by FFT of envelope signals.

E. Loss Function and Anomaly Indicator

In the training stage, the loss function of the proposed model is formulated by combining the reconstruction loss, the sparse regular loss, and the negative log-density of the memory residual. Each item in the loss function is responsible for different parts mentioned above, as shown in the following equation:

$$\mathcal{L}(\theta_e, \theta_d, \theta_m, \theta_h) = \left\|\boldsymbol{X} - \hat{\boldsymbol{X}}\right\|_2^2 - \frac{1}{K}\sum_{k=1}^{K} \log(h(\boldsymbol{V}_{\text{res}})) - \alpha \sum_{m=1}^{M} \hat{a}_m \cdot \log(\hat{a}_m) \tag{13}$$

where α represents the weight of the memory entropy and is set as 0.02 in all experiments of this letter.

According to the definition of surprisal value described in Section II-C and the loss function of the proposed framework, the normal samples tend to yield a small surprisal value in the training stage. Once the early abnormal state of bearing occurs, the surprisal value would increase accordingly. Thus, the surprisal value can be used to indicate the occurrence of anomalies. The reconstruction error and surprisal value are defined as the anomaly indicator to detect the abnormal condition of bearing

$$P = \left\|\boldsymbol{X} - \hat{\boldsymbol{X}}\right\|_2^2 - \frac{1}{K}\sum_{k=1}^{K} \log(h(\boldsymbol{V}_{\text{res}})) \tag{14}$$

$$P_n = \frac{P - \min(P)}{\max(P) - \min(P)}. \tag{15}$$

The normalized operation is applied to transform the anomaly indicator to range [0, 1]. The proposed anomaly indicator can promote the ability of the MRRAE in anomaly detection, which will be demonstrated by the experimental results.

IV. Experimental Verification

To verify the effectiveness of the proposed method in anomaly detection, two sets of run-to-failure experimental

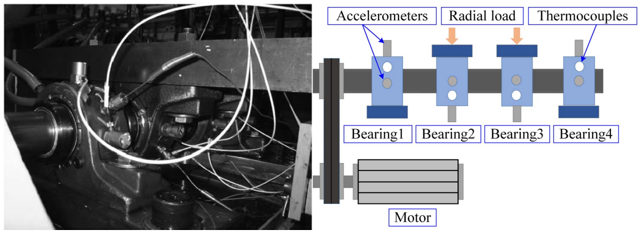

Fig. 3. IMS bearing test rig.

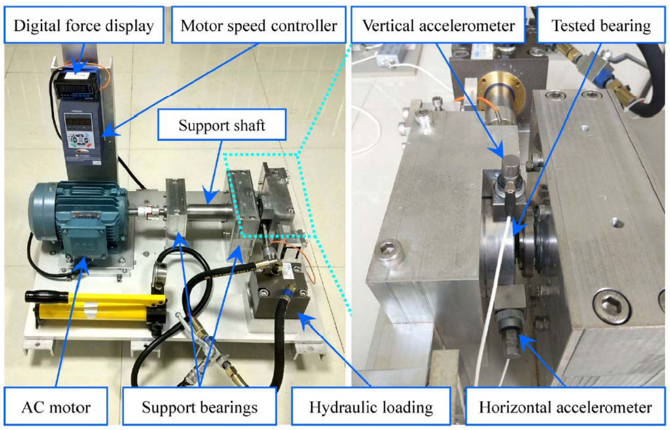

Fig. 4. XJTU-SY bearing test rig.

data set are studied using the proposed anomaly detection architecture.

A. Data Description

1) IMS Bearing Data Sets: This data set is supported by the NSF I/UCR Center for Intelligent Maintenance Systems, University of Cincinnati [54], which has been widely used in the research for machinery condition monitoring. The experimental system is shown in Fig. 3, where Rexnord ZA-2115 double-row bearings are installed on the shaft. The rotation speed is kept constant at 2000 RPM by an ac motor coupled to the shaft via rub belts. A radial load of 6000 lbs is applied onto the shaft and bearing by a spring mechanism [54]. The sampling frequency is set as 20 000 Hz. Three data sets are included in the data packet and only Set No. 2 is analyzed in this letter. A vibration signal with a length of 20 480 data points is collected with a recording interval of 10 min and a total of 984 samples are stored during the bearing's lifetime. At the end of the test-to-failure experiment, outer race failure occurred in bearing 1.
XJTU-SY

2) XJTU-SY Bearing Data Sets: bearing data sets are provided by the Institute of Design Science and Basic Component at Xi'an Jiaotong University and the Changxing Sumyoung Technology Co., Ltd. As shown in Fig. 4, the bearing testbed is composed of an ac induction motor, a motor speed controller, a support shaft, two support bearings, and a hydraulic loading system. The radial force is generated by the hydraulic loading system and applied to the housing of tested bearings, and the rotating speed is set and kept by the speed controller of the ac induction motor [55]. Two accelerometers of type PCB

TABLE I
PARAMETERS OF XJTU-SY TESTED BEARINGS

Parameter	Value	Parameter	Value
Outer race diameter	39.80 mm	Inner race diameter	29.30 mm
Bearing mean diameter	34.55 mm	Ball diameter	7.92 mm
Number of balls	8	Contact angle	0°
Load rating (static)	6.65 kN	Load rating (dynamic)	12.82 kN

352 C33 are positioned at 90° on the housing of the tested bearings, while only the data on the horizontal axis are used in this letter. The type of tested bearings is LDK UER204, and the detailed parameters are given in Table I.

