Measurement Science
and Technology

PAPER • OPEN ACCESS

A novel cross-domain adaption network based on Se-Sk-DenseNet for remaining useful life prediction of rolling bearings under different working conditions

To cite this article: Baosu Guo *et al* 2024 *Meas. Sci. Technol.* **35** 076114

OPEN ACCESS
IOP Publishing
Measurement Science and Technology
Meas. Sci. Technol. 35 (2024) 076114 (11pp)
https://doi.org/10.1088/1361-6501/ad3b2b

A novel cross-domain adaption network based on Se-Sk-DenseNet for remaining useful life prediction of rolling bearings under different working conditions

Baosu Guo[1,2], Hang Li[1], Hao Dong[3], Tianjie Han[1], Yingbing Sun[1,2,*], Jianchang Hou[1], Zhangpeng Jiang[1] and Qing Ni[4,*]

[1] School of Mechanical Engineering, Yanshan University, Qinhuangdao 066004, People's Republic of China
[2] Heavy-duty Intelligent Manufacturing Equipment Innovation Center of Hebei Province, Yanshan University, Qinhuangdao 066004, People's Republic of China
[3] China Astronautics Standards Institute, Beijing 100071, People's Republic of China
[4] School of Mechanical and Mechatronic Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia

E-mail: sunyingbing@ysu.edu.cn and qing.ni@uts.edu.au

Received 5 February 2024, revised 19 March 2024
Accepted for publication 5 April 2024
Published 15 April 2024

Abstract
Effectively predicting the remaining useful life (RUL) of rolling bearings can ensure reliability and safety, minimize machine downtime, and reduce the operation and maintenance costs of enterprises. To solve the problems of data distribution discrepancy caused by different working conditions and the collected signals containing a lot of useless information and noise, a novel cross-domain adaption network (CDAN) is proposed in this study. Firstly, a novel feature extractor, squeeze-and-excitation (Se)-selective kernel (Sk)-DenseNet, is developed to extract useful critical features from the input data and remove the ineffective features by embedding Se and Sk attention blocks; besides, a new objective loss function consist of the RUL loss, the multi-kernel maximum mean discrepancy loss, the contrastive loss, and the Kullback–Leibler divergence loss, is proposed to solve the problem of data distribution shift; finally, the effectiveness and superiority of CDAN are proved on the PHM2012 bearings dataset. The results demonstrate that CDAN can extract deep critical features and achieve the high cross-domain RUL prediction accuracy under different working conditions.

Keywords: Remaining using life, Cross-domain adaption network, Se-Sk-DenseNet, Objective loss function

1. Introduction

Rolling bearing is one of the most essential components in the rotating machinery, which is easy to be failure when running in the complex working conditions. Besides, the damaged bearings will directly influence the service performance of whole mechanical system [1]. Thus, the practical and accurate prognostic and health management (PHM) of rolling bearings can grasp the health status in real-time for fault detection and remaining useful life (RUL) prediction, and it has received great attention nowadays [2–4]. In the study of PHM, RUL means the normal service life of the machinery before the occurrence of the failures [5]. Therefore, RUL prediction aims to guide the replacement strategies of rolling bearings to prevent sudden failures by predicting the health condition, which

* Authors to whom any correspondence should be addressed.

can ensure product safety, improve mechanical operation efficiency and reduce maintenance costs [6, 7].

RUL prediction of rolling bearings can be usually categorized into three approaches: physical model-driven, data-driven, and hybrid approaches [8]. Most of physical model-driven approaches must research the failure mechanism and the damage laws, but it is very difficult to establish the accurate degradation models for complex parts under the complex working conditions [9, 10]. The hybrid approach combines the strengths of physical models and data-driven methods with the aim of improving the accuracy and interpretability of predictions. This approach can take advantage of the *a priori* knowledge and structural information provided by the physical model and calibrate and optimize the model through data-driven methods. However, the design and parameter tuning of hybrid methods can be complex and requires a combination of factors. On the contrary, data-driven approaches can construct the mapping relationship between a large amount of collected monitoring data and corresponding RUL, which do not require building the accurate degradation models [11, 12]. Among these, condition-based maintenance (CBM) is a method of determining maintenance needs by monitoring the condition of equipment. In the case of bearings, CBM involves the use of appropriate sensors to collect data about the bearings, which can include vibration frequency, vibration amplitude, temperature variations, and so on. Data analytics techniques are applied to process and interpret this data, and reasoning is performed to estimate the health and degradation of the bearing. Due to the rapid development of machine learning (ML), data-driven approaches have been widely used to learn degradation laws from historical monitoring data from a large number of rolling bearings under specific working conditions.

Some traditional ML models, for example, restricted Boltzmann machine [13], relevance vector machine [6], auto-encoder [14], artificial neural network (ANN) [15], and support vector machine (SVM) [16] have been applied for RUL prediction. Daroogheh *et al* [17] predicted degradation periods by utilizing particle filters and ANN. Nieto *et al* [18] proposed an RUL prediction model based on the particle swarm optimization (PSO)-SVM which is used for aircraft engines. However, these shallow ML models are challenging to extract deep features for characterizing RUL from non-stationary and non-linear signals [19, 20].

Deep learning (DL) has attracted great attention because of extracting effective deep characteristics from a large amount of the data collected from different sensors by using a lot of hidden layers. The RUL prediction is a typical time-series-related regression problem [21]. DL approaches are mainly divided into recurrent neural network (RNN) [22] and convolutional neural network (CNN) [23]. On the one hand, many sequence networks, including RNN and its versions (long short-term memory, gated recurrent unit, etc.) can capture the dependence relationship between the front and back segments of the time-series data, even the information with a significant time-series span, which are very suitable for the relevant signals generated in bearing degradation [24, 25]. However, the above RNN approaches still have some problems which are very hard to ignore [22]. The subsequent forecast in these models must wait for the previous forecast to complete because of the incapacity for parallel processing, and it will lead to the error accumulation by steps owing to the lower flexibility. In addition, many difficult problems, such as gradient explosion, large memory occupation and gradient disappearance, still exists during model training [26].

On the other hand, as another active branch of DL, the traditional CNN cannot be usually considered available for time-series problem modeling because of the limitation of the convolutional kernel size, resulting in the inability to process the long-term related information. However, many studies have shown that specific CNN can also get excellent results by taking advantage of the strong abilities to process high-dimensional complex data and extract features automatically. Li *et al* [27] proposed a new data-driven RUL prediction approach based on deep CNNs (DCNNs). The original collected normalized data was directly used as the input of the DCNN without the advance prediction and signal processing expertise. The experiments on the popular Commercial Modular Aero-Propulsion System Simulation dataset were carried out to demonstrate the high prediction accuracy and superiority. Besides, Liu *et al* [28] proposed a joint-loss (JL) CNN architecture to capture common characteristics between different relative problems through shared partial parameters and network. A JL function was designed to learn the key characteristics and enhance the generalization of CNN. It can also avoid the risk of the overfitting and reduce the calculation costs. However, the above studies commonly assume that the training and test data have the same distribution. Owing to the change of working conditions and strong noise in the actual industrial environment, there are the obvious distribution discrepancy. This will cause the generalization performance of the methods to be highly affected by the extremely variable working conditions. Thus, new approaches are urgently needed to predict the RUL under different working conditions.

