electronics

MDPI

Article

Robust Detection of Bearing Early Fault Based on Deep Transfer Learning

Wentao Mao [1,2,*,†], Di Zhang [1,†], Siyu Tian [1] and Jiamei Tang [1]

1 School of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China; vencent8692@gmail.com (D.Z.); m.laclac@gmail.com (S.T.); tangjiamei@htu.edu.cn (J.T.)
2 Engineering Lab of Intelligence Business & Internet of Things of Henan Province, Xinxiang 453007, China
* Correspondence: maowt@htu.edu.cn; Tel.: +86-15037301821
† These authors contributed equally to this work.

Received: 14 December 2019; Accepted: 6 February 2020; Published: 13 February 2020

Abstract: In recent years, machine learning techniques have been proven to be a promising tool for early fault detection of rolling bearings. In many actual applications, however, bearing whole-life data are not easy to be historically accumulated, while insufficient data may result in training a detection model that is not good enough. If utilizing the available data under different working conditions to facilitate model training, the data distribution of different bearings are usually quite different, which does not meet the precondition of *independent and identical distribution* (*i.i.d*) and tends to cause performance reduction. In addition, disturbed by the unstable noise under complex conditions, most of the current detection methods are inclined to raise false alarms, so that the reliability of detection results needs to be improved. To solve these problems, a robust detection method for bearings early fault is proposed based on deep transfer learning. The method includes offline stage and online stage. In the offline stage, by introducing a deep auto-encoder network with domain adaptation, the distribution inconsistency of normal state data among different bearings can be weakened, then the common feature representation of the normal state is obtained. With the extracted common features, a new state assessment method based on the robust deep auto-encoder network is proposed to evaluate the boundary between normal state and early fault state in the low-rank feature space. By training a support vector machine classifier, the detection model is established. In the online stage, along with the data batch arriving sequentially, the features of target bearing are extracted using the common representation learnt in the offline stage, and online detection is conducted by feeding them into the SVM model. Experimental results on IEEE PHM Challenge 2012 bearing dataset and XJTU-SY dataset show that the proposed approach outperforms several state-of-the-art detection methods in terms of detection accuracy and false alarm rate.

Keywords: early fault detection; fault diagnosis; state assessment; transfer learning; deep learning

1. Introduction

As an important part of common machinery equipment, rolling bearings are prone to various kinds of faults under complex working conditions like long-term heavy load and strong impact, etc. Faulty bearings will cause the performance deterioration of whole machinery. Successful detection of bearing fault at an initial stage will be helpful to make timely maintenance and avoid serious accident occurrence. Therefore, accurate and reliable detection and diagnosis at the early stage of fault occurrence is considered as a key step of fault prognostic and health management (PHM) [1]. To sum up, the work of fault detection for rolling bearings can be divided into two types: signal analysis-based methods and machine learning-based methods. Please refer to Section 2 for more details. In recent years, the detection methods based on machine learning techniques have obtained

rapid development, as they can solve some shortcomings of traditional signal analysis-based methods, for instance, generally with no need to know prior knowledge of bearing signal in advance. Thus, data-driven intelligent fault detection and diagnosis has received wide attention in the fields of PHM and mechanical manufacturing.

In this paper, we mainly consider this problem: how to improve the accuracy and robustness of early fault detection for bearings on insufficient bearing data. In many applications, bearing whole-life data are hard to be historically accumulated, while insufficient data may result in training a detection model that is not good enough. An intuitive idea is getting help from the available data of the bearings which are of the same manufacturing specifications, even under different working conditions. The ideal goal is to borrow domain information in such data to improve the generalization performance of the detection model on target bearing. However, direct modeling on the data under different working conditions couldn't reach this goal due to a certain deviation of data distribution between the target bearing and training bearings. As most of the statistical machine learning methods work on the precondition of *independent and identical distribution* (*i.i.d*), the deviation of data distribution will certainly reduce the reliability of detection results. According to our observation, some unstable vibration caused by environmental interference or instrumental noise may aggravate such distribution deviation. Therefore, it is necessary to reduce such distribution deviation for improving the detection accuracy and robustness.

We take the bearing dataset of IEEE PHM Challenge 2012 as an example. Figure 1 provides the distribution of probabilistic density and features of normal state data from bearings 1–7 under the first working condition. It is clear that, even these seven bearings have an identical model size and run under the same working conditions, their data distribution of normal state still have an obvious difference, not to mention the bearings under different working conditions. In this phenomenon, the detection model built on the data of some bearings from these seven bearings could not directly apply to the other bearings for detection.

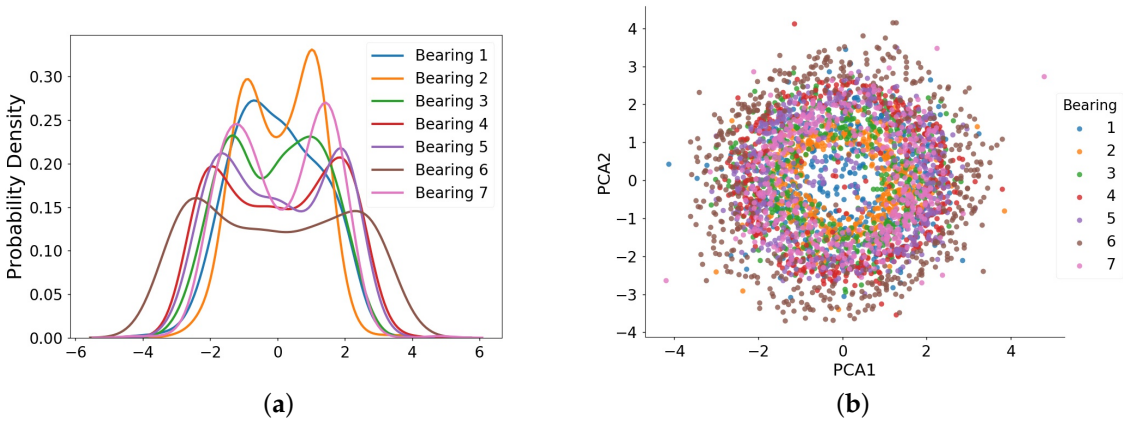


Figure 1. Distribution characteristics of normal state data of seven bearings from IEEE PHM 2012 Challenge dataset with (**a**) probabilistic density distribution and (**b**) feature distribution.

The deviation of data distribution in Figure 1 comes from several reasons. The irregular fluctuation in normal state may be caused by random noise, measurement error, variable working conditions, and other uncertain factors. Consequently, model bias as well as false alarm will be raised, which decreases the robustness of fault detection model. In addition, as the signal of early fault is weak and easy to be concealed by unstable vibration and noisy environment, traditional anomaly detection methods generally have difficulty identifying the early fault state accurately, which also reduces the robustness of the detection model. Therefore, the anti-interference ability of fault detection model needs to be considered more.

Furthermore, we also give another example. Figure 2 shows the root mean square (RMS) curves of vibration signals of the first 300 sample points of bearings 1, 2 and 7 under the first working condition. These 300 sample points can be viewed to be collected in normal state. From Figure 2, the RMS curves of bearings 1, 2, and 7 all fluctuate significantly in the starting normal state. Meanwhile, the RMS curve of bearing 2 has several peaks, which indicates that the signal is frequently disturbed. These interfered signals are not caused by bearing fault itself, but reflect the irregular vibration raised by the impact of external forces on the bearing and interference of the surrounding environment. In this scenario, the state change of bearing data is incapable of being recognized accurately and robustly, and, accordingly, such state assessment results can not support a reasonable early fault detection model. Again, the detection accuracy cannot be guaranteed and the robustness of detection model will be reduced as well.

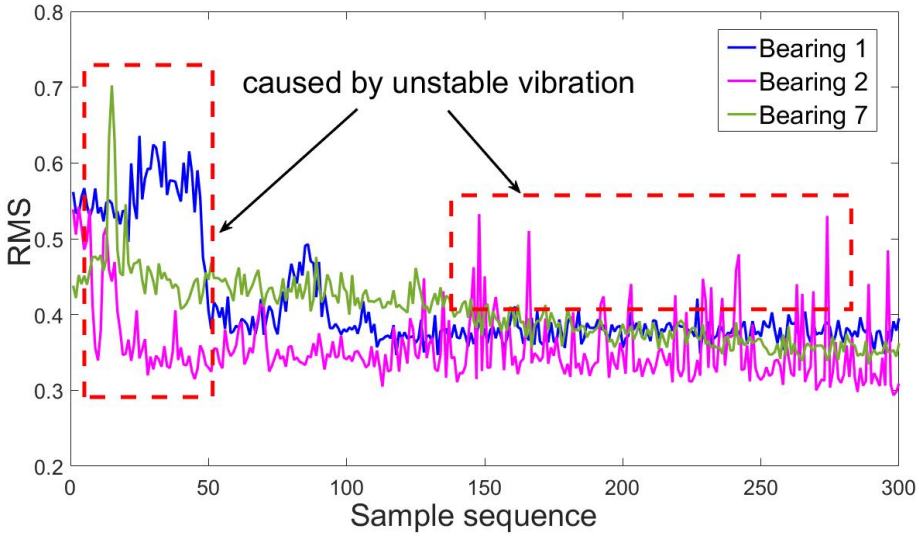


Figure 2. RMS curves of the first 300 sample points of bearings 1, 2 and 7 under the first working condition from IEEE PHM Challenge 2012 dataset.

Based on the analysis mentioned above, we introduce transfer learning to build a fault detection model on the bearing data collected under different working conditions. Transfer learning is machine learning under the shift between training and test distributions. Transfer learning has been successfully applied to bearing fault diagnosis (please see Section 2 for detailed analysis), but, for the problem of early fault detection, there still are some challenges to be solved:

(1) In order to improve detection accuracy for bearing early fault, we need to reduce the data distribution deviation, especially in a normal state. As a result, the detection model built on the bearings data under some working conditions (also called *source domain*) can be dynamically applied to the bearing data under another working condition (called *target domain*).
(2) In order to build an effective online detection model in complex environment and noise interference, it is necessary to find a state assessment method with strong anti-interference ability. Meanwhile, this method should be able to achieve accurate recognition of early fault state on different bearing data. As a result, the robustness of detection model can be improved.

To solve such challenges, this paper presents a new robust detection method for bearing early fault. Two key ideas are adopted: (1) deep transfer learning technique is used to find out the common feature representation of bearings data in source and target domains so as to reduce the impact of data fluctuation and distribution difference; (2) On the basis of (1), outliers in the common feature space need to be further removed in order to realize an accurate state assessment, as the abnormal fluctuation part may still interfere with the common feature. The total technical flowchart is shown in Figure 3.

Specifically, in the stage of feature transfer, a deep auto-encoder (DAE) model with domain adaption is applied to extract the common feature representation in bearing normal state. Then, the robust deep auto-encoder (RDA) algorithm is introduced to conduct state assessment in a low-rank space. After that, the detection model is constructed by training a SVM classifier on the state assessment results. In an online stage, the features of target bearing can be directly extracted by means of the common feature representation. By feeding them into the offline detection model, the occurrence of early fault can then be recognized.

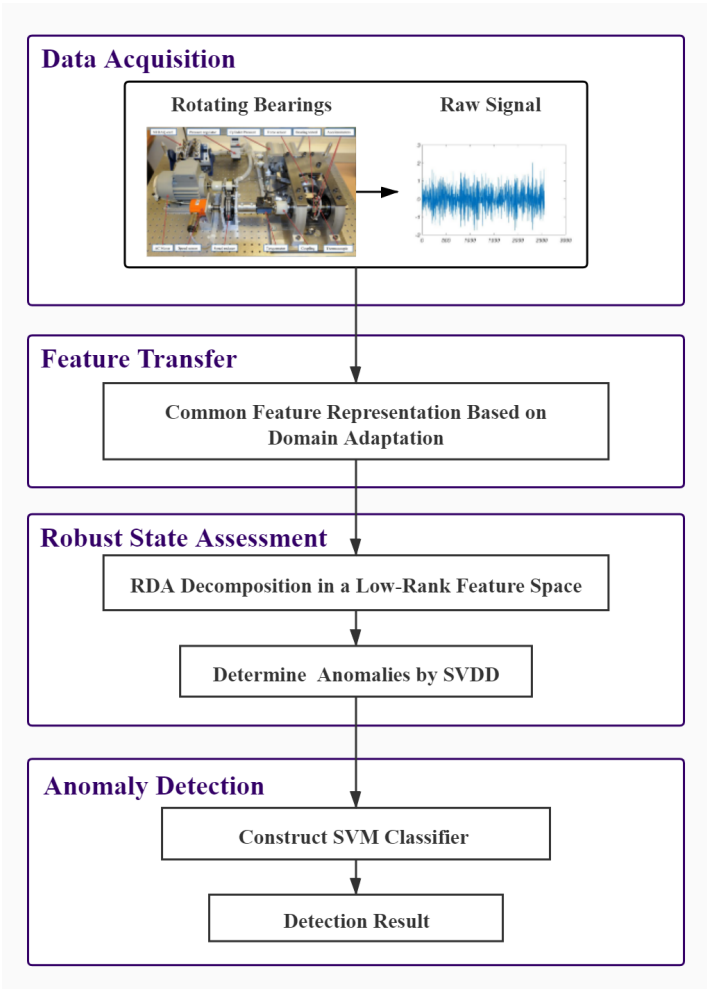


Figure 3. Total flowchart of the proposed robust detection method for bearing early fault.

