ROLLING BEARING FAULT DIAGNOSIS BASED ON MULTI-SCALE ATTENTION AND IMPROVED JOINT DISTRIBUTION ADAPTATION NETWORK

Meng Luo[1], Quanfeng Li[2], Wei Cui[3]*

[1]*Shanghai DianJi University, Shanghai, China*
[2]*Shanghai DianJi University, Shanghai, China*
[3]*Shanghai Yingbeide Intelligent Technology Co., Ltd, Shanghai, China*
**liqf@sdju.edu.cn*

Keywords: ROLLING BEARINGS, FAULT DIAGNOSIS, MULTI-SCALE ATTENTION, PSEUDO-LABEL LEARNING, JOINT DISTRIBUTION ADAPTATION

Abstract

To address the issues of distribution differences caused by complex working conditions in industrial scenarios and the scarcity of labeled data, a novel multi-scale attention and improved joint distribution adaptation network (MA-IJDA) is developed for rolling bearing fault diagnosis. Firstly, a multi-branch convolutional neural network (MACNN) integrated with a channel attention mechanism is constructed as a shared feature extractor. This network sufficiently captures multi-scale transferable fault features through differentiated receptive fields and adaptively enhances key information. Secondly, by fusing softmax confidence and feature space clustering results, a dual-path pseudo-label generation strategy is developed to boost target domain pseudo-labeling reliability. On this basis, an improved joint distribution adaptation (IJDA) mechanism is developed, which employs a joint mean–variance discrepancy (JMVD) metric to synchronously align marginal and conditional distributions, thereby enhancing inter-class separability and cross-domain discriminative capability in the feature space. Extensive cross-working condition experiments on CWRU and JNU bearing datasets verify that MA-IJDA achieves superior diagnostic performance across varying loads and rotational speeds, confirming its exceptional transferability and generalization in complex industrial scenarios.

1 Introduction

As industrial machinery evolves towards automation and intelligence, rotating machinery operates under complex and variable working conditions for extended periods. The operating status is frequently affected by load fluctuations, speed variations, and environmental noise, causing vibration signals to exhibit non-stationary, multi-scale, and highly nonlinear characteristics, which pose significant challenges for fault diagnosis [1]. In recent years, with the rapid application of deep learning in industrial fault diagnosis, convolutional neural networks (CNNs), owing to their end-to-end automatic feature learning capability, have gradually replaced traditional shallow models and become a mainstream approach in intelligent bearing fault diagnosis [2]. Nevertheless, due to the complex and dynamic working environments in industrial sites, the distribution of vibration signals fluctuates over time [3]. Traditional deep models, built on static operating condition data, lack cross-working condition generalization capability, resulting in markedly inadequate adaptability when operating conditions change.

To resolve this, domain adaptation (DA) methods have been extensively utilized to cross-domain intelligent fault diagnosis tasks, aiming to minimize distribution discrepancies inter-domain and enhance transfer performance. Representative works include the domain adversarial neural network (DANN) proposed by Ganin et al. [4], which realizes domain-invariant feature learning through adversarial training between a feature extractor and a domain discriminator. Long et al. [5] introduced a multi-kernel maximum mean discrepancy (MK-MMD) approach to align the mean embeddings of the source and target domains, achieving global distribution alignment. Song et al. [6] introduced a multi-scale subdomain adaptation model that partitions the source and target domains into corresponding subdomains by fault type and employs the local maximum mean discrepancy (LMMD) based on predicted labels to align their conditional distributions. Most of these methods prioritize marginal distribution alignment but overlook conditional distribution disparities. This oversight can cause subdomain misalignment when identical fault features exhibit asymmetric shifts across working conditions. Moreover， target domain pseudo-labels are inevitably prone to noise interference. which can mislead conditional distribution alignment and induce negative transfer., limiting model adaptability in complex cross-working condition scenarios.