Transfer learning (TL) provides a feasible way to solve the distribution discrepancy by obtaining common information between the source and target domains through minimizing the data features [29]. Domain adaption (DA) has gained great attention by reducing the distribution discrepancy of different domains, which can map different domains into a common feature space to learn domain-invariant features through the additional loss terms [30]. Domain adaptive neural networks (DANNs) [31] was proposed to use the maximum mean discrepancy (MMD) as a regularization to reduce the distribution discrepancy. Deep adaption network (DAN) [32] could learn transferable features with statistical guarantees by adding a multi-kernel MMD (MK-MMD) term to the loss function to reduce the distribution discrepancy. Conditional MMD [33], joint MMD [34], weighted MMD [35] were used by embedding of empirical conditional, joint distributions or assigning class-specific weights between the source and target data. Local MMD [36] was also proposed to measure the discrepancy through embedding relevant subdomains in source and target domains. However, these approaches do not have functions that accurately map the sample of the similar distribution unless the target domain is known. In addition, these methods only obtain global domain-invariant features to realize the cross-domain RUL prediction. This will easily lead to the loss of information of local degenerate features, thus affecting the prediction accuracy. Nevertheless, most TL approaches can not ignore the interference of the noise hidden in the original signals which can result in the limited transfer performance. Some noise and redundant information are always hidden in the original vibration signals collected from different working conditions which also cause great difficulties for the RUL prediction. Therefore, it is very necessary to develop an effective method to predict the RUL of rolling bearings under different working conditions.

In this paper, aiming to solve the above problems under different working conditions, a novel cross-DA network for the RUL prediction of rolling bearings is proposed by embedding squeeze-and-excitation (Se) [37] and selective-kernel (Sk) [38] attention blocks into DenseNet (Se-Sk-DenseNet) and designing a new DA architecture. As far as we know, this is the first attempt to skillfully take advantage of Se and Sk attention blocks to weaken the time-series signals unrelated to the actual degradation features, so as to extract the degradation features of key details. Besides, a new DA architecture with the RUL loss, the MK-MMD loss, the contrastive loss, and the Kullback–Leibler (KL) divergence loss is proposed to build a mapping function for comprehensively characterizing the loss function for learning the domain invariant features by minimizing the distribution discrepancy between different domains.

2. CDAN

A novel CDAN is proposed to handle the cross-domain feature distribution discrepancy with a lot of useless information and noise under different working conditions of rolling bearings. The overall procedure of CDAN for the RUL prediction of rolling bearings is presented in figure 1. Firstly, original signals of rolling bearings with a lot of useless information and noise are collected from the source and target domains. Secondly, Se-Sk-DenseNet is proposed to conduct CDAN model training. Among them, as an effective adaptive selection mechanism, Se block is embedded into DenseNet to select essential features from multi-inputs adaptively. Furthermore, Sk block is developed to learn deep features by adaptively adjusting attention weights, so as to further extract critical information and remove useless information. Thirdly, the objective loss function, consist of the RUL loss, the MK-MMD loss, the contrastive loss, and the KL divergence loss, is developed to learn domain invariant features by minimizing the distribution discrepancy between the source and target domains. Finally, the bearing RUL prediction in the target domain can be easily implemented based on the available features extracted from the source domain.

2.1. Deep feature extractor

Degradation data of rolling bearings have two prominent features: (1) the data not only contain the mechanical degradation information but also the noise and useless

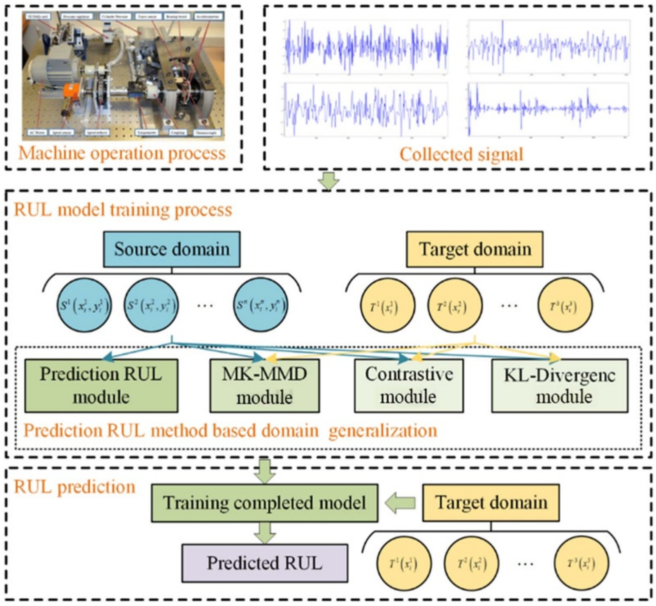

Figure 1. The procedure of CDAN for the RUL prediction of rolling bearings. Machine operation process adopted an experimental platform for accelerated degradation testing of bearings [39].

information; (2) the crucial degradation information caused by the faults may be found in local sequences. Therefore, in order to discover and highlight crucial information, it is very necessary to learn critical areas and restrain useless information. Fortunately, as effective attention mechanisms, Se block and Sk block are reasonably competent for the job, which are dexterously embedded into DenseNet to extract local and global deep degradation features.

2.1.1. DenseNet. DenseNet is a classical DCNN. It solves the problem of residual network destroying data flow through connecting the data of each channel. The network structure of the DenseNet is mainly composed of dense block and transition.

(1) Dense block
It is the most critical module in DenseNet, which extracts useful feature information from the signals. The DenseNet contains multiple dense blocks, and a dense block contains multiple convolution blocks. The input signals flow each convolution block in turn, and then the output of each convolution block is concerted together to flow to the transition. In this way, the feature information of each layer can be reused, and the correlation of the cross-layer information is enhanced.

(2) Transition.
It is mainly used to connect two adjacent dense blocks. It retains the main features, reduces the size of the feature map, and increases the calculation speed, which can prevent overfitting and improve the model generalization performance.

