Knowledge-Based Systems 288 (2024) 111499

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Interpretable physics-informed domain adaptation paradigm for cross-machine transfer diagnosis

Chao He [a,b], Hongmei Shi [a,b,*], Xiaorong Liu [c], Jianbo Li [a,b]

[a] State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing 100044, China
[b] School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China
[c] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China

ARTICLE INFO

Keywords:
Cross-machine diagnosis
Physics-informed AI
Wavelet weights
Traction motor systems
Transfer learning

ABSTRACT

While transfer learning-based intelligent diagnosis has achieved significant breakthroughs, the performance of existing well-known methods still needs urgent improvement, given the increasingly significant distribution discrepancy between source and target domain data from different machines. To tackle this issue, rather than designing domain discrepancy statistical metrics or elaborate network architecture, we delve deep into the interaction and mutual promotion between signal processing and domain adaptation. Inspired by wavelet technology and weight initialization, an end-to-end, succinct, and high-performance physics-informed wavelet domain adaptation network (WIDAN) has been subtly devised, which integrates interpretable wavelet knowledge into the dual-stream convolutional layer with independent weights to cope with extremely challenging cross-machine diagnostic tasks. Specifically, the first-layer weights of a CNN are updated with optimized and informative Laplace or Morlet weights. This approach alleviates troublesome parameter selection, where scaling and translation factors with specific physical interpretations are constrained by the convolution kernel parameters. Additionally, a smooth-assisted scaling factor is introduced to ensure consistency with neural network weights. Furthermore, a dual-stream bottleneck layer is designed to learn reasonable weights to pre-transform different domain data into a uniform common space. This can promote WIDAN to extract domain-invariant features. Holistic evaluations confirm that WIDAN outperforms state-of-the-art models across multiple tasks, indicating that a wide first-layer kernel with optimized wavelet weight initialization can enhance domain transferability, thus validly fostering cross-machine transfer diagnosis.

1. Introduction

Intelligent fault diagnosis has evolved with the advent of industrial big data and artificial intelligence.

Astonishing progress has been made in deep-learning-based methods. However, acquiring a significant volume of labeled real-world fault data is challenging, rendering numerous existing deep learning algorithms invalid in real-world scenarios. However, some public, simulated, and self-collected datasets from testbeds that provide a basis for employing transfer learning are available [1–3]. This approach enables the implementation of fault diagnosis, bridging the gap from laboratory data to real-world applications. Transfer fault diagnosis includes cross-speed, cross-sensor, cross-load, and cross-machine tasks and so forth. When faced with scarce fault samples and constantly changing operations, the cross-machine transfer diagnosis is an excellent solution. This is an important concept in transfer fault diagnosis, which denotes transferring the accumulated annotated data from Device A over to the unannotated data of Device B, for assessing the operational health state of Device B. However, compared with cross-speed, cross-sensor, and cross-load tasks, cross-machine diagnosis is an extremely challenging task because of even larger distribution shifts between datasets with different mechanical equipment. Traditional transfer learning algorithms generally struggle to achieve satisfactory performances. Deep transfer learning with a powerful data-driven expressivity encompasses a wider range of applications in cross-machine diagnosis, including the fine-tuning-based methods, mapping-based methods, and hybrid strategies.

With regards to fine-tuning, Luo et al. [4] developed a modified stacked autoencoder. Zhang et al. [5] gained experience with few-shot learning and simultaneously fine-tuned a high-performance feature encoder. In particular, for long-tailed datasets, Li et al. [6] prioritized the imbalanced samples in the source domain to monitor fan faults, whereas Liu et al. [7] addressed the imbalance in the target domain to develop a transfer residual network based on metadata. Han et al. [8] proposed a network with multiple domain discriminators by incorporating domain adversarial-based ideas.

* Corresponding author at: State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing 100044, China.
E-mail addresses: chaohe@bjtu.edu.cn (C. He), hmshi@bjtu.edu.cn (H. Shi), xr_liu@my.swjtu.edu.cn (X. Liu), jbli@bjtu.edu.cn (J. Li).

https://doi.org/10.1016/j.knosys.2024.111499
Received 20 November 2023; Received in revised form 28 January 2024; Accepted 6 February 2024
Available online 9 February 2024
0950-7051/

Diverse statistical metrics, such as ensemble weighting maximum mean discrepancy [9], multiple kernel local maximum mean discrepancy [10], embedded joint maximum mean discrepancy [11], and maximum mean square discrepancy [12] have been designed for or applied in mapping-based methods. In addition to diminishing the distribution shift of source and target domains, certain statistical metrics consider the intra- and inter-class discriminative features.