The run-to-failure data of 15 rolling element bearings are included in the data packet. Bearing 2_5 data set is analyzed in this article. A vibration signal with a length of 32 768 data points was collected with a recording interval of 1 min and a total of 334 samples were stored during the bearing's lifetime.

3) Data Preprocessing: Vibration signals from the IMS bearing data set and XJTU-SY bearing data set are preprocessed in the same way. The fast Fourier transform (FFT) algorithm is applied to transform the vibration signals into the frequency domain with the length of 2048, as it is widely recognized that the signal in the frequency domain is more fault-sensitive than in the time domain [10]. Based on the procedure described in Section III, the preprocessed signals are picked as the input of the MRRAE model.

The run-to-failure data are divided into the training set and test set according to the following rules: the training set includes only the first 400 normal samples (80 samples in the XJTU-SY bearing data set) and all 984 samples (334 samples in the XJTU-SY bearing data set) are set as the testing set.

B. Performance Evaluation

1) Evaluation Metrics: To evaluate the performance of the proposed method and comparison methods, the commonly used metrics including Precision score, Recall score, F1 score, and Accuracy score are selected as evaluation metrics, which is the standard metric for the anomaly detection task [56]. The confusion matrix of anomaly detection is shown in Table II. The definition of each indicator is presented in the following equation:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}} \tag{16}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{17}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{18}$$

$$\text{F1} = \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}. \tag{19}$$

2) Comparison Methods: To validate the effectiveness of the proposed MRRAE method, several conventional and deep-learning-based anomaly detection methods including

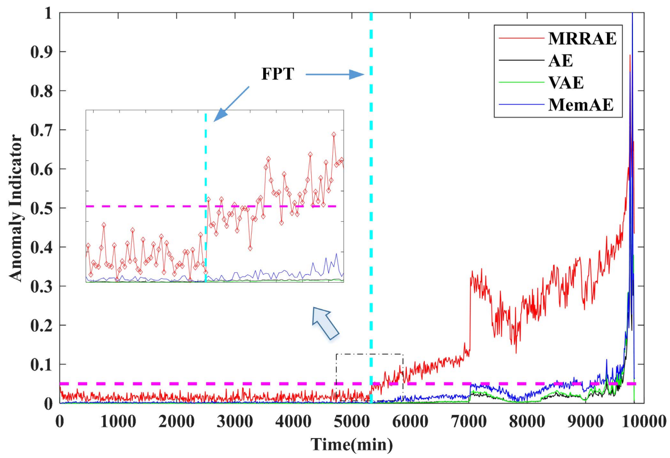

Fig. 5. Anomaly indicator of the IMS bearing data set.

OCSVM [57], isolation forest (IF) [58], kernel density estimation (KDE) [59], AE, variational AE (VAE), and MemAE are implemented for comparisons.

OCSVM and IF are standard and classic methods for anomaly detection. Specifically, 16 manually designed time-domain features in [40] are extracted as input samples of OCSVM and IF. As one of the most popular deep learning methods, AE has been widely applied in dimension reduction and anomaly detection. VAE uses a variational approach for latent representation learning, which results in an additional loss component and a specific estimator for the training algorithm. Besides, it is logical to use MemAE as a comparison method because it is the basis of our proposed method. For a fair comparison with the proposed method, the AE, VAE, and MemAE share similar convolutional encoder–decoder network architecture and parameter setting with the proposed method. Only the reconstruction error is defined as the anomaly indicator in these deep-learning-based bearing abnormal condition detection methods

$$Q = \|X - \hat{X}\|_2^2 \tag{20}$$

$$Q_n = \frac{Q - \min(Q)}{\max(Q) - \min(Q)}. \tag{21}$$

C. Experimental Results

1) IMS Bearing Data Sets: The anomaly indicator of the IMS bearing run-to-failure data calculated by AE, VAE, MemAE and proposed method in the testing stage is shown in Fig. 5, which can be used to reflect the bearing degradation trend. The red line represents the anomaly indicator calculated by the proposed MRRAE method, which is more sensitive than other comparison methods.

It can be seen that the reconstruction error of MemAE is barely larger than that of AE and VAE. The reason is that the unique memory mechanism of MemAE limits and constrains the changes in the latent space during the testing stage and thus causes the reconstruction of the decoder to be close to the normal distribution. When a failure occurs, the reconstruction error tends to increase accordingly.