2.1.2. Se-Dense block. The convolutional layer is usually used to process the input information for obtaining the feature

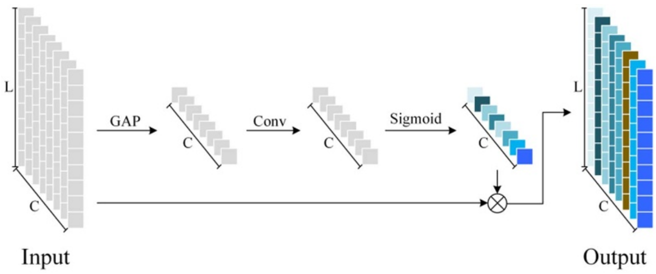

Figure 2. The structure of the Se-Dense block.

map of multiple channels, which contain much useless information and noise. Therefore, the correlation between different channels and the critical input information is calculated to pay attention to the essential information and eliminate the interference of useless information. The precise allocation of each channel weight can fully extract practical features in convolution calculation, thereby improving the non-linear expression ability of the model. Fortunately, the Se block is an excellent way to rule out useless information. As shown in figure 2, the Se block is integrated into dense block to form a new Se-Dense block.

Firstly, global average pooling (GAP) is applied to process the input data for obtaining preliminary global information from all channels. The GAP calculation formula is as follows:

$$F_{\mathrm{GAP}}\left(X_C\right) = \frac{1}{L}\sum_{l=1}^{L} X_C\left(l\right) \tag{1}$$

where X is the input characteristic element, L is the data length, C is the data channel.

Secondly, the 1×1 convolutional layer is used to parameterize the global information. 1×1 convolutional layer can make the parameter calculation amount smaller than the fully connected (FC) layer, which is calculated by:

$$F\left(x\right) = \sigma\left(w \cdot x + b\right) \tag{2}$$

where w is weight, b is biased, x is the global information, and σ is the activation function.

The Sigmoid is usually used as the activation function behind the convolutional layer to obtain the attention weight of each channel by scaling the sequence of global information parameters to the range of [0,1], which is represented as:

$$y = \frac{1}{1 + \mathrm{e}^{-x}} \tag{3}$$

where x is the sequence of global information parameters; y is attention weight.

Finally, the input characteristic element channel data are multiplied with the attention weight sequence to remove some useless information and noise, which can be updated as:

$$y = w_c \otimes X_c \tag{4}$$

where w_c is attention weight sequence; $\otimes$ is the multiplication of corresponding terms.

2.1.3. Sk-Transition block. It is considered that the data obtained from each convolutional block (Conv block) in the dense block are equally important before the original data are transferred to the transition. However, the data of the different Conv blocks have different importance in the actual calculation process. The data weights are calculated to strengthen the valuable information and weaken the useless information. In this way, the dense block can obtain more critical features and improve prediction accuracy. If the weight distribution is unreasonable, the useless information will have a negative effect on the extraction of critical features. Hence, the improved Sk-Transition block is shown in figure 3.

First of all, the data of several Conv blocks are input into the transition. The channel of each Conv block is C, and the number of the Conv block is N. So the total channels are $N \times C$. In order to make the channels of downsampling data 1C, the GAP is chosen to obtain the comprehensive global information, which can be calculated as:

$$y_{\mathrm{total}} = \mathrm{GAP}\left(\sum_{n=1}^{N} x_n^c\right) \tag{5}$$

where y_{total} is the comprehensive global information, x_n^c is the cth channel sequence of the data of the nth Conv block, and GAP is the GAP.

Then the global information of each Conv block is parameterized to get the total global information by the following calculation formula:

$$\left(y_1, y_2, \ldots, y_N\right) = \mathrm{Conv}\left(y_{\mathrm{total}}\right) \tag{6}$$

where y_N is the global information of the nth Conv block, and $(y_1,\ y_2 \ldots y_N)$ is the global information of the Conv blocks.

Afterwards, the softmax function is used to obtain the attention weights of all channels of each Conv block and the sum of the weights of each Conv block is 1, which can be expressed as:

$$\sum_{n=1}^{N} \tau_n^c = 1 \tag{7}$$

where τ_n^c is the nth attention weight of the cth channel. The calculation formula of the softmax function is as follows:

$$\tau_n = \mathrm{softmax}\left(y_n\right) = \frac{y_n}{\sum\limits_{n=1}^{N} \mathrm{e}^{y_n}} \tag{8}$$

where y_n is the global information of the nth Conv block.

Subsequently, in order to obtain the information of different importance, the attention weight and the data channel obtained by the Conv block are correspondly multiplied and added. Through the above operations, useful information is strengthened, and useless information is weakened. The purpose of adding attention mechanism is achieved by the following calculation formula:

$$y^c = \sum_{n=1}^{N} x_n^c \otimes \tau_n^c \tag{9}$$

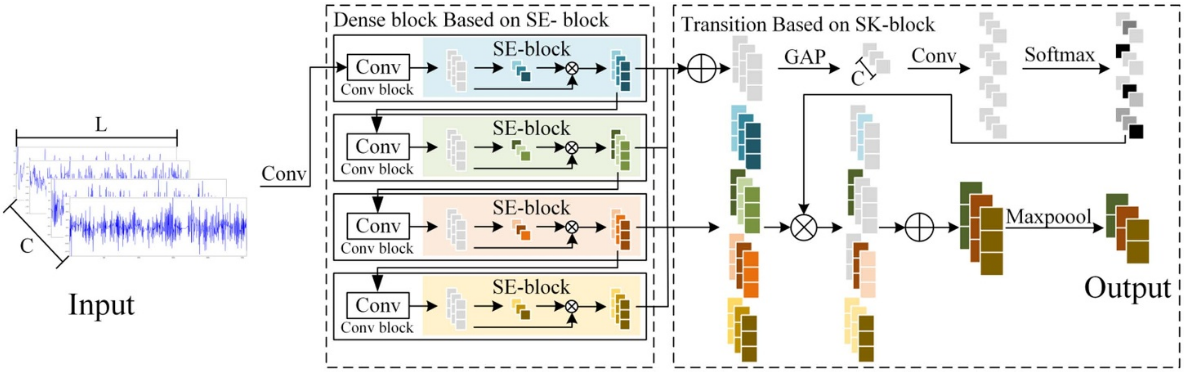

Figure 3. The structure of the improved Sk-Transition block.

where y^c is the sequence of the cth channel.

Finally, the max-pooling layer is used to complete the downsampling operation, and can further strengthen the useful information because of the different importance of the data. The calculation formula of the max-pooling layer is as follows:

$$\text{max-pooling} = \max\left(x_1, x_2, \ldots, x_k\right) \tag{10}$$

where k is the size of the max-pooling kernel.