Furthermore, several scholars have integrated hybrid learning strategies, including few-shot learning [13], meta learning [14], metric learning [15], contrastive learning [16], causality [17], imbalanced learning [18], attention mechanism [19,20], and multi-source domain adaptation [21]. Jang et al. [22] introduced a domain interpolation adaptation network. Wan et al. [23] developed a multi-level domain adaptation network. The investigation of cross-machine diagnosis utilizes multiple state-of-the-art (SOTA) training strategies tailored to different scenarios, to enable a more robust diagnosis for complicated and concrete cross-machine diagnosis tasks. In addition, Bilinear residual network methods [24] and Bilinear neural network methods [25–27] can obtain 100% accurate analytical solution for partial differential equation, which is far more accurate than traditional neural network numerical method. This also provides great inspiration for cross-machine diagnosis.

However, the performances of most purely data-driven algorithms may depend on the quality and quantity of data, making it challenging to train robust models when these conditions are not met. Moreover, statistical metrics of numerous algorithms increase the computational complexity and introduce extra hyper-parameters that generally determine the performance of the algorithms. Searching for the appropriate hyper-parameters is typically time-consuming and labor-intensive. A signal-processing-assisted scheme can alleviate these issues. Therefore, domain adaptation networks (DAN) integrated with signal processing have been introduced into cross-machine diagnosis to promote domain transferability. Ref. [2] highlighted this as a promising research topic. Ref. [1] envisaged the potential of the interpretable physics-informed signal processing to settle it; thus far, such a methodology has not been prevailed in the fault diagnosis domain. Additionally, Kim et al. [28] argued that data pre-processing can decrease the distribution deviation between different datasets, thereby saving training time and improving accuracy. Therefore, they exploited signal processing to transform different datasets into a common pattern space without the prior knowledge of the fault frequency. Refs. [28–31] are ensemble approaches that combine signal processing and DAN. The algorithms mentioned above follow a two-stage strategy [32], and optimizing the hyper-parameters of the signal processing algorithms remains a formidable challenge.

The one-stage strategy considers the hyperparameters of signal processing algorithms as part of the neural networks and automatically updates them, thereby addressing the aforementioned concern. In these schemes, wavelets offer promising prospects as staple and reliable signal processing methods. Ref. [33] combines multiple wavelet kernels such as the Laplace wavelet and the Morlet wavelet, whereas Ref. [34] introduces a modified wavelet kernel. The aforementioned studies concentrated solely on fault-detection tasks on the same machine. For cross-domain tasks, Refs. [35,36] exploited an interpretable wavelet scattering module to construct time-scattering convolutional networks for cross-domain, rather than cross-machine diagnoses. Yue et al. [37] proposed an interpretable multiscale wavelet prototypical network (MWPN) that combines the Laplace wavelet and Morlet wavelet kernel convolution [38] with few-shot learning strategies to tackle cross-component diagnosis. Shang et al. [39] designed a denoising fault-aware wavelet network (DFAWNet) with the Laplace wavelet and Morlet wavelet and investigated its advantages for cross-speed diagnosis. Wavelet scattering, Laplace wavelet and Morlet wavelet have shown excellent performances in wavelet-based interpretable diagnosis. However, the research on whether wavelet weights can decrease intra-and intra-domain discrepancies is scarce.

Recent research [40] has demonstrated the significant role of initialization weights in determining the generalizability of networks. Poor initialization can result in inadequate generalization, which emphasizes the need for a robust initialization approach. However, no specific initialization method has been proposed. In addition, some classical paradigms are widespread in fault diagnosis. A wide kernel in the first convolution layer was empirically verified to facilitate the robustness of deep convolutional neural networks with wide first-layer kernels (WDCNN). Subsequently, seminal SincNet [41] emphasized the significance of the first interpretable bottleneck. In fault identification, WaveKernelNet [38] validated the aforementioned theory, and Ref. [42] highlighted physically interpretable weights.