In addressing these challenges, This study presents a multi-scale attention and improved joint distribution adaptation (MA-IJDA) method for rolling bearing fault diagnosis. A multi-branch convolutional feature extractor with channel attention is constructed for adaptive extraction of multi-scale fault features, and a dual-path pseudo-label strategy is designed to enhance target domain labeling reliability. For distribution alignment, a joint mean–variance discrepancy (JMVD) metric is employed to concurrently synchronize the global (marginal) and fine-grained (conditional) distributions.

Extensive transfer experiments on CWRU and JNU bearing datasets validate the proposed method's superior diagnostic precision and adaptability in complex cross-working condition scenarios.

2. Theoretical Background

2.1 Problem definition of domain adaptation

A domain D comprises of a feature space $\mathcal{X}$ and its corresponding marginal probability distribution $P(X)$, where $X=\{x_i,...x_n\}\subseteq\mathcal{X}$ represents the input sample set. Specifically, the source domain is defined as $D_s=\{x_i^s,y_i^s\}_{i=1}^{n_s}$, which contains n_s labeled samples, drawn from the distribution $P_s(X)$ .The target domain is referred to $D_t=\{x_j^t\}_{j=1}^{n_t}$, which contains n_t unlabeled target domain samples following the distribution $P_t(X)$. A task $\mathcal{T}$ consists of a label space $\mathcal{Y}_s=\{1,2,...,C\}$ and a conditional probability distribution $P(y|x)$, where C is the number of fault categories, $\mathcal{Y}=\{y_i\}$ is the set of labels, and $\mathcal{Y}_s=\mathcal{Y}_t$.

DA methods aim to learn a classifier $f:x\to y$ based on labeled source domain data D_s to predict the labels of target domain data D_t, while minimizing the expected risk on the target domain. To achieve this, it is essential to address both marginal distribution shift $P_s(X)\neq P_t(Y)$ and conditional distribution shifts $P_s(Y|X)\neq P_t(Y|X)$.

2.2 Distribution distance measurement

Mechanical fault diagnosis faces challenges from vibration signals' high variance and non-stationary nature. a single statistical measure (such as the mean or variance) is insufficient to comprehensively characterize the distribution discrepancy across domains. To address this issue, a Joint Mean-Variance Discrepancy (JMVD) metric is proposed in this study, which integrates the Maximum Mean Statistic Discrepancy (MMSD) [7] and the Variance Discrepancy Representation (VDR) [8] to jointly optimize multi-order statistical distributions:

$$\begin{aligned}JMVD^2[D_s,D_t]&=MMSD^2[D_s,D_t]+VDR^2[D_s,D_t]\\&=\left\|E_s[k(x_s,\cdot)\otimes k(x_s,\cdot)]-E_t[k(x_t,\cdot)\otimes k(x_t,\cdot)]\right\|_H^2\\&+\left\|E_s\tau(x_s,\cdot)-E_t\tau(x_t,\cdot)\right\|_{H_1\otimes H_2}^2\end{aligned}\tag{1}$$

Where E_s and E_t represent the expectations over the domains, respectively, $k(\cdot,\cdot)$ denotes the Gaussian kernel function, $\tau(x,\cdot)=\{k(x,\cdot)-E[k(x,\cdot]\}^{\otimes 2}$ represents the mean-removed kernel tensor product, directly reflects the variance information of the data in the RKHS. The symbol denotes the tensor product operation.

3. Multi-scale Attention and Improved Joint Distribution Adaptation Network

The proposed fault diagnosis model MA-IJDA comprises three core modules: a multi-scale feature extraction module, a label generation module, and an improved joint distribution adaptation (IJDA) module. The overall framework is illustrated in Fig. 1. Firstly, parallel multi-scale convolutional layers are employed to extract multi-dimensional fault features under different receptive fields, and a channel attention mechanism is incorporated to dynamically amplifying of critical features. Secondly, a dual-path pseudo-label generation strategy is designed to predict pseudo-labels to target domain samples from both probability confidence and feature space distribution perspectives, effectively improving the reliability of target domain annotations. In addition, the IJDA mechanism, combined with the JMVD metric, is adopted to jointly align the marginal and conditional distributions over the two domains, thereby improving performance across domains and inter-class discriminability.

