



Article

A Novel Domain Adaptation-Based Intelligent Fault Diagnosis Model to Handle Sample Class Imbalanced Problem

Zhongwei Zhang ¹, Mingyu Shao ^{1,*} , Liping Wang ², Sujuan Shao ¹ and Chicheng Ma ¹ 

¹ School of Transportation and Vehicle Engineering, Shandong University of Technology, Zibo 255000, China; zhangzz@sdut.edu.cn (Z.Z.); ssjsdut@sdut.edu.cn (S.S.); machch@sdut.edu.cn (C.M.)

² School of Mathematics and Information Science, Nanjing Normal University of Special Education, Nanjing 210038, China; wlp8631@nuaa.edu.cn

* Correspondence: shaomingyu@sdut.edu.cn

Abstract: As the key component to transmit power and torque, the fault diagnosis of rotating machinery is crucial to guarantee the reliable operation of mechanical equipment. Regrettably, sample class imbalance is a common phenomenon in industrial applications, which causes large cross-domain distribution discrepancies for domain adaptation (DA) and results in performance degradation for most of the existing mechanical fault diagnosis approaches. To address this issue, a novel DA approach that simultaneously reduces the cross-domain distribution difference and the geometric difference is proposed, which is defined as MRMI. This work contains three parts to improve the sample class imbalance issue: (1) A novel distance metric method (MVD) is proposed and applied to improve the performance of marginal distribution adaptation. (2) Manifold regularization is combined with instance reweighting to simultaneously explore the intrinsic manifold structure and remove irrelevant source-domain samples adaptively. (3) The ℓ_2 -norm regularization is applied as the data preprocessing tool to improve the model generalization performance. The gear and rolling bearing datasets with class imbalanced samples are applied to validate the reliability of MRMI. According to the fault diagnosis results, MRMI can significantly outperform competitive approaches under the condition of sample class imbalance.

Keywords: fault diagnosis; samples class imbalance; manifold regularization; maximum variance discrepancy; domain adaptation



Citation: Zhang, Z.; Shao, M.; Wang, L.; Shao, S.; Ma, C. A Novel Domain Adaptation-Based Intelligent Fault Diagnosis Model to Handle Sample Class Imbalanced Problem. *Sensors* **2021**, *21*, 3382. <https://doi.org/10.3390/s21103382>

Academic Editor: Reinaldo Martinez Palhares

Received: 8 April 2021
Accepted: 9 May 2021
Published: 12 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Bearings and gears are vital components and are widely utilized in machinery equipment [1]. In addition, bearing and gear faults are the most common failure mode which may lead to unexpected fatal failures and elevated maintenance costs. Thus, there is a strong demand for intelligent fault diagnosis techniques of bearings and gears to ensure the security and reliability of mechanical equipment [2–4].

For deep learning methods, for instance, deep belief networks [5], sparse filtering [6], and autoencoders (AEs) [7,8], the main assumption is that the datasets applied to train and test the model have the same feature distribution. Unfortunately, the raw vibration signals are usually obtained under variable working cases in practical applications, which show deviation from the assumption [9,10]. As a result, poor performances may be obtained for most machine learning methods. The above issue is often denoted as cross-domain learning.

Within the last decade, DA techniques have been focused on solving the above problem. The source and target domain data show similar but different feature distributions for DA [11]. Many existing DA approaches usually aim to reduce the difference of cross-domain feature distributions, e.g., the distribution adaptation, the instance reweighting, or joint matching (join the distribution adaptation and instance reweighting). The distribution adaptation approaches [12–14] mainly include marginal adaptation (MDA) [15–19], conditional adaptation (CDA) [20], or both [12,21] and are applied for most distribution

adaptation approaches. Lu et al. [15] adapted the marginal distribution by MMD to minimize the distribution discrepancy across domains and introduced MMD into a deep neural network (DNN). Han et al. [21] introduced the joint distribution adaptation (JDA) into a deep transfer network (DTN) to avoid negative adaptation and presented smooth convergence for fault diagnosis in industry applications. The discussion in [12] showed that joint distribution adaptation may obtain better performance in fault diagnosis by reweighting the source instance on the basis of its correlation with the target instance to reduce the cross-domain feature distribution discrepancy [22,23]. Chen et al. [23] developed an unsupervised domain adaptation approach to reduce the domain shifts between the data gathered from the experimental platform and the operating platform of the rotating machine by aligning the features extracted from the two data domains. In addition, some published DA methods tried to join feature reweighting and subspace learning [24,25]. Long et al. [24] reduced the cross-domain distribution discrepancy and achieved good classification results by combining these two learning strategies. However, the above approach matches the sample moments among distinct data distributions and down-weights the irrelevant source domain features, which may perform badly while the data distribution discrepancy across the two domains is rather large, e.g., the sample class imbalance case.

Sample class imbalance denotes a situation where the number of instances in one class is much different from the number of instances in other classes. The class imbalance will lead to a substantially large cross-domain distribution difference and usually exists in many domain adaptation scenarios. Unfortunately, the class imbalance is usually ignored for most DA approaches [12,14]. They usually assume the sample classes are balanced or tackle the sample bias for one domain, which decreases the validity of DA. When the proportion of different classes is substantially imbalanced, distribution adaptation only or independent manifold learning is not enough to obtain good fault classification results. Thus, it is an important challenge to tackle the class imbalanced case in domain adaptation.

To this end, it is very necessary to study the deep information in the marginal distributions [26]. As a distance metric, MVD is very suitable for the class imbalance situation. In addition, manifold regularization can search the intrinsic manifold structure and further exploit the marginal distributions across domains. This motivates us to combine manifold regularization with the MVD, which can further extract effective information by optimizing the manifold consistency underlying marginal distributions and the manifold geometric structure. Moreover, the instance reweighting approach can further reduce the cross-domain difference by down-weighting the irrelevant source domain instances compared with target domain instances.

In recent years, manifold learning has drawn much attention in the field of fault diagnosis [27–29]. Wang et al. [27] applied manifold alignment for cross-domain fault diagnosis and decreased the distributional shift and structural shift at the same time via transforming the fault features into two low-dimension subspaces. Wang et al. [28] applied manifold learning to decrease the dimension of a wave packet envelope matrix to learn the embedded inherent defect characteristics, and reveal the inherent envelope structure of impact impulses without the optimal band selection. Compared with the previous approaches, our work aims to model the manifold regularization, MVD and the instance reweighting techniques in a unified way to solve the class imbalance problem in fault diagnosis.

In this paper, considering the practical defect diagnosis application, a novel DA approach is proposed to handle the class imbalance problems. Firstly, the raw vibration signal under different rotating speed and load conditions are preprocessed by the fast Fourier transformation (FFT) to obtain the frequency spectrum. Then, ℓ_2 -norm regularization is applied for processing the frequency spectrum, which can improve the model generalization performance. Next, manifold regularization is combined with MVD and instance reweighting to simultaneously reduces the cross-domain distribution difference, geometric difference, and the proportion of unrelated source-domain samples, which can obtain the domain-invariant fault features with sufficient transferability. Finally, softmax regression

is applied for predicting the fault types. Moreover, the fault features are normalized by the z-score normalization before fault classification to ensure the robustness of MRMI. The experimental results show that MRMI outperforms baseline DA approaches significantly.

The rest of this paper is organized as follows: In Section 2, DA, MMD, and the softmax regression algorithm are briefly presented. The framework of MRMI is described in Section 3. In Section 4, the validity and robustness of MRMI are validated according to the fault diagnosis experiments. Finally, the conclusions are given in Section 5.

2. Theoretical Background

2.1. Domain Adaptation

As we can see from Figure 1, the categories of data are represented by different shapes, and the labeled training data and test data have an identical data distribution for the traditional intelligent method. By contrast, for domain adaptation, the labeled source domain data $D_s = \{(x_{s_1}, y_{s_1}), \dots, (x_{s_{n_s}}, y_{s_{n_s}})\}$ and unlabeled target domain data $D_t = \{x_{t_1}, \dots, x_{t_{n_t}}\}$ show the different but similar data distributions. The further description of domain adaptation is discussed as follows.

1. \mathcal{X} refers to data space and P denotes a marginal data distribution. Thus, $\{\mathcal{X}, P(X)\}$ denotes that dataset X is drawn from \mathcal{X} and shows the data distribution $P(X)$. For DA, datasets have distinct data spaces and marginal data distributions, i.e., $\mathcal{X}_s \neq \mathcal{X}_t$ and $P_s(X_s) \neq P_t(X_t)$;
2. For the task $\mathcal{T} = \{\mathcal{Y}, f(X)\}$, the prediction function $f(X) = P(Y|X)$ denotes the conditional distribution and $Y \in \mathcal{Y}$. $\mathcal{Y}_s = \mathcal{Y}_t$, $P(Y_s|X_s) = P(Y_t|X_t)$, where \mathcal{Y} is the label spaces, since categories for distinct working conditions are the same.
3. In our research, a transfer function F is used to realize the domain adaptation learning, which satisfies $\mathcal{X}_s \neq \mathcal{X}_t$, $\mathcal{Y}_s = \mathcal{Y}_t$, $P(F(X_s)) = P(F(X_t))$, and $P(Y_s|F(X_s)) = P(Y_t|F(X_t))$.

