Mechanical Systems and Signal Processing 217 (2024) 111544

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

Variance discrepancy representation: A vibration characteristic-guided distribution alignment metric for fault transfer diagnosis

Quan Qian [a,b], Huayan Pu [a,b], Tianjia Tu [a,b], Yi Qin [a,b,*]

[a] State Key Laboratory of Mechanical Transmission for Advanced Equipment, Chongqing University, Chongqing 400044, People's Republic of China
[b] College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044, People's Republic of China

ARTICLE INFO

Keywords:
Distribution alignment metric
Student kernel function
Unlabeled target-domain samples
Fault transfer diagnosis

ABSTRACT

Plenty of maximum mean discrepancy (MMD)-based domain adaptation models have been applied to the fault transfer diagnosis. MMD uses the mean statistic in Hilbert space to measure the distribution discrepancy, whereas it cannot precisely reflect the distribution relationship in some cases. Meanwhile, the working mechanism of MMD needs to be further investigated. To this end, a novel insight into MMD is provided, and its working principle from the perspective of space mapping is theoretically explored. A vibration characteristic-guided distribution alignment metric called variance discrepancy representation (VDR) that can express variance information is proposed to enhance the discrepancy representation ability. The biased and unbiased empirical VDR statistics are developed, and the error bounds between the two statistics and the real distribution discrepancy are derived to ensure the reliability of VDR. Furthermore, a new Student kernel is designed to guarantee the robustness and generalization ability of VDR. Ultimately, the VDR with Student kernel-based fault diagnosis model is constructed. The experimental results on three bearing datasets and the actual vibration signals of wind turbine gearbox show that the proposed method is superior to the typical and advanced fault transfer diagnosis methods, especially the proposed VDR outperforms the current distribution alignment metrics. The related code can be downloaded from https://qinyi-team.github.io/2024/05/Variance-discrepancy-representation.

1. Introduction

Machinery equipment is widely distributed in the manufacturing industry, rail transportation, energy facility, military aviation, and other important engineering fields [1–4]. Since it is usually operated in an unsteady and harsh environment, some key mechanical components, such as the bearing, gear, and shaft, are vulnerable to failure. Thus, precise fault detection and identification are critically important for safe operation [5–7]. Owing to the rapid development of the industrial internet and information measurement technology, plenty of data-driven-based fault diagnosis algorithms have been proposed in the past decade. In particular, deep learning-based intelligent diagnosis methods are mushrooming sharply due to their ability of adaptive feature extraction and free expertise experience [8–10].

However, deep-learning achievements greatly depend on sufficient label data and the same distribution assumption between

* Corresponding author.
E-mail address: qy_808@cqu.edu.cn (Y. Qin).

https://doi.org/10.1016/j.ymssp.2024.111544
Received 4 December 2023; Received in revised form 8 April 2024; Accepted 18 May 2024
Available online 23 May 2024

training and testing datasets, which cannot be achieved in real engineering. Concretely, it will not permit the machines to work in failure or close-to-failure state for a long time, indicating that the collection of fault data is extremely laborious. Furthermore, regular maintenance will also exacerbate the lack of fault data. Owing to the changing work conditions, different mechanical structures, specific noise environments, and other factors, the distribution discrepancy is also inevitable. Thus, an excellent fault diagnosis method, only requiring little or no label data is expected.

In light of the above issues, domain adaptation (DA)-based diagnosis methods offer a promising solution, which can transfer the prior knowledge category of the source domain to the incomplete target domain by removing the domain shift between two domains to improve the diagnosis accuracy of the target domain [11–13]. According to the availability of target-domain label information, DA can be summarized into three categories: unsupervised DA, semi-supervised DA, and supervised DA, where unsupervised DA is the most challenging and popular. The combination of the relationship between label spaces of two domains, closed DA [14], partial DA [15,16], opened DA [17], and other multiple DA tasks [18], are researched and explored. Furthermore, in light of the number of target and source domains, multi-target closed DA [19], multi-source DA [20,21], and several other variants [22] are also generated. Many scholars have started to focus on fine-grained class-wise conditional distribution alignment (CDA) [23] to enhance the ability of domain confusion with marginal distribution alignment (MDA) and CDA. Moreover, many joint distribution alignment (JDA)-based intelligent diagnosis algorithms are presented to better bridge the distribution gap. In addition, some improved JDA mechanisms are built to complete the theory based on Bayesian theory for a higher task score [18,24]. Moreover, aiming to boost intra-class compactness and inter-class separability, discriminative metric learning-based transfer fault diagnosis methods have emerged [25]. Although DA-based diagnosis methods possess a strong improvement ability of diagnosis accuracy, it always needs to be trained from scratch for each new task, which cannot cater to the requirement of real-time diagnosis in practice. Thus, the universal domain generalization (DG)-based methods are proposed to settle this issue [26], where the target-domain data is unavailable during the training phase. Similar to DA tasks, various DG tasks are researched according to the number of data subjects.

Obviously, the key to success in all of the above transfer learning methods greatly relies on an excellent discrepancy measurement metric. The universal metrics can be divided by explicit metric-based distribution discrepancy and implicit metric-based distribution discrepancy. The former mainly includes maximum mean discrepancy (MMD) [27] and correlation alignment (CORAL) [28], which can be directly used to assess the dataset bias. The latter mainly consists of $\mathscr{A}$-distance [29] and $\mathscr{H}\Delta\mathscr{H}$-distance [30], which are indirectly executed by the adversarial mechanism [31,32]. The measuring principle of $\mathscr{A}$-distance and $\mathscr{H}\Delta\mathscr{H}$-distance is based on single-classifier discrimination and dual-classifier XOR operation, respectively.

Compared with the explicit metric-based distribution discrepancy, the adversarial mechanism-based methods have difficulty arriving at the Nash equilibrium, resulting in loss oscillation and unstable task scores. MMD is the most commonly used among explicit distribution alignment metrics due to its outstanding discrepancy representation ability. However, it is based on measuring the mean value of discrepancy, which cannot precisely reflect the distribution discrepancy relationship in some fault diagnosis cases. Specifically, the one-dimensional vibration signals are usually used as an input for achieving the end-to-end diagnosis, and they are approximatively symmetric along the x-axis. Taking Fig. 1 as an instance, the probability density functions (PDFs) of different normal bearing samples from three laboratory bearing datasets are illustrated. It can be observed from the figure that the variance is more significant for reflecting the distribution gap than the mean value. Although MMD can enlarge the distribution discrepancy by the space mapping, the poor discrepancy representation in the low-dimensional space still limits its application.

To deal with the above issues, this paper proposes a new insight into MMD and theoretically explores its working principle via space mapping from low dimension to high dimension. A set of bases in a high-dimensional space that can express the variance information is first constructed. Then, a Hilbert space is spanned via the above bases. Next, based on the kernel trick, a new vibration characteristic-guided distribution alignment metric called variance discrepancy representation (VDR) is proposed to enhance the ability of

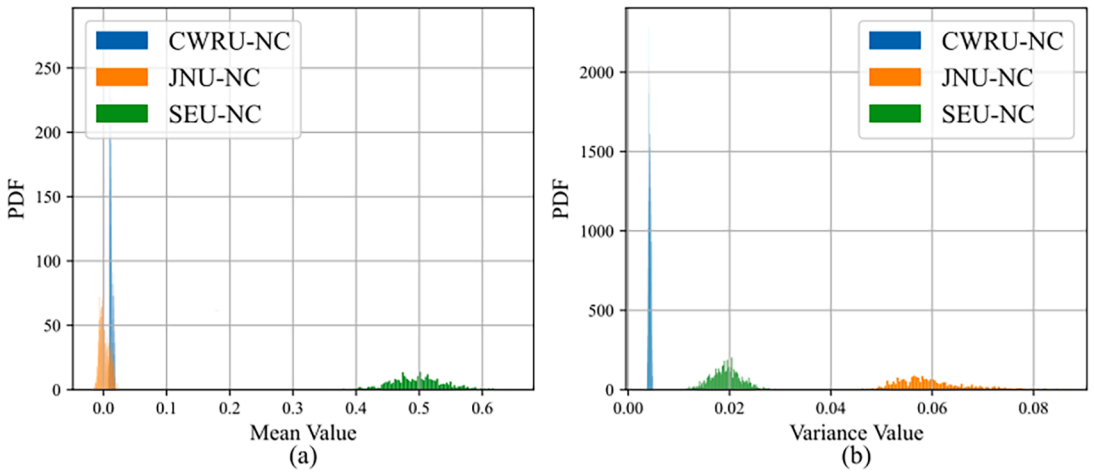

Fig. 1. PDFs under the representation of mean value or variance value on the NC state: (a) the mean representation of different bearing datasets; (b) the variance representation of different bearing datasets. The detailed information on the above datasets is described in Section 4.

discrepancy representation, thereby better achieving the domain confusion. Additionally, considering that the collected monitoring vibration signals are usually subjected to a long-tail distribution and may even mix up some abnormal samples, a new Student kernel function is designed to guarantee the robustness and generalization ability of VDR. Finally, a DA fault diagnosis model based on VDR is constructed to implement the fault transfer diagnosis including the laboratory bearings and the actual wind turbine gearbox. The main contributions and innovations are listed as follows:

1) This paper provides a novel insight into MMD while theoretically exploring the working principle from the perspective of space mapping, laying the foundation for the development of distribution alignment metrics.
2) A vibration characteristic-guided VDR distribution alignment metric based on self-definition reproducing kernel Hilbert space is proposed to enhance the representation ability of distribution discrepancy. Moreover, the biased and unbiased empirical VDR statistics are developed, and the error bounds between the two statistics and the real distribution discrepancy are derived.
3) Considering that the commonly used Gaussian kernel function has poor robustness for the abnormal samples with long-tail distribution, a new Student kernel function is presented to settle the problem in the traditional kernel function.

2. Related works

This paper mainly pays attention to the distribution alignment metrics-based deep transfer diagnosis methods. Firstly, the natural distance, reflecting the distribution discrepancy of two domains ($\mathscr{D}_i, \mathscr{D}_j$), is given as follows [29]:

$$d\left[\mathscr{D}_i, \mathscr{D}_j\right] = 2\sup_{B\in\mathscr{B}}\left|Pr_{\mathscr{D}_i}[B] - Pr_{\mathscr{D}_j}[B]\right| \tag{1}$$

where $\mathscr{B}$ is the set of all measurable subsets under $\mathscr{D}_i$ and $\mathscr{D}_j$. Thus, how to find a metric to maximize the distribution discrepancy between two domains is the core issue. Following the requirement of maximum discrepancy representation, the mainstream distribution alignment metrics are reviewed:

1) $\mathscr{A}$ **distance**: $\mathscr{A}$-distance [29] is derived from the $\mathscr{H}$-divergence, which is devoted to finding a subset A from the entire Borel set $\mathscr{A}$ to arrive at the maximum discrepancy. Given the A via a function set $\mathscr{F}$:

$$A\rightarrow I(f) = \{x \in \mathscr{X} | f(x) = 1, f \in \mathscr{F}\} \tag{2}$$

the $\mathscr{A}$-distance between two domains can be represented as:

$$d_{\mathscr{A}}\left[\mathscr{D}_i, \mathscr{D}_j\right] = 2\sup_{f\in\mathscr{F}}\left|Pr_{\mathscr{D}_i}[I(f)] - Pr_{\mathscr{D}_j}[I(f)]\right| \tag{3}$$

where $I(\cdot)$ is the indicator function. It is evident that Eq. (3) transforms the finding of a subset in Eq. (1) into finding a binary classification function, which greatly simplifies the time complexity. In order to construct a function set possessing a strong representation ability, the neural network was employed to replace it [31].