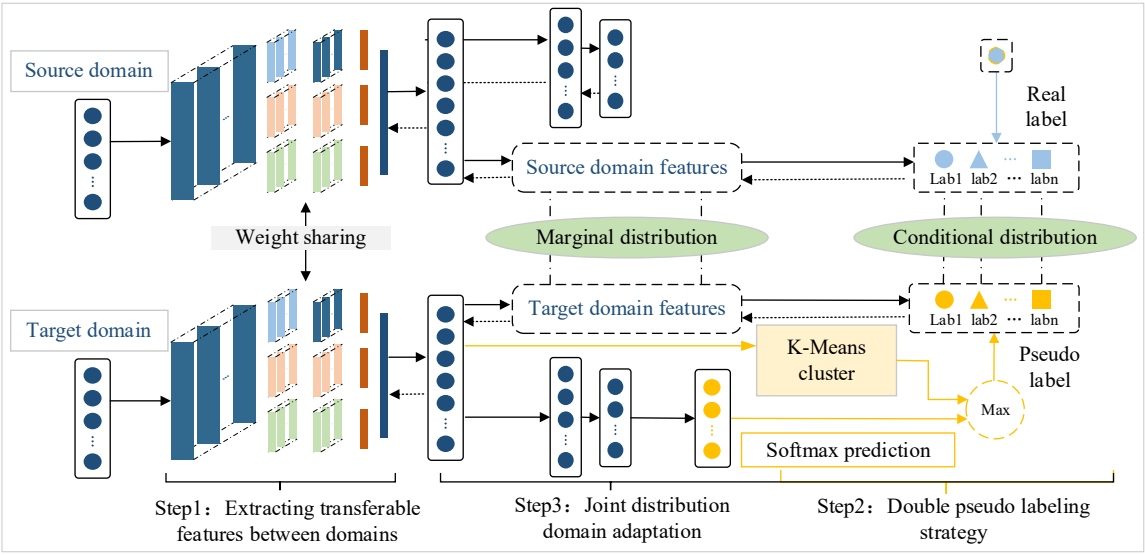

Fig. 1 The network structure of MA-IJDA

3.1 Shared feature extraction network

Rolling bearings typically operate under non-stationary conditions, where the fault features exhibit multi-scale and time-varying complexity. To tackle the inadequate feature extraction ability of traditional CNNs for these signals, An attention mechanism-based multi-scale CNN (MACNN) is designed as the shared feature extraction module., as illustrated in Fig. 2.

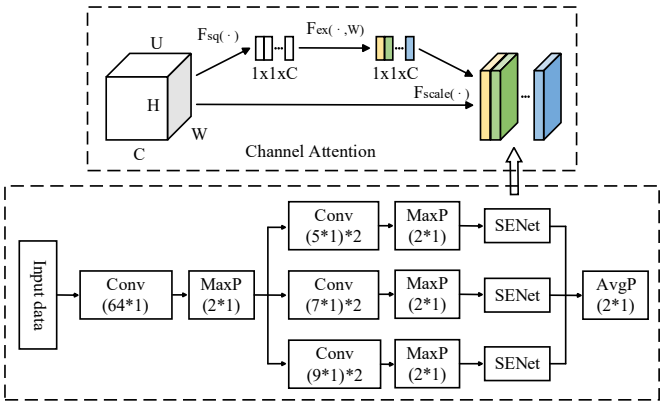

Fig. 2 Architecture of MACNN

Initially, the raw time-series fault signals are fed into a wide convolutional layer to extract global trend features. (adopting the method in [9] to enlarge the receptive field for noise suppression), followed by batch normalization and ReLU activation, and subsequently down sampled via max pooling. On this basis, the multi-scale feature extraction employs parallel convolutional branches with kernel sizes of 5×1, 7×1, and 9×1 to capture fault patterns across different frequency bands and time scales. Subsequently, a squeeze-and-excitation (SENet) channel attention mechanism is introduced, which generates channel-wise statistics through global average pooling, and the resulting statistics are then passed through two fully connected layers and activated by a Sigmoid function to adaptively recalibrate the weights of feature channels. Finally, adaptive pooling is applied to unify the feature dimensions for subsequent joint domain adaptation and classification modules.