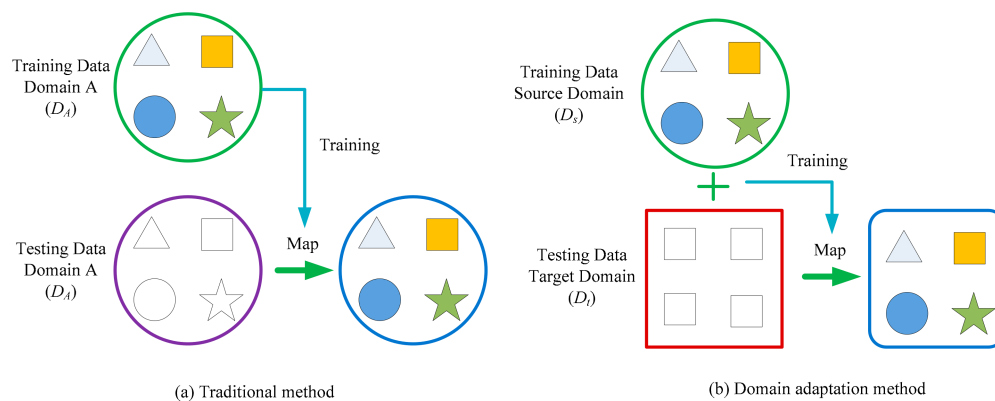


Figure 1. Intelligent learning system [13].

2.2. Maximum Mean Discrepancy

The fundamental challenge for the generalization performance of DA approaches is to decrease the cross-domain distribution discrepancy. Thus, it is vital to minimize the discrepancy between cross-domain probability distributions by formalizing the distinct distribution and proposing effective approaches. Many parametric criteria have been applied to calculate the difference between cross-domain distributions, for instance, KL divergence [30] and Bregman divergence [31]. Nevertheless, as a more difficult density estimation process, the intermediate density estimate aggravates the model's complexity. To solve this non-trivial problem, [32] ignored the intermediate density estimate, proposed a non-parametric divergence-MMD to compute the distance across domains by matching the data to the reproducing kernel Hilbert space (RKHS). Datasets $X = \{x_1, \dots, x_{n_1}\}$ and

$Y = \{y_1, \dots, y_{n_2}\}$ obey the data distributions P and Q , respectively. The cross-domain distance is calculated as follows.

$$Dist(X, Y) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} f(x_i) - \frac{1}{n_2} \sum_{i=1}^{n_2} f(y_i) \right\|_{\mathcal{H}}, \quad (1)$$

where \mathcal{H} represents a universal RKHS [33], $\varphi: X, Y \rightarrow \mathcal{H}$.

2.3. Softmax Regression

The softmax regression (SR) model [34] has been widely used for the supervised learning stages of many domain adaptation approaches. Generally, the predicted labels of SR are multi-class classification instead of binary classification, so SR can be regarded as a generalized case for the logistic regression. In addition, SR is easy to carry out and it has high computing efficiency. To this end, the softmax regression classifier is selected for our research. It should be pointed out that the SR classifier is most suitable under the condition that the corresponding classes are mutually exclusive. Thus, we assume that each fault occurs separately.

The employed dataset is defined to train the softmax regression model, including m samples, that is, $\{(x^1, y^1), \dots, (x^m, y^m)\}$, where $x^{(i)}$ represents the input feature, and the labels consist of $y^{(i)} \in \{1, 2, \dots, k\}$, where k represents the number of health conditions. Furthermore, $p(y^{(i)} = j | x^{(i)})$ represents the probability value for which $x^{(i)}$ pertains to the category j . The probability value of each category is calculated for $x^{(i)}$, and then the output value is identified by selecting the category whose probability value is the maximum. Thus, the output value $h_{\theta}(x^{(i)})$ can be written as:

$$h_{\theta}(x^{(i)}) = \begin{bmatrix} p(y^{(i)} = 1 | x^{(i)}; \theta) \\ p(y^{(i)} = 2 | x^{(i)}; \theta) \\ \vdots \\ p(y^{(i)} = k | x^{(i)}; \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^k \exp(\theta_j^T x^{(i)})} \begin{bmatrix} \exp(\theta_1^T x^{(i)}) \\ \exp(\theta_2^T x^{(i)}) \\ \vdots \\ \exp(\theta_k^T x^{(i)}) \end{bmatrix}, \quad (2)$$

where $\theta_1, \theta_2, \dots, \theta_k$ denote the parameters for the model.

The cost function $J(\theta)$ is displayed as follows:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^k 1\{y^{(i)} = j\} \log \frac{\exp(\theta_j^T x^{(i)})}{\sum_{l=1}^k \exp(\theta_l^T x^{(i)})} + \frac{\lambda}{2} \sum_{i=1}^k \sum_{j=0}^n \theta_{ij}^2, \quad (3)$$

where m represents the sample number, n refers to the n th column of weight matrix θ , k denotes category, λ is the weight decay term.

Generally, the cost function $J(\theta)$ is minimized by:

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{m} \sum_{i=1}^m [x^{(i)} (1\{y^{(i)} = j\} - p(y^{(i)} = j | x^{(i)}; \theta))], \quad (4)$$

$\nabla_{\theta_j} J(\theta)$ represents the partial derivative of $J(\theta)$ w.r.t. θ_j , where $j = 1, 2, \dots, k$.

3. Proposed Framework

In this part, the data preprocessing for MRMI is firstly introduced in Section 3.1. Then, the model structure and the learning algorithm of MRMI are described in Section 3.2. In addition, Table 1 shows the frequently used notations.

Table 1. Notations and descriptions.

Notation	Description	Notation	Description
D_s, D_t	Source/Target domain	\mathbf{X}	Input data matrix
n_s, n_t	Source/Target samples	\mathbf{A}	Alignment matrix
$\mathcal{X}_s, \mathcal{X}_t$	Source/Target data space	\mathbf{L}	Laplacian matrix
k	Subspace bases	\mathbf{M}	MMD matrix
λ, γ	Regularization parameter	\mathbf{G}	Subgradient matrix
\mathbf{Z}	Subspace embedding	\mathbf{K}	Input kernel matrix

3.1. Data Preprocessing

3.1.1. Fast Fourier Transform (FFT)

First of all, FFT is adopted for transforming the original vibration signal into the frequency spectrum. The frequency spectrum can show the discrete frequencies of the constitutive components for the rotating machines [35] and can be good for extracting sensitive defect features that are easily discriminated.

3.1.2. ℓ_2 -norm Regularization

Then, the ℓ_2 -norm regularization is adopted for the frequency spectrum to avoid the overfitting problem. The ℓ_2 -norm regularization can weaken the strong features as much as possible, and highlight the features with smaller values but more characteristics. Thus, it makes the corresponding algorithm more inclined to use all input features, rather than rely heavily on some parts of the input features, which may be very useful to calculate the similarity between two samples by the kernel methods. In general, the form of ℓ_2 -norm can be denoted as $\sqrt{|t_1|^2 + \dots + |t_n|^2}$, where $t = [t_1, t_2, \dots, t_n]$.

f_l^i composes the data matrix, where l represents the row number and i is the column number. First of all, each row is regularized by the ℓ_2 -norm across all the samples.

$$\bar{f}_l = f_l / \|f_l\|_2, \quad (5)$$

Next, each column is regularized by its ℓ_2 -norm. As a result, the features lie on the unit ℓ_2 -ball.

$$\hat{f}^i = \check{f}^i / \|\check{f}^i\|_2, \quad (6)$$

Since the regularized features have been divided by their ℓ_2 -norm across all the samples, it means that the contributions of these features are almost the same.

3.1.3. Data Dimensionality Reduction

As the most commonly used unsupervised linear dimensionality reduction approach, the principal component analysis (PCA) algorithm can map the high-dimensional vectors to the low-dimensional subspaces, and retain as much information as possible about the raw data. Thus, PCA is adopted for the dimensionality reduction of the regularized samples. As a result, the variance of the embedded data is maximized by the transformation matrix $\mathbf{U} \in \mathbb{R}^{m \times k}$.

$$\max_{\mathbf{U}^T \mathbf{U} = \mathbf{I}} \text{tr}(\mathbf{U}^T \mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{U}), \quad (7)$$

where $\text{tr}(\cdot)$ denotes the matrix trace, $\mathbf{X} = [x_1, \dots, x_n] \in \mathbb{R}^{m \times n}$ is the input matrix, $\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T$ represents the centering matrix.

The kernel mapping form $\psi: x \mapsto \psi(x)$ and kernel matrix $\mathbf{K} = \psi(\mathbf{X})^T \psi(\mathbf{X}) \in \mathbb{R}^{n \times n}$ are adopted for converting the data to RKHS. Then, the kernel-PCA is obtained by the representer theorem $\mathbf{V} = \phi(\mathbf{X})\mathbf{A}$.

$$\max_{\mathbf{A}^T \mathbf{A} = \mathbf{I}} \text{tr}(\mathbf{A}^T \mathbf{K} \mathbf{H} \mathbf{K}^T \mathbf{A}), \quad (8)$$

where $A \in \mathbb{R}^{n \times k}$ refers to the transformation matrix. As a result, the subspace embedding is transferred to $Z = A^T K$.