2.1.4. Se-Sk-DenseNet. The proposed Se-Sk-DenseNet can optimize the branch adaptively by adding Se block and Sk block into dense block and transition, respectively. Through the above operations, critical information is strengthened, and useless information is weakened, which can enhance the accuracy of feature extraction. A complete feature extraction process based on Se-Sk-DenseNet is shown in figure 4.

2.2. RUL predictor

The RUL predictor is composed of one FC layer, whose purpose is to obtain the RUL. The predicted RUL can be defined as:

$$\hat{y} = \sigma\left(w \cdot F^{\mathrm{sp}} + b\right) \tag{11}$$

where σ is the rectified linear unit, F^{sp} is the output of the sparse feature selection layer, and w and b denote the weights and biases, respectively.

The extracted features of the source domain by using the Se-Sk-DenseNet feature extractor are transferred to the RUL predictor, which composed of multiple FC layers for outputting the corresponding predicted RUL. Moreover, the RUL predictor and Se-Sk-DenseNet feature extractor are trained in an end-to-end way. The loss between the predicted RUL and the actual RUL is described as follows:

$$L_{\mathrm{RUL}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \tag{12}$$

where $\hat{y}_i$ is the predicted RUL labels, y_i is the actual RUL values, and n is the number of the source samples.

2.3. The proposed DA module

2.3.1. MK-MMD loss term. The primary function of the monitoring module is to extract the critical information in the signals, and the mapping relationship between the information and the RUL is established. But additional module needs to be added to narrow the distribution discrepancy between the source and target domains. MK-MMD is a widely useful way to reduce the distribution distance in TL. The MK-MMD is shown as follows:

$$\mathrm{MK}-\mathrm{MMD}(S,T)^2 = \left\|\frac{1}{m}\sum_{i=1}^{m} f(S) - \frac{1}{n}\sum_{i=1}^{n} f(T)\right\|_{\mathrm{H}}^{2} \tag{13}$$

where S and m are the source domain and the number of data in the source domain, respectively. T and n are the target domain and the number of data in the target domain, respectively. $f(*)$ is to map the domain to the reproducing kernel Hilbert space. The Gaussian kernel function can be transformed into the inner product of the reproducing kernel Hilbert space. The MK-MMD simplified by kernel function is shown as follows

$$\begin{aligned}\mathrm{MK}-\mathrm{MMD}(S,T)^2 = \Bigg\| & \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k\left(S_i, S_j\right) \\ & - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k\left(S_i, T_j\right) \\ & + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k\left(T_i, T_j\right)\Bigg\|\end{aligned} \tag{14}$$

where $k(\cdot)$ is the mapping of the Gaussian kernel function. MK-MMD is the optimal kernel obtained by the combination of multiple linear kernels.

2.3.2. Contrastive loss term. In addition to reducing the distribution distance between the source and target domains in the MK-MMD loss term, it is also necessary to make the model unable to identify the data categories of each group in different working conditions. Minimizing the prediction error of data

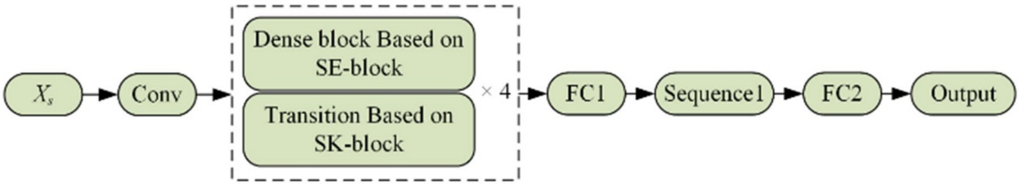

Figure 4. The extraction process of the Se-Sk-DenseNet.

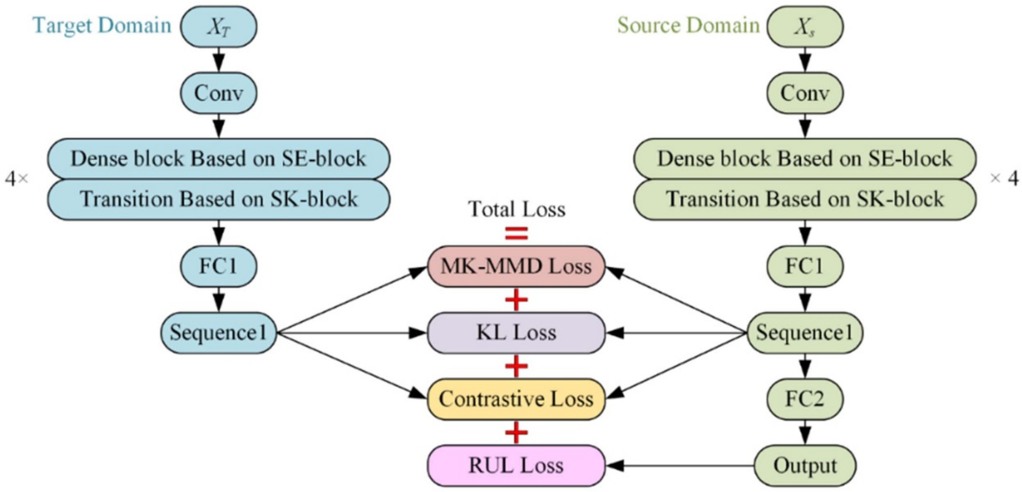

Figure 5. Structure diagram of the complete Se-Sk DenseNet and the proposed loss function.

and maximizing the category error of data is also one of essential ideas in domain generalization. The contrastive loss term expands the category error, which can maximize to study the common information between data under different working conditions. The RUL prediction accuracy is further improved by increasing the generalization performance.

The contrastive loss function proposed by [40] is applied in this paper. The contrastive loss function is shown as follows.

$$\begin{aligned}\mathrm{C}-\mathrm{Loss}\left(S,T,Y\right) &= \left(1-Y\right)\frac{1}{2}\left(D_{S-T}\right)\\ &\quad+\left(Y\right)\frac{1}{2}\left\{\max\left(0,l-D_{S-T}\right)\right\}^2\end{aligned} \tag{15}$$

where Y is the category label. If $Y=0$, S and T belong to the same category; if $Y=1$, S and T belong to different categories. D_{S-T} is the Euclidean distance between S and T. l is the margin to be set. If $Y=0$, the loss is proportional to the Euclidean distance; if $Y=1$, the loss is 0 when the Euclidean distance exceeds l and the loss is inversely proportional to the Euclidean distance when the Euclidean distance is not exceeded l. The purpose of the paper is to identify S and T as the same category, so l always is set to 0.