3.2 Dual path pseudo-labeling strategy

Existing joint distribution alignment methods typically rely on class labels to estimate conditional distribution discrepancies. However, in transfer learning, target domain labels are unavailable during training and must therefore be annotated with pseudo-labels. Most existing methods use softmax classifier outputs for this, but early in training, decision boundaries are unclear, leading to inaccurate pseudo-labels that can misguide distribution alignment and cause prototype shift. To address this, we propose a dual-path pseudo-labeling strategy that integrates softmax prediction with structured prediction, enhancing the accuracy and robustness of pseudo-labels by jointly considering prediction confidence and spatial distribution.

First, based on the trained classifier, The probability of a target domain sample x_t being classified into class y is obtained through Softmax：

$$p_1(y \mid x_t) = \frac{\exp(\theta_y^T f(x_t))}{\sum_{k=1}^{C} \exp(\theta_k^T f(x_t))} \tag{2}$$

Where θ_y denotes the weight parameter corresponding to class y, and $f(x_t)$ represents the feature vector of sample x_t.

A structured prediction method is then applied to refine pseudo-labels by exploiting pairwise similarities within the target domain. Initially, temporary cluster centers for each class are estimated based on the pseudo-labels. These centers serve as initial points for a K-Means clustering process, which iteratively updates the class prototypes to better capture the underlying feature distribution. The updated class probabilities are computed as follows:

$$p_2(y \mid x_t) = \frac{\exp(-\| f(x_t) - \bar{f}_T^y \|)}{\sum_{k=1}^{C} \exp(-\| f(x_t) - \bar{f}_T^k \|)} \tag{3}$$

Where $\bar{f}_T^y$ represents the cluster center of class y in the target domain feature space.

To enhance pseudo-label robustness, the final label for each target sample is assigned based on the class with the highest probability between the two predictions:

$$\hat{y}_t = \arg\max_y p(y \mid x_t) \tag{4}$$

3.3 Improved joint distribution adaptation

Most current fault diagnosis approaches based on transfer learning align the global distribution between the source and target domains at the classification layer using discrepancy metrics, while ignoring the alignment of conditional distributions across classes. This omission may lead to misclassification of samples near the decision boundary. In response to this， an improved joint distribution alignment (IJDA) mechanism is developed to mitigate the discrepancy in feature distributions between domains while promoting better classification performance. The proposed mechanism consists of two components: marginal distribution alignment (CDA) and conditional distribution alignment (MDA), where a joint metric, JMVD, is defined as the alignment criterion to simultaneously optimize both marginal and conditional distributions.

Specifically, the marginal distribution alignment loss is defined as:

$$L_{MDA} = JMVD\left(f\left(x_s\right), f\left(x_t\right)\right) \tag{5}$$

Where JMVD is calculated according to Equation (5). To further align the inter-domain conditional statistics for each class c, the CDA loss is formulated as:

$$L_{CDA} = \sum_{c=1}^{C} \left\| E_{P(x_s \mid y_s^c)} f(x_s) P(y_s = c) - E_{P(x_t \mid y_s^c)} f(x_t) P(y_t = c) \right\|^2 \tag{6}$$

Where the target domain statistics $P(y_t = c)$ are estimated based on the proposed dual-path pseudo-labeling strategy.

Finally, the overall loss function of the IJDA is expressed as:

$$L_{IJDA} = JMVD\left(f(x_s), f(x_t)\right) + \sum_{c=1}^{C} JMVD\left(P(y_s = c) f(x_s), P(y_t = c) f(x_t)\right) \quad (7)$$

3.4 Loss function

To achieve collaborative optimization of domain alignment and classifier discriminability, the objective is to minimize both the joint distribution alignment loss and the classification loss. Typically, cross-entropy loss is applied to labeled source domain samples. To enhance the separability of feature representations, target domain samples with pseudo-labels are additionally incorporated into the classification loss, which is defined as follows:

$$L_W = -\frac{1}{n_s}\sum_{c=1}^{C} y_s^c \log p(y_s^c \mid x_s) - \frac{1}{n_t}\sum_{c=1}^{C} y_t^c \log p(y_t^c \mid x_t) \quad (8)$$

Finally, the overall loss function is defined as:

$$L_{all} = L_W + \lambda L_{IJDA} \quad (9)$$

Where λ is a trade-off hyperparameter.