The main contributions of this paper are as follows:

(1) This paper proposes a robust method of state assessment for rolling bearings. Running with deep transfer learning, this method can accurately identify early fault state on the bearing data under different working conditions. In addition, this method has good anti-interference ability against irregular fluctuation in normal state data. According to our literature survey, the current research about state assessment is seldom concerned about the robustness of assessment results.
(2) This paper proposes a new online detection method for bearing early fault. On the basis of the common feature representation obtained from source and target domains, this method can directly extract representative features for target bearing and identify the occurrence of early faults in real time with a much lower false alarm rate. To our best knowledge, very little research whose focus is robust early fault detection has been found, and there are no other research found about the application of transfer learning on early fault detection of bearings.

The structure of this paper is as follows. Section 2 provides a detailed literature survey of bearing early fault detection methods. Section 3 mainly describes the steps of the proposed method. Section 4 validates the effectiveness of the proposed method on IEEE PHM Challenge 2012 dataset and XJTU-SY dataset, followed by conclusions in the last section.

2. Preliminary Works

Generally speaking, fault detection can be viewed as a pre-step of fault diagnosis. Fault detection is mainly dedicated to detecting the change of system state, while fault diagnosis pays more attention to identifying different fault states such as fault type, crack size, etc. As incipient fault information is easily interrupted by noise, it is hard to determine a specific boundary between normal state and incipient fault state. Generally speaking, there are two kinds of early fault detection methods for rolling bearings: signal analysis-based methods and machine learning-based methods. For signal analysis-based methods, noise cancellation or noise utilization are usually utilized to deal with weak signal, and then time-frequency analysis is performed to extract and compare fault characteristic frequency [1–3]. Although such methods can extract early fault signals from original signals with noise, they still have some drawbacks: (1) If working conditions are varying or unknown, fault characteristic frequency which is used for detection could not be precisely calculated, and, consequently, the reliability and robustness of detection results would be reduced; (2) Improper denoising techniques may weaken the features of early fault, as the signals containing early fault are just of a low signal-noise ratio. As a result, these kinds of methods are perhaps insensitive to early fault and would result in delayed detection results.

In recent years, with quick development of artificial intelligence, machine learning-based methods have received more and more attention. The basic idea of these kinds of methods is utilizing machine learning algorithms to determine the boundary on the varying trend of one or more representative features. These kinds of methods overcome some shortcomings of traditional signal analysis methods such as heavy dependence on prior knowledge, etc. To achieve intelligent detection or diagnosis, these kinds of methods usually include two main steps: (1) feature extraction, and (2) constructing a detection model using machine learning algorithms. For example, Tabrizi et al. [4] used wavelet packet decomposition (WPD) and ensemble empirical mode decomposition (EEMD) to extract feature vectors of vibration signals, and run a support vector machine (SVM) algorithm to assess the state of bearing. Li et al. [5] used supervised local Fisher discriminant analysis to reduce feature dimension, and then used a K-nearest neighbor algorithm to recognize the early fault state. Ocak et al. [6] proposed a bearing fault detection and diagnosis scheme based on a hidden Markov model (HMM). This scheme detects early fault of bearings in an online mode by monitoring the change of probability of HMM model, which is pre-trained under normal state.

It is worth noting that, since the state of incipient fault cannot be recognized in advance, incipient fault detection is essentially an anomaly detection problem in which only one class is available and the anomaly boundary needs to be built in a learning process. Some commonly-used anomaly detection algorithms like support vector data description (SVDD) [7], One-class SVM [8], local outlier factor (LOF) [9], iFOREST [10], etc. have also been applied to early fault detection of bearings. These methods usually use the starting part of normal state data to establish a one-class classification model or build a criterion for determining anomalies. However, as stated in the Introduction section, these methods are generally incapable of tackling the irregular fluctuations in normal state, and consequently, it is easy to arouse false alarm.

In the most recent years, the development of deep learning techniques provides another effective solution for fault feature extraction. As one of the pioneer works, Lei et al. [11] proposed a new local connected deep neural network for fault diagnosis of bearings. This network is stacked by multiple normalized sparse auto-encoder (SAE). Shao et al. [12] proposed an optimized deep belief network (DBN) for fault diagnosis of rolling bearings. This method used a raw time-domain signal to extract directly the representative fault features via DBN. By putting the raw vibration signals into a convolutional neural network, Mao et al. [13] extract deep features of bearing fault with good

representation ability. These works demonstrate the promising performance of deep neural network on fault diagnosis problems in terms of adaptive and automatic feature extraction.

Moreover, fault diagnosis research using deep transfer learning methods have received extensive attention in the past two years. Transfer learning aims to improve prediction performance in a target domain by getting help from the data from source domain. Please note that transfer learning is an algorithm framework rather than a single algorithm. In many realization forms of transfer learning, deep neural network provides a convenient way. As a typical work, Yang et al. [14] proposed a feature-based transfer neural network that uses laboratory bearings diagnostic knowledge to identify the health of real-case bearings. Zhang et al. [15] proposed a deep transfer learning method for fault diagnosis. This method first learns features from a large amount of source data and adjusts the parameters of neural networks accordingly. Second, some parameters are transferred from source task to target task to assist model training on a small amount of target data. Lu et al. [16] used a three-layer SAE network with maximum mean discrepancy (MMD) regularizer to extract features from a raw vibration signal. Here, the MMD regularizer is used to punish the feature difference between training data and test data. By extending the marginal distribution adaptive (MDA) to the joint distribution adaptive (JDA), Han et al. [17] proposed a new fault diagnosis method that can adapt the conditional distribution of unmarked target data by using the discriminant structure in source domain. Through a more accurate distribution matching, this method can get better diagnosis performance.

However, different from the wide application in the field of fault diagnosis, there is no research found about early fault detection based on transfer learning yet. There may be two reasons as follows: (1) For the detection methods based on a binary classifier like SVM, the label information about early fault is much harder to precisely acquire than mature fault. Then, transfer learning could cause model bias unless a robust state assessment method with good anti-inference ability is introduced; (2) For the detection methods based on one-class classifier like SVDD and One-class SVM, only normal state data are available to build a detection model. In this scenario, the classification boundary is rather close to the normal state data (think of the hyper-sphere of SVDD as example), which lacks enough discriminant information about early fault and becomes more sensitive to the irregular fluctuation in normal state. As a result, it is sort of hard for transfer learning to reach good detection performance for early faults. We also notice that deep transfer learning (DTL) has been applied to the problem of remaining useful life (RUL) prediction. For instance, Mao et al. [18] proposed a RUL prediction method based on deep feature representation and transfer learning. By integrating weight transfer and hidden feature learning from historical failure data, this method realizes the prediction of new objects which do not have any supervision information for training. Although early fault detection can be regarded as a preparation process of RUL prediction, this method still can not solve the problems stated above.

Moreover, according to our literature survey, the research of deep learning techniques on early fault detection is still in its infancy. We only find very few works about this topic. As a representative work, Lu et al. [19] proposed an early fault detection method that estimates the occurrence of early fault from a distribution estimator based on a long short-term memory (LSTM) network. It is interesting that, in this paper, an efficient fault alarm strategy is also proposed. However, this work doesn't consider the false alarm rate in a normal state more. Zhao et al. [20] proposed a new bidirectional gated recurrent unit (GRU) network based on local features. The GRU decision model is established to identify early fault with different fault types. Working on deep feature representation, both of the methods can obtain better results than several traditional methods. However, these two methods estimate the fault occurrence merely using the data of target bearing, but they don't focus on extracting accurate early fault features, especially from massive data of auxiliary bearings. Mao et al. [21] noticed the distribution difference between auxiliary bearings and target bearing, and proposed an online fault detection method. Although this method can effectively use auxiliary bearings data to establish detection model, it improves the detection performance mainly by proposing a strategy named self-adaptive deep

feature matching (SDFM), not reducing such distribution difference. Moreover, no matter whether it is these three deep learning-based methods or most traditional anomaly detection algorithms, they are all unable to adapt to the irregular fluctuations of target bearing data which arrive sequentially. As a result, it is prone to reduce the model's robustness and arouse false alarm. According to our literature research, very few articles give an exact solution to improve the robustness of early fault detection model for rolling bearings.

3. Robust Detection Method Based on Deep Transfer Learning

In this section, we propose a robust detection method for incipient fault. As the boundary between normal state and early fault state is generally not obvious, the proposed method firstly conducts robust state assessment on multiple training bearings by applying deep transfer learning. Using the obtained normal state data and early fault state data of training bearings, an offline detection model is established on an SVM classifier. In the online stage, by feeding each data batch of target bearing into the offline SVM classifier sequentially, we can determine whether this data batch contains an early fault or not. Here, the effect of deep transfer learning is to get a common feature representation of training bearings data in a normal state, while the robust state assessment is dedicated to obtaining as accurate a label as possible for training the SVM classifier.

3.1. Preprocessing

As a lot of noise generally exists in original vibration signals, it is necessary to preprocess such a signal in advance. In addition, since the auto-encoder network utilizes an unsupervised learning mode, it is better to preprocess the raw signal to be the data with strong regularity before feature extraction. According to [22], Hilbert–Huang transformation (HHT) has been verified as an effective technique of signal analysis for feature extraction. Therefore, in this paper, we choose HHT to extract a marginal spectrum of the bearing signal as an input of the auto-encoder network. For sake of complete presentation, here we provide a brief explanation [23], as follows:

(1) Decompose the original vibration signal: $x(t) = \sum_{i=1}^{k} c_i(t) + r_k(t)$, where $x(t)$ denotes the original signal, $c_i(t)$ denotes the i-th intrinsic mode function (IMF) component, and $r_k(t)$ denotes the residual term.

(2) Run Hilbert transform for each IMF component:

$$H[x(t)] = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{x(\tau)}{t-\tau} d\tau$$

.

Construct the analytical signal: $C_i^A(t) = c_i(t) + jc_i^H(t) = a_i(t)e^{j\theta_i(t)}$, where $c_i^H(t) = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{c_i(s)}{t-s} ds$, $a_i(t) = \sqrt{c_i^2 + (c_i^H)^2}$, $\theta_i(t) = arctan(c_j^H / c_i)$. The instantaneous frequency is: $\omega = \frac{d\theta(t)}{dt}$.

(3) Construct Hilbert spectrum: $H(\omega, t) = \sum_{i=1}^{n} a_i(t)e^{j\theta_i(t)}$. The final marginal spectrum can be obtained through an integral of Hilbert spectrum: $H(\omega) = \int H(\omega, t)dt$.

3.2. Common Feature Representation Based on Transfer Learning

As a research hotspot in recent years, transfer learning aims to use existing knowledge to solve the problems in the different but related domains. Compared with traditional machine learning methods, transfer learning techniques can solve the problem of inconsistent distribution between training data and test data, while improving the generalization performance on an insufficient amount of training data. In several implementation forms of transfer learning, domain adaptation can map the data from different domains to a same feature space, so that the data from a source domain can be used to

enhance the training performance of a target domain. In model construction of domain adaptation, the maximum mean discrepancy (MMD) distance is often used to construct a new regularizer in loss function in order to make the feature distribution of two domains as identical as possible [24]. The MMD distance between two distributions is defined as [24]:

$$MMD(X,Y) = \left\| \frac{1}{n^x} \sum_{i=1}^{n^x} \phi(x_i) - \frac{1}{n^y} \sum_{j=1}^{n^y} \phi(y_j) \right\|_H^2 \tag{1}$$

where x_i and y_j respectively represent the samples in the source and target domains, and H indicates that the distance is measured in a reproducing kernel Hilbert space (RKHS). By calculating the difference between the mean values of the function on the original samples, the average difference between two distributions is achieved. If this value is small enough, the two distributions are thought to be the same, otherwise, they are considered to be different. By finding a ϕ to maximize the average difference, the common feature space between two domains can be obtained.