2) $\mathscr{H}\Delta\mathscr{H}$ **distance**: Similar to the $\mathscr{A}$-distance, the $\mathscr{H}\Delta\mathscr{H}$-distance [30] also constructs a function set via the neural network. Compared with the single-function discrimination in $\mathscr{A}$-distance, $\mathscr{H}\Delta\mathscr{H}$-distance is based on the XOR operation between dual functions to obtain the maximum discrepancy representation [32]:

$$d_{\mathscr{H}\Delta\mathscr{H}}\left[\mathscr{D}_i, \mathscr{D}_j\right] = 2\sup_{f_1 f_2\in\mathscr{F}}\left|Pr_{\mathscr{D}_i}[I(u)] - Pr_{\mathscr{D}_j}[I(u)]\right| \tag{4}$$

where $I(u)$ denotes:

$$I(u) = \{x \in \mathscr{X} \ominus | u(x) = f_1(x) \oplus f_2(x) = 1, f_1, f_2 \in \mathscr{H}\} \tag{5}$$

3) **CORAL distance**: CORAL [28] is a simple but effective distribution alignment metric obtained by verifying many experimental results, which only measures the covariance discrepancy of different sample dimensions in the original sample space:

$$\mathrm{CORAL}\left(\mathscr{D}_i, \mathscr{D}_j\right) = \frac{1}{4d^2}\left\|\boldsymbol{C}_i - \boldsymbol{C}_j\right\|^2 \tag{6}$$

where $\mathscr{C}_i$and $\mathscr{C}_j$ represent the covariance matrices of $\mathscr{D}_i$ and $\mathscr{D}_j$, respectively.

4) **MMD distance**: Different from the CORAL distance, the MMD distance [27] maps the data points from a low-dimensional sample space into a high-dimensional Hilbert space $\mathscr{H}$ to enhance the ability of discrepancy representation. Then, the mean discrepancy of the two domains is computed in the Hilbert space.

$$\mathrm{MMD}\left[\mathscr{H}, \mathscr{D}_i, \mathscr{D}_j\right] = \sup_{h\in\mathscr{H}}\left|E_{x\in\mathscr{D}_i}[h(x)] - E_{y\in\mathscr{D}_j}[h(y)]\right| \tag{7}$$

Combined with the DA mechanisms or the network structure, other distribution alignment metrics are constructed based on the above four distribution alignment metrics, such as LMMD [33], Homm [34], CMD [35], AHMM [36]. Additionally, it should be noted that Wasserstein distance-based transfer diagnosis methods are similar to $\mathscr{A}$-distance ones, where the only difference is that the function set formed from the neural network indicates the 1-Lipschitz. Thus, this section does not introduce it in detail. Since

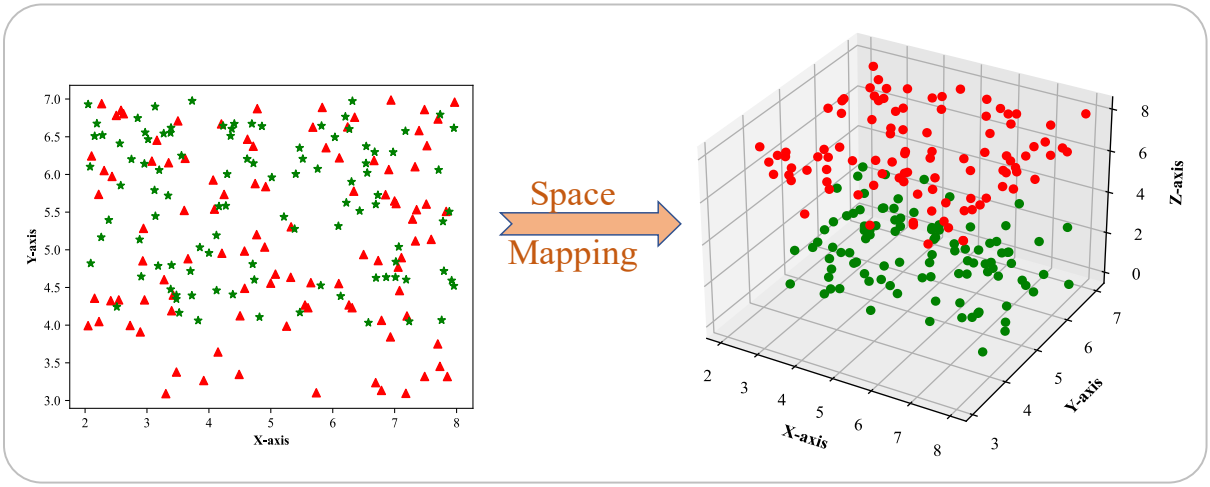

Fig. 2. The instance illustration of space mapping.

randomness is unavoidable during the optimization process of the neural network, the explicit distribution alignment metrics (CORAL, MMD…) are more stable and reliable compared with the implicit ones ($\mathscr{A}$-distance, $\mathscr{H}\Delta\mathscr{H}$-distance…). Although MMD and its variants are the most widely used among them, its discrepancy representation is based on the mean statistic, which still has the risk of poor distribution alignment in some cases, as shown in Fig. 1. To this end, this paper explores a new distribution alignment metric.

3. VDR distribution alignment metric

3.1. Rethinking MMD

The "maximum" word in MMD derives from the space mapping, which maps the data points in low-dimensional sample space into ones in infinite-dimensional Hilbert space, then compute mean discrepancy representation of two domains. In order to visually show the physical meaning of space mapping, space mapping a two-dimensional space to a three-dimensional space is shown in Fig. 2. It is evident that data points in three-dimensional space obtain a better discrepancy representation than two-dimensional space ones via space mapping.

Next, the working principle of space mapping in MMD will be theoretically explored. Let $\mathscr{H}$ first be a Hilbert space and $\mathscr{H}^*$ be its conjugate space, and for each bounded linear function $T \in \mathscr{H}^*$, there is a unique $y_T \in \mathscr{H}$ such that [37]:

$$\begin{aligned} &T(h) = \langle h, y_T \rangle_{\mathscr{H}}, \forall h \in \mathscr{H} \\ &\text{s.t.} \|T\| = \|y_T\| \end{aligned} \tag{8}$$

According to Eq. (8), it can be known that any function h in $\mathscr{H}$ can be represented as the projections on the bases y_T, in which h and y_T can be regarded as the infinite-dimensional vector. Now, how to find a set of bases in $\mathscr{H}$ becomes a key problem. Fortunately, the kernel function $\kappa(x,y)$, possessing a simple mathematical operation property, can satisfy the above requirement via Ref [38], which is:

$$\kappa(x,y) = \langle \kappa(x,\cdot), \kappa(\cdot,y) \rangle = \sum_{i=1}^{\infty} \gamma_i \varphi_i(x) \varphi_i(y) \tag{9}$$

From Eq. (9), we can use the $\kappa(x,\cdot) = \left\{ \sqrt{\gamma_i}\varphi_i(x) \right\}_{i=1,2,\cdots\infty}$ as a set of bases to span a Hilbert space $\mathscr{F} = \text{span}\{\kappa(x,\cdot)|x \in \mathscr{X}\}$, called reproducing kernel Hilbert space (RKHS). Hence, the function $f(x)$ in RKHS can be represented as:

$$f(x) = \langle f, \kappa(x,\cdot) \rangle_{\mathscr{F}} = \sum_{i=1}^{\infty} f_{(i)} \sqrt{\gamma_i} \varphi_i(x), x \in \mathscr{X} \tag{10}$$

Thus, the data points x in the low-dimensional original sample space are mapped into the infinite-dimensional RKHS via the function $\kappa(x,\cdot) : \mathscr{X} \to \mathrm{R}$, which means that the kernel has a variable fixed at x. The kernel function can translate the data points of the low-dimensional sample space as RKHS ones, thereby achieving space mapping.

Given the two domains ($\mathscr{D}_i = \{\mathscr{X}, p(x)\}$, $\mathscr{D}_j = \{\mathscr{Y}, q(y)\}$), where $p(x)$ and $q(y)$ denote the marginal probability distributions, the following formulas can be obtained by reviewing the definition of MMD in Eq. (7):

$$\begin{cases} E_{x\in\mathscr{D}_i}[h(x)] = \left\langle h, \int_{\mathscr{X}} p(x)\kappa(x,\cdot)\mathrm{d}x \right\rangle_{\mathscr{H}} = \langle h, u_p\rangle_{\mathscr{H}} \\ E_{y\in\mathscr{D}_i}[h(y)] = \left\langle h, \int_{\mathscr{Y}} q(y)\kappa(y,\cdot)\mathrm{d}y \right\rangle_{\mathscr{H}} = \langle h, u_q\rangle_{\mathscr{H}} \end{cases} \tag{11}$$

where u_p and u_q are the kernel mean embeddings. Then, Eq. (7) can be rewritten as follows:

$$\begin{aligned} \mathrm{MMD}\left[\mathscr{H}, \mathscr{D}_i, \mathscr{D}_j\right] &= \sup_{\|h\|_{\mathscr{H}}\leqslant 1} \left|\langle h, u_p\rangle_{\mathscr{H}} - \langle h, u_q\rangle_{\mathscr{H}}\right| \\ &= \sup_{\|h\|_{\mathscr{H}}\leqslant 1} \left(\langle h, u_p - u_q\rangle_{\mathscr{H}}\right) = \|h\|_{\mathscr{H}}\|u_p - u_q\|_{\mathscr{H}} \\ &= \|u_p - u_q\|_{\mathscr{H}} \end{aligned} \tag{12}$$

where $\|h\|_{\mathscr{H}}\leqslant 1$ denotes the unit RKHS, and it is used to guarantee that the supremum exists.

According to Eqs. (10–12), it can be seen that MMD obtains the mean discrepancy representation of sample mappings of two domains on RKHS by space mapping. Furthermore, a conclusion can be drawn that the RKHSs spanned by different kernel functions will possess a specific physical meaning, which provides a theoretical guarantee and good guidance in studying the discrepancy representation of other statistics.

3.2. VDR definition

It can be clearly seen from Fig. 1 that the variance representation of fault monitoring signals is more significant than the mean representation. Although MMD can increase linear separability by high-dimensional space mapping, it still suffers the risk of poor discrepancy representation. Thus, a new distribution alignment metric in RKHS, named VDR, is explored.