4. Experimental Validation

4.1 Description of the CWRU bearing dataset

To assess the proposed method's efficacy, experimental studies are conducted based on the Case Western Reserve University (CWRU) bearing dataset. The bearing type used is SKF6205, and the vibration acceleration signals are collected from the drive end at a sampling frequency of 12 kHz. The dataset covers four typical load conditions (0HP, 1HP, 2HP, and 3HP), each containing 10 operating states, including one normal state and nine fault states formed by combining three fault types (inner race, outer race, and ball faults) with three damage sizes (0.18 mm, 0.36 mm, and 0.53 mm). The detailed data distribution is presented in Table 1.

Table 1. Detailed information of CWRU dataset

Name	Working conditions	Fault size(mm)	Healthy
A	0HP 1797rpm	0.18/0.36/0.53	NF/IF/OF/BF
B	1HP 1772rpm	0.18/0.36/0.53	NF/IF/OF/BF
C	2HP 1750rpm	0.18/0.36/0.53	NF/IF/OF/BF
D	3HP 1730rpm	0.18/0.36/0.53	NF/IF/OF/BF

4.2 Results of CWRU cross-condition experiments

To evaluate the performance of the proposed MS-IJDA method for cross-condition rolling bearing fault diagnosis, comparative experiments were conducted against DDC[10], DANN[4], and the backbone network MACNN. The raw time-domain signals were partitioned into segments via a sliding window of length 1024, overlap-ping by 50%, yielding 200 samples for each health state. The networks were trained using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 128. The cross-condition transfer diagnosis accuracies of four models are shown in Table 2.

Table 2. Experimental results based on CWRU (%)

Task	MACNN (base)	DDC	DANN	MS-IJDA
A→B	90.62	95.54	96.58	99.14
A→C	87.47	97.28	96.43	99.92
A→D	80.31	91.96	92.50	99.50
B→A	84.35	98.72	98.32	98.72
B→C	94.57	99.20	99.52	99.85
B→D	92.81	96.14	97.63	99.90
C→A	89.38	95.57	94.67	99.47
C→B	87.97	97.65	97.62	99.95
C→D	90.91	98.28	98.60	99.83
D→A	78.40	87.54	88.17	98.92
D→B	78.95	88.50	93.92	98.98
D→C	81.79	96.70	97.19	99.82
Average	86.46	95.26	95.93	99.50

Results indicate that, the proposed MS-IJDA model achieved the optimal average diagnostic accuracy of 99.50% across the 12 transfer tasks. By simultaneously aligning both marginal and conditional distributions between domains, MS-IJDA effectively reduced the domain discrepancy at both the global and class levels, significantly enhancing cross-condition adaptation performance. The backbone network MACNN achieved an average accuracy of only 86.46%, indicating that relying solely on deep feature extraction without domain adaptation cannot ensure stable and accurate fault diagnosis under significant distribution shifts. DDC, which constrains the marginal distribution via MMD, improved the average accuracy to 95.26%; however, the lack of conditional distribution alignment led to confusion between classes with similar features. DANN, through adversarial training, aligns the marginal distribution but depends entirely on the domain discriminator, failing to lever-age class-structure information, and it is prone to negative transfer.