Based on the above analysis, we modify the loss function of auto-encoder by introducing the strategy of transfer learning. Thus, the auto-encoder network is capable of mapping normal state data of different bearings with an identical model size and under the same working condition to an approximately identical feature space. Specifically, this paper applies the DAE with domain adaptation proposed by [25] to train the loss function of DAE, including the following three terms.

(1) The first one is the reconstruction term of traditional DAE:

$$\text{Loss}_{DAE} = \frac{1}{2n} \|R - X\|_F^2$$

where X denotes the input sample matrix and R indicates the reconstruction feature of DAE, F indicates the Frobenius norm of matrix, and n is the number of samples.

(2) The second one is an MMD regularizer which constrains the distribution discrepancy between normal data of different bearings. We define the symbol C as the combination of multiple auxiliary bearings. The MMD regularizer is defined as:

$$\text{Loss}_{MMD} = \sum_{c=1}^{C} \left\| \frac{1}{n_c} \sum_{p=1}^{n_c} \varphi\left(x_{c,p}\right) - \frac{1}{n'_c} \sum_{q=1}^{n'_c} \varphi\left(x'_{c,q}\right) \right\|_H^2 \tag{2}$$

where x_c and x'_c represent the bearing samples in the source and target domains, respectively. In addition, n_c and n'_c denote the number of samples in the source and target domains.

(3) The third one is the weight regularization term that enhances the representative ability of features extracted from raw data, as follows:

$$\text{Loss}_{weight} = \sum_{\text{k}=1}^{K} \exp\left(-\|W_{\text{k}}\|_F^2 \big/ \sigma\right) \tag{3}$$

where σ is the width parameter, K is the total number of hidden layers, and W_k is the weight matrix of the k-th layer.

By integrating these three terms together, the final loss function of DAE with domain adaptation is:

$$\text{Loss} = Loss_{DAE} + \lambda Loss_{MMD} + \frac{\mu}{2} Loss_{weight} \tag{4}$$

where $\lambda > 0$ and $\mu > 0$ control the trade-off among three terms. Minimizing this loss function can be achieved using a gradient descent algorithm. It is worth noting that, different from [25] which adapts the data of different fault states under different working conditions, here we merely constrain the normal state data of training bearings to have a consistent data distribution for the building detection model. Please refer to the article [25] for the specific network structure.

3.3. Robust State Assessment Method

Based on the common feature representation obtained in the Section 3.2, a detection model needs to be established. Although we can directly train a binary classifier by means of normal state data and mature fault data, the classification model would be not suitable for early fault detection. To improve the discriminant ability of the detection model, it is important to find a method to precisely assess the state of early fault so as to obtain accurate label information. Since the common features cannot remove the irregular fluctuation in original data thoroughly, we need to further consider the robustness of the state assessment method. To get a more robust detection result, it is necessary to eliminate the abnormal points in the common feature space. Traditional state assessment methods are mostly based on singular value decomposition [26], or work directly on RMS value [27]. However, these methods do not consider the negative influence of fluctuations in normal state, so it is easy to generate false alarms.

In this section, we introduce the robust deep auto-encoder network (RDA) to conduct robust state assessment. The RDA is inspired by robust principal component analysis (Robust PCA) [28]. For input sample set X, the RDA decomposes X into two parts $X = L_D + S$, where L_D represents a low-rank common representation part (some linear correlations between rows or columns exist as the training data generally have some structural information inside), S is a sparse matrix (caused by noise and anomalies which are usually sparse). When the RDA is used for anomaly detection, the loss function for X can be considered as the reconstruction error of L_D plus the $L_{2,1}$ norm of S, as follows:

$$\begin{aligned} &\min_{\theta,S} \|L_D - D_\theta(E_\theta(L_D))\|_2 + \lambda\|S\|_{2,1} \\ &s.t.\ \ X - L_D - S = 0 \end{aligned} \tag{5}$$

where E_θ and D_θ respectively represent the encoder and decoder of RDA, θ is the weights of the encoder and decoder, λ is used to control the degree of sparsity in S. The larger the value of λ is, the less the sparsity of solution is, and vice versa. This optimization problem is solved by using the alternating direction method of multipliers (ADMM) algorithm. For the detailed derivation process, please refer to [29].

To sum up, the proposed state assessment method includes two steps:

(1) For the common feature set X of auxiliary bearings extracted by DAE with domain adaptation, we feed them into RDA and calculate L_D, which is the low-rank public representation of auxiliary bearing data. The specific steps to calculate L_D are as follows [29]:

   A. Initialize L_D, S to be zero matrices. Initialize an auto-encoder network with random parameters.
   B. Remove S from X and use the remainder $L_D = X - S$ to train the auto-encoder.
   C. Minimize the reconstruction error by using back-propagation algorithm.
   D. Set L_D to be the reconstruction from the trained autoencoder: $L_D = D_\theta(E_\theta(L_D))$.
   E. Set S to be the difference between X and L_D: $S = X - L_D$.
   F. Optimize S using a proximal operator: $S = prox_{\lambda,l_{2,1}}(S)$.
   G. If the value of S changes less than a pre-defined threshold in two consecutive iterations, return L_D and S, otherwise go to Step B.

(2) For L_D, we train a SVDD model by using the starting part of training data of each auxiliary bearing, and then use the obtained SVDD model to identify the state of each sample in L_D. The SVDD model is a one-class classification algorithm which can detect abnormal samples only using positive samples [30]. SVDD constructs a hyper-sphere which covers as much as target data, and recognizes the sample outside the sphere's boundary as anomaly. The optimization target of SVDD is:

$$\begin{aligned} &\min_{R,a,\xi} \quad R^2 + C\sum_{i=1}^{n}\xi_i \\ &s.t. \quad \|x_i - a\|^2 \le R^2 + \xi_i \end{aligned}$$

where a and R are the center and radius of hyper-sphere, respectively, $\|x_i - a\|$ indicates the distance from the sample x_i to the center a, $\xi_i \geq 0$ is slack variable, C is regularization parameter which makes a trade-off between the hyper-sphere volume and misclassification level. The SVDD model can be optimized by Lagrange multiplier method [30]. For auxiliary bearings, the starting part of offline data can be viewed as positive class data, and the position where anomalies occur can be judged by SVDD. Since a sample in L_D corresponds to an original sample of auxiliary bearings, the state assessment for the auxiliary bearings is conducted.

Different from the current state assessment methods [26,27], the proposed method conduct assessment in a low-rank space which is optimized by RDA. Therefore, the proposed method has good anti-interference ability against irregular fluctuation in normal state data. It is worth noting that RDA is merely used to find the low-rank space and SVDD is used to conduct anomaly detection in this space. The flowchart of the RDA algorithm is also shown in Figure 4.

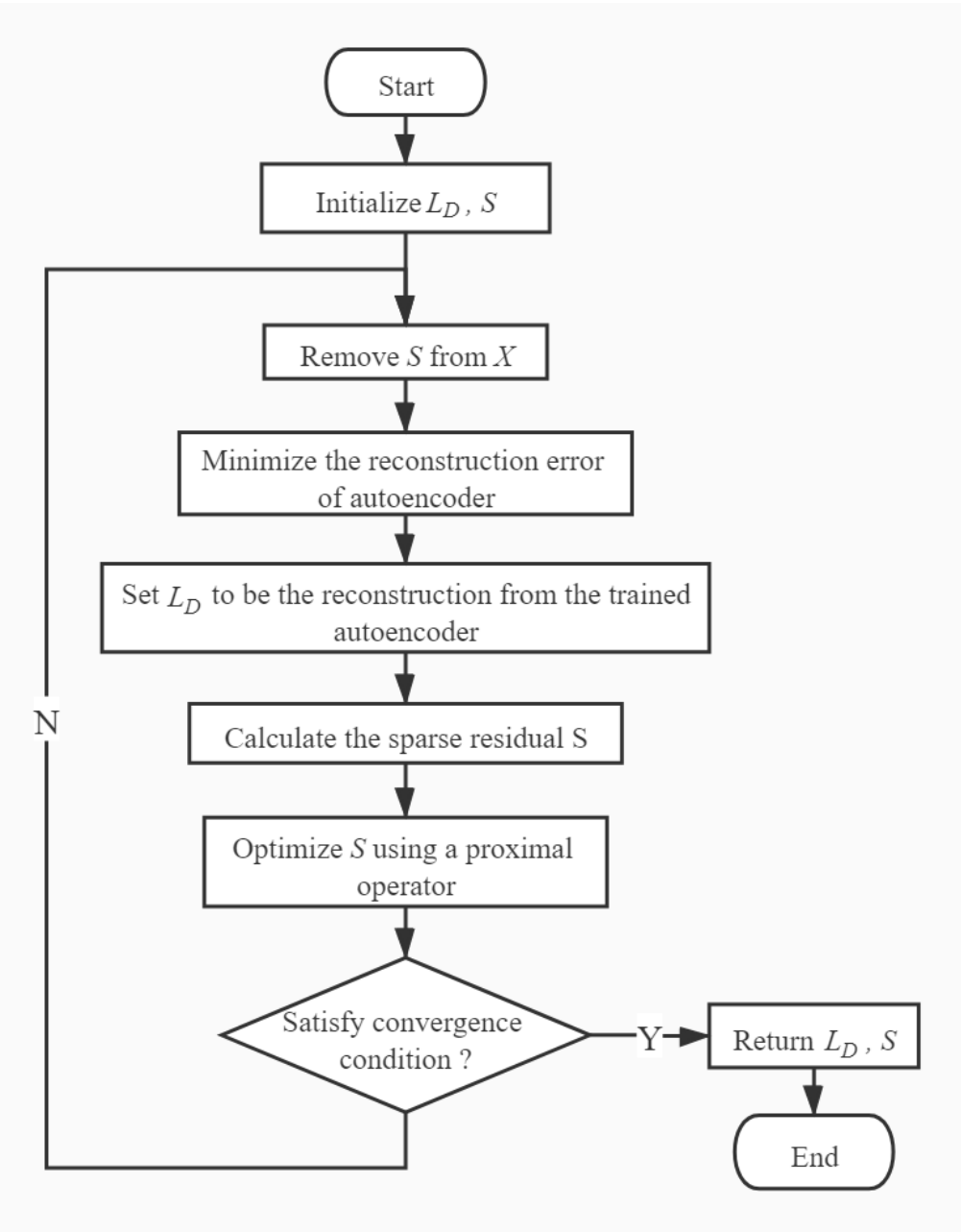


Figure 4. Flowchart of the RDA algorithm.

3.4. Online Detection of Early Fault

Based on the state assessment results on offline auxiliary bearings, we choose the samples with common features in normal state and early fault state to construct a SVM classification model. This classifier model is used to conduct online detection.

In the stage of online detection, the vibration signal of target bearing is collected sequentially. Each time a certain length of signal is taken to extract HHT marginal spectrum. Then, the obtained marginal spectrum data are put into the common feature representation which is established by DAE with domain adaption. Finally, the obtained deep features are fed into the above SVM classifier to detect whether early fault occurs. The flow chart for online detection is shown in Figure 5.

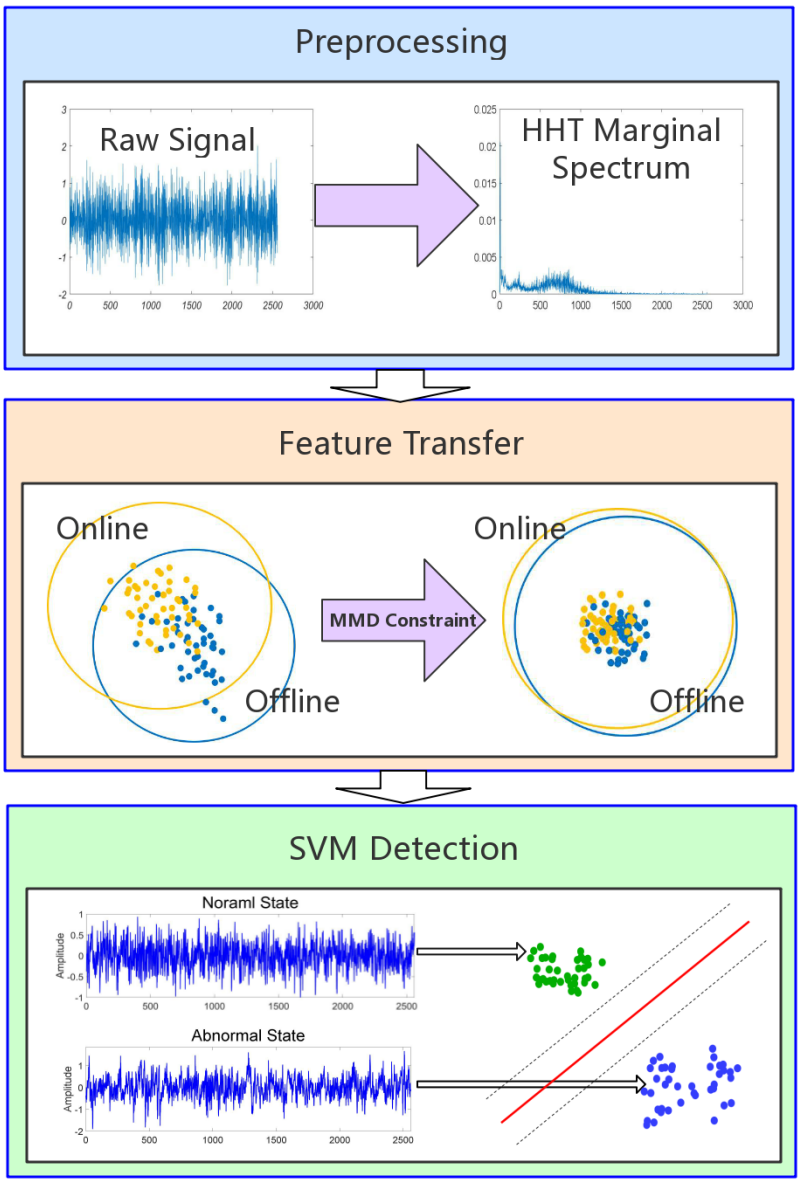


Figure 5. Flow chart of online detection for early fault.