2.3.3. KL divergence loss term. In addition, the difference between probability distribution functions also has a significant influence on the feature extraction and the determination of critical regions. The KL divergence loss function is an effective solution. The concept of KL divergence comes from probability theory and information theory, and the formula is as follows:

$$\begin{aligned}\mathrm{KL}\left[P\left(X\right)\|Q\left(X\right)\right] &= \sum_{x\in X}\left[P\left(x\right)\log\frac{P\left(x\right)}{Q\left(x\right)}\right]\\ &= E_{x\sim P(x)}\left[\log\frac{P\left(x\right)}{Q\left(x\right)}\right]\end{aligned} \tag{16}$$

where $P(X)$ is the actual distribution function, $Q(X)$ is the approximate distribution function used to fit $P(X)$, X is the input value. The KL divergence is asymmetric, which means $\mathrm{KL}([P(X)\|Q(X)])\neq\mathrm{KL}([Q(X)\|P(X)])$. The forward KL divergence is used to characterize the function fitting behavior by finding the mean value. In the calculation process of minimizing forward KL divergence, when $P(X)=0$, $Q(X)$ becomes meaningless. When $P(X)>0$, $Q(X)$ searches for the highest fitting accuracy on the set of values of $P(X)$.

2.3.4. The proposed objective loss function. As shown in figure 5, the total objective loss function $\mathrm{Loss}_{\mathrm{total}}$ is proposed by combining the RUL regression loss term on the source domain (L_{RUL}), the MK-MMD loss term, the contrastive loss term, and the KL divergence loss term in this paper.

$$\begin{aligned}\mathrm{Loss}_{\mathrm{total}} &= L_{\mathrm{RUL}}+\alpha\times\mathrm{MK}-\mathrm{MMD}(S,T)^2\\ &\quad+\beta\times\mathrm{C}-\mathrm{Loss}\left(S,T,Y\right)+\delta\times\mathrm{KL}\left(S,T\right)\end{aligned} \tag{17}$$

whereα, β and γ are positive regular coefficients used to balance the contribution among each term. Therefore, cross-domain regression operation on target domain unlabeled data is carried out to obtain a more accurate RUL value by learning the stronger domain invariant features specified to the targets from the source data.

3. RUL prediction procedure

3.1. Algorithm flow

CDAN is composed of a deep feature extractor, a RUL predictor, and a DA module to achieve the accurate prediction under different working conditions. The detailed prediction process of CDAN is shown in algorithm 1.

Algorithm 1. Training procedure of CDAN.

Input: source domain: $\Im s = \{x_s^{im}, y_s^i\}_{i=1}^{n_s}$, target domain: $\Im t = t\{x_t^{jm}, y_t^j\}_{j=1}^{n_t}$; batch size: n;

Initialize network parameters

Adaptive filtering of obvious non-signal factors of the training set, extracting feature $\tilde{X}$;

Mining features X_{fs} of the source domain through Se-Sk-DenseNet module, and learning features X_{ft} of target domain I adaptively;

Predicting RUL Y_{s} of the source domain and Y_{t} of target domain I through RUL prediction;

Calculate the loss values:

$\mathrm{loss}_{\mathrm{RUL}} = \alpha \mathrm{loss}_{\mathrm{mse}} + \beta \mathrm{loss}_{\mathrm{MK\text{-}MMD}} + \gamma \mathrm{loss}_{\mathrm{con}} + \mu \mathrm{KL}$;

Calculate the current loss value;

Update parameters;

Predicting RUL Y_{t} of source domain II through RUL prediction;

Save parameters, and back propagate;

Predicting RUL of target domain through current parameter model.

3.2. Evaluation metrics

To quantify the experimental results, root mean squared error (RMSE) and mean absolute error (MAE) are used for the evaluation of the proposed method. The RMSE and MAE formulas are shown as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{18}$$

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{19}$$

where m is the number of samples. y_i is the actual RUL value. $\hat{y}_i$ is the RUL value calculated by the model.

Table 1. Settings of some key hyperparameters for the CDAN model.

Hyperparameters	Value
Batch size	16
Epochs	300
Kernel size	3
Learning rate	0.001
Optimizer	Adam

3.3. Hyperparameter setting

The predefined hyperparameters defined by the CDAN structure are mainly batch size, convolution kernel size, learning rate, the maximum number of iterations, etc. For the RUL prediction, these hyperparameters are calculated by cross-validation of experimental datasets, and finally determined after weighing prediction accuracy, calculation speed, and cost. Table 1 lists the final hyperparameter results.

4. Experimental results and discussion

4.1. Description of PHM2012 bearings dataset

The PHM2012 bearings dataset [39] is used for the experiment, which includes three working conditions, and the operating conditions of 17 bearings under the operating conditions of PRONOTIA platform are shown in table 2. The PRONOSTIA platform is shown in figure 6. The horizontal and vertical vibration signals were collected by two accelerometers of Type DYTRAN 3035B. The sampling frequency of the signals were 25.6 kHz, and the signals were recorded every 10 s. In the experiment, with the deepening of the degree of bearing wear, the reaction on the bearing vibration signal is the amplitude increase, in this paper, when the amplitude of the vibration signal is more than 20 g, then the bearing is considered to have reached the rated service life.

4.2. Single working condition RUL prediction experiment of PHM2012 bearings dataset

Due to the different failure positions of each bearing in the bearing dataset, and the using life of the bearings are different, the data labels are set to 100%–0%. The training and test datasets of this comparative experiment are the same as the PHM2012 challenge competition, and are shown in table 3.

The ablation experiments have been conducted to show the effectiveness of Se-Dense block and Sk-Transition. The RMSE and MAE of four models in the B1-3–B1-7 dataset are shown in table 4. The optimal results are indicated in bold in table 4. It can be seen that the results of the B1-3, B1-5, and B1-7 datasets have reached the highest level. Besides, the average of RMSE and MAE are respectively 0.130 and 0.100, which are less than other models.

Table 2. The settings of each operation condition in the dataset.

Operation condition	Radial force	Rotating speed	Dataset
Condition 1	4.0 kN	1800 rmp	B1-1–B1-7
Condition 2	4.2 kN	1650 rmp	B2-1–B2-7
Condition 3	5.0 kN	1500 rmp	B3-1–B3-3

Table 3. The training dataset and test dataset settings.