Via the conclusion in above subsection, a new RKHS $\mathscr{F}_1 \otimes \mathscr{F}_2$, that can reflect the variance information, can be constructed by the following bases $\tau(x,\cdot)$:

$$\begin{aligned} \tau(x,\cdot) &= \left(\kappa(x,\cdot) - E_{x\sim p(x)}\kappa(x,\cdot)\right)^{\otimes 2} \\ &= \left(\kappa(x,\cdot) - E_{x\sim p(x)}\kappa(x,\cdot)\right) \otimes \left(\kappa(x,\cdot) - E_{x\sim p(x)}\kappa(x,\cdot)\right) \end{aligned} \tag{13}$$

Then, based on the setting of Eq. (11), the VDR metric is defined as follows:

$$\begin{aligned} &\mathrm{VDR}\left[\mathscr{F}_1 \otimes \mathscr{F}_2, \mathscr{D}_i, \mathscr{D}_j\right] \\ &= \sup_{\|h\|_{\mathscr{F}_1\otimes\mathscr{F}_2}\leqslant 1} \left(\left\langle h, E_{x\sim p(x)}\tau(x,\cdot) - E_{y\sim q(y)}\tau(y,\cdot)\right\rangle_{\mathscr{F}_1\otimes\mathscr{F}_2}\right) \\ &= \|h\|_{\mathscr{F}_1\otimes\mathscr{F}_2}\left\|E_{x\sim p(x)}\tau(x,\cdot) - E_{y\sim q(y)}\tau(y,\cdot)\right\|_{\mathscr{F}_1\otimes\mathscr{F}_2} \\ &= \left\|E_{x\sim p(x)}\tau(x,\cdot) - E_{y\sim q(y)}\tau(y,\cdot)\right\|_{\mathscr{F}_1\otimes\mathscr{F}_2} \end{aligned} \tag{14}$$

The square VDR is provided to conveniently calculate Eq. (14):

$$\mathrm{VDR}^2\left[\mathscr{F}_1 \otimes \mathscr{F}_2, \mathscr{D}_i, \mathscr{D}_j\right] = \left\|E_{x\sim p(x)}\tau(x,\cdot) - E_{y\sim q(y)}\tau(y,\cdot)\right\|^2_{\mathscr{F}_1\otimes\mathscr{F}_2} \tag{15}$$

According to the theorem that the tensor product of two Hilbert spaces ($\mathscr{H}_1$,$\mathscr{H}_2$) is still another Hilbert space $\mathscr{H} = \mathscr{H}_1 \otimes \mathscr{H}_2$, the following equation can be obtained via the bilinearity of the inner product:

$$\langle \phi_1 \otimes \phi_2, \varphi_1 \otimes \varphi_2\rangle_{\mathscr{H}} = \langle \phi_1, \varphi_1\rangle_{\mathscr{H}_1} \cdot \langle \phi_2, \varphi_2\rangle_{\mathscr{H}_2} \tag{16}$$

where $\phi_1, \varphi_1 \in \mathscr{H}_1$ and $\phi_2, \varphi_2 \in \mathscr{H}_2$. It is evident from the Eq. (16) that the operation property of VDR is also simple like MMD.

3.3. Empirical VDR statistic

1) Biased VDR statistic.

Given m data samples $X : \{x_k\}_{k=1}^m$ in $\mathscr{D}_i$ and n data samples $Y : \{y_k\}_{k=1}^n$ in $\mathscr{D}_j$, the biased VDR statistic is the sum of two V-statistics and a sample average.

$$\mathrm{VDR}_b^2[\mathscr{F}_1 \otimes \mathscr{F}_2, X, Y] = \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m \tau(x_i, x_j) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \tau(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n \tau(x_i, y_j) \tag{17}$$

According to Eq. (16) and $u_x = \frac{1}{m}\sum_{k=1}^m \kappa(x_k,\cdot)$:

$$\begin{aligned}\tau(x_i,x_j) &= \left\langle (\kappa(x_i,\cdot)-u_x)^{\otimes 2}, (\kappa(x_j,\cdot)-u_x)^{\otimes 2} \right\rangle_{\mathscr{F}_1\otimes\mathscr{F}_2} \\ &= \left\langle \kappa(x_i,\cdot)-u_x, \kappa(x_j,\cdot)-u_x \right\rangle^2_{\mathscr{F}_1\otimes\mathscr{F}} \\ &= \left\{ \kappa(x_i,x_j) - \frac{1}{m}\sum_{l=1}^{m}\kappa(x_i,x_l) - \frac{1}{m}\sum_{k=1}^{m}\kappa(x_j,x_k) + \frac{1}{m^2}\sum_{l=1}^{m}\sum_{k=1}^{m}\kappa(x_l,x_k) \right\}^2\end{aligned} \tag{18}$$

Similar to the $\tau(x_i,x_j)$, $\tau(y_i,y_j)$, and $\tau(x_i,y_j)$ can be respectively represented with $u_y = \frac{1}{n}\sum_{l=1}^{n}\kappa(y_k,\cdot)$:

$$\tau(y_i,y_j) = \left\{ \kappa(y_i,y_j) - \frac{1}{n}\sum_{l=1}^{n}\kappa(y_i,y_l) - \frac{1}{n}\sum_{k=1}^{n}\kappa(y_j,y_k) + \frac{1}{n^2}\sum_{l=1}^{n}\sum_{k=1}^{n}\kappa(y_l,y_k) \right\}^2 \tag{19}$$

$$\tau(x_i,y_j) = \left\{ \kappa(x_i,y_j) - \frac{1}{n}\sum_{l=1}^{n}\kappa(x_i,y_l) - \frac{1}{m}\sum_{k=1}^{m}\kappa(x_k,y_j) + \frac{1}{nm}\sum_{l=1}^{n}\sum_{k=1}^{m}\kappa(x_k,y_l) \right\}^2 \tag{20}$$

It can be seen that the calculation of Eqs. (18–20) completely relies on the mathematical property of Eq. (16), which also reveals that why we use the tensor product in Eq. (13) rather than directly define the variance form (dot product).

2) Unbiased VDR statistic.

The unbiased VDR statistic is the sum of two U-statistics and a sample average.

$$\begin{aligned}&\mathrm{VDR}_u^2[\mathscr{F}_1\otimes\mathscr{F}_2, X, Y] = \\ &\frac{1}{(m-1)(m-2)}\sum_{i=1}^{m}\sum_{j\neq i}^{m}\tau(x_i,x_j) + \frac{1}{(n-1)(n-2)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\tau(y_i,y_j) - \frac{2}{(m-1)(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}\tau(x_i,y_j)\end{aligned} \tag{21}$$

It should be noted that Eq. (21) subtracts one degree of freedom, considering the mean value in Eq. (13).

Although the unbiased VDR statistic does not have the systematic error compared with the biased VDR statistic, the following inequation may hold: $\mathrm{VDR}_u^2[\mathscr{F}_1\otimes\mathscr{F}_2, X, Y] < 0$. Thus, when the data samples are relatively sufficient, the biased VDR statistic is recommended to measure the distribution discrepancy between the two domains. Vice versa, when the data samples are limited, the unbiased VDR statistic is nominated to reduce the empirical error.

3) Empirical error bound on biased and unbiased statistics.

Since the observations X and Y are sampled from the true distributions p and q, the error is unavoidable. Thus, the error bound between $\mathrm{VDR}[\mathscr{F}_1\otimes\mathscr{F}_2, X, Y]$ and $\mathrm{VDR}[\mathscr{F}_1\otimes\mathscr{F}_2, p, q]$ should be provided.

Assumed $0\leqslant\kappa(x,y)\leqslant C$, the value range of $\tau(x_i,y_j)$ can be obtained:

$$\begin{aligned}\tau(x_i,y_j) &= \left(\kappa(x_i,y_j) - \kappa(x_i,u_y) - \kappa(u_x,y_i) + \kappa(u_x,u_y)\right)^2 \\ &= \left(\frac{1}{(m-1)(n-1)}\sum_{k\neq i}^{m}\sum_{l\neq j}^{n}\kappa(x_k,y_l)\right)^2 \Rightarrow 0\leqslant\tau(x,y)\leqslant C^2\end{aligned} \tag{22}$$

Then the empirical error bound of biased statistics $\mathrm{VDR}_b[\mathscr{F}_1\otimes\mathscr{F}_2, X, Y]$ can be represented via McDiarmid's theorem [39] and the symmetrization theorem [27] as follows:

$$\begin{aligned}&P_{X,Y}\{|\mathrm{VDR}_b[\mathscr{F}_1\otimes\mathscr{F}_2, X, Y] - \mathrm{VDR}[\mathscr{F}_1\otimes\mathscr{F}_2, p, q]| \\ &\quad > 2\left(C/\sqrt{m} + C/\sqrt{n}\right) + \delta\} \leqslant 2\exp\left(\frac{-\delta^2 mn}{2C^2(m+n)}\right)\end{aligned} \tag{23}$$

where δ is the confidence level.

Via the large deviation bound of U-statistics [40], the empirical error bound of unbiased statistics $\mathrm{VDR}_u[\mathscr{F}_1\otimes\mathscr{F}_2, X, Y]$ can be easily obtained when $m = n$.

$$\begin{aligned}&P_{X,Y}\left\{\left|\mathrm{VDR}_u^2[\mathscr{F}_1\otimes\mathscr{F}_2, X, Y] - \mathrm{VDR}^2[\mathscr{F}_1\otimes\mathscr{F}_2, p, q]\right| \right. \\ &\qquad \left. > \delta\right\} \leqslant 2\exp\left(\frac{-\delta^2\lfloor m/2\rfloor}{8C^4}\right)\end{aligned} \tag{24}$$

where $\lfloor\cdot\rfloor$ represents the floor function.

3.4. Student kernel function

From the definition of empirical VDR statistics in Eq. (17) and Eq. (21), it can be seen that the kernel function $\kappa(x,y)$ is crucial for

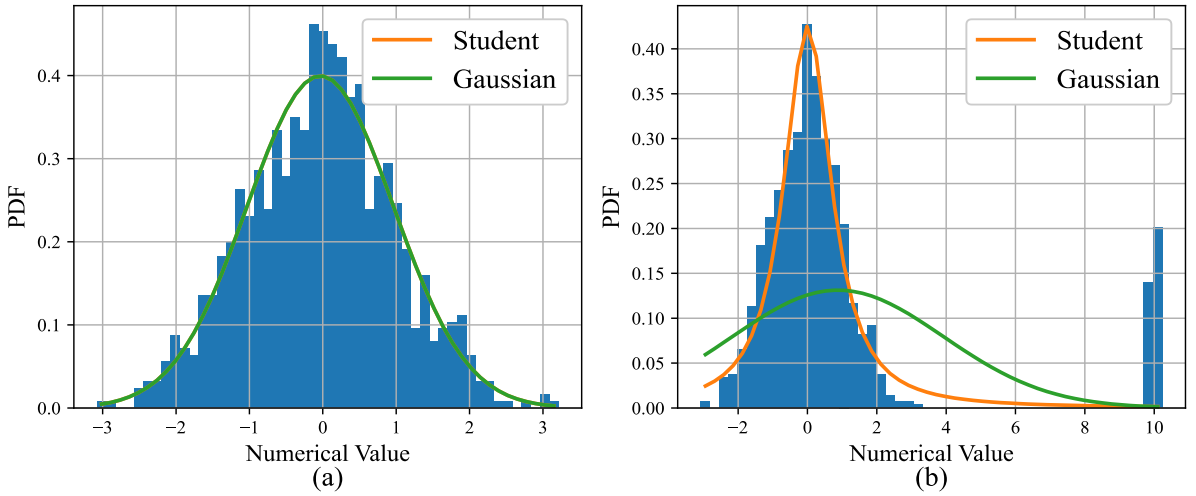

Fig. 3. The distribution fitting of simulated signals: (a) $N(0,1)$; (b) $N(0,1) + N(10,0.1)$.

the discrepancy representation of VDR. The commonly used Gaussian function has poor robustness for the abnormal samples and long-tail distribution. Unfortunately, the situation always exists in practice due to noise disturbance, artificial operation error, and other factors.

