

Fault Diagnosis of Gearbox with a Transferable Deep Neural Network

Greg Smith^{1,*}, Masayoshi Shibatani²

¹ Department of computer engineering, Mount San Antonio College, Walnut, CA, USA

² Department of Mechanical Engineering, Mie University, Mie, Japan

Abstract— In the past years, various intelligent machine learning and deep learning algorithms have been developed and widely applied for gearbox fault detection and diagnosis. However, the real-time application of these intelligent algorithms has been limited, mainly due to the fact that the model developed using data from one machine or one operating condition has serious diagnosis performance degradation when applied to another machine or the same machine with a different operating condition. The reason for poor model generalization is the distribution discrepancy between the training and testing data. This paper proposes to address this issue using a deep learning based cross domain adaptation approach for gearbox fault diagnosis. Labelled data from training dataset and unlabeled data from testing dataset is used to achieve the cross-domain adaptation task. A deep convolutional neural network (CNN) is used as the main architecture. Maximum mean discrepancy is used as a measure to minimize the distribution distance between the labelled training data and unlabeled testing data. The study proposes to reduce the discrepancy between the two domains in multiple layers of the designed CNN to adapt the learned representations from the training data to be applied in the testing data. The proposed approach is evaluated using experimental data from a gearbox under significant speed variation and multiple health conditions. An appropriate benchmarking with both traditional machine learning methods and other domain adaptation methods demonstrates the superiority of the proposed method.

Index Terms—Domain adaptation; Deep learning; Maximum mean discrepancy; Gearbox; Fault Detection

I. INTRODUCTION

GEARS are critical mechanical components in modern world industrial mechanical systems. They have widespread use in wind generation, aerospace, machining and other engineering machinery [1]. However, rugged working environments and harsh operating conditions can lead to mechanical shutdown, considerable economical loses and even casualties. Additionally, the closed operating environments make it hard to identify failures in time. Therefore, due to these requirements, strategies for fault detection and diagnosis of

gears, have been widely researched to ensure operational safety of systems.

Various intelligent algorithms have been proposed for gearbox fault diagnosis in the past decade, such as K-Nearest neighbor [2], SOM [3], Artificial Neural Networks [4], Support vector machine [5] etc. Most of these classical frameworks include four steps: a) data preprocessing and segmentation, b) feature extraction, c) feature selection, and d) model building and validation [1]. However, following this procedure is time-consuming and requires an in-depth knowledge of the system and characteristics of the faults. Recently, deep learning based fault detection methods have drawn attention due to their high-level feature representation, higher classification accuracy and eliminating the need for domain knowledge [6-7]. In general, both the classical algorithms and deep learning algorithms require supervised retraining as the operational condition changes significantly. During the training phase, these classification algorithms use characteristic fault features extracted from raw time domain measurement signals as an input and map them to pre-defined labels (such as pitting, eccentricity, abrasion) as an output. However, the basic assumption for their implementation is that the data distribution of the training and testing data has to be roughly the same [8]. However, in practical real world situations, gears are expected to operate under variable operating conditions owing to different load/speed requirements. Under such circumstances, the diagnostic/classification models (source classifier) trained on features extracted from one operating condition (source/training data) may not work reliably for the same set of features extracted from the same gearbox operating at a different load/speed condition (target/testing data). The unreliability corresponds to the wrong labelling of fault features in testing data. This inherent shortcoming of the source classifier model of not being applicable to the target data is due to a phenomenon of domain shift [9]. The concept of domain shift is illustrated in Fig. 1 and corresponds to the shift in the distribution of features.

The above mentioned inherent limitation of the machine learning methods i.e. domain shift, allows the trained model to perform reliable fault diagnosis on testing data under similar

* Corresponding Author:

Email address: gsmith23@student.mtsac.edu

operating condition. Since in practical situations the working/operating conditions may change over time, the trained model is not able to adapt itself to the new condition. Training a model from scratch for every new condition or retraining a model after maintenance operations is not only time consuming, but requires a large amount of training data. Though, highly desirable, collecting such huge amount of labelled data (especially faulty data) may not be feasible from industrial equipment to make the built models adaptable to every possible condition. Therefore, to deal with this problem of domain shift, recently domain adaptation techniques have attracted much attention for addressing these problems. Since, source and target domain data distributions are closely related but are shifted due to different statistics, domain adaptation techniques attempt to transfer knowledge from well-labeled source domain data to unlabeled target domain data.

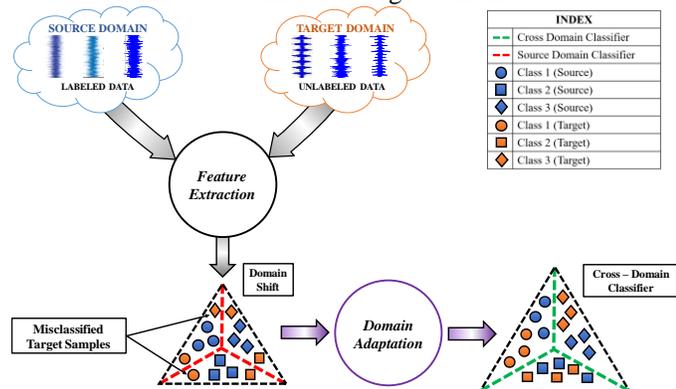


Figure 1 Illustrative example of Domain shift and the domain adaptation technique

Domain adaptation or transfer learning techniques enable intelligent models built on labelled data under one operating condition to be able to classify unlabeled data from another operating condition. In general, these approaches can be classified as: (a) instance-based, (b) parameter-based, (c) relation-based and (d) feature-based TL [10]. These techniques have aroused large amounts of interest and research in recent years in the area of image classification, natural language processing, object recognition, and feature learning [11-13]. Recently, deep transfer learning techniques have attracted considerable attention from researchers, since deep learning methods could capture more hidden knowledge in the process of feature extraction in hierarchical structures and have well data adaptability in domain adaptation tasks. Deep transfer learning methods possess strong capabilities in domain-invariant feature learning and offer flexibility in the integration of the distribution differences in multi-domains. Hence, a number of techniques based on deep transfer learning have been reported for fault diagnosis of rotating machinery components [14-18]. The principle idea is to obtain a feature representation between the labelled source domain data and the semi-labeled/unlabeled target domain data and attempt to reduce the distribution discrepancy between different both domains using a distance metric.