TABLE II
CONFUSION MATRIX OF ANOMALY DETECTION

Confusion matrix		True Class: Positive	True Class: Negative
Predicted Class	Positive	True positive(TP)	False posi-tive(FP)
	Negative	False negative(FN)	True nega-tive(TN)

TABLE III
DETECTION RESULTS OF THE IMS BEARING DATA SET

Methods	Accuracy	Precision	Recall	F1
OCSVM	0.8648	0.7723	1.0	0.8715
IF	0.5681	0.5148	1.0	0.6797
KDE	0.5335	0.4956	1.0	0.6627
AE	0.5681	1.0	0.0576	0.1090
VAE	0.5721	1.0	0.0665	0.1247
MemAE	0.6291	1.0	0.1907	0.3202
MRRAE	**0.9797**	**1.0**	0.9557	**0.9773**

Although there has been some improvement from the perspective of reconstruction error, MemAE is still not sensitive to the change point of the bearing state. The anomaly indicator calculated by the proposed MRRAE method changes significantly at 533 points, since a regression model in the residual space is constructed from the perspective of Bayes surprise, which is very sensitive to the change point.

To prove the superiority of the proposed method, the evaluation metrics described in (16)–(19) is used to evaluate the performance of different methods. The class label is set according to the degradation trend of bearing. The signals in 1–532 samples are set with positive class label and the signals in 533–984 samples are set with negative class label, since the 533 is recognized as the FPT [8] of run-to-failure data sets.

In the process of classification, 0.05 is set as the threshold to distinguish normal and anomaly class, which means the samples with anomaly score greater than 0.05 are considered as negative class and those less than 0.05 are considered as positive class. Thus, the classification results of the comparison and the proposed method can be obtained through this classification strategy.

After that, the Accuracy Precision score, Recall score, F1 score, and Accuracy score are calculated based on the comparison of the classification results and labels, which can be used to describe the performance of different methods quantitatively.

The evaluation metrics mentioned above are shown in Table III; the proposed MRRAE method shows good performances on anomaly detection with an accuracy score of 0.9797, which means very few samples were misclassified. The precision score of the proposed method is 1. The comprehensive index F1 score of the proposed method can reach 0.9773, which is the maximum value among all methods. Notably, the recall score of the proposed method is slightly smaller than that of OCSVM, IF, and KDE, which is for the reason that the anomaly indicator calculated by the MRRAE

TABLE IV
DETECTION RESULTS OF THE XJTU-SY BEARING DATA SET

Methods	Accuracy	Precision	Recall	F1
OCSVM	0.7847	0.7491	1.0	0.8566
IF	0.6490	0.6469	1.0	0.7856
KDE	0.7965	0.7596	1.0	0.8634
AE	0.8142	1.0	0.7110	0.8311
VAE	0.8053	1.0	0.6972	0.8216
MEMAE	0.8083	1.0	0.7018	0.8248
MRRAE	**0.9351**	0.9083	**1.0**	**0.9520**

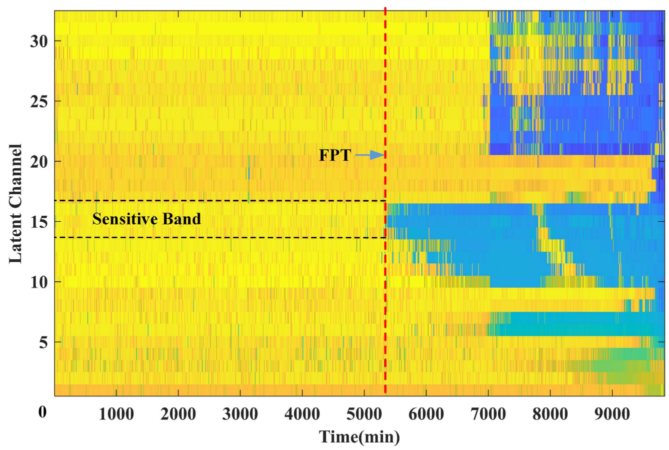

Fig. 6. Probability density distribution of IMS latent memory residual representation.

model has a short fluctuation around 533. However, compared with these methods comprehensively, the proposed method has a higher recognition accuracy, which means the regression model in the residual space shows a better recognition performance.

To further analyze the degradation process of the bearing, the probability density distribution of the latent memory residual representation $[\log(h(\boldsymbol{V}_{\text{res}}))]$ is collected during the testing stage and presented in Fig. 6, where the FPT [8] is marked by red dashed line. Information of memory residual distribution is represented by different colors, which refers to the 2-D form of the time–frequency spectrum. In this way, the latent channel with a sudden change in probability density distribution can be detected visibly. As shown in Fig. 6, 14–16 channels in latent memory residual space change obviously, which means these channels lead to the change in the probability density distribution. The frequency band corresponding to 14–16 channels can be determined through back mapping operation based on the shared-weights architecture and translation invariance characteristics of convolutional network. The sampling frequency of the IMS bearing data set is 20 000 Hz, and the latent dimension of the proposed model is set as 32. Thus, the Fourier support 4062–5000 Hz is considered as the sensitive band.