3.2. Model Structure and Learning Algorithm of MRMI

In this section, the model framework of MRMI is firstly presented, and then the corresponding learning algorithm is introduced.

3.2.1. MRMI Model

The proposed MRMI is realized by minimizing the listed complementary objective functions:

- (1) The MVD term for minimizing the discrepancy between the marginal probability distributions P_s and P_t ;
- (2) The $\ell_{2,1}$ -norm structured sparsity regularization term for reweighting the source domain instances by structured sparsity;
- (3) The manifold regularization for maximizing the manifold consistency between P_s and P_t .

The prediction function $f = w^T \phi(x)$ is applied for classification, where w denotes a parameter of the classifier. The final objective function of MRMI is summarized as follows.

$$f = \arg \min_{f \in \mathcal{H}_K} D_{f,K}(P_s, P_t) + \lambda \|T\|_{2,1} + \gamma M_{f,K}(P_s, P_t), \quad (9)$$

where \mathcal{H}_K denotes a set of f in the kernel space, K represents the kernel function which is calculated by ϕ , so $\langle \phi(x_i), \phi(x_j) \rangle = K(x_i, x_j)$. In addition, the raw feature vector is projected into a Hilbert space \mathcal{H} by the mapping function $\phi: \mathcal{X} \mapsto \mathcal{H}$ [26]. T denotes the feature transformation to adapt different domains, $\|T\|_{2,1}$ represents the $\ell_{2,1}$ -norm of T . $D_{f,K}(P_s, P_t)$ denotes the discrepancy for P_s and P_t , and $M_{f,K}(P_s, P_t)$ represents the manifold regularization which can extract more information from P_s and P_t . λ represents the regularization parameter which is employed for trading off instance reweighting and feature matching. γ refers to positive regularization parameters. Each term in Equation (9) is interpreted in the following discussion.

(1) MVD Term

While the distribution discrepancy across domains is rather large, the MMD algorithm performs badly for marginal distribution adaptation as MMD mainly regards the first-order statistics. By contrast, MVD simultaneously regards the first-order and second-order statistics, which shows better performance of marginal distribution adaptation and can bridge the cross-domain discrepancy more effectively than MMD. In addition, the deviation of cross-domain data distribution is reduced while the variance difference is decreased. For MRMI, we introduce MVD for the feature matching to further decrease the distribution difference.

In general, we can obtain the sample variance S^2 by:

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}, \quad (10)$$

where n represents the size of the sample, S denotes the standard deviation for the sample, and \bar{x} is the average value.

In addition, the sample variance can be transferred into the other form DU.

$$DU = E(U^2) - [E(U)]^2, \quad (11)$$

where U denotes a vector of sample, EU represents the expectation.

Let \mathbf{Z}_i represents the i th sample of the subspace embedding, we can obtain:

$$E(\mathbf{Z}_i) = \sum_{i=1}^n f(\mathbf{Z}_i) \mathbf{Z}_i, \quad (12)$$

where $f(\mathbf{Z}_i)$ denotes the probability value of the i -th sample.

The probability value that every sample occurred is assumed to be equal. As a result, Equation (12) can be calculated by:

$$E(\mathbf{Z}_i) = \frac{\sum_{i=1}^n \mathbf{Z}_i}{n}, \quad (13)$$

Kernel-PCA is applied to obtain the k dimension embedding for MVD. Then, the corresponding empirical mathematical expectations are computed by joining Equations (8) and (11).

$$\begin{aligned} & \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} (\mathbf{A}^T \mathbf{k}_i)^2 - \frac{1}{n_t} \sum_{j=n_s+1}^{n_s+n_t} (\mathbf{A}^T \mathbf{k}_j)^2 \right\|_{\mathcal{H}}^2 + \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbf{A}^T \mathbf{k}_i + \right. \\ & \left. \frac{1}{n_t} \sum_{j=n_s+1}^{n_s+n_t} \mathbf{A}^T \mathbf{k}_j \right\|_{\mathcal{H}}^2 \times \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbf{A}^T \mathbf{k}_i - \frac{1}{n_t} \sum_{j=n_s+1}^{n_s+n_t} \mathbf{A}^T \mathbf{k}_j \right\|_{\mathcal{H}}^2 = \text{tr}(\mathbf{A}^T \mathbf{K}_1 \mathbf{M} \mathbf{K}_1^T \mathbf{A}) - \\ & \text{tr}(\mathbf{A}^T \mathbf{K} \mathbf{M} \mathbf{K}^T \mathbf{A}) * \text{tr}(\mathbf{A}^T \mathbf{K} \mathbf{M}_1 \mathbf{K}^T \mathbf{A}), \end{aligned} \quad (14)$$

where $\mathbf{K}_1 = \psi(\mathbf{X}^2)^T \psi(\mathbf{X}^2) \in \mathbb{R}^{n \times n}$, \mathbf{M} and \mathbf{M}_1 are both the MVD matrix, which can be computed as follows

$$\mathbf{M}_{ij} = \begin{cases} \frac{1}{n_s n_s}, x_i, x_j \in D_s \\ \frac{1}{n_i n_t}, x_i, x_j \in D_t \\ -\frac{1}{n_s n_t}, \text{otherwise} \end{cases}, \mathbf{M}_{1ij} = \begin{cases} \frac{1}{n_s n_s}, x_i, x_j \in D_s \\ \frac{1}{n_i n_t}, x_i, x_j \in D_t \\ \frac{1}{n_s n_t}, \text{otherwise} \end{cases}, \quad (15)$$

where n_s and n_t represent the samples of the source and target domain, respectively.

(2) The $\ell_{2,1}$ -norm Structured Sparsity Regularization Term

Nevertheless, only applying the MVD term for minimizing $D_{f,K}(P_s, P_t)$ is not enough to obtain representative features, because there are some irrelevant and redundant source instances. To this end, it is very necessary to down-weight the irrelevant source instances to further decrease domain discrepancy. In this section, we employ instance reweighting by the $\ell_{2,1}$ -norm structured sparsity regularization to down-weight the irrelevant source instances in the instance space. The $\ell_{2,1}$ -norm regularization is applied to induce *row-sparsity* in matrix A . Owing to *row-sparsity*, each row of the transformation matrix A can be regarded as an instance which intrinsically facilitates the instance reweighting. Instance reweighting regularization can be constructed in the following way [24].

$$\|\mathbf{A}_s\|_{2,1} + \|\mathbf{A}_t\|_F^2, \quad (16)$$

where $\mathbf{A}_s \mathbf{A}_{1:n_s,:}$ represents the source domain transformation matrix, and $\mathbf{A}_t \mathbf{A}_{n_s+1:n_s+n_t,:}$ denote the target domain one. It should be noted that $\ell_{2,1}$ -norm regularization is only employed to reweight the source domain instances with their correlation to the target ones. When Equation (16) is minimized, Equation (9) will be maximized, which means that the irrelevant and redundant source instances are down-weighted adaptively in a novel subspace embedding $\mathbf{Z} = \mathbf{A}^T \mathbf{K}$. As a result, the robustness of MRMI is improved for the domain discrepancy resulting from irrelevant source instances.

(3) Manifold Regularization Term

The MVD term and the $\ell_{2,1}$ -norm structured sparsity regularization term can reduce the domain discrepancy in \mathcal{H} and the instance space, respectively. However, they only match the cross-domain sample moments and down-weight the irrelevant source domain features, which may perform badly when feature distribution discrepancy across domains

is rather large, e.g., the class imbalance problem. Thus, manifold regularization is induced for researching the intrinsic manifold structure and further exploiting the information from P_s and P_t to learn better functions. Generally, the unlabeled target domain data may reveal the potential and hidden information, such as sample variances. According to the *manifold assumption* [36], the conditional distributions $\mathcal{Q}_s(y_s|x_s)$ and $\mathcal{Q}_t(y_t|x_t)$ are similar, when data points $x_s, x_t \in \mathcal{X}$ are close to each other in the geometry structure. After smoothing the geodesic, manifold regularization is calculated as

$$M_{f,K}(P_s, P_t) = \sum_{i,j=1}^{n_s+n_t} (f(x_i) - f(x_j))^2 W_{ij} = \sum_{i,j=1}^{n_s+n_t} f(x_i) L_{ij} f(x_j), \quad (17)$$

where W represents the graph affinity matrix, and L denotes the normalized graph Laplacian matrix. In addition, W is formulated as [37]

$$W_{ij} = \begin{cases} \cos(x_i, x_j), & \text{if } x_i \in \mathcal{N}_p(x_j) \vee x_j \in \mathcal{N}_p(x_i) \\ 0, & \text{otherwise} \end{cases}, \quad (18)$$

where $\mathcal{N}_p(x_i)$ refers to the p -nearest neighbors. L can be calculated as $L = I - D^{-1/2} W D^{-1/2}$ [26].