3.5. Process of the Proposed Method

From the perspective of algorithm execution, the proposed method can be divided as an offline stage and an online stage. Note that the DAE with domain adaptation and the RDA both run in the

offline stage, and the online stage only covers the detection process for target bearing which certainly belongs to the target domain. The implementation flow of the proposed method is shown in Figure 6. For sake of better understanding, we also list each step as follows.

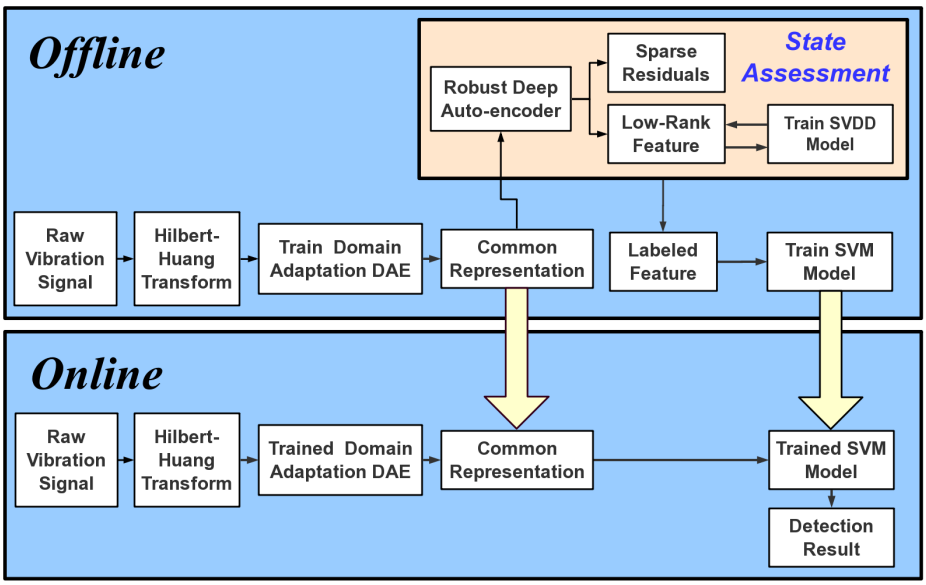


Figure 6. Flow chart of the proposed method.

3.5.1. Offline Stage

The offline stage mainly covers extracting common feature representation, running robust state assessment and building an SVM classifier:

Step 1. Extract marginal spectrum by HHT from raw vibration signals in source and target domains.

Step 2. Feed the marginal spectrum data into the DAE with domain adaptation to obtain common features of training bearings.

Step 3. On the basis of the obtained common features, the robust state assessment method is applied to to specify the location of early fault occurrence in a low-rank space found by RDA. The results are used to add label for the training samples of auxiliary bearings, i.e., setting the samples before the location as positive while the samples after this location as anomalous.

Step 4. Train a SVM classification model using the labeled samples in Step 3. This classifier serves as the final detection model.

3.5.2. Online Stage

In an online stage, there are three steps to conduct online detection:

(1) Extract marginal spectrum by HHT from online data batch of test bearing .
(2) Feed the marginal spectrum data into the DAE with domain adaptation trained in offline stage to extract the common features.
(3) Put the common features to the SVM classifier trained in the offline stage and then obtain detection results.

4. Experimental Results

In order to verify the effect of the method proposed in this paper, simulation experiments are carried out on IEEE PHM Challenge 2012 dataset and XJTU-SY bearing dataset, respectively. The programming environment used in this paper is Python 3.6, and the computer used in the

experiment is configured as i5-7300HQ processor with 8 GB of memory. All data are linearly normalized to [−1,+1] before processing.

4.1. Dataset Description

4.1.1. IEEE PHM Challenge 2012 Dataset

The dataset of IEEE PHM Challenge 2012 was collected from the test platform named PRONOSTIA, which is shown in Figure 7. This experimental platform can provide the experimental data of the whole life cycle of rolling bearings from normal to fault. The dataset contains bearing data under three different working conditions. In the first condition, the engine speed is 1800 rpm and the load is 4000 N. In the second condition, the engine speed is 1650 rpm and the load is 4200 N. In the third working condition, the speed is 1500 rpm and the load is 5000 N.

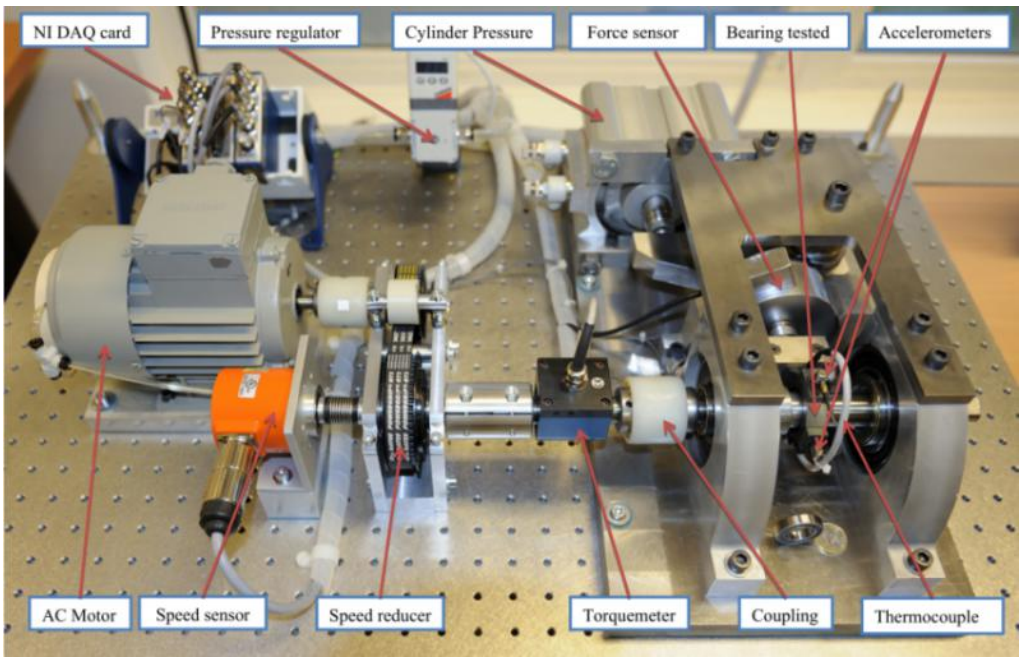


Figure 7. PRONOSTIA test platform [31].

4.1.2. XJTU-SY Bearing Dataset

The bearing accelerated degradation test platform used in the experiment was designed by the joint laboratory and manufactured by shengyang technology. The bearing testbed is shown in Figure 8. This platform can carry out accelerated degradation experiments of various kinds of rolling bearings or sliding bearings, and obtain the monitoring data of the whole life cycle of bearings. The experimental object of the data set was LDK UER204 rolling bearing. Three kinds of experimental working conditions were designed, and five bearings were tested in each working condition. In the first condition, the engine speed is 2100 rpm and the load is 12 kN. In the second condition, the engine speed is 2250 rpm and the load is 11 kN. In the third working condition, the speed is 2400 rpm and the load is 11 kN.

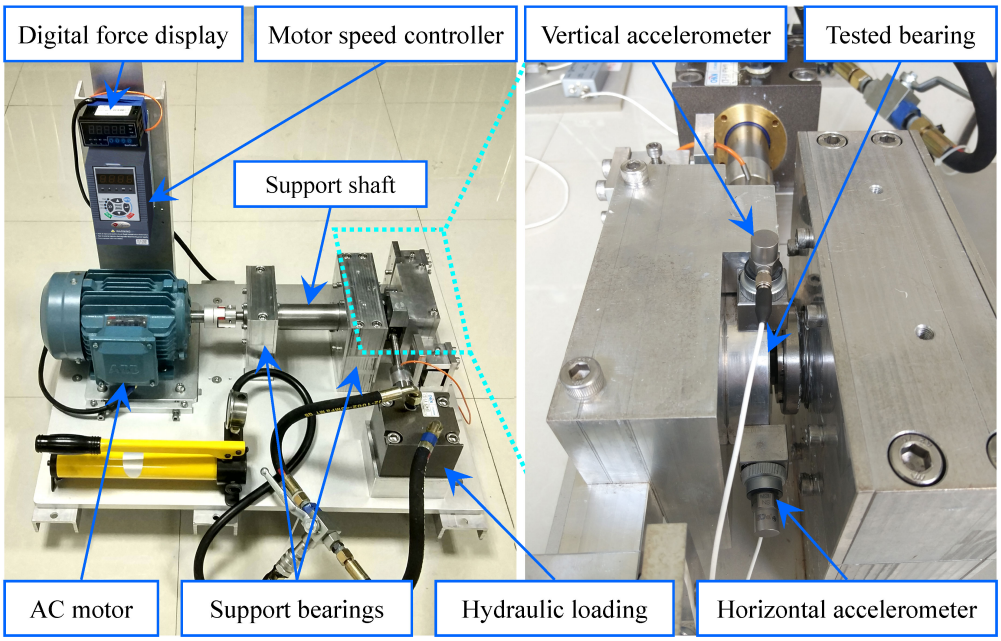


Figure 8. XJTU-SY bearing accelerated degradation testbed [32].

4.2. Experimental Setup

4.2.1. Experiment 1

For the IEEE PHM Challenge 2012 dataset, the bearings 1, 2, 3, 4, 5, 6, and 7 under the first condition are selected as source domain, and the bearings 1, 2, and 6 under the second condition are as target domain. The experimental settings are shown in Table 1. In this experiment, we randomly choose one bearing in the target domain as test bearing, and the other bearings in the source domain and target domain are selected as training bearings.

Table 1. Setting of experiment 1.

Bearing in the Experiment 1	Actual Bearing
Source 1	Bearing1-1
Source 2	Bearing1-2
Source 3	Bearing1-3
Source 4	Bearing1-4
Source 5	Bearing1-5
Source 6	Bearing1-6
Source 7	Bearing1-7
Target 1	Bearing2-1
Target 2	Bearing2-2
Target 3	Bearing2-6

4.2.2. Experiment 2

For the XJTU-SY dataset, the bearings 1, 2, 3, and 4 under the first condition are selected as source domain, and the bearings 1, 2, and 3 under the second condition are as target domain. The experimental settings are shown in Table 2. The same as Experiment 1, we randomly choose one bearing in the target domain as test bearing, and the other bearings in the source domain and target domain are selected as training bearings.

Table 2. Setting of experiment 2.

Bearing in the Experiment 2	Actual Bearing
Source 1	Bearing 1-1
Source 2	Bearing 1-2
Source 3	Bearing 1-3
Source 4	Bearing 1-4
Target 1	Bearing 2-1
Target 2	Bearing 2-2
Target 3	Bearing 2-3

4.3. Experiment 1

4.3.1. Preprocessing

In this section, the HHT method is used to preprocess the raw vibration signal. Taking the Target 1 as test bearing, we plot the raw time signal and the corresponding HHT marginal spectrum in three health conditions, as shown in Figure 9. It is obvious that the HHT marginal spectrum changes largely from normal state to severe fault state, which indicates that HHT marginal spectrum is sensitive to the change trend of time signal and beneficial to further deep feature extraction.