The vibration signal of 533 samples is processed by a band-pass filter with the sensitive band determined by back mapping operation. After that, the envelope spectrum of the filtered signal is calculated and the result is depicted in Fig. 7, in which the outer race fault characteristic frequency and its harmonics can be obviously observed.

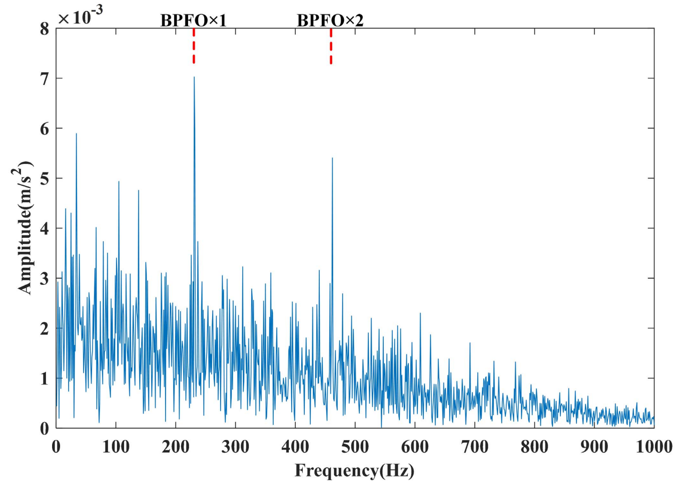

Fig. 7. Envelope spectrum of the sensitive band.

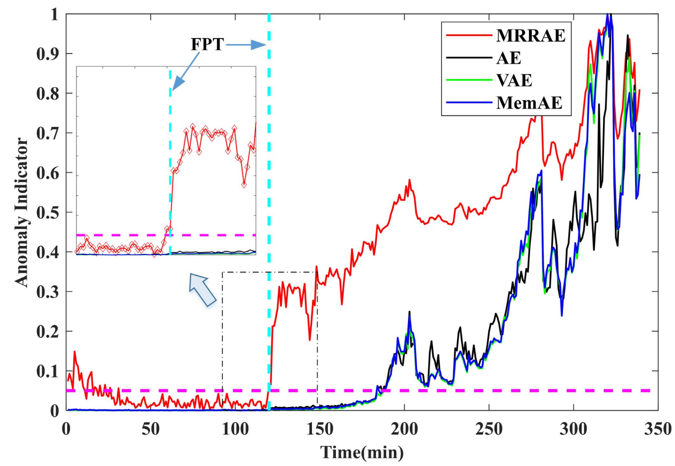

Fig. 8. Anomaly indicator of the XJTU-SY bearing data set.

2) XJTU-SY Bearing Data Sets: The anomaly indicator of the XJTU-SY bearing data sets obtained by AE, VAE, MemAE, and the proposed method is depicted in Fig. 8. The red line represents the anomaly indicator calculated by the proposed MRRAE method, which changes obviously at around sample 122. On the contrary, the AE, VAE, and MemAE method are not sensitive to the change point of the bearing state.

The evaluation metrics is also used to evaluate the performance of different methods on the XJTU-SY bearing data set. The signals in 1–121 samples are set with positive class label and the signals in 122–334 samples are set with negative class label, since 122 is recognized as the FPT of this data set. Same as the process of classification on the IMS data set, 0.05 is set as the threshold to distinguish between normal and anomaly class and the classification results of comparison and the proposed method can be obtained.

The result of the XJTU-SY bearing data set is shown in Table IV, the proposed MRRAE method also shows good anomaly detection performance with an accuracy score of 0.9351. The Recall score of the proposed method is 1. The comprehensive index F1 score of the proposed method is 0.9520, which is the maximum value among all methods. Notably, the precision score of the proposed method is slightly

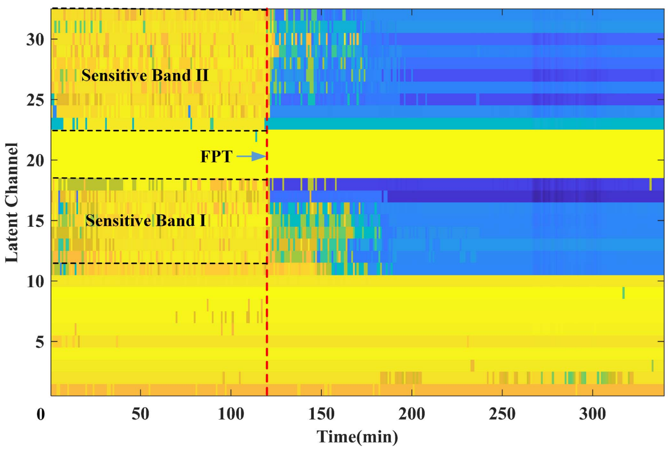

Fig. 9. Probability density distribution of XJTU-SY latent memory residual representation.