Maximizing the consistency of the intrinsic manifold structure can be used to further explore the marginal data distributions via regularizing (14) with (17), and the discriminative hyperplanes across domains can be substantially matched. According to the representer theorem, the manifold regularization can be rewritten as

$$M_{f,K}(P_s, P_t) = \text{tr}(A^T K L K^T A), \quad (19)$$

Above all, by combining Equations (14), (16) and (19), the final objective function is obtained as follows:

$$\min_{A^T K H K^T A = I} \text{tr}(A^T K_1 M K_1^T A) - \text{tr}(A^T K M K^T A) * \text{tr}(A^T K M_1 K^T A) + \lambda (\|A_s\|_{2,1} + \|A_t\|_F^2) + \gamma (\text{tr}(A^T K L K^T A)), \quad (20)$$

by regarding A as the adaptation matrix throughout the rest of this article to emphasize its functionality. It provides great convenience for the implementation and deployment of MRMI because a principled dimensionality reduction procedure is applied.

(4) Construct the Softmax Regression Classifier

In the process of classification, z-score normalization can eliminate the influence of dimension on classification results to develop the classification accuracy. Moreover, the learning rate and the efficiency of dealing with the optimal solution in the process of back propagation can be optimized via z-score normalization. Hence, it is adopted for processing the input data for the classifier. In other words, the training data T_r and the testing data T_t are computed by $T_r = F(Z_S)$ and $T_t = F(Z_T)$, where $Z_S = A_s^T K_s$ and $Z_T = A_t^T K_t$. Z-score normalization is formulated as follows:

$$F(X) = \frac{X - \bar{X}}{\sigma}, \quad (21)$$

where X denotes invariant feature subspace Z_S or Z_T in the finite domain, \bar{X} refers to the average value of X , σ is the standard deviation. After carrying out z-score normalization, the rescaled subspace $F(X)$ with a standard normal distribution is acquired.

Then, the probability value $p(y^{(i)} = j | T_t^{(i)})$ corresponding to each category j is calculated by Equation (2), then the fault category is predicted by selecting the j with maximum value. Finally, the classification performance of MRMI is obtained by comparing the consistency between the predicted fault type and the real one.

3.2.2. Learning Algorithm

By the constrained optimization theory, $\Phi = \text{diag}(\phi_1, \dots, \phi_k) \in \mathbb{R}^{k \times k}$ is adopted as the Lagrange multiplier for Equation (20). Thus, the Lagrange function is derived as:

$$F = \text{tr}(A^T K_1 M K_1^T A) - \text{tr}(A^T K M K^T A) * \text{tr}(A^T K M_1 K^T A) + \lambda (\|A_s\|_{2,1} + \|A_t\|_F^2) + \gamma \text{tr}(A^T K L K^T A) + \text{tr}((I - A^T K H K^T A) \Phi), \quad (22)$$

Let $\frac{\partial F}{\partial A} = 0$, generalized eigen-decomposition is approximately calculated as:

$$\left(K_1 M K_1^T - (K M K^T) * (K M_1 K^T) + \lambda G + \gamma K L K^T \right) A = K H K^T A \Phi, \quad (23)$$

As $\|A_s\|_{2,1}$ refers to a non-smooth function, the subgradient is computed as $\frac{\partial (\|A_s\|_{2,1} + \|A_t\|_F^2)}{\partial A} = 2GA$, where G represents a diagonal subgradient matrix which consists of the i -th element as below:

$$G_{ii} = \begin{cases} \frac{1}{2\|a^i\|}, & x_i \in D_s, a^i \neq 0 \\ 0, & x_i \in D_s, a^i = 0 \\ 1, & x_i \in D_t \end{cases}, \quad (24)$$

In the next step, matrix A is reduced to k smallest eigenvectors by (23). Nevertheless, the subgradient matrix G and adaptation matrix A are not known in advance. To overcome this deficiency, the parameters are optimized alternately by updating one parameter while fixing the other one.

For better interpretation, the structure of MRMI is shown in Figure 2.

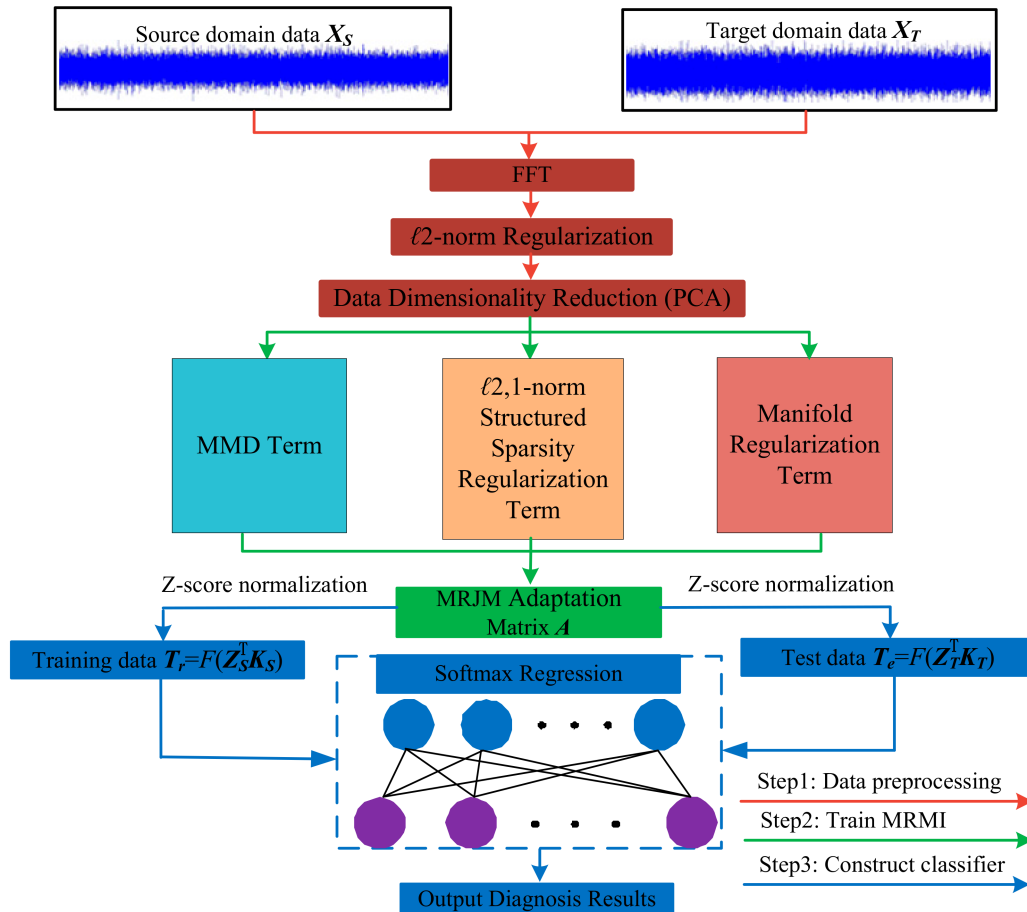


Figure 2. The framework of MRMI.

4. Experiment Results and Analysis

4.1. Case 1: Bearing Fault Diagnosis

4.1.1. Experimental Setup and Data Description

A rolling bearing dataset offered by Case Western Reserve University was employed to validate the performance of MRMI in this part [38]. It was acquired by accelerometers installed in the driving position of the motor and includes the normal (Nor) and faulty data. Furthermore, the faulty data consist of a single-point fault at the inner bearing race (FI), the outer race (FO), and the roller (FR). Each defect type of the faulty dataset contains three fault levels, i.e., 0.18, 0.36, 0.54 mm. Therefore, there are 10 health types obtained for the rolling bearing dataset in this section. The vibration signals were acquired under four loads (0, 1, 2, 3 hp). In addition, the sampling frequency was fixed as 12 kHz. In addition, we select the four motor loads as the four scenarios for domain adaptation. To simulate the situation of class imbalance, Table 2 shows the sample distribution for all domain adaptation tasks.

Table 2. Rolling bearing dataset with sample class imbalance distribution.

Fault Location	Nor	Roller			Inner Ring			Outer Ring			Total
Category Labels	1	2	3	4	5	6	7	8	9	10	
Fault Size (mm)	0	0.18	0.36	0.54	0.18	0.36	0.54	0.18	0.36	0.54	
A (load 0)	100	30	20	10	30	20	10	30	20	10	280
B (load 1)	100	10	15	10	10	15	10	10	15	10	205
C (load 2)	100	50	50	50	30	30	30	20	20	20	400
D (load 3)	100	100	100	100	100	100	100	100	100	100	1000

In Table 2, the vibration data collected with load 0, 1, 2, 3 hp are chosen as the DA scenarios A, B, C, D, respectively. The numbers of experimental samples for source and target domains are distinct from each other for different DA scenarios. In DA task “B→D”, B represents the labeled source domain dataset which includes 205 experimental samples collected under load 1 hp, while D denotes the unlabeled target domain dataset which contains 1000 experimental samples collected with load 3 hp. Therefore, the data distributions of these two domains are imbalanced.