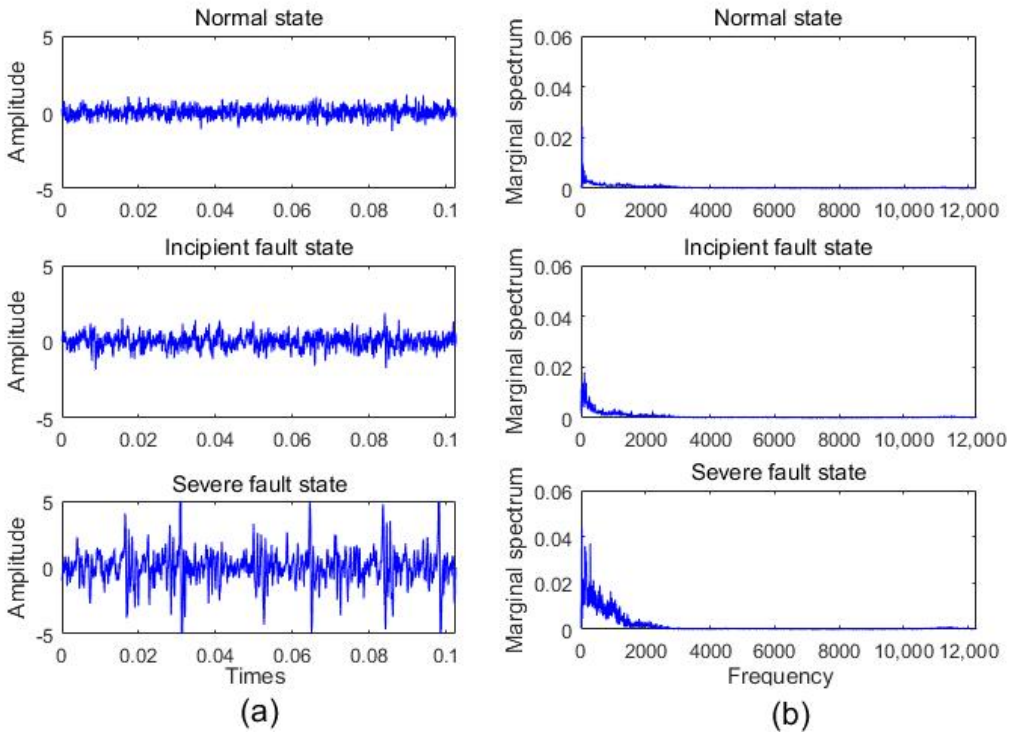


Figure 9. The (**a**) raw time signal and (**b**) HHT marginal spectrum of the bearing 1 from IEEE PHM Challenge 2012 dataset.

4.3.2. Extraction of Common Feature Representation by Transfer Learning

In this section, we adapt the normal state data of training bearings to a common distribution. We also choose the Target 1 as the test bearing. The first 500 samples of seven source domain bearings are taken as normal state samples, and the first 100 samples of two target domain bearings (i.e., Target 2 and Target 3) are also taken to train the DAE with domain adaptation. The determination of network structure and parameters gets reference from [25]. The parameter λ and μ in Equation (4) are set 0.001 and 0.0001, respectively. The structure of encoder is set [2558, 1200, 600], which means

the input dimension is 2558 (from HHT marginal spectrum) and two hidden layers are with 1200 and 600 neurons. The probability density distribution of these nine training bearings in three feature spaces is shown in Figure 10. To visualize such distribution, we first reduce each feature representation to one dimension by using principal component analysis (PCA).

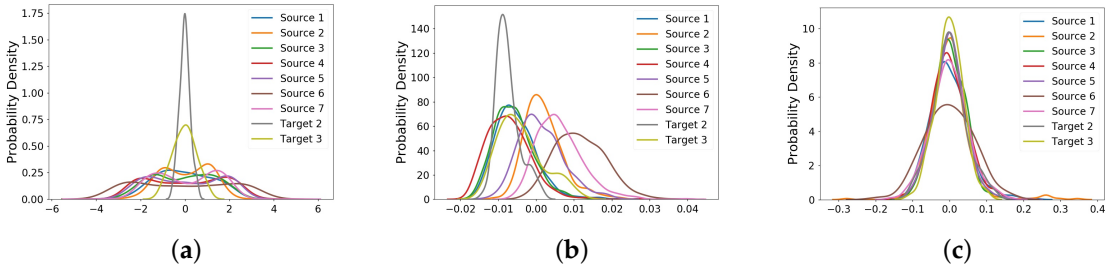


Figure 10. Probability density distribution of nine training bearings with (**a**) raw time signal, (**b**) HHT marginal spectrum feature, and (**c**) deep feature extracted by DAE with domain adaption.

From Figure 10, the probability density distribution of nine training bearings is obviously different in raw time domain, but after the common feature mapping, the probability density distribution of all bearings tends to be consistent, approximately in accordance with the same distribution. These comparative results show that the DAE with domain adaptation is able to map the data of different bearings to a common feature subspace, which eliminates the phenomenon of inconsistent distribution in normal state.

In addition, to further verify the effect of transfer learning on feature extraction, we plot the feature distribution before and after domain adaptation, as shown in Figure 11. In addition, for the sake of illustration, we use PCA to visualize the distribution.

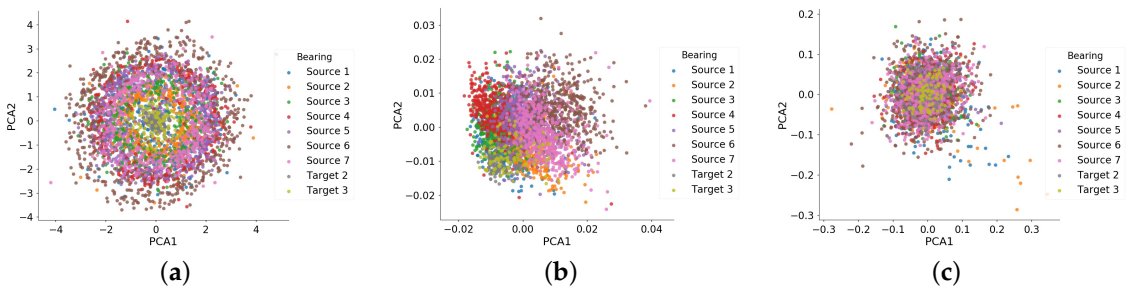


Figure 11. Feature distribution of nine training bearings with (**a**) raw time signal, (**b**) HHT marginal spectrum, and (**c**) the common features extracted by domain adaptation DAE.

It can be seen from Figure 11a that features of different bearings are gathered into rings with different sizes, which indicates that the spatial data distribution in a normal state of nine training bearings are obviously different in time domain. After HHT, the features are gathered towards the center, but some exceptional clusters still exist. After domain adaptation as shown in Figure 11c, the samples of different bearings interweave together in the obtained feature space. In this scenario, the features by domain adaptation DAE have an approximately identical distribution and are suitable for further state assessment.

It is worth noting that Figures 10 and 11 only provide the feature distribution of offline training bearings. In order to verify the effect of transfer learning on target bearing, we choose Target 1 and Target 2 as the test bearing, respectively, and obtain the corresponding features by putting their data directly into the DAE model established in the offline stage. The probability density distribution of Target 1 and Target 2 in three feature spaces are shown in Figures 12 and 13, respectively. To make an intuitive comparison, we also plot the distribution of nine auxiliary bearings (blue line).

From Figures 12 and 13, we find that transfer learning can provide a feature distribution between target domain bearings and source domain bearings which is more identical than the distribution of raw signal and HHT marginal spectrum. As a result, the detection for target domain bearings can get prior information directly from source domain bearings, with no model bias which is generally caused by the inconsistent data distribution between training and test data.

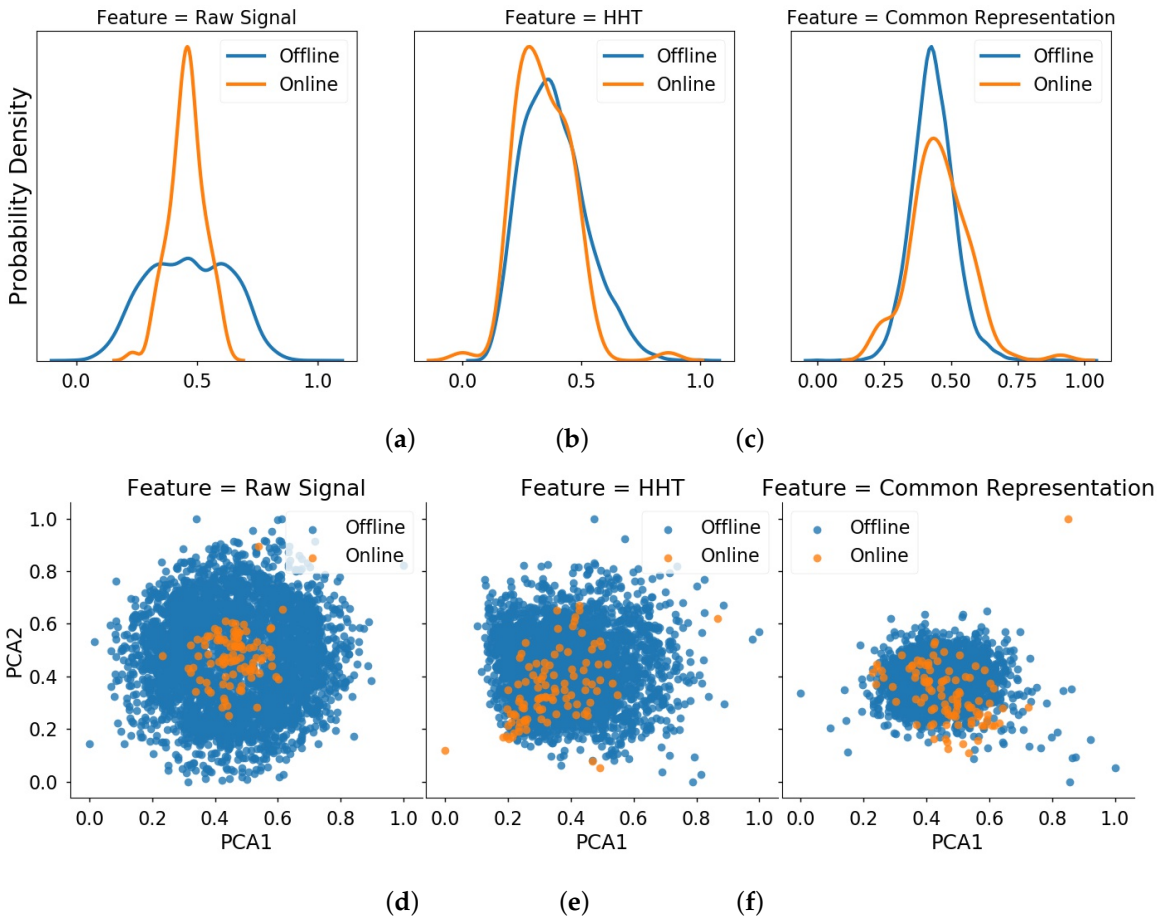


Figure 12. Comparative distribution of Target 1 and nine training bearings, with (**a**–**c**) the probability density distribution of raw time signal, HHT marginal spectrum, the common deep features by domain adaptation DAE, and (**d**–**f**) the feature distribution corresponding to (**a**–**c**), respectively. Please note that the legend "Online" denotes the distribution of the target bearing (Target 1) and "Offline" denotes the distribution of nine training bearings (similar to Figures 10 and 11).

4.3.3. State Assessment Using RDA

In this section, we use the common features obtained in the Section 3.2 to establish a state assessment model. To reduce the negative effect of irregular fluctuation in normal state data, the RDA algorithm is used to get a low-rank feature sub-space. Based on this sub-space, the SVDD algorithm is adopted to identify the state of each sample. Then, the normal state and early fault state on training bearings can be determined.

Here, we take Experiment 1 as an example. After obtaining a set of common features X via transfer learning, we put X into RDA to calculate the low-rank representation L_D and sparse residual S. The hidden layer units of RDA are set [600, 300, 150] and the parameter λ is set 0.2. Figure 14 shows the decomposition results of RDA on Sources 1 and 3. For sake of illustration, we choose the first feature from the results of RDA.