Guo et al. [14] proposed a 1-D-deep convolutional transfer learning method to extract domain-invariant features from raw

data and used domain adaptation to maximize the domain recognition error and minimize the probability distribution distance. The effectiveness of their method was validated on transfer tasks between three different bearing datasets: CWRU [19], IMS [20], and Railway Locomotive [21]. Lu et al. [15] used Deep Neural Network for extracting features from the spectrum instead of the raw data and only used labelled source domain data and normal category data from the target domain to accomplish DA tasks. Authors evaluated their method on bearing and gearbox datasets [22], demonstrating potential of the proposed method in solving DA problems in the fault detection and diagnosis field. Wen et al. [16] proposed a DA architecture based on a three-layer sparse auto-encoder for automatic feature learning from raw vibration signal and minimizing the maximum mean discrepancy (MMD) term between the features from training and testing domains. The proposed DA approach was evaluated on CWRU bearing dataset and reported a higher prediction accuracy compared to other methods. In [17], representative images of time-domain acoustic emission signals were used in a CNN based transfer learning model for predicting labels of target domain dataset in low-speed bearing applications. Cao et al. [18] converted raw time domain vibration signal into gray scaled images and used them as an input to the CNN model to develop a deep CNN based transfer learning method for gearbox fault diagnosis. In their approach, the first 'n' layers of the CNN network are trained on the source dataset and fine-tuned on the target transfer tasks. Meanwhile the last (m-n) layers are trained using data from the new task.

In view of the above discussion, to the best of the authors' knowledge there is no study on the effectiveness of CNN-based cross domain adaptation for gearbox fault diagnosis under significant variations in operating speed using vibration measurement signals. Hence, this paper proposes a 1-D deep cross-domain adaptation technique based on CNN for gearbox fault diagnosis. CNN would be adopted to automatically learn rich knowledge from the raw time-domain vibration data without any preprocessing of the source/target domain for reliable domain adaptation. The presented method attempts to reduce the domain discrepancy by minimizing the difference between features distributions from labelled source domain and unlabeled target domain using multi-kernel MMD within multiple layers of the CNN network. The main contributions of this work can be summarized as follows:

- To improve upon the generalization ability for gearbox fault diagnosis, a unified cross-domain adaptation approach is proposed to classify unlabeled target domain data by simultaneously minimizing cross-entropy loss and multi-layered multi-kernel MMD (obtained from CNN) loss between labelled source and unlabeled target datasets. It is demonstrated that it is not only the fully connected layer (last layer) but in fact multiple layers that contribute to the domain biases through the CNN network. Attempt has been made to demonstrate the necessity of the application of minimizing distribution discrepancy in multiple CNN layers using MMD.
- The proposed method is validated on experimental data

consisting of raw vibration measurement signals from five different gear health conditions. The transfer tasks are conducted with regards to significant variation in the

$\{x_{Ti}\}$ represents the target domain with unlabeled testing data; where x_{Ti} represents a data sample in χ_T . It must be noted that in this study $\chi_S \neq \chi_T$ and $P(X_S) \neq P(X_T)$ and thus the central

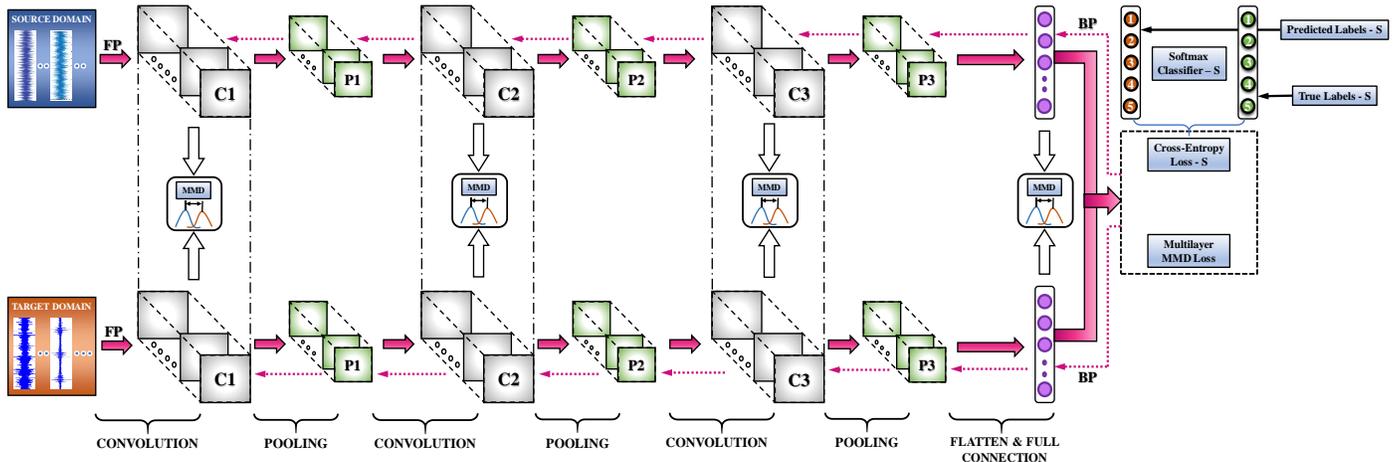


Figure 2 Proposed CNN-CDA model framework

operating speed of the gearbox.

- The obtained results are compared and benchmarked against conventional machine learning methods, traditional shallow transfer learning methods, CNN without domain adaptation and CNN with domain adaptation using only features from the last layer of the network.
- Network architecture and tuning parameters are investigated to provide high classification accuracy in gearbox fault diagnosis under significant speed variations.

The rest of the paper is organized as follows. Section II gives details of some preliminaries. The domain adaptive CNN-MMD model is detailed in Section III. In Section IV, details are provided corresponding to the experimental test rig and experimental plan. Section V presents the results and discussion. Section VI presents conclusions and future work.

II. PRELIMINARIES

A. Domain Shift Problem

Fig. 1 depicts the generalization performance of classifiers in practical real world industrial applications where the performance of the source domain classifier is poor as the training samples (source domain) used for building the model cannot be generalized well to the testing samples (target domain). In other words, a classifier trained on the source domain data would not be directly applied to the target domain. This happens because the data from the source and target domains follow different data distributions. This phenomenon is known as domain shift problem.