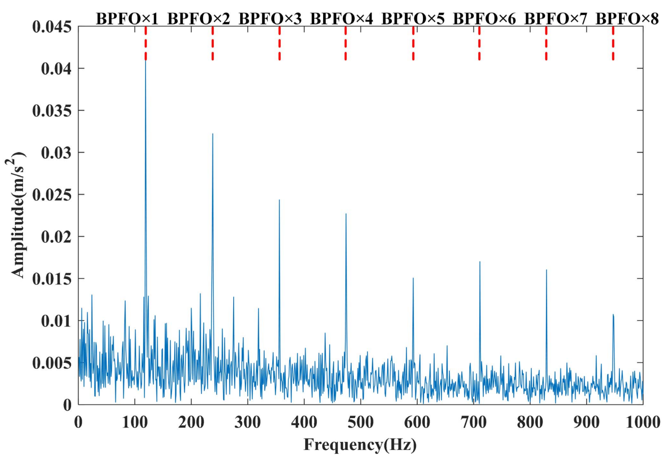

Fig. 10. Envelope spectrum of sensitive band I.

smaller than that of AE, VAE, and MEMAE, which is for the reason that the anomaly indicator calculated by the MRRAE model fluctuates at the beginning of the data set and several values are slightly greater than 0.05. However, compared with these methods comprehensively, the proposed method also shows a better recognition performance than other comparison methods on the XJTU-SY bearing data set.

As shown in Fig. 9, 12–18 channels and 23–32 channels in the probability density distribution of latent memory residual space change obviously. The sampling frequency of the XJTU-SY bearing data set is 25 600 Hz, and thus the Fourier supports 4400–7200 Hz and 8800–12 800 Hz are considered as sensitive bands for further analysis. The vibration signal of 122 samples in the XJTU-SY data set is processed by band-pass filters with the sensitive bands.

Then, the envelope spectrum of the filtered signals is calculated and the results are depicted in Figs. 10 and 11, respectively, while the outer race fault characteristic frequency and its harmonics can be observed, especially in the envelope spectrum of the filtered signal corresponding to the sensitive band I. Besides detecting anomaly, the proposed method can further analyze the anomaly in combination with the probability density distribution of the latent memory residual representation.

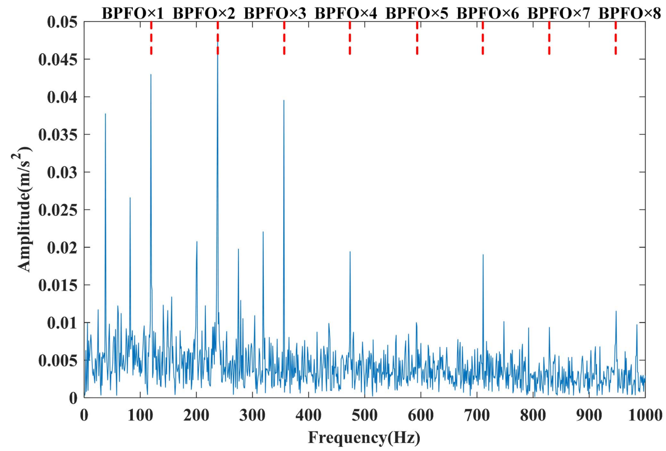

Fig. 11. Envelope spectrum of sensitive band II.

In general, MRRAE achieves superior performance compared with the conventional and deep-learning-based anomaly detection methods, which proves the effectiveness of the proposed module in bearing condition monitoring. Furthermore, the proposed method pays close attention to the special structure of bearing vibration signal and provides a new way for explaining the decision-making processes of deep neural networks.

V. Conclusion

An innovative deep learning model, namely, MRRAE, is developed for the health management of rolling element bearings, which improves the performance of the AE-based unsupervised anomaly detection algorithms. The proposed MRRAE is an unsupervised learning model driven only by normal data, which means the parameter of the proposed model is only determined by normal data in the training stage. The convolutional AE and memory module are used to obtain latent representation of the vibration signals. The memory module is trained to record the prototypical elements of normal patterns. The output of the memory part thus tends to be close to a normal sample. A parametric density estimator is designed to calculate the probability density distribution of the memory residual via an autoregressive procedure. The reconstruction errors and surprisal values are used to indicate the abnormal condition of bearing. The validity and feasibility of the proposed method are verified on two run-to-failure experimental data sets. The proposed MRRAE model shows better anomaly detection performances than other comparison methods.

Furthermore, this research interprets the decision-making processes from the perspective of autoregressive density estimation. The probability density distribution calculated by autoregressive estimator is able to locate the sensitive band which leads to the change in anomaly indicators. Besides anomaly detection of bearing condition, the proposed method can diagnose the incipient fault directly with the help of envelop demodulation.