In order to better illustrate the problem, the distribution fitting of simulated signals is plotted in Fig. 3, where the signals are simulated by the Gaussian distribution $N(0,1)$, and the abnormal interference is simulated by $N(10,0.1)$. It can be observed from Fig. 3 (b) that the Student distribution is more robust than the Gaussian distribution when the signals exit the interference.

To this end, the Student kernel function is designed and motivated by the Student distribution, which is:

$$\kappa(x,y) = \frac{\Gamma((d+1)/2)}{\sqrt{d}\Gamma(d/2)}\left(1 + \|x-y\|^2/d\right)^{-\frac{d+1}{2}} \tag{25}$$

where the $d > 0$ is the degree of freedom and $\Gamma(\cdot)$ is the gamma function.

$$\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\mathrm{d}t, x > 0 \tag{26}$$

It is evident that $\kappa(x,y)$ can be precisely calculated when d is fixed.

To prove that Eq. (25) possesses the properties of the kernel function, that is, it satisfies Mercer's theorem [38], the following lemma is provided:

Lemma: $\frac{\Gamma((d+1)/2)}{\sqrt{d}\Gamma(d/2)}\left(1 + \|x-y\|^2/d\right)^{-\frac{d+1}{2}}$ is a kernel function when d is the positive integer.

Proof: Due to the complexity of proving Mercer's theorem directly, we will use the following two properties [41] of kernel functions for indirect proof.

①If $\kappa(x,y)$ is a kernel function, then for any constant $\gamma \geqslant 0$, $\gamma \cdot \kappa(x,y)$ is also a kernel function.

②If $\kappa(x,y)$ is a kernel function, then for any positive integer n, $\kappa(x,y)^n$ is also a kernel function.

First, Eq. (25) is rewritten as:

$$\begin{aligned}\kappa(x,y) &= \frac{\Gamma((d+1)/2)}{\sqrt{d}\Gamma(d/2)}\left(1 + \|x-y\|^2/d\right)^{-\frac{d+1}{2}} = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)}d^{d/2}\left(d + \|x-y\|^2\right)^{-\frac{d+1}{2}} \\ &= \frac{\Gamma((d+1)/2)}{\Gamma(d/2)}d^{d/2}\left(\sqrt{d + \|x-y\|^2}\right)^{-(d+1)}\end{aligned} \tag{27}$$

Given $\Theta(d) = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)}d^{d/2}$, Eq. (27) is transformed as:

$$\kappa(x,y) = \Theta(d)\left(\sqrt{d + \|x-y\|^2}\right)^{-(d+1)} \tag{28}$$

where it can be seen that $\Theta(d)$ is a positive constant when d is fixed. Thus, the Student kernel function $\kappa(x,y)$ can be considered as a transformation of the inverse multivariate quadratic kernel function [42] through the properties ① and ②.

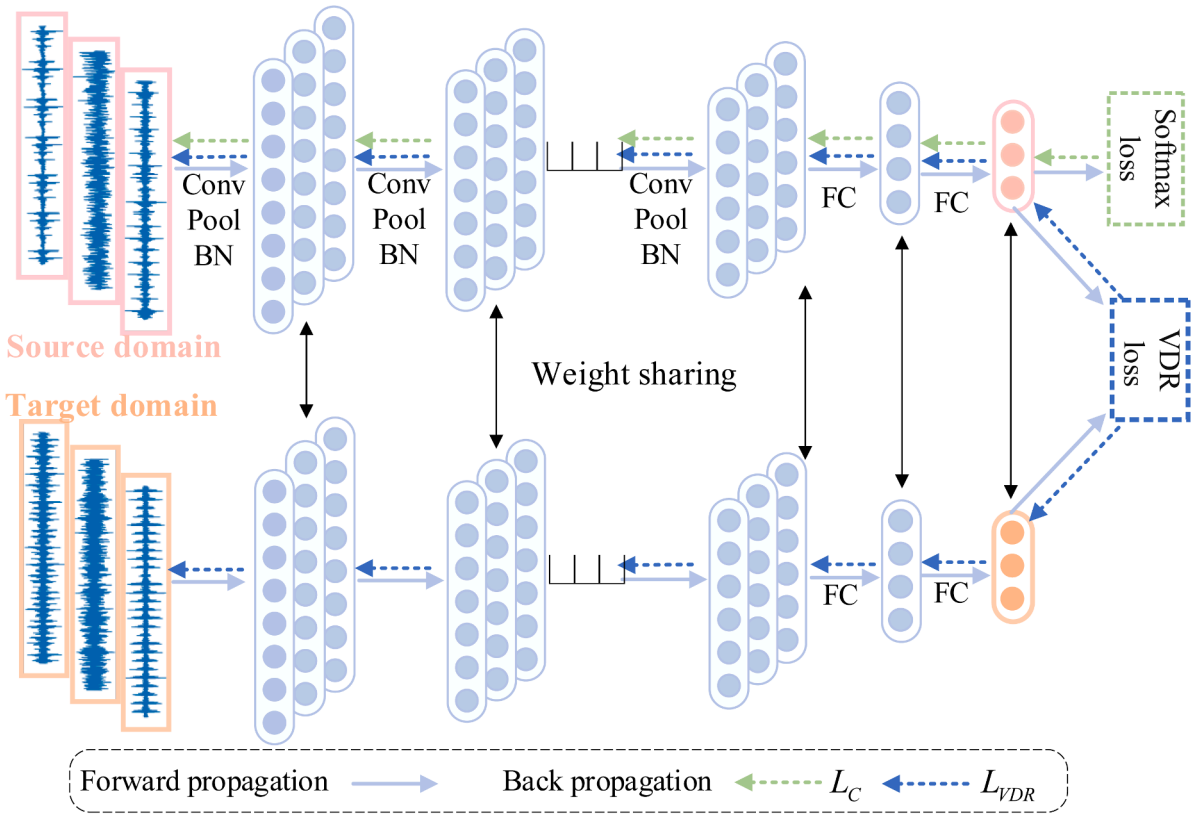

Fig. 4. VDR-based diagnosis model.

$$\left(\sqrt{d+\|x-y\|^2}\right)^{-1} \Rightarrow^{②} \left(\sqrt{d+\|x-y\|^2}\right)^{-(d+1)} \Rightarrow^{①} \Theta(d)\left(\sqrt{d+\|x-y\|^2}\right)^{-(d+1)} \tag{29}$$

3.5. VDR-based model

As shown in Fig. 4, the VDR-based DA diagnosis model employs a one-dimensional convolutional neural network as the backbone to extract fault features. Note that the other network structure can also be utilized. In order to save the computing resources, all the optimization losses are placed at the last fully connected layer.

The study mode of this paper is the common setting, i.e., unsupervised domain adaptation, where only the source domain $\mathscr{D}_S = \{x_S^i, y_S^i\}_{i=1}^{n_S}$ is labeled, and the target domain $\mathscr{D}_T = \{x_T^i\}_{i=1}^{n_T}$ is unlabeled. The Softmax cross-entropy loss L_C works on the source domain to obtain discriminative features via supervised learning:

$$L_C = -\frac{1}{n_S}\sum_{i=1}^{n_S}\sum_{c=1}^{C} I(y_S^i = c)\log(\widehat{y}_S^i) \tag{30}$$

where n_S, C, y_S, $\widehat{y}_S$ respectively denote the sample number, category number, true sample label and predicted sample label in source domain, $I(\cdot)$ is an indictor function. The proposed biased VDR distribution alignment metric is used as the DA loss L_{VDR} to remove the distribution discrepancy between the two domains:

$$L_{VDR} = \mathrm{VDR}_b^2[\mathscr{F}_1 \otimes \mathscr{F}_2, \mathscr{D}_S, \mathscr{D}_T] \tag{31}$$

Then, Adam optimizer is employed to update the network parameters Θ:

$$\Theta^{(t+1)} \leftarrow \Theta^{(t)} - \varepsilon\left(\frac{\partial L_C}{\partial \Theta^{(t)}} + \lambda\frac{\partial L_{VDR}}{\partial \Theta^{(t)}}\right) \tag{32}$$

where ε and λ respectively denote the learning rate and trade-off parameter.

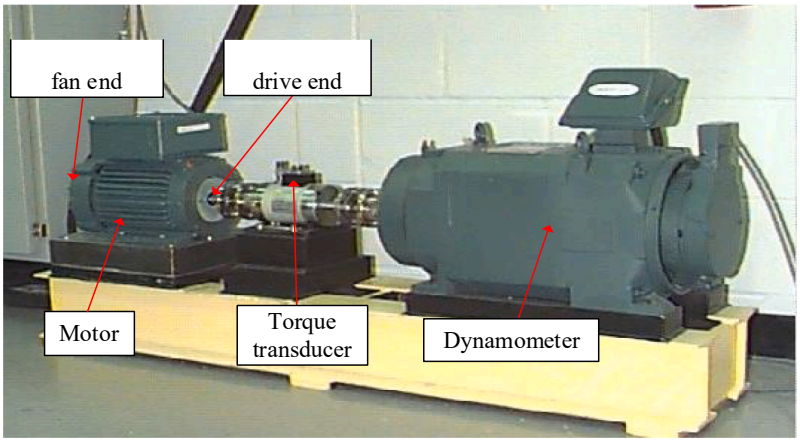

Fig. 5. CWRU bearing testbed.

4. Experiment study

4.1. Testbed description

1) CWRU bearing dataset.

CWRU bearing dataset [43] is collected by Case Western Reserve University, which is the most commonly used standard benchmark dataset in the field of fault diagnosis. As shown in Fig. 5, the testbed consists of a motor, a torque transducer and a dynamotor. The monitored raw vibration signals are gathered on the drive end and fan end, where four health conditions are simulated, including normal condition (NC), inner race fault (IF), ball fault (BF), and outer race fault (OF). There are two types of sampling frequencies (12 000 Hz and 48 000 Hz) on the drive end. All fan-end bearing signals are collected at 12 000 Hz. There are four load types, which include 0 hp, 1 hp, 2 hp, and 3 hp. In this paper, the drive-end bearing is used to implement the transfer diagnosis experiment.

2) JNU bearing dataset.