To explicitly describe the problem to be addressed, let's introduce a basic mathematical notation for the cross-domain adaptation task. Let χ be a feature space, X be a particular sample, and $P(X)$ be a marginal probability distribution. $D = \{\chi, P(X)\}$ be a domain, where $X = \{x_1, x_2, \dots, x_n\} \in \chi$ and x_i is the i^{th} feature term. $D_S = \{x_{Si}, y_{Si}\}$ represents the source domain with labelled training data; $x_{Si} \in \chi_S$ represents a data sample and $y_{Si} \in Y_S$ represents the corresponding label of x_{Si} . $D_T =$

idea is on adapting a deep learning model trained on D_S (with labelled fault information) to a data belonging to D_T (with unlabeled fault information). After the domain adaptation, we expect that the model trained on D_S would be able to recognize health conditions from data samples belonging to D_T . This study follows some basic assumptions:

- The number and type of health labels are same in both the source domain and target domain, i.e. $Y_S = Y_T$.
- While model training phase, labeled samples are available from source domain, however only unlabeled samples are available from target domain
- The source and target domains are related to each other and only differ in their respective data probability distributions.

B. Convolutional Neural Network

The architectural framework of CNN is depicted in Fig. 2. In general, the architecture majorly consists of three layers: 1) convolutional layer; 2) pooling layer; and 3) fully-connected (FC) layer. The convolutional layer contains a number of filters which are convolved with the raw data. The main function of the convolutional layer is to identify important features or feature maps (set of weights) from the given set of input data. The idea behind using a number of filters is that different filters would detect/identify a different set of features when convolved with the input data. Assuming $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$, where N denotes the length of a sequential signal input, the convolution operation can be interpreted as a multiplication operation between a filter kernel $w, w \in R^N$, and a concatenation vector representation $x_{i:i+D-1}$, which is given as:

$$x_{i:i+D-1} = x_i \oplus x_{i+1} \oplus \dots \oplus x_{i+D-1} \quad (1)$$

where $x_{i:i+D-1}$ represents window of D length sequential signal starting from index i to index $i+D-1$ and \oplus concatenates each data samples into longer embedding. The final convolution operation becomes:

$$z_i = \phi(w^T x_{i:i+D-1} + b) \quad (2)$$

where ϕ is the activation function, b is the bias term and z_i is the learned feature of the filter kernel w . Hence, feature map of the j^{th} filter can be denoted as

$$z_j = [z_j^1, z_j^2, \dots, z_j^{N-D+1}] \quad (3)$$

with corresponding lengths of $x_{1:D}, x_{2:D+1}, \dots, x_{N-D+1:N}$.

The extracted features are then passed on to the next layer. i.e. the pooling layer (usually follows a convolution layer) which attempts to reduce the spatial size of the representation from the first layer. It can be seen as a down-sampling layer (lower resolution representation) which reduces the feature dimensions of the input. Max-pooling function is applied with pooling length g as:

$$h_j = [h_j^1, h_j^2, \dots, h_j^s] \quad (4)$$

$$h_j^t = \max(z_j^{(t-1)g+1}, z_j^{(t-1)g+2}, \dots, z_j^{tg}) \quad (5)$$

where h_j is a s -dimensional vector, which is the output of pooling layer applied to the j^{th} feature map. After several alternating convolutional and pooling layers, the fully-connected layers are followed by one or several fully-connected layers. Finally, the result of fully-connected layers is input to SoftMax or Sigmoid function classifier.

C. Maximum Mean Discrepancy

Maximum Mean discrepancy (MMD) is a measure of the difference between two probability distributions from their samples. In this paper, MMD is employed to measure the discrepancy of labelled samples according to the categories. Given two probability distributions P and Q on \mathcal{X} , MMD is defined as [23]:

$$MMD(\mathcal{F}, P, Q) = \sup_{f \in \mathcal{F}} (E_{x_1 \sim P}[f(x_1)] - E_{x_2 \sim Q}[f(x_2)]) \quad (6)$$

where \mathcal{F} is a class function $f: \mathcal{X} \rightarrow \mathcal{H}$. \mathcal{H} denote Reproducing Kernel Hilbert Space (RKHS). $\mathbf{X}^1 = \{x_i^{(1)}\}_{i=1}^{n_1}$ and $\mathbf{X}^2 = \{x_i^{(2)}\}_{i=1}^{n_2}$ are data vectors drawn from data space D_1 and D_2 , respectively. Based on the fact that f is in the unit ball in a universal RKHS, Eq. (1) can be rewritten as:

$$D(\mathbf{X}^1, \mathbf{X}^2) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \varphi(x_i^{(1)}) - \frac{1}{n_2} \sum_{i=1}^{n_2} \varphi(x_i^{(2)}) \right\|_{\mathcal{H}} \quad (7)$$

where $\varphi(\cdot): \mathcal{X} \rightarrow \mathcal{H}$ is referred to the feature space map. Practical evaluation of MMD is done by employing the kernel method, which aids in evaluating the distance between the distributions of high level learned features between different domains via:

$$D_k(\mathbf{X}^1, \mathbf{X}^2) = \left[\frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \langle \varphi(x_i^{(1)}), \varphi(x_j^{(1)}) \rangle + \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \langle \varphi(x_i^{(2)}), \varphi(x_j^{(2)}) \rangle + \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \langle \varphi(x_i^{(1)}), \varphi(x_j^{(2)}) \rangle \right]^{\frac{1}{2}} \quad (8)$$

$$= \left[\frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} k(x_i^{(1)}, x_j^{(1)}) + \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} k(x_i^{(2)}, x_j^{(2)}) + \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} k(x_i^{(1)}, x_j^{(2)}) \right]^{\frac{1}{2}}$$

where $D_k(\mathbf{X}^1, \mathbf{X}^2)$ is the unbiased estimation of $D_k(\mathbf{X}^1, \mathbf{X}^2)$. $k(\cdot, \cdot)$ is a kernel function to compute the inner product in a higher-dimensional space, i.e., $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$.