References

[1] D. Zhao, J. Li, W. Cheng, and W. Wen, "Compound faults detection of rolling element bearing based on the generalized demodulation algorithm under time-varying rotational speed," *J. Sound Vib.*, vol. 378, pp. 109–123, Sep. 2016.

[2] Y. Hu, W. Bao, X. Tu, F. Li, and K. Li, "An adaptive spectral kurtosis method and its application to fault detection of rolling element bearings," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 3, pp. 739–750, Mar. 2020.

[3] X. Huang, G. Wen, L. Liang, Z. Zhang, and Y. Tan, "Frequency phase space empirical wavelet transform for rolling bearings fault diagnosis," *IEEE Access*, vol. 7, pp. 86306–86318, Jun. 2019.

[4] L. Cui, X. Wang, H. Wang, and J. Ma, "Research on remaining useful life prediction of rolling element bearings based on time-varying Kalman filter," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 6, pp. 2858–2867, Jun. 2020.

[5] M. Kordestani, M. Saif, M. E. Orchard, R. Razavi-Far, and K. Khorasani, "Failure prognosis and applications—A survey of recent literature," *IEEE Trans. Rel.*, early access, Sep. 17, 2019, doi: 10.1109/TR.2019.2930195.

[6] H. Wang, J. Xu, R. Yan, and R. X. Gao, "A new intelligent bearing fault diagnosis method using SDP representation and SE-CNN," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2377–2389, May 2020.

[7] X. Jin, Y. Sun, Z. Que, Y. Wang, and T. W. S. Chow, "Anomaly detection and fault prognosis for bearings," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 9, pp. 2046–2054, Sep. 2016.

[8] N. Li, Y. Lei, J. Lin, and S. X. Ding, "An improved exponential model for predicting remaining useful life of rolling element bearings," *IEEE Trans. Ind. Electron.*, vol. 62, no. 12, pp. 7762–7773, Dec. 2015.

[9] M. Rezamand, M. Kordestani, R. Carriveau, D. S.-K. Ting, M. E. Orchard, and M. Saif, "Critical wind turbine components prognostics: A comprehensive review," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9306–9328, Dec. 2020.

[10] J. Wu, Z. Zhao, C. Sun, R. Yan, and X. Chen, "Fault-attention generative probabilistic adversarial autoencoder for machine anomaly detection," *IEEE Trans. Ind. Informat.*, vol. 16, no. 12, pp. 7479–7488, Dec. 2020.

[11] L. Ruff, R. A. Vandermeulen, and N. Gornitz, "Deep one-class classification," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 4393–4402.

[12] W. Liu *et al.*, "Towards visually explaining variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8642–8651.

[13] C. Liu and K. Gryllias, "A semi-supervised support vector data description-based fault detection method for rolling element bearings based on cyclic spectral analysis," *Mech. Syst. Signal Process.*, vol. 140, Jun. 2020, Art. no. 106682.

[14] X. Zhu, Y. Zhang, and Y. Zhu, "Bearing performance degradation assessment based on the rough support vector data description," *Mech. Syst. Signal Process.*, vol. 34, nos. 1–2, pp. 203–217, Jan. 2013.

[15] S. Yin, X. Zhu, and C. Jing, "Fault detection based on a robust one class support vector machine," *Neurocomputing*, vol. 145, pp. 263–268, Dec. 2014.

[16] J. Saari, D. Strömbergsson, J. Lundberg, and A. Thomson, "Detection and identification of windmill bearing faults using a one-class support vector machine (SVM)," *Measurement*, vol. 137, pp. 287–301, Apr. 2019.

[17] M. Zeng, Y. Yang, S. Luo, and J. Cheng, "One-class classification based on the convex hull for bearing fault detection," *Mech. Syst. Signal Process.*, vol. 81, pp. 274–293, Dec. 2016.

[18] D. Gong *et al.*, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[19] A. Zimek, E. Schubert, and H. P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Stat. Anal. Data Mining*, vol. 5, pp. 363–387, Oct. 2012.

[20] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 481–490.

[21] Z. Chen and W. Li, "Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1693–1702, Jul. 2017.

[22] X. Zhao, M. Jia, and Z. Liu, "Fault diagnosis framework of rolling bearing using adaptive sparse contrative auto-encoder with optimized unsupervised extreme learning machine," *IEEE Access*, vol. 8, pp. 99154–99170, 2020.

[23] X. Zhao, M. Jia, and Z. Liu, "Semisupervised deep sparse auto-encoder with local and nonlocal information for intelligent fault diagnosis of rotating machinery," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[24] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 1, pp. 136–144, Jan. 2019.

[25] B. Han, X. Wang, S. Ji, G. Zhang, S. Jia, and J. He, "Data-enhanced stacked autoencoders for insufficient fault classification of machinery and its understanding via visualization," *IEEE Access*, vol. 8, pp. 67790–67798, 2020.