First of all, the data preprocessing process is conducted for the rolling bearing dataset. As a result, the spectra of original vibration signals are obtained by fast Fourier transformation (FFT). Then, the time-domain samples with 1200 sample lengths are converted to 600 length samples in the frequency domain.

4.1.2. Experimental Results

(1) Comparison Methods

To validate the effectiveness of manifold regularization-based joint matching (MRMI), several successful domain adaptation approaches are selected as the baseline methods. The details of these baseline approaches are described as follows.

1. Deep neural network (DNN)-based DA approach (DAFD) [35], which combines MMD with DNN to extract the domain-invariant features;
2. Geodesic flow kernel (GFK) [15,18], which represents a typical DA approach;
3. Transfer joint matching (TJM) [24], which introduces feature matching into instance reweighting;
4. Adaptation regularization-based transfer learning (ARTL) [26], which combines JDA with manifold regularization;
5. Domain-adversarial neural networks (DANNs) [39], which develop a novel representation learning method for DA.

(2) Setup of the Algorithm

To provide a relatively fair environment for comparison, the hyperparameter space is empirically searched to select the best parameter settings. For reducing the randomness of the experiments, we carry out 15 trials of experiments to every DA task, then calculate the average classification accuracy to evaluate the performance for each approach. Moreover, the SR classifier is adopted for predicting the fault types of the target domain for all these domain adaptation methods.

For all the baseline approaches, the optimum dimension of the subspace is obtained by searching $\{10, 20, \dots, 200\}$ and the optimum value is selected by searching $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. In addition, the structure of the neural network is $\{600, 1000, 10\}$ for DAFD [37], and the size of the hidden layer for DANN is set as 200.

The proposed method contains only three model parameters: subspace dimension k , regularization parameters λ and γ . Empirical analysis of parameter sensitivity will be discussed in a later section. According to the parameter selection of the baseline approaches, the parameters of MRMI are set as $k = 50$, $\lambda = 1$, $\gamma = 10$ and the linear kernel is employed for MRMI.

In this paper, the diagnosis accuracy for the unlabeled data of the target domain is employed as the performance evaluation index, which has been applied in numerous published studies [40–42].

$$CA = \frac{|x : x \in \mathcal{D}^t \wedge \text{prediction}(y) = \text{true}(y)|}{|x : x \in \mathcal{D}^t|} \times 100\%, \quad (25)$$

(3) Results

For the experiment in this section, 12 DA scenarios are selected: $A \rightarrow B$, $A \rightarrow C$, $A \rightarrow D$, $B \rightarrow A$, $B \rightarrow C$, $B \rightarrow D$, $C \rightarrow A$, $C \rightarrow B$, $C \rightarrow D$, $D \rightarrow A$, $D \rightarrow B$, and $D \rightarrow C$. The experimental results of MRMI and all baseline methods are illustrated in Table 3. The result shows that the average accuracy of DA task $a \rightarrow b$ is distinct from $b \rightarrow a$, e.g., the classification result of $A \rightarrow D$ is 99.50% for MRMI but is 96.86% for scenario $D \rightarrow A$.

As we can see from the results listed in Table 3, MRMI yields the best diagnosis accuracy and robustness and outperforms the other four listed compared approaches in most (11 out of 12) domain adaptation scenarios. This indicates that more transferable and robustness fault features could be extracted for MRMI. Furthermore, we can draw several observations as follows.

Firstly, the proposed method performs worse than GFK in the scenario $D \rightarrow A$. For GFK, the final diagnosis result for all 12 domain adaptation tasks can reach 90.91% which is the highest accuracy compared with the other baseline approaches and is 8.55% less than MRMI. The smooth transmission of the object datasets can be guaranteed by mapping the global GFK into a low dimension representation, thus, good diagnosis performance can be obtained. Nevertheless, GFK performs worse in DA scenarios $A \rightarrow D$ and $B \rightarrow D$, which indicates that only applying the geodesic flow distance to correct the distribution mismatch is not enough when the cross-domain discrepancy is rather large.

Secondly, DAFD combines MMD with DNN to extract the domain-invariant features. However, DAFD performs worse than MRMI, which highlights that MVD can bridge the cross-domain difference more effectively than MMD. The reason is that MVD simultaneously regards the first-order and second-order statistics to minimize the marginal distribution mismatch. In addition, from the results, we also observe that only adopting the marginal distribution adaptation is not enough to reduce the cross-domain conditional distribution difference. Therefore, the average classification accuracy for DAFD is under 80%, which performs worse than ARTL and TJM.

Thirdly, TJM combines instance reweighting with MMD in a principled dimensionality reduction process to reduce the cross-domain discrepancy. In addition, TJM aims to build a novel feature representation. It is invariant to distribution discrepancy and irrelevant source instances. Thus, TJM performs well when the cross-domain distribution difference is rather large. However, MMD mainly regards the first-order statistics, and while the

distribution discrepancy across domains is rather large, the MMD algorithm performs badly for marginal distribution adaptation. As a result, the average classification accuracy is still 16.06% lower than the proposed approach, which indicates that information of P_s and P_t needs to be further explored to extract more representative transferable features for TJM.

Table 3. The classification results (%) on class imbalanced rolling bearing dataset.

Source Domain	Method	Target Domain			
		A	B	C	D
A	DAFD	-	81.46 ± 1.21	83.25 ± 0.36	71.80 ± 0.60
	GFK	-	96.10 ± 0.15	88.25 ± 0.00	78.20 ± 0.60
	TJM	-	88.78 ± 0.49	76.00 ± 0.25	95.6 ± 1.10
	ARTL	-	95.12 ± 0.38	87.25 ± 0.25	89.50 ± 0.35
	DANN	-	96.09 ± 0.25	89.50 ± 0.17	73.90 ± 0.43
	MRMI	-	99.61 ± 0.20	99.60 ± 0.05	99.50 ± 0.20
	B	DAFD	75.71 ± 1.05	-	79.25 ± 0.83
GFK		92.50 ± 0.36	-	89.00 ± 0.50	76.10 ± 1.40
TJM		80.54 ± 0.18	-	85.00 ± 0.00	76.00 ± 0.15
ARTL		77.86 ± 0.95	-	83.75 ± 0.63	73.00 ± 1.20
DANN		96.07 ± 0.46	-	93.25 ± 0.14	65.70 ± 0.29
MRMI		99.64 ± 0.25	-	100.00 ± 0.00	99.80 ± 0.05
C		DAFD	74.29 ± 0.78	82.93 ± 1.35	-
	GFK	89.29 ± 0.26	95.7 ± 0.31	-	93.20 ± 0.40
	TJM	80.76 ± 0.54	84.88 ± 0.00	-	79.00 ± 0.25
	ARTL	88.93 ± 0.44	80.98 ± 0.36	-	93.60 ± 0.60
	DANN	97.14 ± 0.18	96.59 ± 0.31	-	90.90 ± 0.78
	MRMI	98.86 ± 0.35	100.00 ± 0.00	-	99.95 ± 0.05
	D	DAFD	76.43 ± 0.28	76.20 ± 1.35	71.25 ± 0.75
GFK		97.00 ± 0.50	97.5 ± 0.26	98.10 ± 0.65	-
TJM		95.00 ± 0.36	78.54 ± 0.56	80.75 ± 0.50	-
ARTL		93.93 ± 0.48	89.76 ± 0.18	92.00 ± 0.58	-
DANN		91.79 ± 0.84	95.61 ± 0.36	91.00 ± 0.19	-
MRMI		96.86 ± 0.05	99.75 ± 0.25	100.00 ± 0.00	-

Fourthly, MRMI significantly outperforms ARTL, which is a state-of-the-art DA approach based on JDA and manifold regularization. ARTL only matches the features without reweighting source instances. As a result, when cross-domain distribution discrepancy is larger, some source instances which are irrelevant to the target instances will always be contained in the feature-matching subspace. Thus, compared with ARTL, the performance boost of 12.32% can be achieved for MRMI.

Finally, the average accuracy for DANN can reach 89.80%, which performs worse than the proposed approach on the whole. In particular, for the DA tasks A→D and B→D, the accuracies of DANN can only reach 73.90% and 65.70%, respectively. This indicates that the performance of DANN decreases dramatically when the cross-domain discrepancy is substantially large.

4.1.3. Effectiveness Analysis

(1) Feature Distribution

The distribution of features drawn by GFK and MRMI for domain adaptation scenario B→C is displayed in Figure 3. It can be seen from Figure 3 that the abscissa denotes a total of 400 samples and the amount of samples contained in different fault types is imbalanced. Furthermore, the ordinate represents the dimensions of each sample and different colors refer to the different amplitude sizes. According to the feature distributions extracted by GFK, many defect features are identified. However, some fault features still perform similarly. For MRMI, the discrepancies among distinct defect feature distributions are more obvious

which makes the fault category easier to be distinguished. Thus, MRMI can extract more discriminative and representative features and obtain better classification performance.