Though some other researches in the literature have also adopted MMD for domain adaptation tasks, but these studies attempt to minimize the distribution discrepancy only in the fully connected layer of CNN. However, studies in [24] pointed out that only fully connected layer MMD evaluation may not be sufficient to minimize the distribution discrepancy between the source and target domains. Other layers of the CNN architecture are also equally susceptible to domain shift as well [25]. Therefore, for more effective cross domain adaptation we consider both the representation layers (convolutional layers of the CNN architecture) and the classifier (fully connected layer) for multi-kernel MMD evaluation. Thus, collective distribution discrepancy loss is thus defined as,

$$L_{MMD} = \sum_{i \in L} \sum_{k \in K} MMD(A^i, B^i) \quad (9)$$

where K denotes the kernel set, A^i and B^i denote the output of the i^{th} layer in the CNN architecture for the source and target samples respectively. $MMD(A^i, B^i)$ represents the multi-kernel MMD evaluated between the source and target domains i^{th} layer in the CNN.

III. PROPOSED METHOD

This section presents the details corresponding to the proposed methodology. Fig. 3 describes the individual steps to conduct cross domain fault diagnosis in three phases:

- In the first phase, raw time-domain vibration data is collected from the gearbox corresponding to all health labels. The data is partitioned into two sets (a) source domain data (labels are available for the given health conditions) and (b) target domain data (no labels are available for the given health conditions). The target domain data is further partitioned into two sets, where one half is used in training the CNN algorithm and the other half is used as a testing data to test the trained CNN algorithm.
- In the second phase, a CNN architecture is constructed and both the source domain and target domain data is passed through the designed network to extract high level feature representations. At this point, two losses are evaluated per epoch of the CNN architecture training phase: (a) cross entropy loss, which is evaluated at the output of the Softmax layer for the labelled source domain target data and (b) MMD loss. MMD loss is evaluated between the outputs of the corresponding convolutional layers and the fully connected layer of the source and target domain data. The standard Softmax regression loss can be described as follows:

$$L_{classification}(\theta, x_s, y_s) = \frac{1}{n_s} \sum_{i=1}^{n_s} \sum_{k=1}^K 1\{y_{si} = k\} \log(y_{sik}) \quad (10)$$

where, $1\{\cdot\}$ is a binary indicator function that detects whether the i th training pattern returns 1 if the condition is true, and 0 if the condition is false; indicating whether it belongs to the k th category. K is the number of categories, $\theta = [\theta_1, \theta_2, \dots, \theta_K]^T$ denotes the Softmax model parameters and y_{sik} is the predicted output probability distribution for i th observation belonging to class k .

By integration the Categorical Cross Entropy Loss (Eq. (10)) with Distribution Discrepancy Loss (Eq. (9)), the overall objective function becomes:

$$\min L = \alpha L_{MMD} + \beta L_{classification} \quad (11)$$

where α and β are the regularization parameters. The default value of α is '1' and the value of β is '0' for

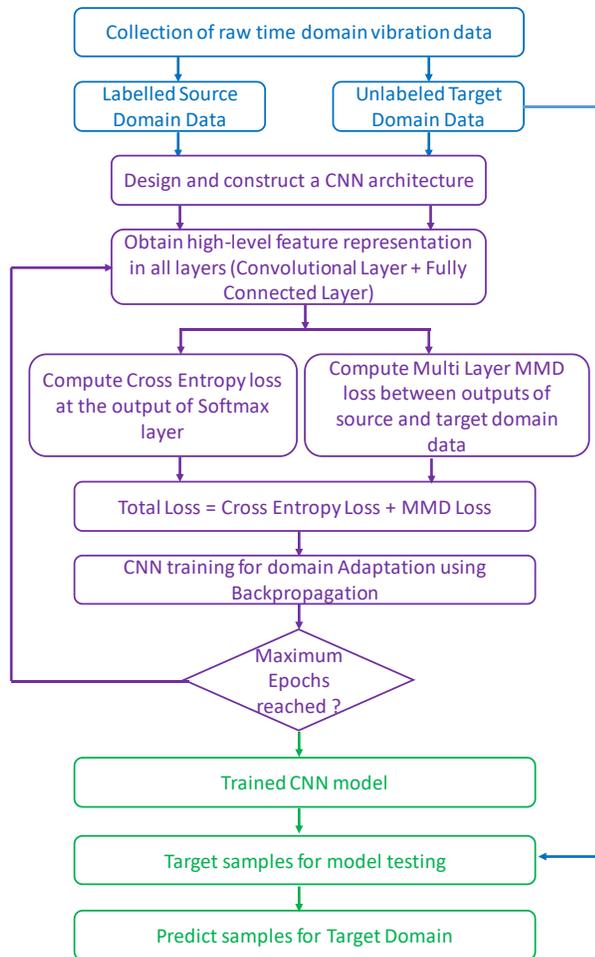


Figure 3 The proposed framework for CNN based cross domain adaptation

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Test Bench

The experimental platform (Gearbox prognostic simulator (GPS) from Spectraquest [26]) is shown in Fig. 3. The experimental platform consists of two confronted electrical

motors (10 Hp, three-phase, induction motors having two pairs of poles) where one acts as drive motor and the other as load motor. The position of accelerometers and scheme of the gear disposition inside the monitored gearbox is depicted in Fig. 4. The monitored gearbox is composed of four spur gears and two PCB accelerometers (Model 60811A11 Industrial ICP) installed on the gearbox capture vibration in X and Y direction. Data was recorded using a computer with a National Instruments acquisition card (NI 4472 series) at a sampling rate of 50kS/sec.

B. Dataset Description

The maximum speed that the GPS test bench can reach is 1500 rpm. The minimum selected speed is 500 rpm. One more intermediary speed, 1000 rpm was selected for the domain adaptation task. The focus is set on analyzing different speeds especially in the no-load condition. Faulty gears were inserted in the monitored gearbox of the GPS. During the experiments, only one gear is the subject of the test, the one inserted in the position of the gear 1 of the first gearbox (the 32 teeth gear in Fig. 3).