Similar to the CWRU dataset, the JNU bearing dataset [44] collected by Jiangnan University also consists of NC, IF, BF, and OF. The structure sketch of bearing testbed is illustrated in Fig. 6, which includes a motor, a testing bearing, two support bearings and a loading system. The vibration signals under three input speeds are simulated on testing bearing, which include 600 r/min, 800 r/min, and 1000 r/min. The sampling frequency is 50 000 Hz.

3) SEU bearing dataset.

The SEU bearing dataset [45] is gathered from Southeast University via the dynamic drivetrain simulator. As shown in Fig. 7, the testbed consists of a motor, a motor controller, a planetary gearbox, a parallel gearbox, a brake and a brake controller. The accelerometer is placed on shell of planetary gearbox to collect the monitoring signals. The load information, including 0 V and 2 V, is simulated by controlling the voltage of the brake controller. The input speeds are 1200 r/min and 1800 r/min. Similarly, four health conditions (NC, IF, BF, OF) in the bearing dataset are simulated.

4.2. Implementation details

Six cross-bearing transfer tasks in Table 2 are built via the three bearing datasets to verify the effectiveness of the proposed VDR, in which the detailed task information is listed in Table 1. According to Table 1, the cross-bearing transfer tasks comprehensively reflect the cross-load and cross-speed transfer tasks. Thus, it is challenging for the VDR. In order to ensure that the fault information exists in the data samples, the data length of each sample is set as 3072. 1000 samples of each healthy state for three bearing datasets are

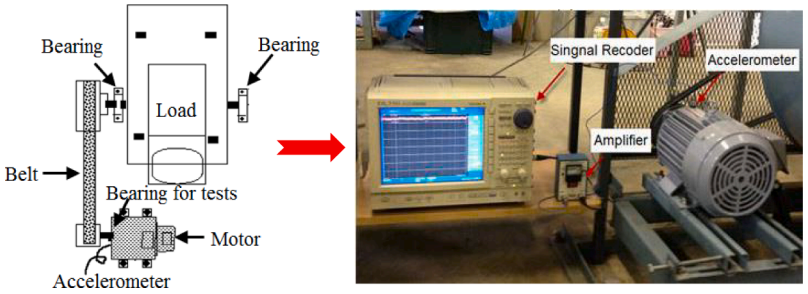

Fig. 6. JNU bearing testbed.

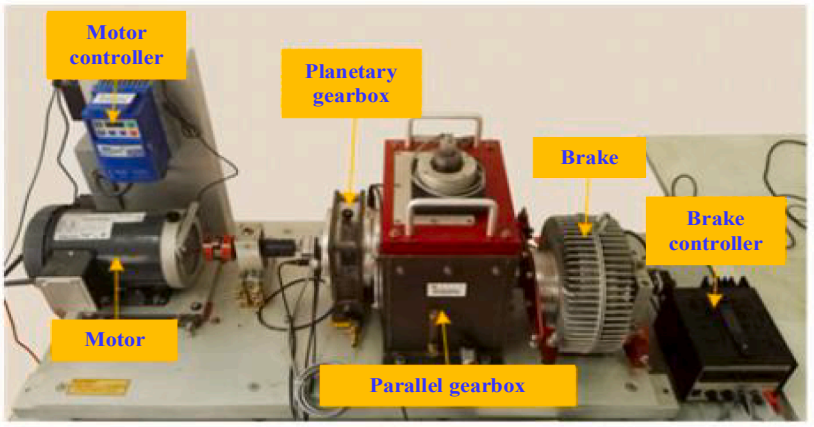

Fig. 7. SEU bearing testbed.

Table 1
Detailed information of three bearing dataset.

Name	Bearing	Health Condition	Speed	Load	Sampling Frequency
A	CWRU	NC\IF\BF\OF	1797 r/min	0 hp	48000 Hz
B	JNU	NC\IF\BF\OF	1000 r/min	/	50000 Hz
C	SEU	NC\IF\BF\OF	1800 r/min	2 V	5120 Hz

produced via the sliding window sampling. All raw vibration signals are directly used as the input for the fault diagnosis without any expertise or experience.

Several famous distribution alignment metrics-based intelligent diagnosis models are used as comparative methods to evaluate the effectiveness and superiority of VDR, which includes the $\mathscr{A}$ distance-based DANN [31], $\mathscr{H}\Delta\mathscr{H}$ distance-based MCD [32], CORAL distance-based DCORAL [46], and MMD distance-based DDC [47] with Gaussian kernel. Besides, the CNN backbone is employed as comparative benchmark. In order to guarantee fairness, the backbone network and optimization setting of all methods are kept the same. Furthermore, the VDR (Gaussian)-based diagnosis model with the Gaussian kernel function is constructed as the ablation experiment to validate the efficacy of the Student kernel function. It should be noted that the default kernel function of VDR is the Student kernel. The selection criterion of trade-off parameter between classification loss and VDR loss is to make them have the same order of magnitude via the binary search.

4.3. Experimental results

Each cross-bearing transfer task is reduplicated five times to ensure the credibility of diagnosis results. The comparative experimental results of six transfer tasks, including the mean value and standard deviation, are shown in Table 2. Benefited from the excellent capability of discrepancy representation, it can be observed that the average diagnosis accuracy of VDR (Gaussian) and VDR (Student) models is significantly higher than other comparative models, in which the improvements arrive at 24.5 % and 43.76 %, respectively. Additionally, in contrast to the VDR (Gaussian), the VDR (Student) has a 19.26 % higher average diagnosis accuracy, revealing the superiority of the proposed Student kernel function. Since the domain confusion in DANN and MCD is achieved by the adversarial training between neural networks, their stabilities on diagnosis accuracy are worse than explicit distribution alignment metrics-based, i.e., a bigger variance occurs.

Table 2
Experimental results of six cross-machine tasks.

Methods	Cross-machine transfer diagnosis tasks (%)						Average
	A → B	B → A	A → C	C → A	B → C	C → B	
CNN	30.13 ± 3.67	40.59 ± 1.83	22.46 ± 2.49	30.47 ± 1.18	19.28 ± 3.00	16.86 ± 2.27	26.63
DDC	33.92 ± 2.50	53.08 ± 1.55	62.45 ± 3.04	55.09 ± 2.46	21.24 ± 2.20	22.11 ± 1.94	41.31
DCORAL	47.97 ± 1.32	58.75 ± 1.55	39.14 ± 1.76	48.78 ± 2.01	18.14 ± 1.12	13.53 ± 0.96	37.72
DANN	49.75 ± 2.08	39.89 ± 3.24	41.76 ± 4.76	41.48 ± 4.76	39.12 ± 6.37	49.92 ± 4.80	43.65
MCD	28.04 ± 1.29	42.49 ± 4.07	24.02 ± 4.22	24.01 ± 2.65	30.77 ± 3.31	31.75 ± 3.61	30.18
VDR(Gaussian)	49.22 ± 1.10	57.10 ± 1.52	92.60 ± 2.16	**96.28 ± 2.46**	72.02 ± 4.21	41.69 ± 1.48	68.15
VDR(Student)	**80.93 ± 3.27**	**97.44 ± 2.51**	**93.73 ± 2.55**	95.38 ± 1.16	**86.17 ± 2.01**	**70.78 ± 2.86**	**87.41**

4.4. Analysis and discussion

The phenomenon that the diagnosis performance of VDR (Gaussian) and VDR (Student) is almost unanimous in A → C and C → A transfer tasks compared with the obvious distinctions in other tasks can be clearly observed in Table II. All PDFs of three bearing datasets under the variance representation are illustrated in Fig. 8 to explore the reason. According to Fig. 8, the PDF of IF healthy condition of JNU bearing dataset is inclined to a long-tail distribution (Student distribution). The Student distribution is able to focus on the sample body by ignoring the outlier samples. However, the Gaussian distribution is extremely difficult to fit the long-tail distribution. Therefore, the designed Student kernel can better display the distribution discrepancy between source-domain and target-domain samples compared with the commonly used Gaussian kernel. An inference can be drawn that the above phenomenon is the result of the non-Gaussian form of IF in the JNU bearing dataset.

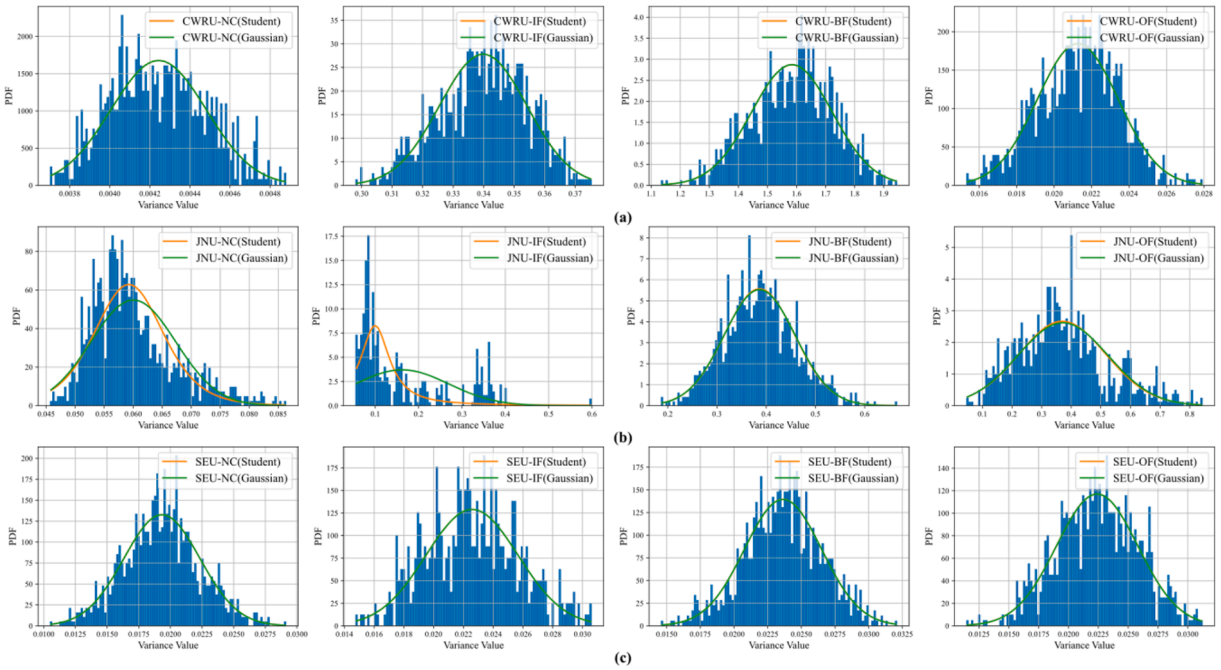

Fig. 8. The PDFs of all health conditions under the variance representation: (a) CWRU; (b) JNU; (c) SEU.