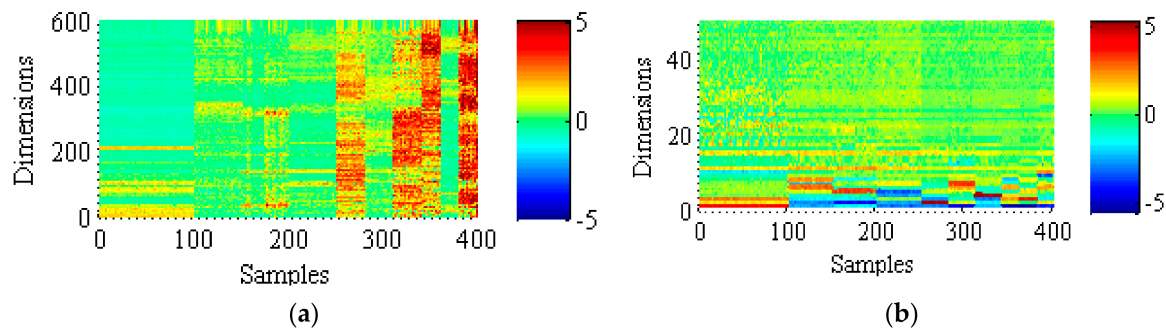


Figure 3. Feature distributions of the unlabeled target domain data based on the learned transferable features (DA task B→C): (a) GFK; (b) MRMI.

(2) Discussion for MRMI

MRMI greatly outperforms the baseline approaches mainly by introducing ℓ_2 -norm regularization, manifold regularization, MVD, and instance reweighting. Several single-factor-based experiments are executed to further study the contributions of these components for MRMI individually, and the experimental results are depicted in Figure 4. To further show the effectiveness of the components of the proposed model, the results of the ablation study for MRMI are summarized in Table 4. Based on the ablation study, it can be seen that the average diagnosis accuracy of MRMI without manifold regularization (MR) can reach 97.28%, which is 2.18% lower than MRMI. This indicates that inducing manifold regularization can obtain a 2.18% transfer improvement comparing with MRMI without MR. For the proposed method, when we do not apply manifold regularization and MVD, the average classification accuracy is 93.85%. This result means that the contribution of MVD to the diagnosis accuracy of MRMI is 3.43%. For MRMI, when we do not apply manifold regularization, MVD, and ℓ_2 -norm, the average classification accuracy is 89.41%. This result means that only inducing ℓ_2 -norm can bring a 4.44% accuracy improvement for the proposed method. When k-nearest neighbor (kNN) is applied as the classifier for MRMI, the final diagnosis result is 0.94% lower than the proposed approach which can reach 98.52%. Notably, the accuracy for task D→A of MRMI with kNN is only 91.2%, which indicates the bad robustness of the kNN classifier in this experiment. Thus, the softmax regression classifier-based MRMI can obtain better diagnostic performance than the kNN classifier-based one.

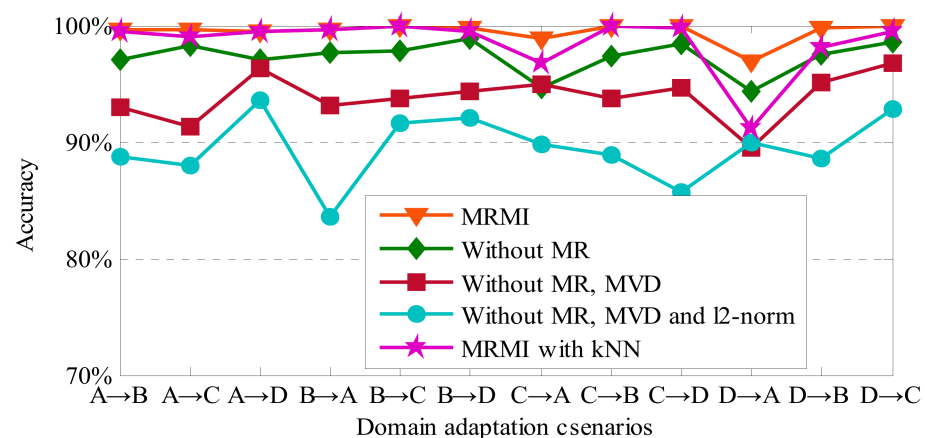


Figure 4. Classification results of single-factor experiments for MRMI.

Table 4. An ablation study for MRMI: Performances are evaluated on rolling bearing dataset.

Model	MR	MVD	ℓ_2 -norm	Instance Reweighting	Softmax Regression	KNN	Average Accuracy (%)
MRMI	✓	✓	✓	✓	✓		99.46
Without MR		✓	✓	✓	✓		97.28
Without MR, MVD			✓	✓	✓		93.85
Without MR, MVD, ℓ_2 -norm				✓	✓		89.41
MRMI with KNN	✓	✓	✓	✓		✓	98.52

Moreover, according to the experiment results, it is necessary to join ℓ_2 -norm regularization, manifold regularization, MVD, and instance reweighting to guarantee the effectiveness and robustness of MRMI while the distribution difference is rather large.

(3) Confusion Matrix

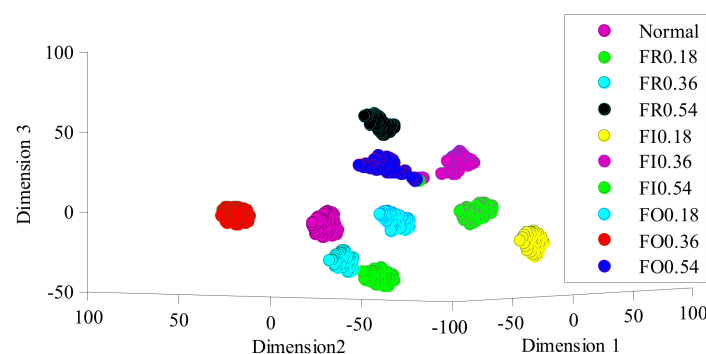
To further study the fault diagnosis effectiveness of MRMI, the confusion matrix of the classification results for DA scenario B→D is displayed in Figure 5. In Figure 5, the rows represent the actual defect types, and the columns stand for the predicted defect types. As we can see from Figure 5, the misclassification issue mainly happens for the defect types of FI 0.36 and FI 0.54. In detail, only one sample of FI 0.36 and one sample of FI 0.54 are misclassified to FO 0.54, thus the classification accuracy of 99.8% is finally obtained for domain adaptation task B→D.

Nor	100	0	0	0	0	0	0	0	0	0
FR0.18	0	100	0	0	0	0	0	0	0	0
FR0.36	0	0	100	0	0	0	0	0	0	0
FR0.54	0	0	0	100	0	0	0	0	0	0
FI0.18	0	0	0	0	100	0	0	0	0	0
FI0.36	0	0	0	0	0	99	0	0	0	1
FI0.54	0	0	0	0	0	0	99	0	0	1
FO0.18	0	0	0	0	0	0	0	100	0	0
FO0.36	0	0	0	0	0	0	0	0	100	0
FO0.54	0	0	0	0	0	0	0	0	0	100
	Nor	FR0.18	FR0.36	FR0.54	FI0.18	FI0.36	FI0.54	FO0.18	FO0.36	FO0.54

Figure 5. Confusion matrix of the fault diagnosis results for DA task B→D.

(4) Feature Visualization

In this section, we execute the t-SNE [43] algorithm to transform the 100-dimension feature vector into a map with 3 dimensions to estimate the ability to learn representative features for MRMI. For instance, the visualization maps of MRMI for DA task B→C is built, and the results are depicted in Figure 6. We can see that most fault features with the same labels are concentrated in the corresponding cluster and different clusters are separated from each other [37]. Thus, MRMI is verified to show strong feature learning ability.

**Figure 6.** Visualization maps of the learned features of DA task B→C.

(5) Parameter Sensitivity

In this part, sensitivity analysis on representative DA tasks $A \rightarrow D$, $B \rightarrow A$, and $C \rightarrow B$ is employed for evaluating the effectiveness and selection of the parameters for MRMI due to space limitation. The classification results with respect to varied parameters k , λ , and γ are displayed in Figure 7. First of all, we implement MRMI with varied values of $k \in [10, 100]$, and the other parameters are fixed as $\lambda = 1$ and $\gamma = 10$. According to the results shown in Figure 7a, stable classification performances can be obtained when subspace dimension k is larger than 50. Thus, we select $k \in [50, 100]$ for MRMI. Then, the proposed approach with varying values of $\lambda \in [1, 10]$ is executed when $k = 50$ and $\gamma = 10$. From Figure 7b, robust diagnosis accuracies can be gained with $\lambda \in [3, 6]$. Finally, varying values of regularization parameter $\gamma \in [1, 10]$ are implemented for MRMI with the other parameter settings of $k = 50$ and $\lambda = 1$. As we can see from Figure 7c, stable diagnosis performance is obtained when γ is larger than 7. Therefore, the optimum regularization parameter γ is set as $\gamma \in [7, 10]$.