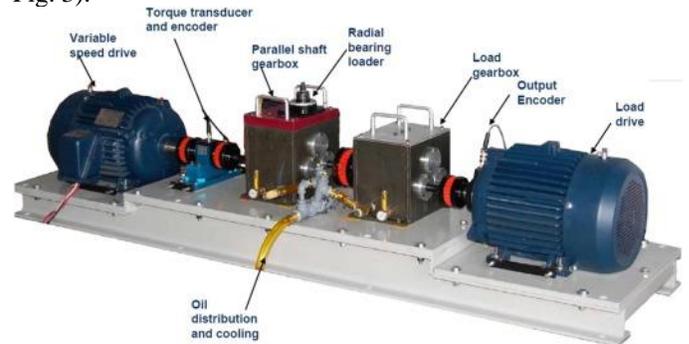


Figure 4 Detail of the gear prognostics simulator test rig [26].

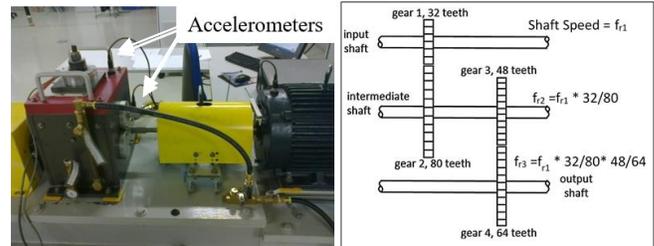


Figure 5 (a) Position of Accelerometers (b) Scheme of the gear disposition inside the monitored gearbox.

The detailed information of the dataset and the transfer tasks is presented in Tables 1. N_{source} and N_{target} samples for each health condition under one speed are selected as the source and target domain data, respectively. In this paper, we assume $N_{source} = N_{target}$ for simplicity. The raw data sequence is divided into N_{input} sequential points (by default, $N_{input} = 1000$ is used in this paper).

All reported experimental results are averaged by 10 trials to reduce the effect of randomness. Our implementation is carried out in the Tensorflow platform in Windows operating system on Intel(R) Xeon(R) W-2145 CPU @ 3.7 GHz Processor running at 32.0 GB RAM and GPU parallel computing

(NVIDIA GeForce TX 2080 Ti) is employed to accelerate the computing.

TABLE I
TRANSFER TASKS CONDUCTED IN THIS PAPER

Transfer task	Source Domain (Speed)	Target Domain (Speed)
T ₅₀₀₋₁₀₀₀	500	1000
T ₅₀₀₋₁₅₀₀	500	1500
T ₁₀₀₀₋₅₀₀	1000	500
T ₁₀₀₀₋₁₅₀₀	1000	1500
T ₁₅₀₀₋₅₀₀	1500	500
T ₁₅₀₀₋₁₀₀₀	1500	1000

V. RESULTS AND DISCUSSION

A. Network architecture

Fig. 2 shows the structure of the proposed CNN network for gearbox fault diagnosis. In the considered CNN architecture, three stacked 1D convolutional layers are designed for feature extraction. For convenience, all layers are supposed to share the same configuration. Zero-padding operation is implemented to keep the feature map dimension from changing. In order to decrease the data dimension while keeping significant spatial information, each convolutional layer is followed by a max-pooling layer. Further, batch normalization is used after each convolutional layer. At last, after passing the data through the three stacked 1-D convolution pooling operations, the obtained feature representations are flattened and passed to a FC layer. To avoid overfitting, Dropout layer with rate of 0.5 is used. In the last stage, a Softmax layer is adopted to predict the various gearbox health conditions. In this design of deep architecture, rectified linear units (ReLU) activation functions have been adopted to avoid problems of vanishing gradients in the training process.

Due to the importance of network structure, we first investigate the effect of a number of filters and size of kernels in each convolutional layer with variability in the terms of input size of source domain and target domain data on the overall cross domain testing accuracy. The transfer task T₁₀₀₀₋₁₅₀₀ is used for illustration and the results are depicted in Fig. 6(a). In general, larger N_{input} is generally expected to provide higher diagnosis performance and the same is validated by the results of Fig. 6(a). Further, as show in Fig. 6(b) larger filter size and a large number of convolutional filters in each layer improve the learning capability of CNN; increasing these parameters provides significant improvements in testing accuracies. Overall, higher values of number of filters and size of kernels in each convolutional layer and higher N_{input} improves network performance.

Batch size is another tuning parameter that along with the number of tuning parameters that may have a significant effect the diagnosis accuracy. Figure 6(b) shows the achieved results for different tuning parameters. For the considered dataset, low batch size leads to lower diagnosis performance and the variance in testing accuracy is also high. Further, a low number of neurons in the Fully connected layer negatively impacts the prediction results. Consequently, based on the results of Fig. 6, the following parameters were selected for the final diagnosis model; $N_{input}=1000$, the number of filters in each convolutional layer =30, size of kernels in each convolutional layer =30, batch

size =16 and number of neurons in fully connected layer =512. Based on the final CNN model parameters, the confusion matrix classification accuracy for task T₁₀₀₀₋₁₅₀₀ is shown in Figure 7. It can be observed that two classes ‘eccentricity’ and ‘pitting’ are misclassified and all other classes are correctly classified.

B. Domain Adaptation Results

In order to present a complete evaluation of the proposed method and to demonstrate its effectiveness and superiority, the obtained results are compared with the following approaches:

- Traditional machine learning approaches: SVM, kNN and LDA methods have been considered for comparison purposes. Both time domain and frequency domain features are extracted from the raw time domain vibration data from the both the source domain and target domain respectively. Machine learning models are trained on source domain dataset and target domain data is directly used as testing data.
- Other transfer learning approaches: Both new and well explored feature based transfer learning approaches, JDA[27], BDA[28], EasyTL[29], GFK[30] have been considered for benchmarking purposes.