[26] M. Roy, S. K. Bose, B. Kar, P. K. Gopalakrishnan, and A. Basu, "A stacked autoencoder neural network based automated feature extraction method for anomaly detection in on-line condition monitoring," Oct. 2018, *arXiv:1810.08609*. [Online]. Available: http://arxiv.org/abs/1810.08609

[27] J. Chen, J. Li, W. Chen, Y. Wang, and T. Jiang, "Anomaly detection for wind turbines based on the reconstruction of condition parameters using stacked denoising autoencoders," *Renew. Energy*, vol. 147, pp. 1469–1480, Mar. 2020.

[28] F. Xu, F. Yang, X. Fan, Z. Huang, and K. L. Tsui, "Extracting degradation trends for roller bearings by using a moving-average stacked auto-encoder and a novel exponential function," *Measurement*, vol. 152, Feb. 2020, Art. no. 107371.

[29] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture IE*, vol. 2, pp. 1–18, Dec. 2015.

[30] S. Khobahi and M. Soltanalian, "Model-aware deep architectures for one-bit compressive variational autoencoding," Nov. 2019, *arXiv:1911.12410*. [Online]. Available: http://arxiv.org/abs/1911.12410

[31] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," Dec. 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114

[32] S. Zhang, F. Ye, B. Wang, and T. G. Habetler, "Semi-supervised learning of bearing anomaly detection via deep variational autoencoders," Dec. 2019, *arXiv:1912.01096*. [Online]. Available: http://arxiv.org/abs/1912.01096

[33] M. Hemmer, A. Klausen, H. V. Khang, K. G. Robbersmyr, and T. I. Waag, "Health indicator for low-speed axial bearings using variational autoencoders," *IEEE Access*, vol. 8, pp. 35842–35852, 2020.

[34] Y. Kawachi, Y. Koizumi, and N. Harada, "Complementary set variational autoencoder for supervised anomaly detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2366–2370.

[35] F. Di Mattia, P. Galeone, M. De Simoni, and E. Ghelfi, "A survey on GANs for anomaly detection," Jun. 2019, *arXiv:1906.11632*. [Online]. Available: http://arxiv.org/abs/1906.11632

[36] T. Schlegl, P. Seeboeck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Information Processing in Medical Imaging*. Cham, Switzerland: Springer, 2017, pp. 146–157.

[37] H. Zenati, F. C. Sheng, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient GAN-based anomaly detection," Feb. 2018, *arXiv:1802.06222*. [Online]. Available: https://arxiv.org/abs/1802.06222

[38] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comp. Vis. (ACCV)*, 2019, pp. 622–637.

[39] Y. Lyu, Z. Han, J. Zhong, C. Li, and Z. Liu, "A generic anomaly detection of catenary support components based on generative adversarial networks," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2439–2448, May 2020.

[40] W. Jiang, Y. Hong, B. Zhou, X. He, and C. Cheng, "A GAN-based anomaly detection approach for imbalanced industrial time series," *IEEE Access*, vol. 7, pp. 143608–143619, Sep. 2019.

[41] X. Li, W. Zhang, and Q. Ding, "Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism," *Signal Process.*, vol. 161, pp. 136–154, Aug. 2019.

[42] H. Wang, Z. Liu, D. Peng, and Y. Qin, "Understanding and learning discriminant features based on multiattention 1DCNN for wheelset bearing fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 5735–5745, Sep. 2020.

[43] B. Wang, Y. Lei, N. Li, and W. Wang, "Multi-scale convolutional attention network for predicting remaining useful life of machinery," *IEEE Trans. Ind. Electron.*, early access, Jun. 25, 2020, doi: 10.1109/TIE.2020.3003649.

[44] Z. Chen, M. Wu, R. Zhao, F. Guretno, R. Yan, and X. Li, "Machine remaining useful life prediction via an attention-based deep learning approach," *IEEE Trans. Ind. Electron.*, vol. 68, no. 3, pp. 2521–2531, Mar. 2021.

[45] Y. Qin, D. Chen, S. Xiang, and C. Zhu, "Gated dual attention unit neural networks for remaining useful life prediction of rolling bearings," *IEEE Trans. Ind. Informat.*, early access, Jun. 2, 2020, doi: 10.1109/TII.2020.2999442.

[46] H. Zhang, Q. Zhang, S. Shao, T. Niu, and X. Yang, "Attention-based LSTM network for rotatory machine remaining useful life prediction," *IEEE Access*, vol. 8, pp. 132188–132199, 2020.

[47] H. Liu, Z. Liu, W. Jia, and X. Lin, "Remaining useful life prediction using a novel feature-attention-based end-to-end approach," *IEEE Trans. Ind. Informat.*, vol. 17, no. 2, pp. 1197–1207, Feb. 2021.

[48] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun, "Predicting deeper into the future of semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2017, pp. 648–657.