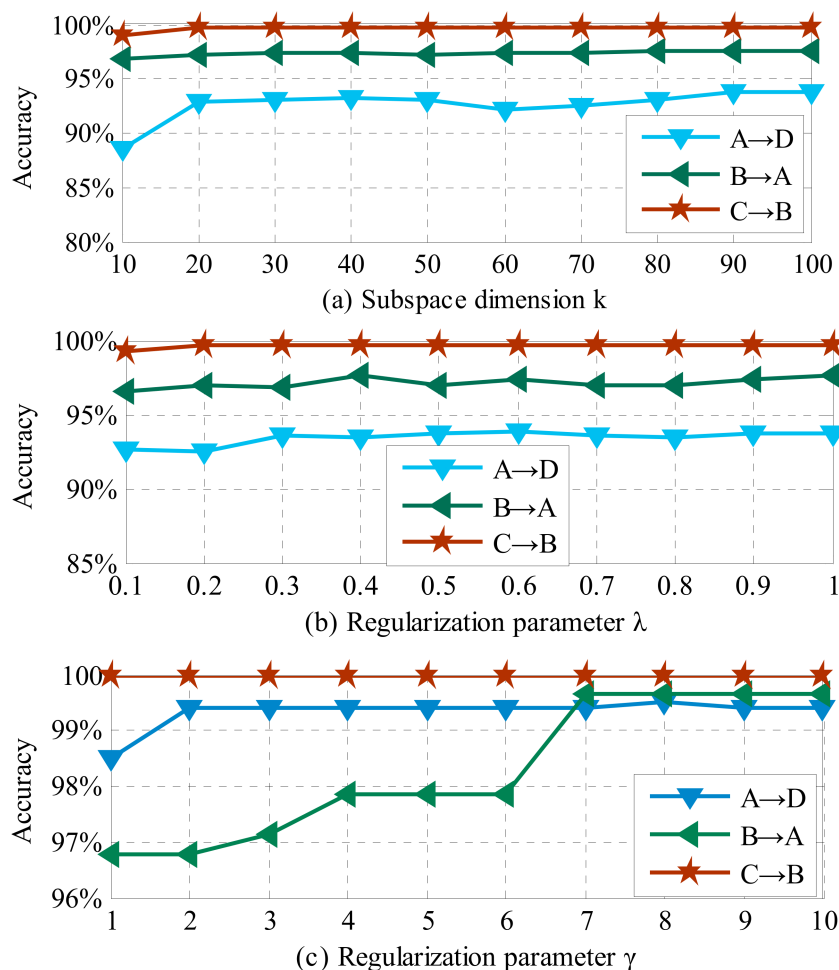


Figure 7. Parameter sensitivity for MRMI on rolling bearing datasets.

4.2. Case 2: Gear Fault Diagnosis

4.2.1. Experimental Setup and Data Description

To further verify the effectiveness of MRMI, a gear dataset with different loads provided by a specially designed gearbox platform is adopted in this part [44]. The raw signals of gears were collected by the sensors installed on the fixed plate of the driving end. Four types of gear fault are considered for the gear fault diagnosis experiment: (1) Single wheel pitting fault; (2) single pinion wear fault; (3) compound fault of pinion wear and wheel pitting; (4) compound fault of pinion wear and wheel teeth broken. We define the normal

state and these four kinds of faults as Type 1 to Type 5, respectively. In addition, the raw vibration signal was acquired with three distinct loads which were denoted as dataset A, B, and C, respectively.

The same as the rolling bearing experiment in case 1, a class imbalanced dataset is adopted for the gear fault diagnosis experiment. The distribution of each dataset is illustrated in Table 5. In addition, the original samples of each dataset are selected alternately to avoid overlap between samples. Then, FFT is employed for preprocessing the raw data. Finally, the time-domain sample containing 1200 datapoints is converted to the frequency-domain sample containing 600 data points.

Table 5. Gear dataset with sample class imbalanced distribution.

Fault Type	Type 1	Type 2	Type 3	Type 4	Type 5	Total
Category Labels	1	2	3	4	5	
Dataset A	100	40	30	20	10	200
Dataset B	50	15	10	15	10	100
Dataset C	100	100	100	100	100	200

4.2.2. Experimental Results

In this experiment, the compared methods and their corresponding parameter selection method are the same as those of the experiment in case 1. Furthermore, six DA scenarios are adopted for empirical evaluation: B→A, B→C, C→A, C→B, A→B, and A→C. The fault classification results for MRMI and the compared methods are displayed in Figure 8.

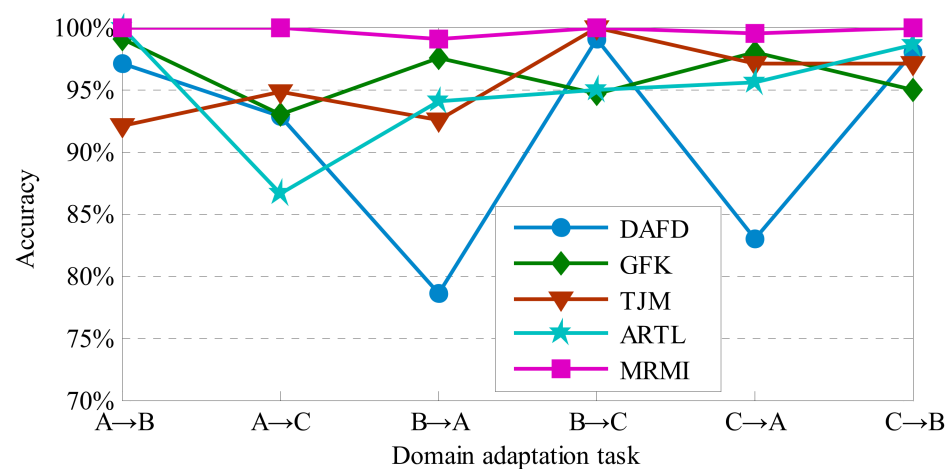


Figure 8. The diagnosis results for sample class imbalanced gear dataset.

It can be seen from Figure 8 that MRMI significantly outperforms the listed baseline approaches in all the domain adaptation scenarios.

Since only six fault types are included in each gear dataset, and the discrepancy between cross-domain distributions is small, higher classification levels can be gained for the DA approaches. Thus, the diagnosis results of all the approaches depicted in Figure 7 are all over 90%. The mean classification result of the six DA tasks can reach 99.75%, and a 3.57% diagnosis performance improvement is acquired for MRMI in comparison to GFK which can obtain the best diagnosis performance among all baseline methods. In general, DAFD performs worse than the other baseline approaches, especially in the DA scenarios B→A and C→A. TJM and ARTL can acquire good classification results, and their mean accuracies are only 4.2% and 4.8% lower than that of MRMI, respectively. Moreover, the robustness of MRMI also performs better than the other compared methods according to the diagnosis results. All in all, the classification results of the gear dataset prove the effectiveness and robustness of MRMI.

5. Conclusions

This study develops a new MRMI method for mechanical fault diagnosis in a class imbalance environment. MRMI joins manifold regularization, MVD, and instance reweighting to handle the class imbalance problem. In addition, ℓ_2 -norm regularization is employed for improving the generalization ability of MRMI. The proposed method is tested on two sample class imbalanced vibration datasets. The classification results show that MRMI can effectively extract more transferable features and significantly outperform the other four baseline domain adaptation approaches while the distribution discrepancy across domains is rather large. Thus, MRMI is a robust and effective DA model for cross-domain mechanical fault diagnosis problems. In the near future, MRMI could be extended to other related fields, such as online health monitoring.