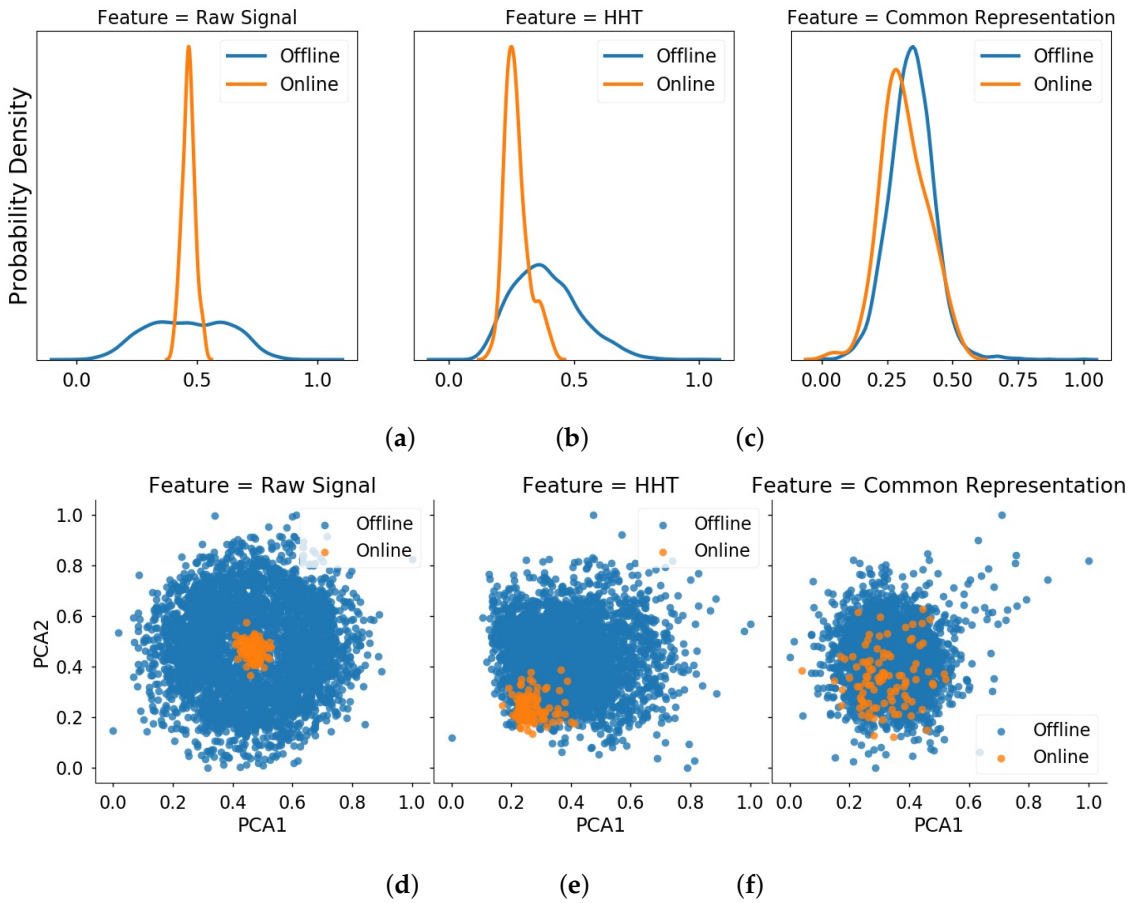


Figure 13. Comparative distribution of Target 2 and nine training bearings, with (**a**–**c**) the probability density distribution of raw time signal, HHT marginal spectrum, the common deep features by domain adaptation DAE, and (**d**–**f**) the feature distribution corresponding to (**a**)–(**c**), respectively. Please note that the legend "Online" denotes the distribution of the target bearing (Target 2) and "Offline" denotes the distribution of nine training bearings. (similar to Figures 10 and 11).

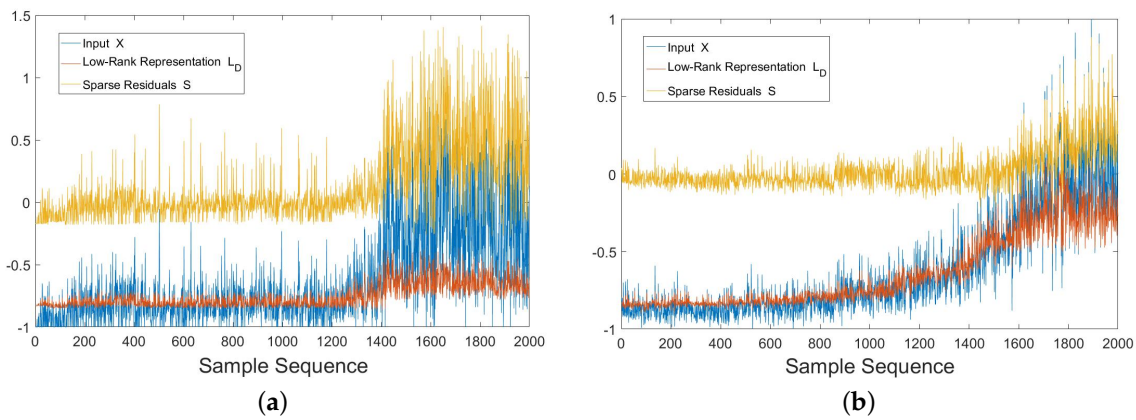


Figure 14. Decomposition results of the first feature extracted by RDA on (**a**) Source 1 and (**b**) Source 3.

From Figure 14, there are many fluctuations in raw input data X. After the decomposition of RDA, it is clear that many sparse and sharp components are separated to S, while the essential representation part L_D of X becomes more stable in normal state and change drastically at some points (considered as early fault occurrence). The results in Figure 14 indicates that RDA can find the essential representation part L_D of X by means of the sparse $L_{2,1}$ norm penalty. Compared to X, L_D is more stable because

outliers in the common feature space have been removed. Then, a more reliable state assessment result of bearing data can be achieved based on L_D.

For L_D, we take the first 500 samples of each training bearing (considered as normal state data) to construct a SVDD model. The output of SVDD model is then used to assess the state for the remaining samples of L_D. In this experiment, the toolbox of SVDD [33] is adopted. Gaussian kernel is used where the kernel parameter is set 0.001. In addition, the regularization parameter is set 1. We provide the results of state assessment for each training bearing in Table 3. Here, Target 1 is chosen as test bearing.

Table 3. Results of state assessment on training bearings with Target 1 chosen as test bearing.

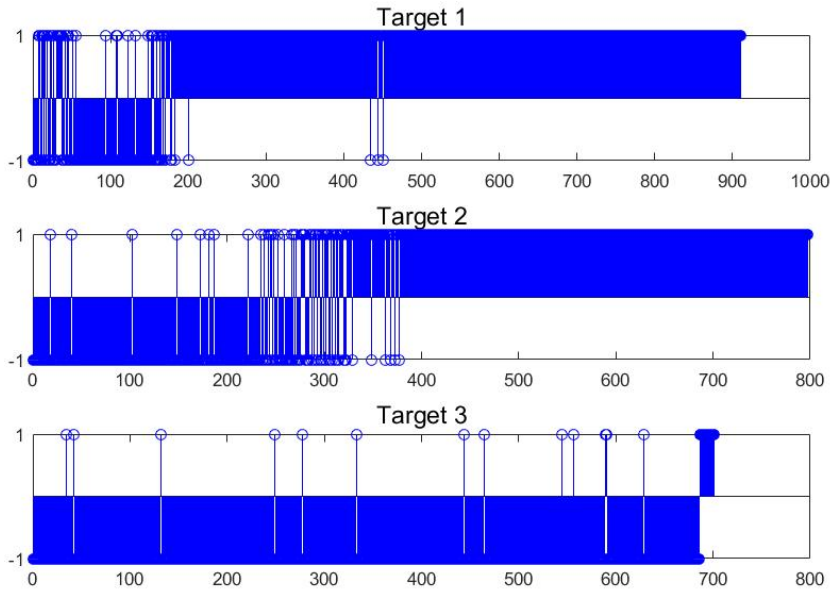
Training Bearing	Normal State Period	Fault State Period
Source 1	[1–1405]	[1405–2803]
Source 2	[1–826]	[827–871]
Source 3	[1–1174]	[1175–2375]
Source 4	[1–1087]	[1088–1428]
Source 5	[1–2443]	[2444–2463]
Source 6	[1–1590]	[1591–2448]
Source 7	[1–2212]	[2213–2559]
Target 2	[1–255]	[255–797]
Target 3	[1–688]	[688–701]

4.3.4. Comparative Results of Online Detection

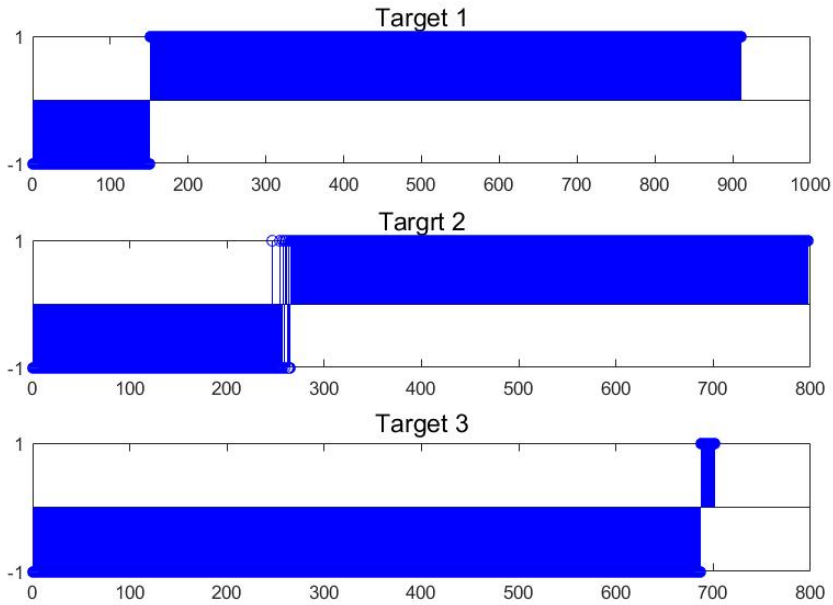
In this section, we construct a SVM classifier based on the results of state assessment listed in Table 3. When training the SVM model, all the normal state data in Table 3 are used as positive samples, and all the fault state data are used as negative samples. RBF Gaussian kernel is used, and the regularization and kernel parameters are set to 1 and 0.002, respectively. Please note that Table 3 is only for the detection on Target 1 which is chosen as test bearing. Besides Target 1, we also choose bearings 2 and 3 as test bearing respectively. Due to space limitation, we would not provide the results of state assessment on training bearings just like Table 3. The results of anomaly detection on the test bearings are shown in Figure 15. For comprehensive comparison, we also provide the results of SVDD based on HHT marginal spectrum in which we take the first 500 samples of each training bearing to construct the SVDD model. In addition, the regularized parameter is set to 1 and the kernel parameter is set to 0.001.

In this experiment, we choose the following alarm strategy: only five successive anomalies can trigger alarm. The point where such anomalies appear is defined the occurrence location of early fault. Moreover, false alarm is defined as the anomalies which appear before this location.

From Figure 15b, it is clear that the proposed method doesn't raise any false alarm in the initial part of normal state on three bearings. However, from Figure 15a, SVDD tends to get many more anomalies that are considered as false alarms. Especially on Target 1, SVDD has a few of anomalies in the starting part. This phenomenon is obviously caused by the run-in period of the bearing in which the collected vibration signal is non-stationary. If the detection model is not robust enough, it is easy to generate false alarm. Moreover, even after the occurrence location around the point 160 on Target 1, SVDD still identifies several anomalies in normal states. This comparison indicates that the robustness of the proposed method has been significantly improved.



(**a**)

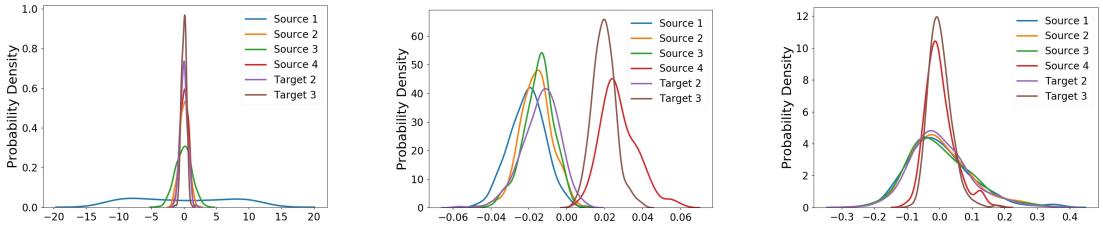


(**b**)

Figure 15. Detection results on Targets 1, 2, and 3 of Experiment 1 by using (**a**) SVDD and (**b**) the proposed method, where the label 1 and −1 represent normal state and anomaly state, respectively.