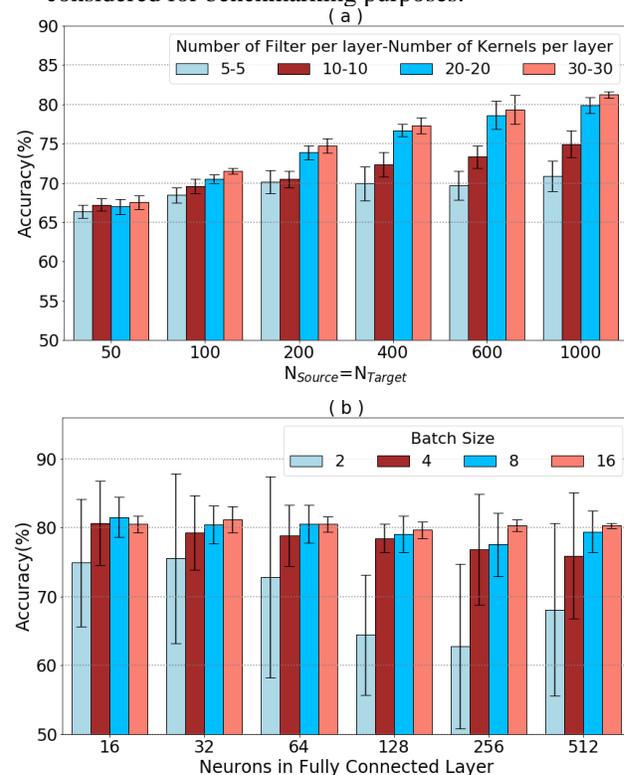


Figure 6 Impact of different CNN architectures on testing accuracy for task T₁₀₀₀₋₁₅₀₀

		Confusion Matrix						
Output Class		Healthy	Eccentric	Pitting	Chipped Tooth	Missing Tooth		
		Healthy	400 20.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100%
Eccentric	0 0.0%	59 2.9%	21 1.1%	0 0.0%	0 0.0%	73.8%	26.2%	
Pitting	0 0.0%	341 17.1%	379 18.9%	0 0.0%	0 0.0%	52.6%	47.4%	
Chipped Tooth	0 0.0%	0 0.0%	0 0.0%	400 20.0%	0 0.0%	100%	0.0%	
Missing Tooth	0 0.0%	0 0.0%	0 0.0%	0 0.0%	400 20.0%	100%	0.0%	
		100% 0.0%	14.8% 85.3%	94.8% 5.2%	100% 0.0%	100% 0.0%	81.9% 18.1%	
		Healthy	Eccentric	Pitting	Chipped Tooth	Missing Tooth		
		Target Class						

Figure 7 The performance of the final model for task $T_{1000-1500}$

- CNN without domain adaptation: CNN based feature extraction without cross domain adaptation (only the classification loss in Eq. (11) is considered) i.e. CNN model trained on source domain data is directly used for testing (Without DA).
- Single layer MMD based CNN with domain adaptation: CNN based feature extraction with cross domain adaptation in just fully connected layer.

Results from all 6 transfer tasks, i.e. $T_{500-1000}$, $T_{1000-500}$, $T_{1000-1500}$, $T_{1500-1000}$, $T_{500-1500}$ and $T_{1500-500}$, are presented in Fig. 8.

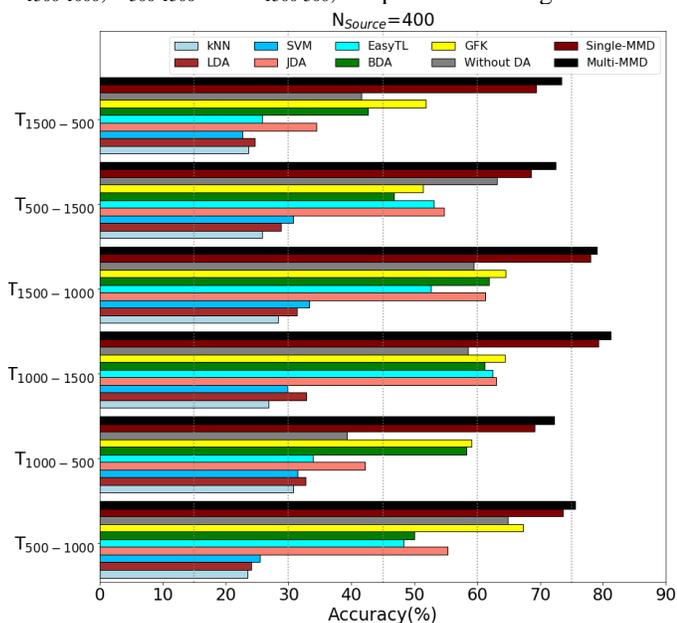


Figure 8 Analysis of the six transfer tasks using different domain adaptation methods. $N_{Source} = 400$ is used.

It can be observed that the models trained on traditional machine learning models (SVM, kNN, LDA) using source domain data when directly applied to target domain data depict poor generalization. This happens because of the dissimilarity in the feature distributions of source and target domain datasets. Further, it can be observed that all representative methods for

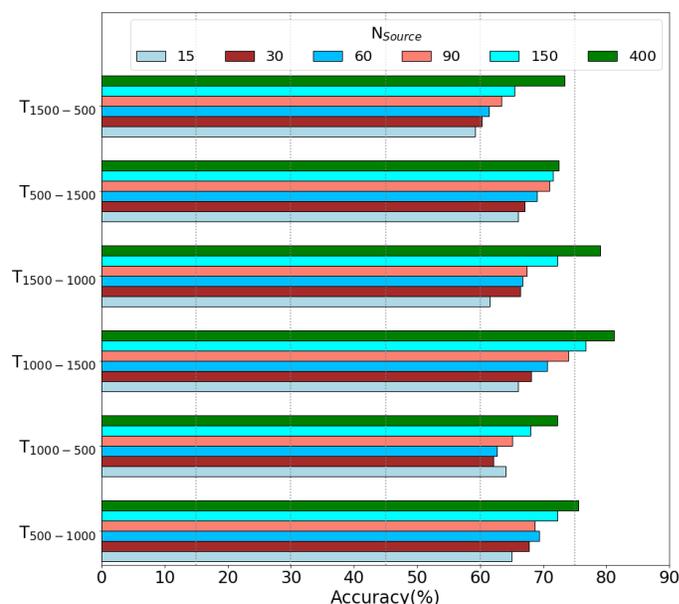


Figure 9 Analysis of the six transfer tasks using increasing amounts of labelled source domain data.

domain adaptation techniques perform significantly better than the traditional machine learning algorithms. In some tasks, the Without DA method achieves reasonably well testing diagnosis results. For example, the Without DA method performs better than EasyTL approach in most transfer tasks. Single layer MMD based CNN with domain adaptation approach demonstrates larger cross-domain diagnosis improvements in comparison to the traditional machine learning methods and other transfer learning approaches. In contrast, the proposed multi-layered MMD evaluation based CNN cross domain adaptation methodology achieves the highest testing diagnosis performance among all techniques in different tasks. In addition, the results of the proposed methodology are almost bidirectional in nature. It implies, that the testing accuracies when T_{500} is transferred to T_{1000} or when T_{1000} is transferred to T_{500} are almost close by. It must be noted that the higher testing accuracies in tasks, such as $T_{1000-1500}$, $T_{1500-1000}$ indicates the inherent closeness by nature of the respective source and target domains.