[49] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," Sep. 2016, *arXiv:1609.03499*. [Online]. Available: http://arxiv.org/abs/1609.03499

[50] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: A noise-assisted data analysis method," *Adv. Adapt. Data Anal.*, vol. 1, no. 1, pp. 1–41, 2009.

[51] I. Daubechies, Y. Wang, and H.-T. Wu, "ConceFT: Concentration of frequency and time via a multitapered synchrosqueezed transform," Jul. 2015, *arXiv:1507.05366*. [Online]. Available: http://arxiv.org/abs/1507.05366

[52] C. Mishra, A. K. Samantaray, and G. Chakraborty, "Rolling element bearing defect diagnosis under variable speed operation through angle synchronous averaging of wavelet de-noised estimate," *Mech. Syst. Signal Process.*, vols. 72–73, pp. 206–222, May 2016.

[53] S. Dong, G. Wen, Z. Zhang, Y. Yuan, and J. Luo, "Rolling bearing incipient degradation monitoring and performance assessment based on signal component tracking," *IEEE Access*, vol. 7, pp. 45983–45993, Mar. 2019.

[54] H. Qiu, J. Lee, J. Lin, and G. Yu, "Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics," *J. Sound Vib.*, vol. 289, nos. 4–5, pp. 1066–1090, Feb. 2006.

[55] B. Wang, Y. Lei, N. Li, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Trans. Rel.*, vol. 69, no. 1, pp. 401–412, Mar. 2020.

[56] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative study of anomaly detection schemes," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, May 2003, pp. 25–36.

[57] B. Scholkopf, R. Williamson, A. Smola, J. S. Taylor, and J. Platt, "Support vector method for novelty detection," in *Proc. Neural Inf. Process. Syst.*, 2000, pp. 582–588.

[58] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2008, pp. 413–422.

[59] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, Sep. 1962.

Xin Huang received the B.S. and M.S. degrees in mechanical engineering from Xinjiang University, Urumqi, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree in mechanical engineering with the Department of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China.

His research interests include mechanical signal processing, mechanical system fault diagnosis, and prognosis.

Guangrui Wen received the B.S., M.S., and Ph.D. degrees in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 1998, 2001, and 2006, respectively.

He was a Post-Doctoral Fellow at Xi'an Shaangu Power Company, Ltd., Xi'an, from 2008 to 2010. He has authored two books and more than 80 articles and holds more than 20 patents. His research interests include mechanical system fault diagnosis and prognosis, and mechanical equipment life cycle health monitoring and intelligent maintenance.

Dr. Wen is a member of the Chinese Mechanical Engineering Society (CIME) and the Chinese Society for Vibration Engineering (CSVE).

Shuzhi Dong received the B.S. degree from Southwest Jiaotong University, Chengdu, China, in 2015. He is currently pursuing the Ph.D. degree in mechanical engineering with the Department of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China.

His research interests are fault diagnosis and health monitoring of mechanical equipment, mainly related to deep learning and other artificial intelligence methods.

Haoxuan Zhou received the B.S. degree from Southwest Petroleum University, Chengdu, China, in 2016, and the M.S. degree from the Kunming University of Science and Technology, Kunming, China, in 2019. He is currently pursuing the Ph.D. degree in mechanical engineering with the Department of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China.

His main research direction is mechanical signal processing, rotating machinery fault diagnosis, and mechanical system health monitoring based on industrial big data.

Zihao Lei received the B.S. degree in mechanical engineering from Southwest Jiaotong University, Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree in mechanical engineering with the Department of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China.

His current research is focused on transfer learning, deep learning, machinery condition monitoring, intelligent fault diagnosis, and prognosis.

Zhifen Zhang received the B.S. and M.S. degrees in materials processing engineering from the Lanzhou University of Technology, Lanzhou, China, in 2007 and 2010, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2015.

She was a Lecturer with the Department of Mechanical Engineering, Xi'an Jiao Tong University, Xi'an, China, from 2015 to 2019. Since October 2019, she has been an Associate Professor at the same department. She has authored more than ten magazine articles including published in the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, MECHANICAL SYSTEMS AND SIGNAL PROCESSING and the *Journal of Materials Processing Technology*. Her research interests include intelligent manufacturing of laser shocking peening and information fusion.

Xuefeng Chen (Senior Member, IEEE) received the Ph.D. degree from Xian Jiaotong University, Xian, China, in 2004.

He is currently a Professor of mechanical engineering with Xian Jiaotong University. His current research interests include finite-element method, mechanical system and signal processing, diagnosis and prognosis for complicated industrial systems, smart structures, aeroengine fault diagnosis, and wind turbine system monitoring.

Dr. Chen was a recipient of the National Excellent Doctoral Dissertation of China in 2007, the Second Award of Technology Invention of China in 2009, the National Science Fund for Distinguished Young Scholars in 2012, and a Chief Scientist of the National Key Basic Research Program of China (973 Program) in 2015.