4.4. Experiment 2

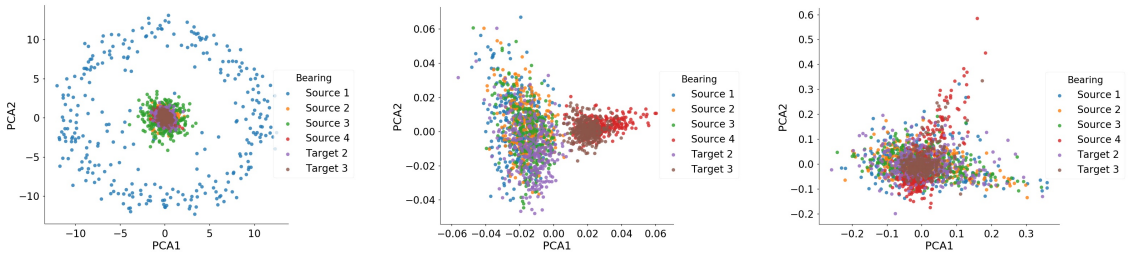
For the XJTU-SY dataset, the experimental steps are the same as the ones in Experiment 1. Due to space limitation, here we briefly provide some important results, with no detailed description of experimental process. The settings of network structure and hyper-parameters are almost identical to Experiment 1. First, we choose Target 1 in Table 2 as a test bearing while the other six bearings in source domain and target domain are as training bearings. The probability density distribution of these six training bearings in three feature spaces is shown in Figure 16.



(**a**) (**b**) (**c**)

Figure 16. Probability density distribution of six training bearings with (**a**) raw time signal, (**b**) HHT marginal spectrum feature, and (**c**) deep feature extracted by domain adaptation DAE.

Obviously, Figure 16 provides a similar phenomenon about common feature mapping with Figure 10, which indicates again that transfer learning is beneficial to extract common features for the bearings data under different working conditions. We also plot the feature distribution before and after transfer learning, as shown in Figure 17. It is clear that the domain adaptation DAE is capable of extracting a set of deep features with consistent distribution on different bearings.



(**a**) (**b**) (**c**)

Figure 17. Feature distribution of six training bearings with (**a**) raw time signal, (**b**) HHT marginal spectrum, and (**c**) the common features extracted by domain adaptation DAE.

We further check the effect of transfer learning on test bearing, we choose Target 1 and Target 2 in Table 2 as the test bearing, and plot their probability density distribution in three feature spaces in Figures 18 and 19, respectively. The blue line is the distribution of six training bearings for comparison, and the meaning of the legend "Online" and "Offline" are same as Figures 12 and 13. Similar to Experiment 1, transfer learning can extract a more identical feature representation between two domains than the other two methods, which is definitely helpful to improve the detection performance for the bearings under different working conditions.

Figure 20 provide the decomposition results of RDA on Source 1. Obviously, through low-rand decomposition, some outliers can be recognized and eliminated, which is regarded as beneficial for getting reliable state assessment results. Due to space limitation, here we won't provide the detailed numerical results about state assessment.

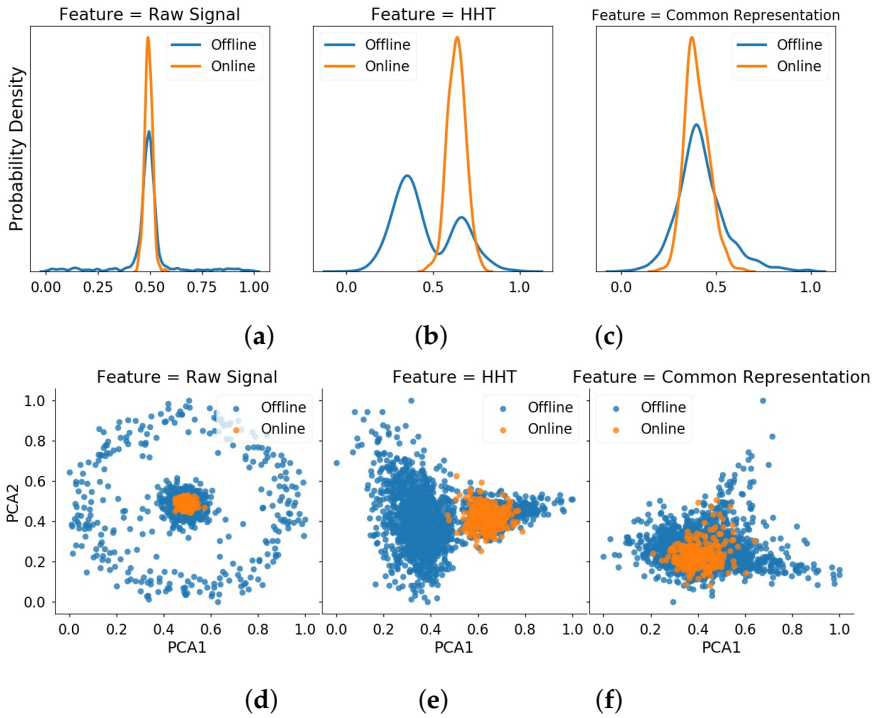


Figure 18. Comparative distribution of Target 1 and six training bearings, with (**a**–**c**) the probability density distribution of raw time signal, HHT marginal spectrum, the common deep features by domain adaptation DAE, and (**d**–**f**) the feature distribution corresponding to (**a**–**c**), respectively.

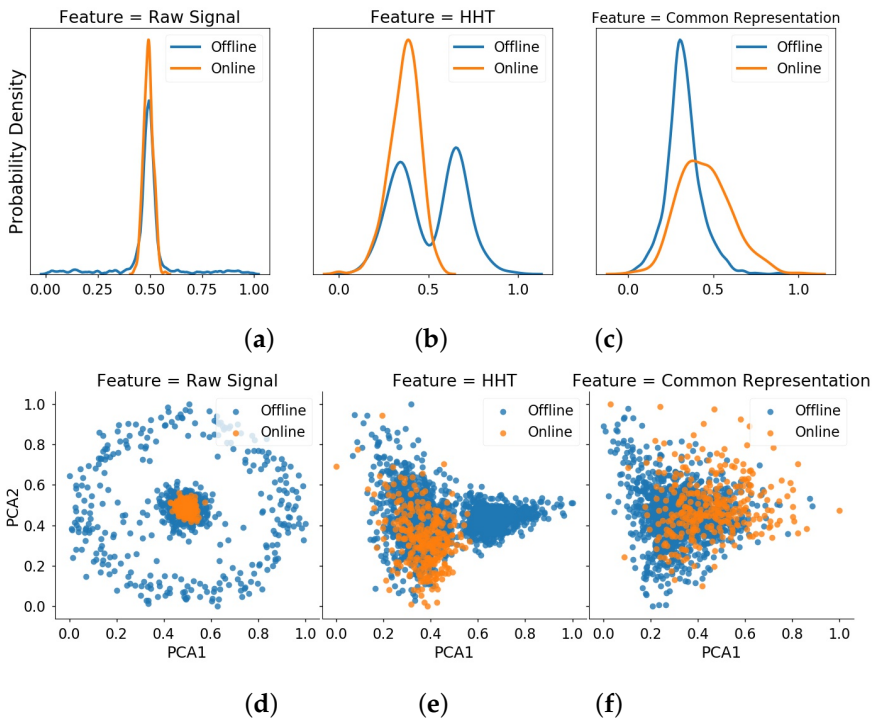


Figure 19. Comparative distribution of Target 2 and six training bearings, with (**a**–**c**) the probability density distribution of raw time signal, HHT marginal spectrum, the common deep features by domain adaptation DAE, and (**d**–**f**) the feature distribution corresponding to (**a**–**c**), respectively.

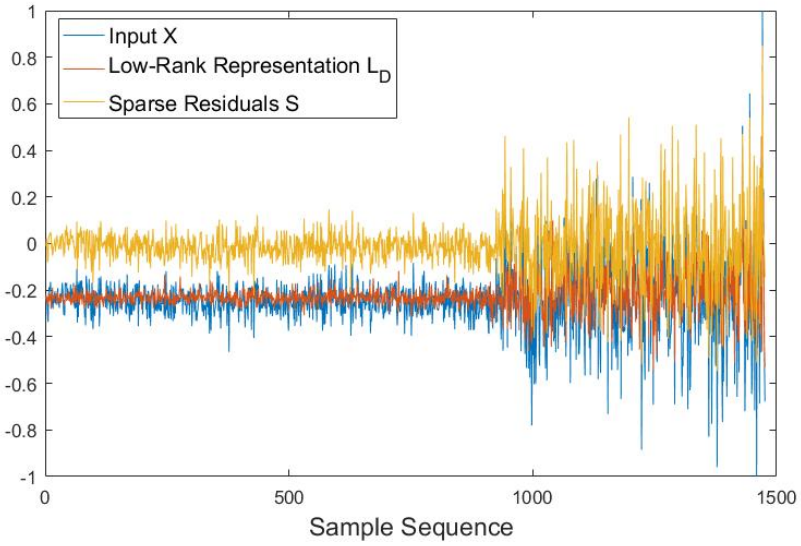


Figure 20. Decomposition results of the first feature extracted by RDA from Source 1 of Experiment 2.

Taking Target 1, Target 2, and Target 3 as test bearing respectively, we provide the results of anomaly detection on these bearings are shown in Figure 21. The alarm strategy is also the same as Experiment 1, i.e., five successive anomalies indicate incipient fault occurrence, and the anomalies before the occurrence locations are regarded as false alarm. From Figure 21b, the proposed method gets very similar comparative results like Figure 15b. Specifically, the most interesting phenomenon is also the very low number of false alarms. Even if few anomalies appear in the normal state on Target 2, they are still much less than the number of false alarms by SVDD (as shown in Figure 21a). According to our observation, the degradation process of Target 2 has many irregular fluctuations in the normal state that is why SVDD has so many false alarms. Obversely, the proposed method presents a much more robust effect in early fault detection, which demonstrates the effectiveness of deep transfer learning and robust state assessment.

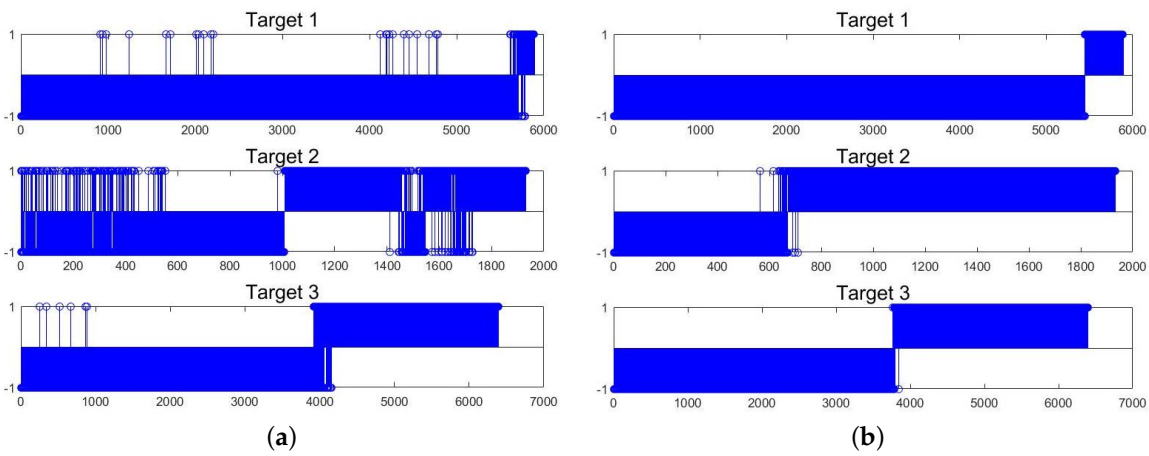


Figure 21. Detection results on Targets 1, 2, and 3 of Experiment 2 by using (**a**) SVDD and (**b**) the proposed method, where the label 1 and −1 represent normal state and anomaly state, respectively.

5. Comparative Experiment

In order to further verify the effectiveness of the proposed method, we compare four anomaly detection methods that are widely used in the field of early fault diagnosis and detection. These methods include LOF [9], One-class SVM [8], SVDD [7], and iFOREST [10]. Besides HHT marginal spectrum (HHT-MS), there are two typical features used in the comparative experiment,

i.e., RMS and Kurtosis value of a raw time signal. By combining the three features and the four methods, we have 12 methods of combination for comparison.

In addition, we also compare two state-of-the-art methods, BEMD+AMMA [34] and SDFM [21]. BEMD+AMMA which utilizes bandwidth EMD is viewed as the state-of-the-art incipient fault diagnosis method based on signal analysis. SDFM is viewed as the newest work about incipient fault detection with deep learning. Considering that our method is the application of transfer learning, we also compared two typical transfer learning algorithms, namely, transfer componet analysis (TCA) [35] and geodesic flow kernel (GFK) [36]. These two algorithms both achieve good performance of the problem of domain adaptation just like ours. For these two algorithms, the input is HHT marginal spectrum and the detection model is SVDD. Then, we call these two methods as TCA+SVDD and GFK-SVDD.