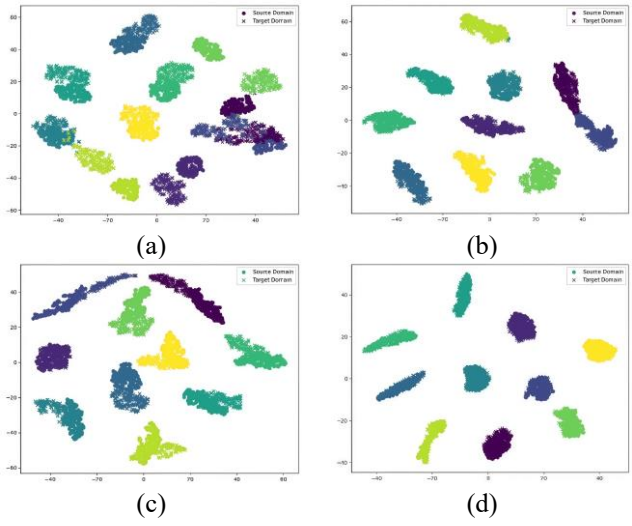

Fig. 3 t-SNE visualization of CWRU dataset (Task B→C). (a) MACNN, (b) DDC, (c) DANN, (d) MS-IJDA

To further intuitively showcase the strengths of MS-IJDA model in cross-domain feature alignment and class discrimination, t-distributed Stochastic Neighbor Embedding (t-SNE) was employed to reduce the dimensionality of the outputs from of the feature extractor. Taking the B→C transfer task as an example, Fig. 3 illustrates the two-dimensional visualization results of MACNN, DDC, DANN, and MS-IJDA. MS-IJDA generates more compact and well-separated clusters for samples of the same class, with the least mixing between source and target samples of the same class and the clearest inter-class boundaries. This confirms its superior performance in joint distribution alignment and classification discrimination.

4.3 Fan bearing experimental verification

To further validate the effectiveness of the proposed method, experiments were conducted on a centrifugal fan bearing fault test platform developed by Jiangnan University (JNU). Bearing vibration signals were sampled at 50 kHz to construct a dataset under three rotational speeds: 600 rpm, 800 rpm, and 1000 rpm, includes four fault types: normal, inner race fault, outer race fault, and ball fault. Different combi-nations of rotational speeds were designed as distinct transfer tasks. The detailed data distribution is presented in Table 3.

Table 3. Detailed information of JNU dataset

Name	Speed	Healthy
E	600rpm	NF/IF/OF/BF
F	800rpm	NF/IF/OF/BF
G	10000rpm	NF/IF/OF/BF

The same four comparison methods as in the previous subsection were employed, and six cross-condition transfer scenarios were designed according to the three rotational speeds. The diagnostic performance of various methods applied to the fan bearing dataset is shown in Table 4 and Fig. 4.

Table 4. Experimental results based on JNU (%)

Task	MACNN (base)	DDC	DANN	MS-IJDA
E→F	90.31	95.62	97.16	99.42
E→G	87.56	93.25	94.87	99.31
F→E	81.38	90.06	92.64	97.37
F→G	89.06	95.31	98.59	99.62
G→E	84.58	94.68	91.30	96.34
G→F	87.40	97.87	96.85	99.65
Average	86.72	94.15	95.24	98.62

As indicated, the performance of MACNN is notably inferior to that of domain adaptation-based approaches in terms of diagnostic accuracy, demonstrating the necessity of domain adaptation for cross-condition fault diagnosis. Both DDC and DANN only align the marginal distributions without considering intra-class distance, which is prone to misclassification of samples adjacent to the decision boundary. In contrast, the proposed method effectively reduces the probability of misclassification near decision boundaries by jointly aligning marginal and conditional distributions. It achieves an average diagnostic accuracy of 98.62%, further confirming its strong adaptability and robustness under complex working conditions.