Fig. 9 depicts the results for a lower amount of the labeled source domain data used for training for all tasks. The cross-domain diagnosis performance is observed to be showing improvements with an increase in the amount of data. The findings are consistent with the general notion about deep learning methodologies that a higher amount of samples used in training phase typically lead to improvement in network performance. However, a comparison between other benchmarked approaches (Fig. 8 with $N_{source} = 400$) and results from Fig. 9 for $N_{source} > 60$ highlights that the proposed multi layered MMD evaluation based CNN cross domain adaptation methodology performs better even with low number of labelled source domain samples. The results from Fig. 9 indicate the ability of the proposed method to deal with overfitting problem; a common issue in situations of insufficient training data. Therefore, the proposed method can be extended to other real-world industrial applications which precisely have low availability of labeled data in practice.

C. Visualization of learned representation

This section presents visualization of high-dimensional features in various layers of the designed CNN architecture. t-SNE algorithm was used for visualizing the high-dimensional data representation. Samples are mapped from the original feature space into a 2-D space map. Fig. 10 depicts the 2D plots of learned representations in the FC layer using the proposed method and ‘Without-DA’ approach for both domain (source and target) for task $T_{1000-1500}$. Plots on the left side of Fig. 10 (Without Domain Adaptation) clearly highlight, that generally, distribution discrepancy between domains exists in all the layers in the network. The learned feature representations from the same health conditions cluster together, but they do so independently in both domains. This scenario necessitates the proposed idea of domain adaptation.

In contrast, the features with cross domain adaptation are well separated for three health conditions (Healthy, missing tooth and chipped tooth) as they cluster together samples from both domains practically overlap each other. For these 3 conditions (Healthy, missing tooth and chipped tooth), the data samples in both domains that belong to the same health class group and overlap in the region of the feature space. Further, this phenomenon is more pronounced in the higher layers than the lower ones. This is because lower layers extract generic features and the higher layers extract level abstract task-specific features. Therefore, the clustering and overlapping of the source domain and target domain datasets improve with each successive layer.

I. CONCLUSIONS

This paper proposes a deep learning based gearbox fault detection and diagnosis approach based on domain adaptation. The distribution discrepancy between labelled source domain and unlabeled target domain data is minimized based on the collective minimization of the cross entropy loss and multi kernel maximum mean discrepancy (MMD) measure. The proposed approach is evaluated on experimental data from a gearbox test rig under significant speed variations and the influence of network architecture and tuning parameters of the CNN model are comprehensively investigated to evaluate the performance of gearbox fault diagnosis. It is demonstrated that the proposed methodology of integrating MMD loss from multiple layers further improves the domain adaptation performance, and outperforms the methodology with single layer MMD and other compared approaches as well. Further, the cross-domain diagnosis performance by the proposed method are still significant when a smaller labeled training dataset is used, thus highlighting the ability of the proposed method to deal with overfitting problem which is a common issue encountered while implementing intelligent algorithms in situations of insufficient training data.

In the current approach, the authors assume of having a balanced training dataset, which might not be the case in actual industrial applications. Thus, in future studies the authors will focus on extending the use of the proposed methodology to imbalanced training dataset.

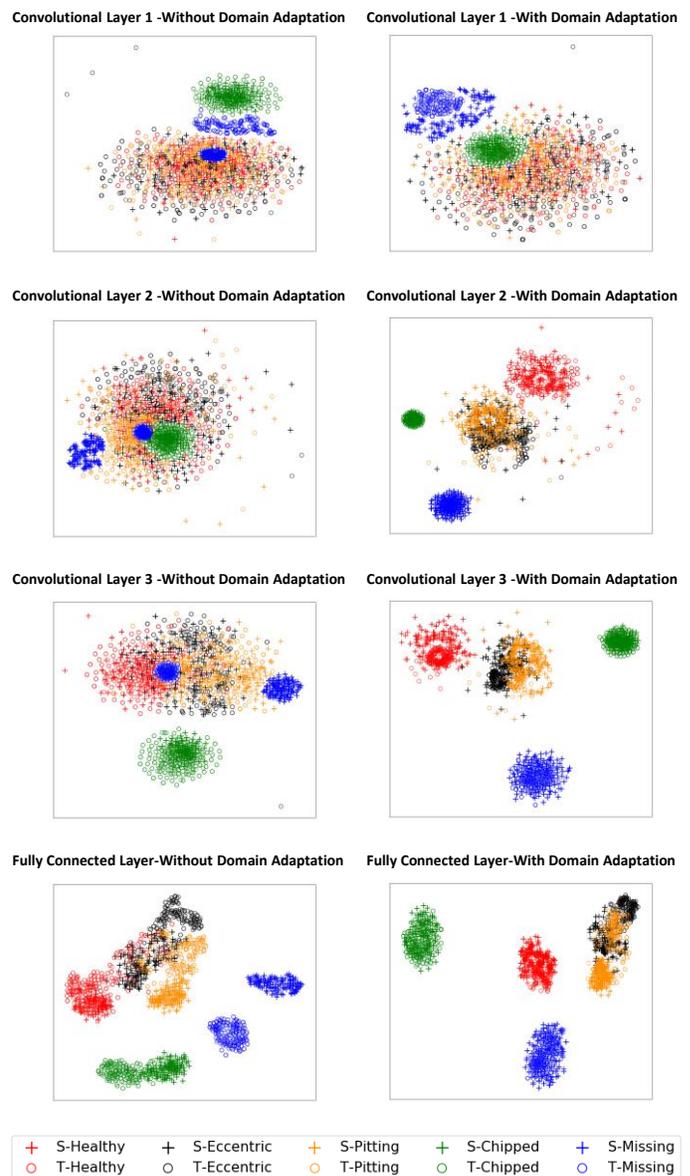


Figure 10 Feature visualization in the various layers for $T_{1000-1500}$ task.