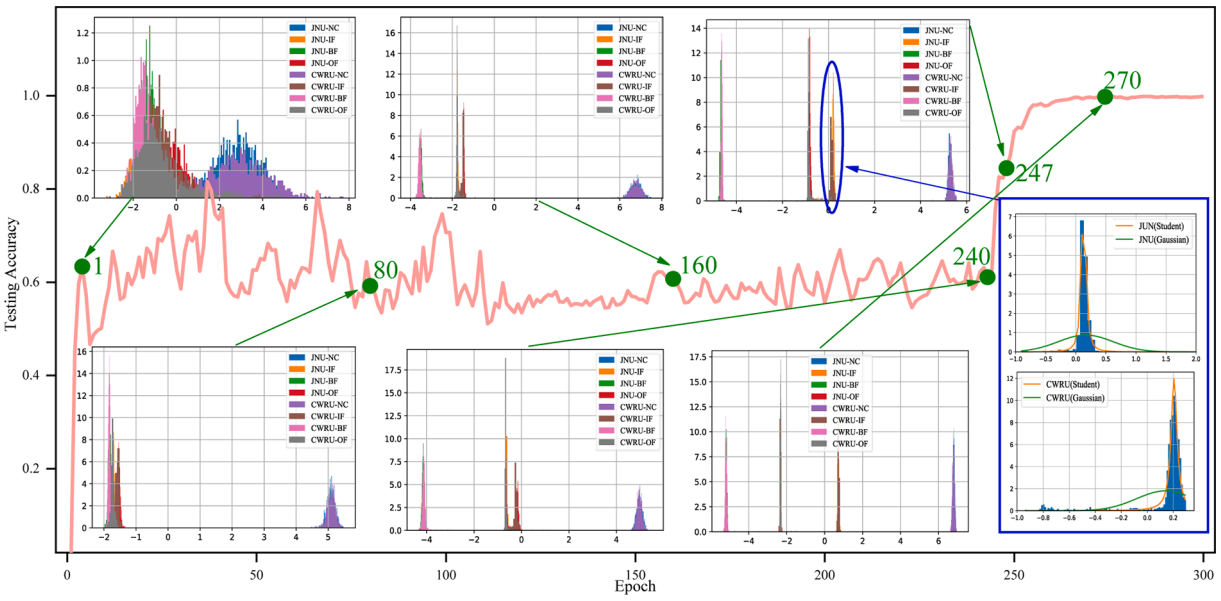

Fig. 9. The distribution changes with the curve of test accuracy under the JNU → CWRU transfer task.

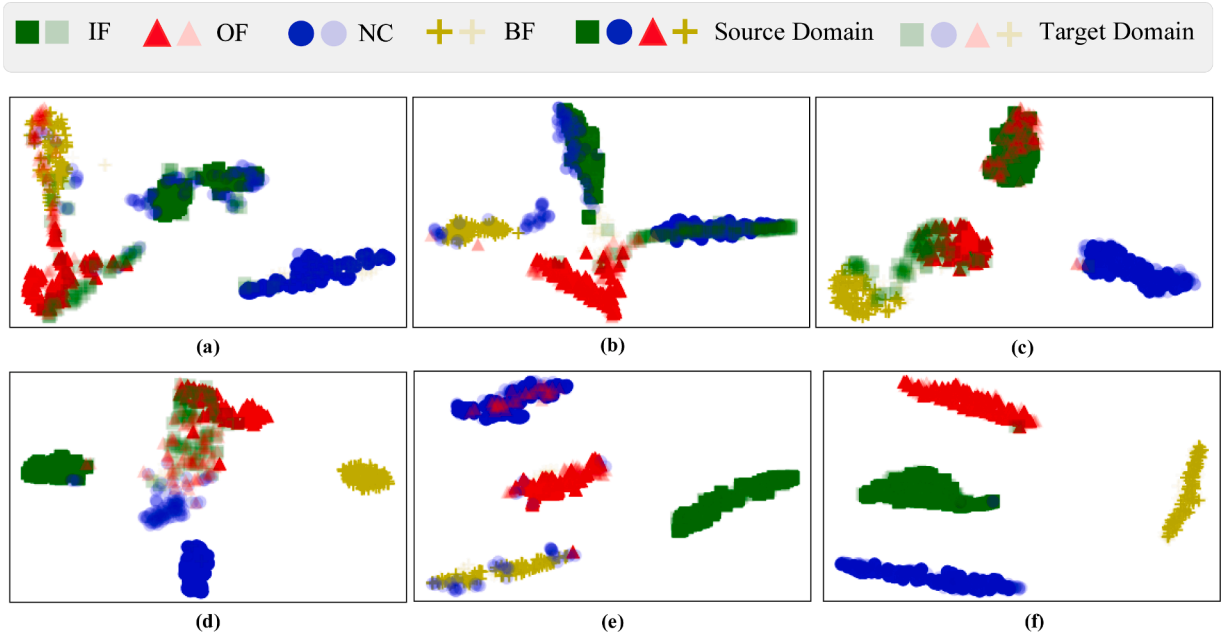

Fig. 10. The t-SNE mappings of the features of all health conditions extracted by six diagnosis methods on the JNU → SEU transfer task: (a) DDC; (b) DCORAL; (c) DANN; (d) MCD; (e) VDR(Gaussian); (f) VDR(Student).

In order to show the working mechanism of VDR (Student) during the training process, the distribution changes with the curve of testing accuracy are plotted in Fig. 9 by taking the JNU → CWRU transfer task as an instance. There are mainly three stages to be analyzed and discussed:

① At the early stage (epoch: 0–80): The testing accuracy rapidly increases to about 60 % and then fluctuates around 60 %; only the healthy state NC is separable between the source and target domains, whereas the probability distributions of the rest three faults exhibit extensive overlap; the distribution ranges of all categories are wide.

② At the intermediate stage (epoch: 80–240): The testing accuracy still fluctuates around 60 %; the distribution ranges of all categories have been narrowed, indicating the smaller intra-class distances; the BF is gradually far away from the IF and OF, while the NC continues to maintain a highly separable state.

③ At the later stage (epoch: 240–270): A sudden improvement in diagnosis accuracy from 60 % to 100 % can be observed, and the feature distributions of all categories rapidly become separable. In order to clearly illustrate the reason of rapid change, the probability distribution of IF features is enlarged in the epoch of 247. It can be seen from local zoom-in graph that the distribution of IF features is matched with that of the original sample in Fig. 8 (b). Therefore, it can be concluded that the long-tail distribution in IF and OF features causes the difficulty and slowness of their class-wise distribution confusion at the intermediate stage. Finally, in the epoch of 270, the OF and IF features of two domains also quickly become separable with the help of proposed VDR and Student kernel.

The t-SNE mappings from all diagnosis methods are displayed in Fig. 10 by taking the JNU → SEU transfer task as an instance to intuitively evaluate the extraction ability of domain-invariant and discriminative features of VDR. First, all features from the source domain possess a significant separability in all diagnosis methods. However, the domain confusion abilities of the six methods are significantly different. Observing Fig. 10 (a)-(d), it can be seen that there are many overlapping points among different categories, and the miserable alignment performance of contrast models may be mainly attributed to their poor discrepancy representations. In contrast with Fig. 10(a)-(d), the VDR (Gaussian) has a more apparent decision boundary on the target domain. Unfortunately, there are still some error point matchings, mainly focusing on OF and NC. In summary, Fig. 10 (f) possesses the minimum intra-class distance and the maximum inter-class distance compared to other methods. The comparative results further reveal that the proposed VDR (Student) has a strong capacity for fault transfer diagnosis.

4.5. Application into the cross-machine fault diagnosis of actual 2 MW wind turbine gearbox bearings

In order to further evaluate the effectiveness of proposed VDR distribution alignment metric, the actual 2 MW wind turbine gearbox bearing (WTGB) vibration signals are used to implement the laboratory-actual transfer diagnosis tasks with the above laboratory bearing datasets, which includes A → WTGB, WTGB → A, B → WTGB, WTGB → B, C → WTGB, and WTGB → C. The transmission structure of wind turbine is demonstrated in Fig. 11, which mainly consists of a planetary gear transmission chain and a two-stage parallel shaft gear transmission chain. The WTGB dataset also includes four healthy states (NC, IF, BF, and OF), occurring in high-speed shaft, and the sampling frequency is 25.6 kHz. Different from the foregoing three bearing datasets, where their artificial-machining fault shapes are usually regular, the shapes of WTGB bearing dataset are irregular due to the slow degradation and

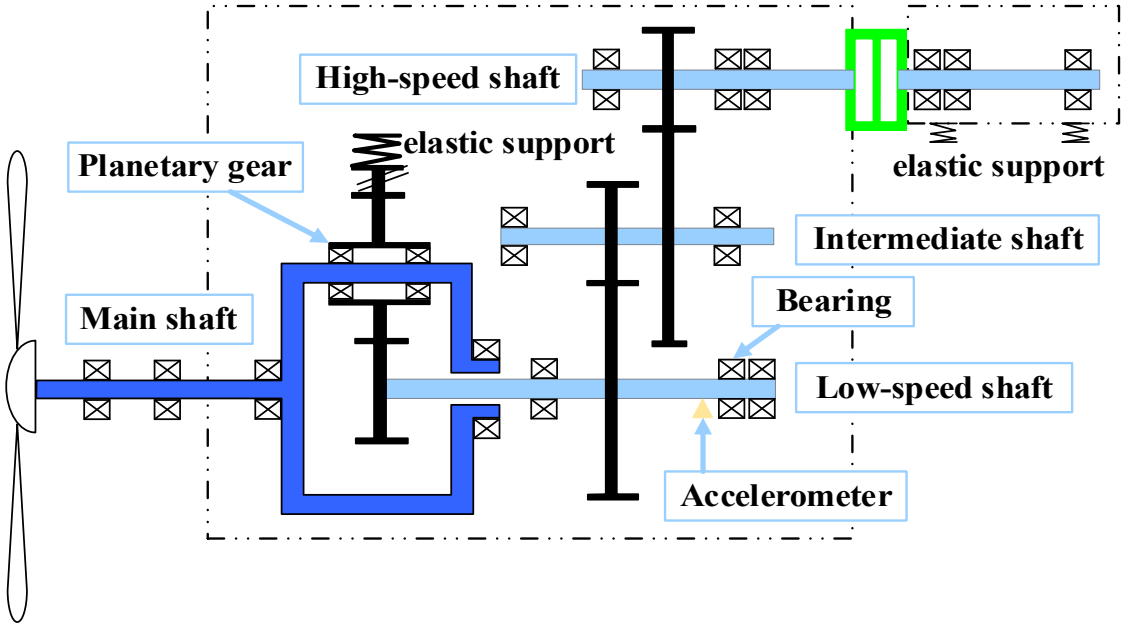

Fig. 11. The transmission structure of the actual 2 MW wind turbine.

Table 3
Experimental results of six laboratorial-actual transfer diagnosis tasks.