Author Contributions: Conceptualization, Z.Z. and M.S.; methodology, Z.Z.; software, Z.Z. and L.W.; validation, Z.Z., C.M., and S.S.; formal analysis, Z.Z. and L.W.; investigation, Z.Z. and M.S.; resources, Z.Z. and C.M.; data curation, Z.Z.; writing—original draft preparation, Z.Z.; writing—review and editing, Z.Z. and M.S. 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 number 11702162, Natural Science Foundation of Shandong Province, China, grant number ZR2018LE014, Natural Science Foundation of Shandong Province, China, grant number ZR2020MA057 and Natural Science Foundation of Shandong Province, China, grant number ZR2020MA060.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Onel, M.; Kieslich, C.A.; Guzman, Y.A.; Pistikopoulos, E.N. Big data approach to batch process monitoring: Simultaneous fault detection and diagnosis using nonlinear support vector machine-based feature selection. *Comput. Chem. Eng.* **2018**, *115*, 46–63. [[CrossRef](#)]
2. Alexakos, C.T.; Karnavas, Y.L.; Drakaki, M.; Tzifettas, I.T. A combined short time fourier transform and image classification transformer model for rolling element bearings fault diagnosis in electric motors. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 228–242. [[CrossRef](#)]
3. Tighiz, L.; Nasab, M.A.; Yang, H.; Addeh, A. An intelligent system based on optimized ANFIS and association rules for power transformer fault diagnosis. *ISA Trans.* **2020**, *103*, 63–74. [[CrossRef](#)]
4. Adam, G. Fault diagnosis of single-phase induction motor based on acoustic signals. *Mech. Syst. Signal. Pr.* **2019**, *117*, 65–80. [[CrossRef](#)]
5. Li, M.; Tang, Z.; Tong, W.; Li, X.J.; Wang, L.Z. A multi-level output-based DBN model for fine classification of complex geo-environments area using ziyuan-3 TMS imagery. *Sensors* **2021**, *21*, 2089. [[CrossRef](#)]
6. Zennaro, F.M.; Chen, K. Towards understanding sparse filtering: A theoretical perspective. *Neural Netw.* **2018**, *98*, 154–177. [[CrossRef](#)]
7. Gabriel, S.M.; Enrique, L.D.; Viviane, M.; Marcio, C.M. Deep variational auto-encoders: A promising tool for dimensionality reduction and ball bearing elements fault diagnosis. *Struct. Health Monit.* **2019**, *18*, 1092–1128. [[CrossRef](#)]
8. Principi, E.; Damiano, R.; Squartini, S.; Piazza, F. Unsupervised electric motor fault detection by using deep autoencoders. *IEEE-CAA J. Automatic.* **2019**, *6*, 441–451. [[CrossRef](#)]
9. Viola, J.; Chen, Y.Q.; Wang, J. Fault face: Deep convolutional generative adversarial network (DCGAN) based ball-bearing failure detection method. *Inf. Sci.* **2021**, *542*, 195–211. [[CrossRef](#)]
10. Zhang, R.; Tao, H.Y.; Wu, L.F.; Guan, Y. Transfer learning with neural networks for bearing fault diagnosis in changing working conditions. *IEEE Access* **2017**, *5*, 14347–14357. [[CrossRef](#)]
11. Azamfar, M.; Li, X.; Lee, J. Intelligent ball screw fault diagnosis using a deep domain adaptation methodology. *Mech. Mach Theory* **2020**, *151*, 103932. [[CrossRef](#)]
12. Xu, Z.; Huang, D.; Sun, G.X.; Wang, Y.C. A fault diagnosis method based on improved adaptive filtering and joint distribution adaptation. *IEEE Access* **2020**, *8*, 159683–159695. [[CrossRef](#)]
13. An, Z.H.; Li, S.M.; Wang, J.R.; Xin, Y.; Xu, K. Generalization of deep neural network for bearing fault diagnosis under different working conditions using multiple kernel method. *Neurocomputing* **2019**, *352*, 42–53. [[CrossRef](#)]
14. Tong, Z.; Li, W.; Zhang, B.; Jiang, F.; Zhou, G.B. Bearing fault diagnosis under variable working conditions based on domain adaptation using feature transfer learning. *IEEE Access* **2018**, *6*, 76187–76197. [[CrossRef](#)]

15. Wei, J.R.; Liang, J.; He, R.; Yang, J.F. Learning discriminative geodesic flow kernel for domain adaptation. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
16. Singh, J.; Azamfar, M.; Ainapure, A.; Lee, J. Deep learning-based cross-domain adaptation for gearbox fault diagnosis under variable speed conditions. *Meas. Sci. Technol.* **2020**, *31*. [CrossRef]
17. Liu, Y.X.; Li, W.; Zhang, Y.Z.; Coleman, S.; Chi, J.N. Joint transfer component analysis and metric learning for person re-identification. *Electron. Lett.* **2018**, *54*, 821–823. [CrossRef]
18. Zhang, Z.W.; Chen, H.H.; Li, S.M.; An, Z.H.; Wang, J.R. A novel geodesic flow kernel based domain adaptation approach for intelligent fault diagnosis under varying working condition. *Neurocomputing* **2020**, *376*, 54–64. [CrossRef]
19. Makigusa, N.; Naito, K. Asymptotic normality of a consistent estimator of maximum mean discrepancy in Hilbert space. *Stat. Probab. Lett.* **2020**, *156*, 108596. [CrossRef]
20. Satpal, S.; Sarawagi, S. Domain adaptation of conditional probability models via feature subsetting. In Proceedings of the 11th European Conference on Principle and Practice of Knowledge Discovery in Databases (PKDD), Warsaw, Poland, 17–21 September 2007; pp. 224–235.
21. Han, T.; Liu, C.; Yang, W.G.; Jiang, D.X. Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application. *ISA Trans.* **2020**, *97*, 269–281. [CrossRef] [PubMed]
22. Paul, A.; Rottensteiner, F.; Heipke, C. Iterative re-weighted instance transfer for domain adaptation. *ISPRS Annals of the Photogrammetry. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 339–346. [CrossRef]
23. Chen, J.H.; Wang, J.; Zhu, J.X.; Lee, T.H.; Silva, C.D. Unsupervised cross-domain fault diagnosis using feature representation alignment networks for rotating machinery. *IEEE-ASME Trans. Mech.* **2020**. [CrossRef]
24. Long, M.S.; Wang, J.M.; Ding, G.G.; Sun, J.G.; Yu, P.S. A Transfer joint matching for domain adaptation. In Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–29 June 2014; pp. 1410–1417.
25. Razzak, I.; Saris, R.A.; Blumenste, M.; Xu, G.D. Integrating joint feature selection into subspace learning: A formulation of 2DPCA for outliers robust feature selection. *Neural Netw.* **2020**, *121*, 441–451. [CrossRef]
26. Long, M.S.; Wang, J.M.; Ding, G.G.; Pan, J.L.; Yu, P.S. Adaptation regularization: A general framework for transfer feature learning. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1076–1089. [CrossRef]
27. Wang, X.G.; Jie, R.; Liu, S. Distribution adaptation and manifold alignment for complex processes fault diagnosis. *Knowl.-Based Syst.* **2018**, *156*, 100–112. [CrossRef]
28. Wang, J.; He, Q.B. Wavelet packet envelop manifold for fault diagnosis of rolling bearings. *IEEE Trans. Instrum. Meas.* **2016**, *65*, 2515–2526. [CrossRef]
29. Ibañez, R.; Abisset-Chavanne, E.; Aguado, J.V.; Gonzalez, D.; Cueto, E.; Chinesta, F. A manifold learning approach to data-driven computational elasticity and inelasticity. *Arch. Comput. Method Eng.* **2018**, *25*, 47–57. [CrossRef]
30. Lu, Q.G.; Jiang, B.B.; Harinath, E. Fault diagnosis in industrial processes by maximizing pairwise Kullback-Leibler divergence. *IEEE Trans. Control Syst. Technol.* **2019**, 1–6. [CrossRef]
31. Okuno, A.; Shimodaira, H. Hyperlink regression via Bregman divergence. *Neural Netw.* **2020**, *126*, 362–383. [CrossRef]
32. Borgwardt, K.M.; Gretton, A.; Rasch, M.J.; Kriegel, H.P.; Schölkopf, B. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* **2006**, *22*, 49–57. [CrossRef]
33. Nitin, K.; Suyash, P.A. Semi-supervised robust mixture models in RKHS for abnormality detection in medical images. *IEEE Trans. Image Process.* **2020**, *29*, 4772–4787. [CrossRef]
34. Zhang, Y.Z.; Ji, X.Q.; Peng, L.Y.; Liang, X.P.; Xu, Q.W. Fault diagnosis for power transformer using stacked auto-encoders and Softmax regression. *China Sci.* **2018**, *13*, 2694–2699.
35. Lu, W.N.; Liang, B.; Cheng, Y.; Meng, D.S.; Yang, J.; Zhang, T. Deep model based domain adaptation for fault diagnosis. *IEEE Trans. Ind. Electron.* **2017**, *64*, 2296–2305. [CrossRef]
36. Hao, Z.H.; Ma, S.W.; Chen, H.; Liu, J.J. Dataset denoising based on manifold assumption. *Math. Probl. Eng.* **2021**, 1–14. [CrossRef]
37. Zhang, Z.W.; Chen, H.H.; Li, S.M.; An, Z.H. A novel unsupervised domain adaptation based on deep neural network and manifold regularization for mechanical fault diagnosis. *Meas. Sci. Technol.* **2020**, *31*. [CrossRef]
38. Loparo, K. Case Western Reserve University Bearing Data Center. 2013. Available online: <http://cseggroups.case.edu/bearingdatacenter/pages/12k-drive-end-bearing-fault-data> (accessed on 12 January 2021).
39. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Franois, L.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35. [CrossRef]
40. Wang, X.X.; He, H.B.; Li, L.S. A hierarchical deep domain adaptation approach for fault diagnosis of power plant thermal system. *IEEE Trans. Ind. Inform.* **2019**, *15*, 5139–5148. [CrossRef]
41. Li, X.; Zhang, W.; Ding, Q.; Sun, J.Q. Multi-Layer domain adaptation method for rolling bearing fault diagnosis. *Signal Process.* **2019**, *157*, 180–197. [CrossRef]
42. Duan, L.X.; Tsang, L.W.; Xu, D. Domain transfer multiple kernel learning. *IEEE Trans. Pattern. Anal.* **2012**, *34*, 465–479. [CrossRef]
43. Agis, D.; Pozo, F. A frequency-based approach for the detection and classification of structural changes using t-SNE. *Sensors* **2019**, *19*, 5097. [CrossRef] [PubMed]
44. Jiang, X.X.; Li, S.M.; Wang, Y. A novel method for self-adaptive feature extraction using scaling crossover characteristics of signals and combining with LS-SVM for multi-fault diagnosis of gearbox. *J. Vibroeng.* **2015**, *17*, 1861–1878.