References

- [1] Lee J, Azamfar M, Singh J. A blockchain enabled Cyber-Physical System architecture for Industry 4.0 manufacturing systems. *Manufacturing Letters*. 2019 Apr 1;20:34-9.
- [2] D.Wang, "K-nearest neighbors based methods for identification of different gear crack levels under different motor speeds and loads: Revisited," *Mech. Syst. Signal Process.*, vols. 70/71, pp. 201–208, 2016.
- [3] G. Cheng, Y.–L. Cheng, L.-H. Shen, J.-B. Qiu, and S. Zhang, "Gear fault identification based on Hilbert–Huang transform and SOM neural network," *Measurement*, vol. 46, no. 3, pp. 1137–1146, 2013.
- [4] Azamfar M, Singh J, Bravo-Imaz I, Lee J. Multisensor data fusion for gearbox fault diagnosis using 2-D convolutional neural network and motor current signature analysis. *Mechanical Systems and Signal Processing*. 2020 Oct 1;144:106861.
- [5] F. Chen, B. Tang, R. Chen, A novel fault diagnosis model for gearbox based on wavelet support vector machine with immune genetic algorithm, *Measurement*. 46 (2013) 220–232.
- [6] Singh J, Azamfar M, Li F, Lee J. A systematic review of machine learning algorithms for PHM of rolling element bearings:

- fundamentals, concepts, and applications. *Measurement Science and Technology*. 2020 Apr 28.
- [7] F. Cheng, J. Wang, L. Qu, W. Qiao, "Rotor-current-based fault diagnosis for DFIG wind turbine drivetrain gearboxes using frequency analysis and a deep classifier", *IEEE Trans. Ind. Appl.*, vol. 54, no. 2, pp. 1062-1071, Mar./Apr. 2018
 - [8] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2005.
 - [9] Azamfar M, Li X, Lee J. Intelligent ball screw fault diagnosis using a deep domain adaptation methodology. *Mechanism and Machine Theory*. 2020 Sep 1;151:103932.
 - [10] S. J. Pan, Q. Yang, "A survey on transfer learning", *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.
 - [11] Y. Zhu et al., "Heterogeneous transfer learning for image classification", *Proc. 25th AAAI*, 2011.
 - [12] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong. "Crosslanguage knowledge transfer using multilingual deep neural networks with shared hidden layers," *ICASSP*, 2013.
 - [13] L. A. Alexandre, "3D object recognition using convolutional neural networks with transfer learning between input channels", *Proc. 13th Int. Conf. Intell. Autonomous Syst.*, pp. 889-898, 2014.
 - [14] L. Guo, Y. Lei, S. Xing, T. Yan, N. Li, *Deep Convolutional Transfer Learning Network : A New Method for Intelligent Fault Diagnosis of Machines With Unlabeled Data*, *IEEE Trans. Ind. Electron.* 66 (2019) 7316–7325.
 - [15] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, T. Zhang, "Deep model based domain adaptation for fault diagnosis", *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296-2305, Mar. 2017.
 - [16] L. Wen, L. Gao, X. Y. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis", *IEEE Trans. Syst. Man Cybern. Syst.*.
 - [17] J. Hasan, M.M.M. Islam, J. Kim, Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions. *Measurement*. 138 (2019) 620–631.
 - [18] P.E.I. Cao, S. Zhang, J. Tang, Preprocessing-Free Gear Fault Diagnosis Using Small Datasets With Deep Convolutional Neural Network-Based Transfer Learning, *IEEE Access*. 6 (2018) 26241–26253.
 - [19] Lee J, Azamfar M, Singh J, Siahpour S. Integration of digital twin and deep learning in cyber-physical systems: towards smart manufacturing. *IET Collaborative Intelligent Manufacturing*. 2020 Apr 9;2(1):34-6.
 - [20] H. Qiu, J. Lee, J. Lin, G. Yu, Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics, *J. Sound Vib.* 289 (2006) 1066–1090.
 - [21] Y. Lei, *Intelligent fault diagnosis and remaining useful life estimation of rotating machinery*, Butterworth-Heinemann, 2016.
 - [22] Singh, Jaskaran, et al. "Deep learning-based cross-domain adaptation for gearbox fault diagnosis under variable speed conditions." *Measurement Science and Technology* 31.5 (2020): 055601.
 - [23] K.M. Borgwardt, A. Gretton, M.J. Rasch, H. Kriegel, A.J. Smola, B. Scho, Integrating structured biological data by kernel maximum mean discrepancy, *Bioinformatics*. 22 (2006) 49–57.
 - [24] Lu, Weining, et al. "Deep model based domain adaptation for fault diagnosis." *IEEE Transactions on Industrial Electronics* 64.3 (2016): 2296-2305.
 - [25] Aljundi, R. and Tuytelaars, T., 2016, October. Lightweight unsupervised domain adaptation by convolutional filter reconstruction. In *European Conference on Computer Vision* (pp. 508-515). Springer, Cham.
 - [26] Spectraquest Gearbox fault simulator <<http://spectraquest.com/prognostics/details/gps/>>, (n.d.).
 - [27] M. Long, J. Wang, G. Ding, J. Sun, P.S. Yu, Transfer Feature Learning with Joint Distribution Adaptation, in: *Proc. IEEE Int. Conf. Comput. Vis.*, 2013: pp. 2200–2207.
 - [28] J. Wang, Y. Chen, S. Hao, W. Feng, Z. Shen, Balanced Distribution Adaptation for Transfer Learning, in: *2017 IEEE Int. Conf. Data Min.*, 2017: pp. 1129–1134.
 - [29] J. Wang, Y. Chen, H. Yu, M. Hunag, Q. Yang, Easy Transfer Learning By Exploiting Intra-domain Structures, *ArXiv Prepr. ArXiv1904.01376*. (2019).
 - [30] Singh J, Darpe AK, Singh SP. Bearing damage assessment using Jensen-Rényi Divergence based on EEMD. *Mechanical Systems and Signal Processing*. 2017 Mar 15;87:307-39.