Methods	Laboratorial-actual transfer diagnosis tasks (%)						Average
	A → WTGB	WTGB → A	B → WTGB	WTGB → B	C → WTGB	WTGB → C	
CNN	29.96 ± 1.81	31.74 ± 2.13	22.03 ± 0.99	25.37 ± 3.56	23.03 ± 1.60	52.85 ± 1.24	30.83
DDC	40.34 ± 3.02	33.44 ± 1.78	44.49 ± 3.86	40.76 ± 3.18	65.06 ± 6.94	47.53 ± 6.33	45.27
DCORAL	31.30 ± 3.70	29.83 ± 4.77	31.50 ± 3.54	37.84 ± 3.82	60.47 ± 8.10	59.30 ± 5.07	41.71
DANN	25.28 ± 2.14	31.11 ± 3.75	37.13 ± 2.37	40.19 ± 6.46	53.73 ± 6.17	48.72 ± 7.26	39.36
MCD	23.15 ± 3.03	25.91 ± 1.28	29.15 ± 2.44	33.77 ± 3.44	37.37 ± 2.01	33.68 ± 3.92	30.50
VDR	**80.02 ± 2.64**	**74.67 ± 3.53**	**81.81 ± 4.97**	**77.56 ± 3.44**	**83.34 ± 3.69**	**81.90 ± 5.17**	**79.88**

complex working condition. Besides, the speed and load of WTGB continuously and nonlinearly vary along with the wind condition, and the samples of these fault types are collected in different working conditions. Thus, these six transfer tasks are more challenging than those in Table 2.

The experimental results are listed in Table 3, in which the experiment implementation details are the same as those in section 4.2. It can be clearly observed from Table 3 that the diagnosis performance of VDR outperforms other fault transfer diagnosis models, and its average accuracy arrives at 79.88 %. However, the average accuracy is 7.53 % lower than that in Table 2 due to the following factors, such as the irregular fault shape, the complex signal components, the varying working condition, and the big distribution discrepancy between WTGB and laboratory bearing datasets. The comparative diagnosis results again verify that the proposed VDR possesses a stronger generalization ability than other typical and advanced models in the cross-machine fault transfer diagnosis. In order to further evaluate the effectiveness of our proposed Student kernel, we take the CWRU → WTGB and JNU → WTGB transfer tasks as the examples to plot the Fig. 12, in which the Student kernel and Gaussian kernel are respectively applied to the DDC and VDR. It can be seen from Fig. 12 that the Student kernel can help DDC and VDR improve the diagnostic accuracy in both transfer tasks, which shows the superiority of Student kernel over Gaussian kernel.

4.6. Application into the cross-speed fault diagnosis of actual 3 MW wind turbine gearbox

As shown in Fig. 13, the transmission structure of the actual 3 MW wind turbine gearbox is comprised of a two-stage planetary gear transmission and a one-stage parallel-axis gear transmission, whose whole speed ratio is 109.64. There are two faults in this gearbox, including second gear ring (SGR) and high-speed shaft gear (HSG). The SGR is from wind turbines labeled by F34 and F47 in a wind farm, and the HSG is from wind turbines labeled by F14 and F38. The sampling frequencies for SGR and HSG are 12.8 and 25.6 kHz respectively. For containing a fault period at least and reducing the computational burden, the sample dimension is set as 20000. Besides, the sample number of each category is 200. As the working condition is continually varying, we select two speed ranges, namely, “700–750 rpm” and “1200–1250 rpm”, to form the cross-speed diagnosis tasks for different wind turbines. In the transfer

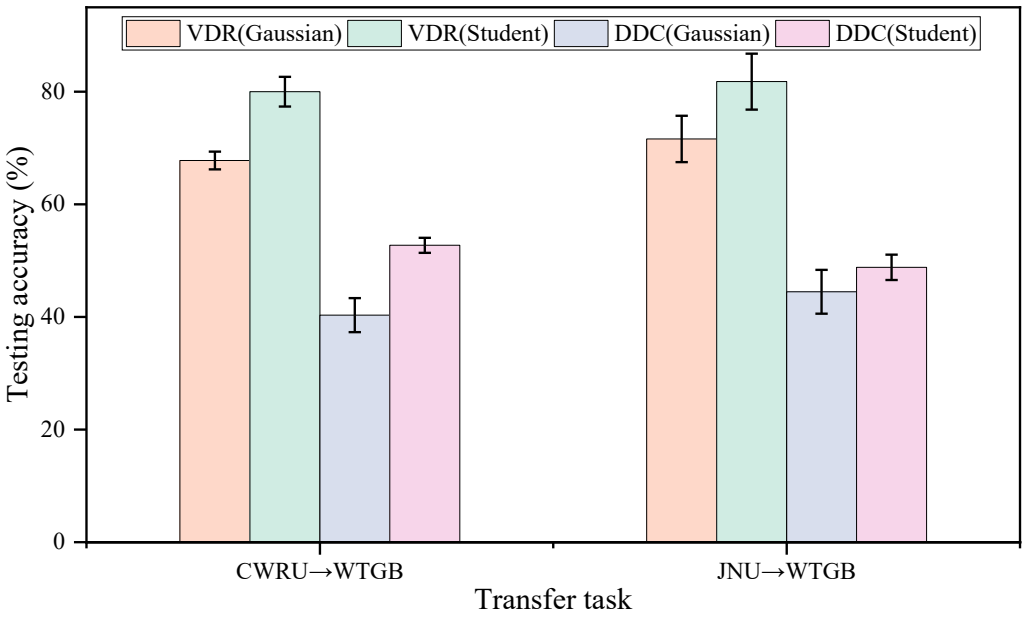

Fig. 12. The experimental results of Student kernel and Gaussian kernel on two transfer diagnosis tasks.

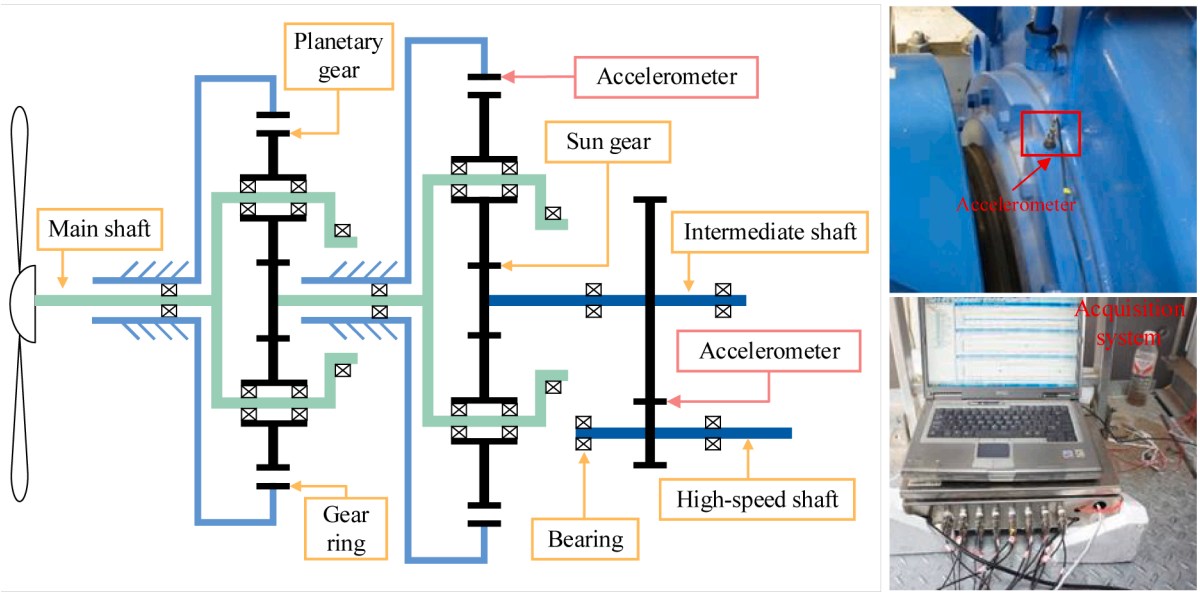

Fig. 13. The transmission structure of the actual 3 MW wind turbine.

tasks, the SGR and HSG samples in "700–750 rpm" are respectively from F34 and F14, and those in "1200–1250 rpm" are respectively from F47 and F38.

The experimental results of two cross-speed transfer diagnosis tasks in 3WM wind turbine are illustrated in Fig. 14. First, even though these tasks are a binary classification problem, it can be seen from the figure that the accuracies of most of diagnosis methods are around 80 %. Particularly, the accuracy of the CNN backbone-based diagnosis method is only about 50 %. This reflects that the actual cross-speed diagnosis case is also challenging like the cross-bearing case in Section 4.4. The proposed VDR realizes around 93 % average accuracy, which significantly outperforms other diagnosis methods. In summary, the comprehensive experimental results further prove the effectiveness and advantage of proposed VDR distribution alignment metric and Student kernel.

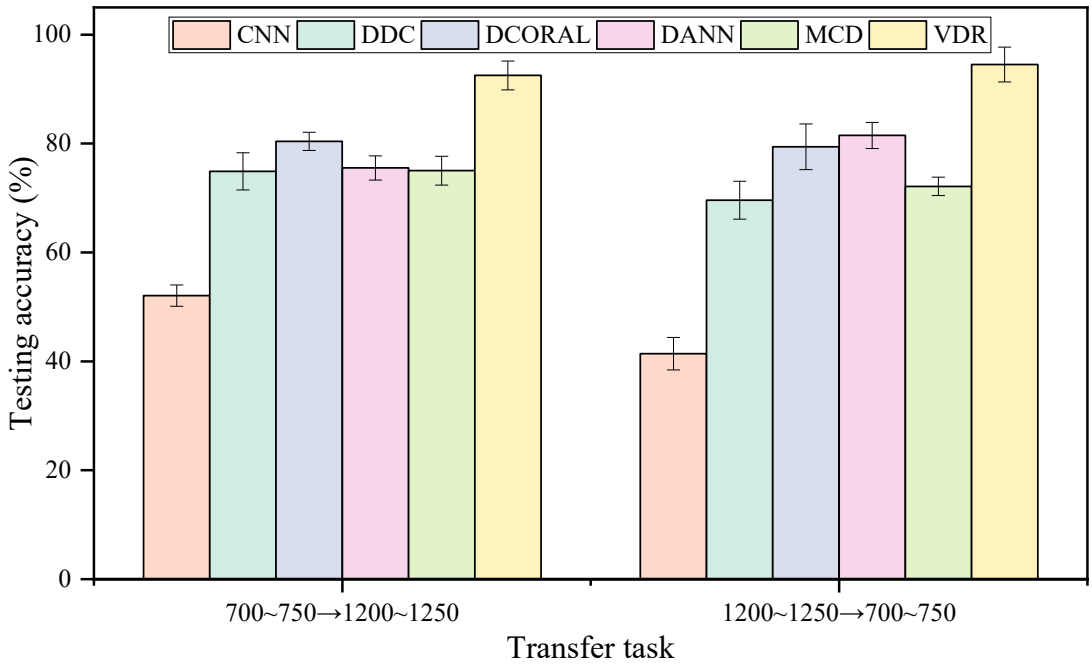

Fig. 14. The experimental results of two cross-speed transfer diagnosis tasks.

4.7. Limitation and further work

The proposed VDR metric with Student kernel is only verified by the closed transfer diagnosis tasks. In practical engineering, the label space relationship between source domain and target domain is unpredictable, which may also be open-set and partial-set. On the other hand, the excellent ability of discrepancy representation in VDR mainly depends on the variance statistic. However, the variance statistic may not be appropriate for other monitoring signals such as current, voltage, temperature, etc. These two limitations greatly constrain the engineering value and application potential of VDR. In future work, we will design a more universal distribution alignment metric tailored to the characteristics of other monitoring signals, and combine the proposed metric with the metric based on mean value for enhancing the discrepancy representation ability in fault transfer diagnosis tasks.