Results from prior gains have contributed significantly to this field. However, some issues still warrant further attention.

1. There is a gap in existing research on the application of wavelet weights to decrease intra- and intra-domain discrepancy and to extract discriminative fault features.
2. For transfer learning in fault diagnosis, current research only has a comprehensive understanding about the usage of a larger convolutional kernel for the first-layer CNN, but lacks a unified standard for the initialization of weights.
3. Current research on the impact of interpretable learning (a signal processing-driven method) in facilitating cross-machine transfer remains limited in depth and scope. Wavelet-based investigations require an additional layer to be connected to neural networks and lack a detailed discussion of the link between interpretable wavelet weights and domain transferability, particularly concerning auxiliaries for cross-machine transfer diagnosis.
4. Wavelet-based methods fail to account for the influence of sampling frequency and ignore the relationship between wavelet parameters and convolution parameters, leading to a large gap between the range of values for wavelet weights and the universal weights following a normal distribution. This decreases the trainability.

To address these challenges, reasonable first-layer weight initialization can continuously optimize neural networks in a more generalized and robust direction. Concurrently, different domain data must be matched with reasonable weights to pre-transform them into a uniform common space. Wavelet signal processing technology can extract the generic features that can minimize the distribution discrepancy between the source and the target domains; however, its parameter selection is challenging. Based on prior efforts, we propose that interpretable weight initialization based on signal mechanisms can boost domain transferability and facilitate the learning of domain-invariant features. Therefore, WIDAN, which expands upon previous efforts by exploring the potential value of reasonable initialization of the first-layer weights in deep models, while extending the scope to inherently challenging cross-machine diagnosis, has been proposed. It highlights a novel paradigm, namely a wide first-layer kernel initialized with physically interpretable weights, that can validly foster cross-machine diagnosis.

From an application perspective, this study is devoted to the cross-machine diagnosis of traction motor systems in high-speed trains [43], incorporating interpretable wavelet weights. To the best of our knowledge, this work is the first time to apply the wavelet weight initialization with interpretability and consistency with the other weights to more challenging cross-machine diagnostic scenarios. This innovative approach remains an unexplored topic.

This study highlights the following key contributions.

1. A physics-informed cross-machine diagnosis network that utilizes optimized wavelet weights to initialize the first-layer weights is proposed. This approach enhances domain transferability from a signal-processing-based perspective. This systematic aggregation leverages signal processing reliability to

promote the robustness of data-driven models, while alleviating the troublesome parameter selection of signal processing algorithms utilizing the powerful expressive capability of data-driven models.
2. A smooth-augmented wavelet kernel that introduces a smoothing factor to regulate the consistency of the scale factor (s) and the translation factor (u) is devised as wavelet weights. This addresses the gradient vanishing issue of the model. The convolution kernel size and the output channel number are tied to the parameters s, u, and t (time), endowing the convolution parameters with a certain wavelet physical significance, which in turn extracts more valuable and informative features.
3. In view of the sampling frequency, a dual-stream architecture with unshared weights is designed to transform data from different machines into a same common space using optimized, non-shared wavelet weights in order to facilitate domain transferability.
4. A paradigm for the first layer of the transfer learning fault diagnosis is designed using a wide kernel and optimized wavelet weight initialization. It can enhance the domain transferability and further validly promote cross-machine diagnosis.

The WIDAN framework, exhibiting the remarkable superiority and extrapolation capability, can serve as the basis feature extractor. This framework has been validated in a cross-machine fault diagnosis using four datasets from different sources. Section 2 presents the preliminaries of the study. Smooth-augmented wavelet kernel (SWK) is scrutinized through a detailed explanation in Section 3. Section 4 incorporates experiments and analyses to demonstrate the superior performance of the WIDAN. Section 5 presents conclusions and recommendations for advanced research.