No.	Training dataset	Test dataset
S1	B1-1, B1-2	B1-3, B1-4, B1-5, B1-6, B1-7

Table 4. The RMSE and MAE of four models in the B1-3–B1-7 dataset.

model		B1-3	B1-4	B1-5	B1-6	B1-7	Average
DenseNet	RMSE	0.166	0.064	0.348	0.244	0.139	0.192
	MAE	0.148	**0.052**	0.276	0.175	0.113	0.153
Se-DenseNet	RMSE	0.159	0.208	0.185	0.172	0.117	0.168
	MAE	0.140	0.127	0.140	**0.110**	0.094	0.122
Sk-DenseNet	RMSE	0.177	0.081	0.243	0.206	0.126	0.167
	MAE	0.154	0.063	0.178	0.156	0.101	0.130
Se-Sk-DenseNet	RMSE	0.158	0.109	0.121	0.152	0.110	**0.130**
	MAE	**0.124**	0.086	**0.089**	0.112	**0.090**	**0.100**

As shown in figure 6, the overall trend of experiments conducted through the Se-Sk-DenseNet model shows a monotonous downward trend, which indicates that the actual degradation characteristics of bearings can be learned. When the error between the predicted value and the real value gradually increases, the model can timely draw closer to the predictive trending, indicating that the Se-Sk-DenseNet model can mine the relationship between hidden features and the amount of front and rear degradation with insignificant degradation features, so as to correct the model's prediction error. This is due to the Se and Sk attention block added in the DenseNet. Through two adaptive optimizations of the weight coefficients of different channels, the data weight of each time will be adjusted according to its importance, so that the prediction results will approach the actual values. In addition, figures 6(a) and (d) show that Se-Sk-DenseNet can significantly shorten the results of the predicted value and the true value in the late stage of bearing degradation compared with DenseNet in the prediction of B1-5. This means that the model can constrain non critical areas and strengthen the key information in the bearing degradation process, and strengthen the representation ability of the network.

4.3. Variable operation condition RUL prediction experiment of PHM2012 bearings dataset

The RUL prediction method based on Se-Sk-DenseNet has achieved high accuracy in the single working condition. However, the prediction method should be designed to cope with the multi-working conditions of the bearing. The data distribution of multi-working condition signals is usually different, and the single-working condition model has not learned the new degradation law in different working conditions, which leads to poor monitoring accuracy. In order to verify the validity of the bearing RUL prediction method based on the proposed CDAN in this paper, comparative experiments under multi-working conditions are designed. The settings of the training and test datasets are shown in table 5. Two of the three conditions are assumed to be selected and the RUL prediction results are transferred from A to B. The training dataset contains A with labeled and B without labeled, and the test dataset contains the remaining B data. In addition to the CDAN being trained, five models are used as comparison methods, respectively, Se-Sk-DenseNet without DA, CNN with DA, Miao's sparse domain adaption network (SDAN), transferable CNN and a TL method based on bidirectional gated recurrent unit. The RMSE and MAE of six models in the M1–M6 are shown in table 6, with the best results highlighted in bold.

It can be seen from table 6 that the bearing RUL prediction method based on the CDAN is better than other methods on the whole. The prediction impact of Se-Sk-DenseNet is poor, while the effect of CDAN will be significantly improved. The RMSE of the proposed CDAN is reduced by 37.8%, 12.6%, 15.5%, 40.1% and 24.2%, respectively, compared with the other methods. Therefore, the proposed DA model is an effective way to solve the RUL prediction problem under multi-working conditions. The MK-MMD loss term, the contrastive loss term, and the KL divergence loss term all play an important role in reducing the distribution discrepancy

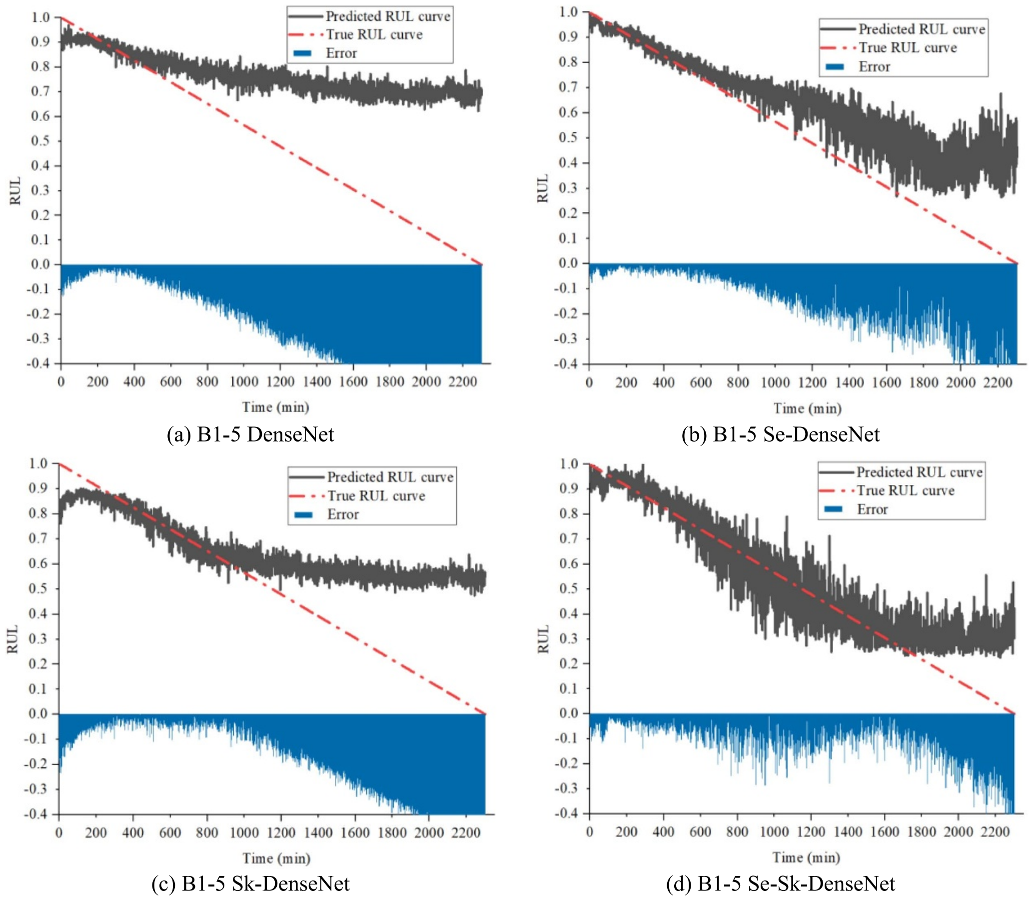

Figure 6. The predicted and true RUL curves of the two models.

Table 5. The training dataset and test dataset settings.

No.	Domain adaption	Training dataset	Test dataset
M1	Condition 1 → Condition 2	Labeled: B1-1, B1-2. Unlabeled: B2-1, B2-2	B2-3, B2-4, B2-5, B2-6, B2-7
M2	Condition 1 → Condition 3	Labeled: B1-1, B1-2. Unlabeled: B3-1, B3-2	B3-3
M3	Condition 2 → Condition 1	Labeled: B2-1, B2-2. Unlabeled: B1-1, B1-2	B1-3, B1-4, B1-5, B1-6, B1-7
M4	Condition 2 → Condition 3	Labeled: B2-1, B2-2. Unlabeled: B3-1, B3-2	B3-3
M5	Condition 3 → Condition 1	Labeled: B3-1, B3-2. Unlabeled: B1-1, B1-2	B1-3, B1-4, B1-5, B1-6, B1-7
M6	Condition 3 → Condition 2	Labeled: B3-1, B3-2. Unlabeled: B2-1, B2-2	B2-3, B2-4, B2-5, B2-6, B2-7

Table 6. The RMSE and MAE of four models in theM1–M6 dataset.