5. Conclusions

In this work, a new insight was introduced into MMD, and its working principle was theoretically explored from the perspective of space mapping. Then, a new VDR distribution alignment metric that can express variance information was proposed to improve the ability of discrepancy representation and achieve domain confusion. The empirical biased and unbiased VDR statistics and its error bounds with true distribution discrepancy were provided for real application under the limited samples. Additionally, a new Student kernel function was proposed to help the VDR to eliminate the interference of abnormal samples, obtaining excellent robustness. The VDR with Student kernel-based model can arrive at an average diagnosis accuracy of over 87 % in six cross-bearing transfer tasks. Lastly, the proposed method is successfully applied to diagnose the faults of actual wind turbine gearbox bearings in virtue of the laboratory bearing datasets. The comprehensive experimental results and analyses prove that the proposed VDR has a stronger transfer diagnosis performance than other well-known distribution alignment metrics. It is noteworthy that VDR can be easily extended into other DA tasks, such as multi-source DA, opened DA, and partial DA.

CRediT authorship contribution statement

Quan Qian: Writing – review & editing, Methodology, Conceptualization. **Huayan Pu:** Validation, Formal analysis. **Tianjia Tu:** Visualization, Data curation. **Yi Qin:** Writing – review & editing, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

The work described in this paper was supported by the National Natural Science Foundation of China (no. 52175075).

References

[1] M. Feng, Z. Hua, G. Qingshan, K. Hon, A novel energy evaluation approach of machining processes based on data analysis, Energy Sources Part A 45 (2023) 4789–4803.
[2] T. Yu, W. Chen, G. Junfeng, H. Poxi, Intelligent detection method of forgings defects detection based on improved EfficientNet and memetic algorithm, IEEE Access 10 (2022) 79553–79563.
[3] J.M. Fordal, P. Schjølberg, H. Helgetun, T.Ø. Skjermo, Y. Wang, C. Wang, Application of sensor data based predictive maintenance and artificial neural networks to enable Industry 4.0, Advances in Manufacturing 11 (2023) 248–263.
[4] Y. Qin, Q. Li, S. Wang, P. Cao, Dynamics modeling of faulty planetary gearboxes by time-varying mesh stiffness excitation of spherical overlapping pittings, Mech. Syst. Sig. Process. 210 (2024) 111162.
[5] C. Han, W. Lu, H. Wang, L. Song, L. Cui, Multistate fault diagnosis strategy for bearings based on an improved convolutional sparse coding with priori periodic filter group, Mech. Syst. Sig. Process. 188 (2023) 109995.
[6] Z. Lei, P. Zhang, Y. Chen, K. Feng, G. Wen, Z. Liu, R. Yan, X. Chen, C. Yang, Prior knowledge-embedded meta-transfer learning for few-shot fault diagnosis under variable operating conditions, Mech. Syst. Sig. Process. 200 (2023) 110491.
[7] Y. Zhang, J. Ji, Z. Ren, Q. Ni, F. Gu, K. Feng, K. Yu, J. Ge, Z. Lei, Z. Liu, Digital twin-driven partial domain adaptation network for intelligent fault diagnosis of rolling bearing, Reliab. Eng. Syst. Saf. 234 (2023) 109186.
[8] J. Sun, X. Gu, J. He, S. Yang, Y. Tu, C. Wu, A robust approach of multi-sensor fusion for fault diagnosis using convolution neural network, Journal of Dynamics, Monitoring and Diagnostics (2022) 103–110.
[9] A.A.T. Anvar, H. Mohammadi, A novel application of deep transfer learning with audio pre-trained models in pump audio fault detection, Comput. Ind. 147 (2023) 103872.
[10] Q. Qian, Y. Wang, T. Zhang, Y. Qin, Maximum mean square discrepancy: A new discrepancy representation metric for mechanical fault transfer diagnosis, Knowl.-Based Syst. 276 (2023) 110748.
[11] Y. Ma, J. Yang, L. Li, Gradient aligned domain generalization with a mutual teaching teacher-student network for intelligent fault diagnosis, Reliab. Eng. Syst. Saf. 239 (2023) 109516.
[12] S. Liu, H. Jiang, Z. Wu, Z. Yi, R. Wang, Intelligent fault diagnosis of rotating machinery using a multi-source domain adaptation network with adversarial discrepancy matching, Reliab. Eng. Syst. Saf. 231 (2023) 109036.
[13] X. Yu, Z. Zhao, X. Zhang, X. Chen, J. Cai, Statistical identification guided open-set domain adaptation in fault diagnosis, Reliab. Eng. Syst. Saf. 232 (2023) 109047.
[14] Y. Xiao, H. Shao, S. Han, Z. Huo, J. Wan, Novel joint transfer network for unsupervised bearing fault diagnosis from simulation domain to experimental domain, IEEE/ASME Trans. Mechatron. 27 (2022) 5254–5263.
[15] Q. Qian, Y. Qin, J. Luo, S. Wang, Partial Transfer Fault Diagnosis by Multiscale Weight-Selection Adversarial Network, IEEE/ASME Trans. Mechatron. 27 (2022) 1539–1548.
[16] T. Han, J. Wang, C. Peng, X. Wu, X. Geng, L. Zhang, M. Jiang, F. Zhang, Novel adaptive loss weighted transfer network for partial domain fault diagnosis, ISA Trans. 145 (2024) 362–372.
[17] P. Panareda Busto, J. Gall, Open set domain adaptation, Proceedings of the IEEE international conference on computer vision, 2017, pp. 754-763.
[18] Q. Qian, Y. Qin, J. Luo, D. Xiao, Cross-machine transfer fault diagnosis by ensemble weighting subdomain adaptation network, IEEE Trans. Ind. Electron. 70 (2023) 12773–12783.
[19] M. Deng, A.-D. Deng, Y. Shi, M. Xu, Correlation regularized conditional adversarial adaptation for multi-target-domain fault diagnosis, IEEE Trans. Ind. Inf. 18 (2022) 8692–8702.
[20] Y. Ding, P. Ding, X. Zhao, Y. Cao, M. Jia, Transfer learning for remaining useful life prediction across operating conditions based on multisource domain adaptation, IEEE/ASME Trans. Mechatron. 27 (2022) 4143–4152.
[21] Q. Qian, J. Luo, Y. Qin, Adaptive Intermediate Class-Wise Distribution Alignment: A Universal Domain Adaptation and Generalization Method for Machine Fault Diagnosis, IEEE Trans. Neural Networks Learn. Syst. (2024).
[22] J. Li, R. Huang, Z. Chen, G. He, K.C. Gryllias, W. Li, Deep continual transfer learning with dynamic weight aggregation for fault diagnosis of industrial streaming data under varying working conditions, Adv. Eng. Inf. 55 (2023) 101883.
[23] M. Long, J. Wang, G. Ding, J. Sun, P.S. Yu, Transfer Feature Learning with Joint Distribution Adaptation, Proceedings of the 2013 IEEE International Conference on Computer Vision, 2013, pp. 2200-2207.
[24] K. Zhao, H. Jiang, K. Wang, Z. Pei, Joint distribution adaptation network with adversarial learning for rolling bearing fault diagnosis, Knowl.-Based Syst. 222 (2021) 106974.
[25] Q. Qian, Y. Qin, J. Luo, Y. Wang, F. Wu, Deep discriminative transfer learning network for cross-machine fault diagnosis, Mech. Syst. Sig. Process. 186 (2023) 109884.
[26] Q. Qian, J. Zhou, Y. Qin, Relationship Transfer Domain Generalization Network for Rotating Machinery Fault Diagnosis Under Different Working Conditions, IEEE Trans. Ind. Inf. 19 (2023) 9898–9908.
[27] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, The, J. Mach. Learn. Res. 13 (2012) 723–773.
[28] B. Sun, J. Feng, K. Saenko, Return of Frustratingly Easy Domain Adaptation, Proceedings of the AAAI conference on artificial intelligence, 2016, pp. 2058-2065.
[29] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, Analysis of representations for domain adaptation, Adv. Neural Inf. Proces. Syst. 19 (2006) 1–8.
[30] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J.W. Vaughan, A theory of learning from different domains, Mach. Learn. 79 (2010) 151–175.
[31] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-Adversarial Training of Neural Networks, J. Mach. Learn. Res. 17 (2017) 2096.
[32] K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3723–3732.
[33] Y. Zhu, F. Zhuang, J. Wang, G. Ke, J. Chen, J. Bian, H. Xiong, Q. He, Deep subdomain adaptation network for image classification, IEEE Trans. Neural Networks Learn. Syst. 32 (2020) 1713–1722.
[34] C. Chen, Z. Fu, Z. Chen, S. Jin, Z. Cheng, X. Jin, X.-S. Hua, Homm, Higher-order moment matching for unsupervised domain adaptation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 3422–3429.
[35] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, S. Saminger-Platz, Central moment discrepancy (cmd) for domain-invariant representation learning, arXiv preprint arXiv:.08811, (2017).

[36] Y. Feng, J. Chen, S. He, T. Pan, Z. Zhou, Globally localized multisource domain adaptation for cross-domain fault diagnosis with category shift, IEEE Trans. Neural Networks Learn. Syst. 34 (2023) 3082–3096.
[37] M. Reed, Methods of modern mathematical physics: Functional analysis, Elsevier, 2012.
[38] J. Mercer, functions of positive and negative type, and their connection the theory of integral equations, Philosophical transactions of the royal society of London, Series a, Containing Papers of a Mathematical or Physical Character 209 (1909) 415–446.
[39] J. Hammersley, A generalization of McDiarmid's theorem for mixed Bernoulli percolation, Mathematical Proceedings of the Cambridge Philosophical Society, Cambridge University Press, 1980, pp. 167-170.
[40] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Am. Stat. Assoc. 58 (1963) 13–30.
[41] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[42] C.A. Micchelli, Algebraic aspects of interpolation, Proc. Symp. Appl. Math. (1986) 81–102.
[43] K.A. Loparo, Bearings vibration data set case western reserve university, Available at https://engineering.case.edu/bearingdatacenter.
[44] K. Li, X. Ping, H. Wang, P. Chen, Y.J.S. Cao, Sequential Fuzzy Diagnosis Method for Motor Roller Bearing in Variable Operating Conditions Based on Vibration Analysis 13 (2013) 8013–8041.
[45] S. Shao, S. McAleer, R. Yan, P. Baldi, Highly accurate machine fault diagnosis using deep transfer learning, IEEE Trans. Ind. Inf. 15 (2018) 2446–2455.
[46] B. Sun, K. Saenko, Deep CORAL: Correlation Alignment for Deep Domain Adaptation, European Conference on Computer Vision (2016) 443–450.
[47] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell, Deep domain confusion: Maximizing for domain invariance, arXiv preprint arXiv:1412.3474, (2014).