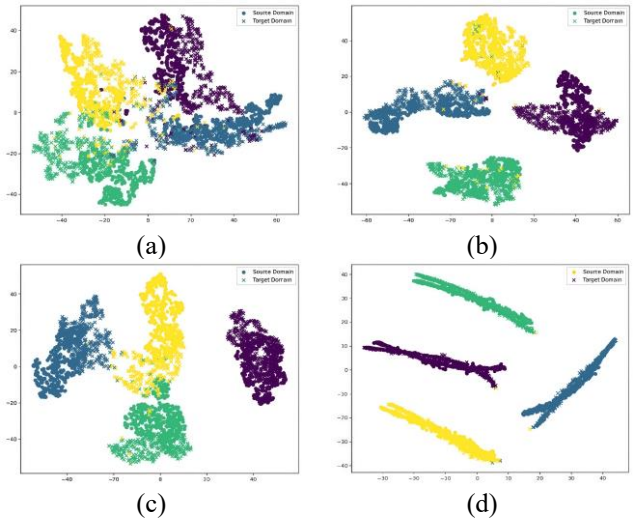

Fig. 4 t-SNE visualization of JNU dataset (Task E→G).
(a) MACNN, (b) DDC, (c) DANN, (d) MS-IJDA

4 Conclusion

In response to the challenges of feature distribution inconsistency and the lack of labels in the target domain for cross-condition bearing fault diagnosis, this paper presents a novel diagnostic framework, MA-IJDA, which integrates a Multi-Scale Attention Convolutional Neural Network (MACNN) with an Improved Joint Distribution Adaptation (IJDA) mechanism. The MACNN employs parallel multi-scale convolutional branches and a channel attention module to effectively integrate fault features at different time-frequency scales, generating robust and highly discriminative feature representations. To handle the unlabeled target domain samples, a dual pseudo-labeling strategy combining softmax prediction and structured clustering prediction is designed to enhance pseudo-label quality. The IJDA mechanism, which uses the Joint Mean and Variance Distance (JMVD) as the distribution distance metric, further improves the intra-class compactness and inter-class separability of the feature space. Our proposed method demonstrates superior performance and robust-ness in cross-domain fault diagnosis scenarios, as evidenced by extensive experiments on two publicly available bearing datasets.

5 References

[1] E. Zio, “Prognostics and Health Management (PHM): Where are we and where do we (need to) go in theory and practice,” Reliability Engineering & System Safety, vol. 218, p. 108119, 2022.

[2] P. Liang, W. Wang, X. Yuan, S. Liu, L. Zhang, and Y. Cheng, “Intelligent fault diagnosis of rolling bearing based on wavelet transform and improved ResNet under

noisy labels and environment," Engineering Applications of Artificial Intelligence, vol. 115, p. 105269, 2022.

[3] Z. Zhao et al., "Applications of unsupervised deep transfer learning to intelligent fault diagnosis: A survey and comparative study," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1–28, 2021.

[4] Y. Ganin et al., "Domain-adversarial training of neural networks," Journal of machine learning research, vol. 17, no. 59, pp. 1–35, 2016.

[5] J. Li, Z. Ye, J. Gao, Z. Meng, K. Tong, and S. Yu, "Fault transfer diagnosis of rolling bearings across different devices via multi-domain information fusion and multi-kernel maximum mean discrepancy," Applied Soft Computing, vol. 159, p. 111620, 2024.

[6] X. Song, W. Sun, G. Liu, et al., "Deep Subdomain Adaptive Network for Motor Rolling Bearing Fault Diagnosis Across Working Conditions," Transactions of China Electro-technical Society, vol. 39, no. 1, pp. 182-193, 2024.

[7] Z. Lv, S. Dong, S. Zhu, et al., "Cross-Condition Rolling Bearing Fault Diagnosis Based on Multi-Source Domain Deep Domain Adaptation," Machine Tool & Hydraulics, vol. 52, no. 20, pp. 230-238, 2024.

[8] Q. Qian, H. Pu, T. Tu, and Y. Qin, "Variance discrepancy representation: A vibration characteristic-guided distribution alignment metric for fault transfer diagnosis," Mechanical Systems and Signal Processing, vol. 217, p. 111544, 2024.

[9] X. Zhang, J. Shang, G. Yu, et al., "Bearing Fault Diagnosis Based on Attention Mechanism and Multi-Scale Convolutional Neural Network," Journal of Jilin University (Engineering Edition), vol. 54, no. 10, pp. 3009-3017, 2024.

[10] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep Domain Confusion: Maximizing for Domain Invariance," Dec. 10, 2014, arXiv: arXiv:1412.3474. doi: 10.48550/arXiv.1412.3474