	Se-Sk-DenseNet		CDACNN		Miao's SDAN [41]		TCNN [42]		TBiGRU [43]		CDAN	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
1 → 2	0.293	0.237	0.297	0.229	0.398	—	0.33	0.31	**0.17**	**0.15**	0.274	0.218
1 → 3	0.246	0.187	0.212	0.159	0.064	—	0.31	0.29	**0.15**	**0.13**	0.187	0.146
2 → 1	0.383	0.224	0.284	0.210	0.353	—	0.24	0.22	0.23	0.21	**0.218**	**0.167**
2 → 3	0.278	0.231	0.244	0.200	**0.098**	—	0.27	0.25	0.22	0.20	0.191	**0.152**
3 → 1	0.698	0.290	0.246	0.205	0.291	—	0.60	0.58	0.64	0.62	**0.243**	**0.184**
3 → 2	0.311	0.222	0.291	0.217	0.421	—	0.54	0.52	0.40	0.38	**0.260**	**0.202**
Average	0.368	0.232	0.262	0.203	0.271	—	0.382	0.362	0.302	0.282	**0.229**	**0.178**

between different domains. The MK-MMD loss term has excellent performances to reduce the data distribution distance between the source domain and the target domain. What's more, if the difference between the predicted value of 3 → 1, 2 → 1 and the real value is too large, the predicted value will be corrected in time, which depends on the contrastive loss term. The contrastive module can extract hidden common features when the source domain and target domain are quite different.

In addition, the KL divergence loss term makes the CDAN to have stronger generalization performance.

5. Conclusions

Rolling bearings are one of the most important key components in rotating machinery and equipment. Due to the special characteristics of rotating machinery, bearing failure often occurs, therefore, the remaining life prediction and degradation assessment of rolling bearings can predict the failure of rolling bearings in advance, which can provide an effective guarantee for the safety and reliability of rotating machinery and equipment. This paper focuses on the existing problems of residual life prediction, takes rolling bearings as the research object, and carries out in-depth research and improvement of the method of residual life prediction. In this paper, the RUL prediction of the rolling bearings under different working conditions with a lot of useless information and noise is studied. Two extensive experiments are conducted to confirm the effectiveness of the proposed method. The conclusions can be summarized as follows:

(1) A improved deep feature extractor, Se-Sk-DenseNet is proposed to learn domain invariant features from multiple input signals by embedding Se and Sk blocks for weakening the signals unrelated to the real degradation features and extracting the degradation features of critical details. The ablation experimental results under the single working condition show the effectiveness of Se-Sk-DenseNet.
(2) A new objective loss function, composed of the RUL loss, the MK-MMD loss, the contrastive loss, and the KL divergence loss, is proposed to reduce the distribution discrepancy of different domains. Among them, the contrastive loss can maximize the common information of different categories of data and the KL divergence loss can reduce the difference of probability distribution functions between source domain and target domain. These are beneficial supplements to the MK-MMD loss term.
(3) A novel CDAN called CDAN for the RUL estimation of rolling bearings based on Se-Sk-DenseNet is proposed under different working conditions.it can successfully reduce the distribution discrepancy of the features from multiple input signals of rolling bearings with a lot of noise under different working conditions. Experimental results under multi-working conditions show that the proposed method has the advantages of strong generalization and high accuracy, and is superior to other methods.

However, a major limitation of this work is that the effectiveness of the proposed technique has only been tested using laboratory data, its validity in industrial applications requires further work where a continuous full-life degradation data may not be the case. This constitutes to a future research direction.

Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

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.

Conflict of 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.

Funding

The authors gratefully acknowledge the financial support of the National Natural Science Foundation of China (No. 52175488), Central Guiding Local Science and Technology Development Fund Projects (No. 236Z1808G), Higher Education Science and Technology Plan of Hebei Provincial Department of Education (No. BJ2021045), and the Yantai Science and Technology Plan Project (2023ZDCX012).

ORCID iDs

Yingbing Sun https://orcid.org/0000-0002-7510-2314
Qing Ni https://orcid.org/0000-0001-7537-3934

References

[1] Li W and Deng L 2023 A hybrid model-based prognostics approach for estimating remaining useful life of rolling bearings *Meas. Sci. Technol.* **34** 105012
[2] Li Y, Wang X, Zheng J, Feng K and Ji J C 2023 Bi-filter multiscale-diversity-entropy-based weak feature extraction for a rotor-bearing system *Meas. Sci. Technol.* **34** 065011
[3] Xu Y, Chen Y, Zhang H, Feng K, Wang Y, Yang C and Ni Q 2023 Global contextual feature aggregation networks with multiscale attention mechanism for mechanical fault diagnosis under non-stationary conditions *Mech. Syst. Signal Process.* **203** 110724
[4] Feng K, Ji J C, Ni Q and Beer M 2023 A review of vibration-based gear wear monitoring and prediction techniques *Mech. Syst. Signal Process.* **182** 109605
[5] Jiang G, Duan Z, Zhao Q, Li D and Luan Y 2024 Remaining useful life prediction of rolling bearings based on TCN-MSA *Meas. Sci. Technol.* **35** 025125
[6] Tang J, Zheng G, He D, Ding X, Huang W, Shao Y and Wang L 2021 Rolling bearing remaining useful life prediction via weight tracking relevance vector machine *Meas. Sci. Technol.* **32** 024006
[7] Mao W, Zhang W, Feng K, Beer M and Yang C 2024 Tensor representation-based transferability analytics and selective

transfer learning of prognostic knowledge for remaining useful life prediction across machines *Reliab. Eng. Syst. Saf.* **242** 109695

[8] Zhu J, Chen N and Shen C 2020 A new data-driven transferable remaining useful life prediction approach for bearing under different working conditions *Mech. Syst. Signal Process.* **139** 106602

[9] Nguyen K T P and Medjaher K 2019 A new dynamic predictive maintenance framework using deep learning for failure prognostics *Reliab. Eng. Syst. Saf.* **188** 251–62

[10] Aremu O O, Hyland-Wood D and McAree P R 2019 A relative entropy Weibull-SAX framework for health indices construction and health stage division in degradation modeling of multivariate time series asset data *Adv. Eng. Inform.* **40** 121–34