All 16 methods for comparison are listed as follows:

1. HHT-MS + One-class SVM
2. RMS + One-class SVM
3. Kurtosis + One-class SVM
4. HHT-MS + SVDD
5. RMS + SVDD
6. Kurtosis + SVDD
7. HHT-MS + LOF
8. RMS + LOF
9. Kurtosis + LOF
10. HHT-MS + iFOREST
11. RMS + iFOREST
12. Kurtosis + iFOREST
13. BEMD + AMMA [34]
14. SDFM [21]
15. TCA [35] + SVDD
16. GFK [36] + SVDD

In this paper, two estimate metrics are used to evaluate the detection methods: detection location and false alarm number. Here, the detection location is the occurrence location of early fault determined by a certain threshold or criteria. We will give the specific threshold and criteria setting for each method in the following part.

For SVDD and One-class SVM, we use cross-validation strategy to determine the optimal hyper-parameters. For LOF, the value of K is set 10. For iFOREST, the number of trees is set 100, and each tree has 256 samples. Moreover, for SVDD, One-class SVM, and LOF, we take the first 500 samples of bearings as normal data to train a detection model. For iFOREST, the maximum number of segmentation times for the first 500 samples is taken as the threshold. An abnormal sample will be determined if the maximum number of the segmentation times in successive ten samples is less than the threshold.

For all methods, early fault is considered as occurring if five anomalies appear successively in the detection results. In particular, considering LOF is not sensitive to the early fault as it has almost no anomalies in normal state, we reduce the threshold from five successive samples to two successive samples. For BEMD + AMMA, it does not involve the issue of false alarm because this method directly matches the fault characteristic frequency.

Due to space limitation, we choose one bearing from the datasets of IEEE PHM Challenge 2012 and XJTU-SY, respectively, and provide the comparative results in Table 4. Specifically, for PHM Bearing1-1, the methods 1, 14 and our method all get earlier detection location than others. However, our method has only 14 false alarms while the false alarm numbers of the methods 1 and 14 are 138 and 42, respectively. As far as the false alarm number is concerned, method 7 is the least, but its detection location is too lagging, which indicates that this method is insensitive to early fault. The reason may be

that the fault threshold determined by LOF is relatively high, leading to a seriously delayed detection result and a small number of false alarms. Similar results can be found on the XJTU-SY dataset. HHT + LOF and RMS + LOS also get small numbers of false alarms, but their detection results are much delayed. This phenomenon indicates that the detection performance of LOF heavily relies on its threshold. For the XJTU-SY dataset, the method 7, 14 and our method have more earlier detection location than others, while the false alarm number of the method 7 is the least. Although the detection result of the method proposed in this paper is not as early as that of method 14, the false alarm number of the method proposed in this paper is 21 less than that of method 14, and the detection result is not delayed too much. Consequently, our method can be considered to be more sensitive to bearing early fault and with better robustness as well. To sum up, our method can determine early fault in earlier time than other methods with less number of false alarms. Therefore, our method is more suitable for online detection of rolling bearings.

Table 4. Comparison of early fault detection results on the datasets of IEEE PHM Challenge 2012 and XJTU-SY.

Method	PHM Bearing1-1		XJTU-SY Bearing1-1	
	Detection Result	False Alarm	Detection Result	False Alarm
1. HHT + One-class SVM	1410	138	967	52
2. RMS + One-class SVM	1640	27	945	19
3. Kurtosis + One-class SVM	2152	117	1023	117
4. HHT + SVDD	1525	116	958	93
5. RMS + SVDD	1735	20	959	20
6. Kurtosis + SVDD	1642	58	1163	134
7. HHT + LOF	2050	4	944	5
8. RMS + LOF	2023	52	1275	8
9. Kurtosis + LOF	2381	65	1372	33
10. HHT + iFOREST	1556	82	1041	25
11. RMS + iFOREST	2336	69	961	31
12. Kurtosis + iFOREST	2057	159	1257	93
13. BEMD + AMMA	1900	–	1130	–
14. SDFM	1374	42	930	27
15. TCA + SVDD	1427	33	986	28
16. GFK + SVDD	1573	20	1146	17
17. Our method	1401	14	937	6

6. Conclusions

In this paper, a new detection method for bearing early fault is proposed based on deep transfer learning. This work can be viewed as a combination of two methods: domain adaptation DAE and RDA. Aiming at robustness of early fault detection, this method can get a low false alarm rate, especially in the scenario of online detection. From the experimental results, the following conclusions can be drawn:

(1) Deep transfer learning works well to extract a common feature representation for different auxiliary bearings, which is vital in online detection.
(2) State assessment can be achieved in a low-rank subspace by RDA, which will eliminate the negative effect of signal fluctuation and bring the robustness of detection model.
(3) The proposed method is suitable for online detection with earlier detection location and less number of false alarms, as it can reduce the inconsistent data distribution between auxiliary bearings and target bearing.

In our next work, we plan to study theoretically the structure of deep neural network with better robustness and generalization ability. Ideally speaking, we can add a regularizer into RDA to constrain

the data distribution of common features. Moreover, an incremental learning strategy and network structure for online detection will be considered more.

Author Contributions: Conceptualization, W.M. and D.Z.; Methodology, W.M.; Software, D.Z.; Validation, W.M., and J.T.; Formal analysis, W.M.; Investigation, W.M.; Resources, W.M.; Data curation, W.M.; Writing—original draft preparation, W.M.; Writing—review and editing, S.T.; Visualization, S.T.; Supervision, J.T.; Funding acquisition, W.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China Grant No. U1704158 and China Postdoctoral Science Foundation Special Support Grant No. 2016T90944.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mao, W.; Feng, W.; Liang, X. A novel deep output kernel learning method for bearing fault structural diagnosis. *Mech. Syst. Signal Process.* **2019**, *117*, 293–318. [CrossRef]
2. Mao, W.; He, L.; Yan, Y.; Wang, J. Online sequential prediction of bearings imbalanced fault diagnosis by extreme learning machine. *Mech. Syst. Signal Process.* **2017**, *83*, 450–473. [CrossRef]
3. Wang, H.; Chen, J.; Dong, G. Feature extraction of rolling bearing's early weak fault based on EEMD and tunable Q-factor wavelet transform. *Mech. Syst. Signal Process.* **2014**, *48*, 103–119. [CrossRef]
4. Tabrizi, A.; Garibaldi, L.; Fasana, A. Early damage detection of roller bearings using wavelet packet decomposition, ensemble empirical mode decomposition and support vector machine. *Meccanica* **2015**, *50*, 865–874. [CrossRef]
5. Li, F.; Wang, J.; Chyu, M. Weak fault diagnosis of rotating machinery based on feature reduction with Supervised Orthogonal Local Fisher Discriminant Analysis. *Neurocomputing* **2015**, *168*, 505–519. [CrossRef]
6. Ocak, H.; Loparo, K. A new bearing fault detection and diagnosis scheme based on hidden Markov modeling of vibration signals. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001; pp. 3141–3144.
7. Tao, X.; Chen, W.; Du, B.; Xu, Y. A novel model of one-class bearing fault detection using SVDD and genetic algorithm. In Proceedings of the 2007 2nd IEEE Conference on Industrial Electronics and Applications, Harbin, China, 23–25 May 2007; pp. 802–807.
8. FernáNdez-Francos, D.; MartíNez-Rego, D.; Fontenla-Romero, O. Automatic bearing fault diagnosis based on one-class ν-SVM. *Comput. Ind. Eng.* **2013**, *64*, 357–365. [CrossRef]
9. Ma, H.; Hu, Y.; Shi, H. Fault detection and identification based on the neighborhood standardized local outlier factor method. *Ind. Eng. Chem. Res.* **2013**, *52*, 2389–2402. [CrossRef]
10. Wang, Z.; Zhang, Q.; Xiong, J. Fault diagnosis of a rolling bearing using wavelet packet denoising and random forests. *IEEE Sens. J.* **2017**, *17*, 5581–5588. [CrossRef]
11. Jia, F.; Lei, Y.; Guo, L. A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. *Neurocomputing* **2018**, *272*, 619–628. [CrossRef]
12. Shao, H.; Jiang, H.; Zhang, X. Rolling bearing fault diagnosis using an optimization deep belief network. *Meas. Sci. Technol.* **2015**, *26*, 115002. [CrossRef]
13. Mao, W.; He, J.; Tang, J. Predicting remaining useful life of rolling bearings based on deep feature representation and long short-term memory neural network. *Adv. Mech. Eng.* **2018**, *10*, 1687814018817184. [CrossRef]
14. Yang, B.; Lei, Y.; Jia, F. An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings. *Mech. Syst. Signal Process.* **2019**, *122*, 692–706. [CrossRef]
15. Zhang, R.; Tao, H.; Wu, L. Transfer learning with neural networks for bearing fault diagnosis in changing working conditions. *IEEE Access* **2017**, *5*, 14347–14357. [CrossRef]
16. Wen, L.; Gao, L.; Li, X. A new deep transfer learning based on sparse auto-encoder for fault diagnosis. *IEEE Trans. Syst. Man Cybern. Syst.* **2017**, *49*, 136–144. [CrossRef]
17. Han, T.; Liu, C.; Yang, W. Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application. *arXiv* **2018**, arXiv:1804.07265.
18. Mao, W.; He, J.; Zuo, M. Predicting remaining useful life of rolling bearings based on deep feature representation and transfer learning. *IEEE Trans. Instrum. Meas.* **2020**, In press. [CrossRef]

19. Lu, W.; Li, Y.; Cheng, Y. Early fault detection approach with deep architectures. *IEEE Trans. Instrum. Meas.* **2018**, *67*, 1679–1689. [CrossRef]
20. Zhao, R.; Wang, D.; Yan, R. Machine health monitoring using local feature-based gated recurrent unit networks. *IEEE Trans. Ind. Electron.* **2017**, *65*, 1539–1548. [CrossRef]
21. Mao, W.; Chen, J.; Liang, X. A New Online Detection Approach for Rolling Bearing Incipient Fault via Self-Adaptive Deep Feature Matching. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 443–456. [CrossRef]
22. Jia, F.; Lei, Y.; Lin, J. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mech. Syst. Signal Process.* **2016**, *72*, 303–315. [CrossRef]
23. Huang, N.; Shen, Z.; Long, S. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **1998**, *454*, 903–995. [CrossRef]
24. Borgwardt, K.; Gretton, A.; Rasch, M. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* **2006**, *22*, 49–57. [CrossRef]
25. Lu, W.; Liang, B.; Cheng, Y. Deep model based domain adaptation for fault diagnosis. *IEEE Trans. Ind. Electron.* **2016**, *64*, 2296–2305. [CrossRef]
26. Mao, W.; Liu, Y.; Ding, L. Imbalanced Fault Diagnosis of Rolling Bearing Based on Generative Adversarial Network: A Comparative Study. *IEEE Access* **2019**, *7*, 9515–9530. [CrossRef]
27. Mao, W.; He, J.; Li, Y. Bearing fault diagnosis with auto-encoder extreme learning machine: A comparative study. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2017**, *231*, 1560–1578. [CrossRef]
28. Wright, J.; Ganesh, A.; Rao, S. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Adv. Neural Inf. Process. Syst.* **2009**, *2080*–2088.
29. Zhou, C.; Paffenroth, R. Anomaly detection with robust deep autoencoders. In Proceedings of KDD 17: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 Augest 2017; pp. 665–674.
30. Tax, D.; Duin, R. Support vector data description. *Mach. Learn.* **2004**, *54*, 45–66. [CrossRef]
31. Nectoux, P.; Gouriveau, R.; Medjaher, K. PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In Proceedings of IEEE International Conference on Prognostics and Health Management, PHM'12, Denver, CO, USA, 18–21 June 2012; pp. 1–8.
32. Wang, B.; Lei, Y.; Li, N. A Hybrid Prognostics Approach for Estimating Remaining Useful Life of Rolling Element Bearings. *IEEE Trans. Reliab.* **2019**, 1–12. [CrossRef]
33. Chang, C.; Lin, C. LIBSVM: A library for support vector machines. *ACM trans. Intel. Syst. Technol. (TIST)* **2011**, *2*, 27. [CrossRef]
34. Li, Y.; Xu, M.; Liang, X. Application of bandwidth EMD and adaptive multiscale morphology analysis for incipient fault diagnosis of rolling bearings. *IEEE Trans. Ind. Electron.* **2017**, *64*, 6506–6517. [CrossRef]
35. Pan, S.J.; Tsang, I.W.; Kwok, J.T. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* **2010**, *22*, 199–210. [CrossRef] [PubMed]
36. Gong, B.; Shi, Y.; Sha, F. Geodesic flow kernel for unsupervised domain adaptation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2066–2073.