[11] Ding P, Jia M and Wang H 2021 A dynamic structure-adaptive symbolic approach for slewing bearings' life prediction under variable working conditions *Struct. Health Monit.* **20** 273–302

[12] Feng K, Ji J C, Zhang Y, Ni Q, Liu Z and Beer M 2023 Digital twin-driven intelligent assessment of gear surface degradation *Mech. Syst. Signal Process.* **186** 109896

[13] Zou Y, Zhang Y and Mao H 2021 Fault diagnosis on the bearing of traction motor in high-speed trains based on deep learning *Alex. Eng. J.* **60** 1209–19

[14] Chen D, Qin Y, Wang Y and Zhou J 2021 Health indicator construction by quadratic function-based deep convolutional auto-encoder and its application into bearing RUL prediction *ISA Trans.* **114** 44–56

[15] Kumar H S, Pai P S, Sriram N S and Vijay G S 2013 ANN based evaluation of performance of wavelet transform for condition monitoring of rolling element bearing *Procedia Eng.* **64** 805–14

[16] Tran V T, Thom Pham H, Yang B-S and Tien Nguyen T 2012 Machine performance degradation assessment and remaining useful life prediction using proportional hazard model and support vector machine *Mech. Syst. Signal Process.* **32** 320–30

[17] Daroogheh N, Baniamerian A, Meskin N and Khorasani K 2017 Prognosis and health monitoring of nonlinear systems using a hybrid scheme through integration of PFs and neural networks *IEEE Trans. Syst. Man Cybern. Syst.* **47** 1990–2004

[18] Nieto P G, García-Gonzalo E, Sánchez Lasheras F and de Cos Juez F J 2015 Hybrid PSO–SVM-based method for forecasting of the remaining useful life for aircraft engines and evaluation of its reliability *Reliab. Eng. Syst. Saf.* **138** 219–31

[19] Cheng Y, Hu K, Wu J, Zhu H and Lee C K M 2022 A deep learning-based two-stage prognostic approach for remaining useful life of rolling bearing *Appl. Intell.* **52** 5880–95

[20] Wang B, Lei Y, Li N and Yan T 2019 Deep separable convolutional network for remaining useful life prediction of machinery *Mech. Syst. Signal Process.* **134** 106330

[21] Ye R and Dai Q 2018 A novel transfer learning framework for time series forecasting *Knowl.-Based Syst.* **156** 74–99

[22] Yu Y, Si X, Hu C and Zhang J 2019 A review of recurrent neural networks: LSTM cells and network architectures *Neural Comput.* **31** 1235–70

[23] Gu J *et al* 2018 Recent advances in convolutional neural networks *Pattern Recognit.* **77** 354–77

[24] Zhang J, Wang P, Yan R and Gao R X 2018 Long short-term memory for machine remaining life prediction *J. Manuf. Syst.* **48** 78–86

[25] Yan H, Qin Y, Xiang S, Wang Y and Chen H 2020 Long-term gear life prediction based on ordered neurons LSTM neural networks *Measurement* **165** 108205

[26] Yuan M, Wu Y and Lin L 2016 Fault diagnosis and remaining useful life estimation of aero engine using LSTM neural network *2016 IEEE/CSAA Int. Conf. on Aircraft Utility Systems (AUS)* (IEEE) pp 135–40

[27] Li X, Ding Q and Sun J-Q 2018 Remaining useful life estimation in prognostics using deep convolution neural networks *Reliab. Eng. Syst. Saf.* **172** 1–11

[28] Liu R, Yang B and Hauptmann A G 2020 Simultaneous bearing fault recognition and remaining useful life prediction using joint-loss convolutional neural network *IEEE Trans. Ind. Inform.* **16** 87–96

[29] Gligic L, Kormilitzin A, Goldberg P and Nevado-Holgado A 2020 Named entity recognition in electronic health records using transfer learning bootstrapped neural networks *Neural Netw.* **121** 132–9

[30] Rozantsev A, Salzmann M and Fua P 2019 Beyond sharing weights for deep domain adaptation *IEEE Trans. Pattern Anal. Mach. Intell.* **41** 801–14

[31] Ghifary M, Kleijn W B and Zhang M 2014 Domain adaptive neural networks for object recognition *PRICAI 2014: Trends in Artificial Intelligence* (https://doi.org/10.1007/978-3-319-13560-1_76)

[32] Long M, Cao Y, Wang J and Jordan M I 2015 Learning transferable features with deep adaptation networks *Int. Conf. on Machine Learning*

[33] Long M, Wang J, Ding G, Sun J and Yu P S 2013 Transfer feature learning with joint distribution adaptation *2013 IEEE Int. Conf. on Computer Vision (ICCV)* (IEEE) pp 2200–7

[34] Long M, Zhu H, Wang J and Jordan M I 2017 Deep transfer learning with joint adaptation networks *Proc. 34th Int. Conf. on Machine Learning* vol 70 (PMLR) pp 2208–17

[35] Yan H, Ding Y, Li P, Wang Q, Xu Y and Zuo W 2017 Mind the class weight bias: weighted maximum mean discrepancy for unsupervised domain adaptation *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE) pp 945–54

[36] Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F and Vaughan J W 2010 A theory of learning from different domains *Mach. Learn.* **79** 151–75

[37] Hu J, Shen L, Albanie S, Sun G and Wu E 2020 Squeeze-and-excitation networks *IEEE Trans. Pattern Anal. Mach. Intell.* **42** 2011–23

[38] Li X, Wang W, Hu X and Yang J 2019 Selective kernel networks *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE) pp 510–9

[39] Nectoux P, Gouriveau R, Medjaher K, Ramasso E, Morello B, Zerhouni N and Varnier C 2012 PRONOSTIA: an experimental platform for bearings accelerated degradation tests *IEEE Int. Conf. on Prognostics and Health Management, PHM'12*

[40] Hadsell R, Chopra S and LeCun Y 2006 Dimensionality reduction by learning an invariant mapping *2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition—Volume 2 (CVPR'06)* vol 2 (IEEE) pp 1735–42

[41] Miao M, Yu J and Zhao Z 2022 A sparse domain adaption network for remaining useful life prediction of rolling bearings under different working conditions *Reliab. Eng. Syst. Saf.* **219** 108259

[42] Cheng H, Kong X, Chen G, Wang Q and Wang R 2021 Transferable convolutional neural network based remaining useful life prediction of bearing under multiple failure behaviors *Measurement* **168** 108286

[43] Cao Y, Jia M, Ding P and Ding Y 2021 Transfer learning for remaining useful life prediction of multi-conditions bearings based on bidirectional-GRU network *Measurement* **178** 109287