2. Preliminary knowledge

2.1. Background knowledge of traction motor of high-speed trains

The "Beijing Jiaotong University NTN Traction Motor Bearing Test Bench" has a set of specialized bearing testing equipment in China that can conduct durability tests on traction motor bearings for high-speed trains. The salient attributes of the test bench include the integration of cylindrical roller bearings and deep groove ball bearings in pairs, possessing the same bearing configuration as the real-world electric motors, in pairs. The main shaft conducts bearing tests, and the experimental system encompasses superior levels of automation and failsafe protective mechanisms, facilitating unattended operation for prolonged durations. The vibration accelerations of the test bearings are monitored and recorded throughout the process.

According to the function, the layout of the test bench is included two mainshafts, motor, control cabinet, test systems, oil supply systems, cooling system, oil-gas atomizers and auxiliary equipment. Fig. 1 is a schematic diagram of the spindle structure used for traction motor bearing testing. The spindle is equipped with tested bearings and devices that can withstand radial loads and axial loads, bearing inner-ring temperature telemetry units, and other components. The components inside the sleeve differed when test bearings of different sizes are installed.

2.2. Practical scenario-based problem statement

This study focused on the cross-machine diagnosis of traction motors in high-speed trains by utilizing interpretable wavelet weights. Obtaining large-scale traction motor-bearing data annotation is expensive and time consuming; however, public, synthetic, and self-collected data, which provide a basis for tackling unsupervised cross-machine diagnosis, are available.

Described in Fig. 2, the aim of interest is to resolve cross-machine diagnosis dilemma. Wavelet weights are leveraged to extract the transferable feature mapping from the strongly supervised source-domain $\mathscr{D}_s : \mathscr{X}^s = \{(x_i^s, y_i^s)|i = 1, 2, \ldots, n_s\}$ to the unsupervised target-domain $\mathscr{D}_t : \mathscr{X}^t = \{x_j^t|j = 1, 2, \ldots, n_t\}$. Source-domain gleans three data sources, including public, self-collected, and synthetic, each respectively following the distribution of P_{s_I} and with corresponding to $Y^{s_I} \in \mathbb{R}^R$. Similarly, target-domain gleans from the traction motor of high-speed trains, following the distribution of P_{s_J}.

2.3. Wavelet convolution kernel (WCK)

WCK is from WaveKernelNet. Continuous wavelet convolutional layer is formed based on the convolution operation, and a response is:

$$y_l = \delta(k_l x + b_l), \tag{1}$$

where k_l and b_l means weights and bias, x is raw signals, and δ is an activation function.

In time-domain, wavelet basic dictionary $\psi_{u,s}(t)$ is:

$$\psi_{u,s}(t) = \psi(\frac{t-u}{s}), \tag{2}$$

where $\psi(\cdot)$ denotes wavelet basis function, t is time, s is scaling factor, u is translation factor, s and u are learnable parameters. WCK k is constructed as follows:

$$k = \psi_{u,s}(t). \tag{3}$$

3. WIDAN transfer diagnostic model

3.1. Pipeline of WIDAN diagnostic method

The framework WIDAN is illustrated in Fig. 3. For the source and target domains, the first-layer weights possess specificity, without weight sharing. With the assistance of smooth-augmented wavelet weights without weight sharing, the dual-stream bottleneck layer transforms the data in the source and target domains into a common feature representation space.

The intermediate bottleneck and the top classifiers share weights, drawing upon a typical WDCNN with batch normalization (BN) and a ReLU activation function. The extracted features are fed into a metric function to pursue implicit domain alignment. Finally, a softmax activation function is applied to output the fault category. During the testing process, the test set from the target domain is fed into a well-trained target-domain model to assess the performance of the initialization of wavelet weights.

Apparently, WIDAN elaborately devises a feature space transformation layer that exploits wavelet weights without employing complicated statistical metrics.

3.2. Smooth-augmented wavelet kernel (SWK) and interpretable analysis

The input channel is N_i, output channel is N_k, sampling frequency is f, and convolution kernel size is K.

As illustrated in Fig. 4, Kaiming and Xavier initialization exhibit several key features.

- Condition I: The unit coordinate axes of the filters corresponds to a single values.
- Condition II: The initialized values follow a distribution between −1.0 and 1.0.

The SWK (originating from WIDAN), which is a generic variant of WCK (originating from WaveKernelNet), is devised to address the limitations of WCK. As observed in Table 1, WCK empirically sets parameters to fixed values, which not only ignores the discrepancy in

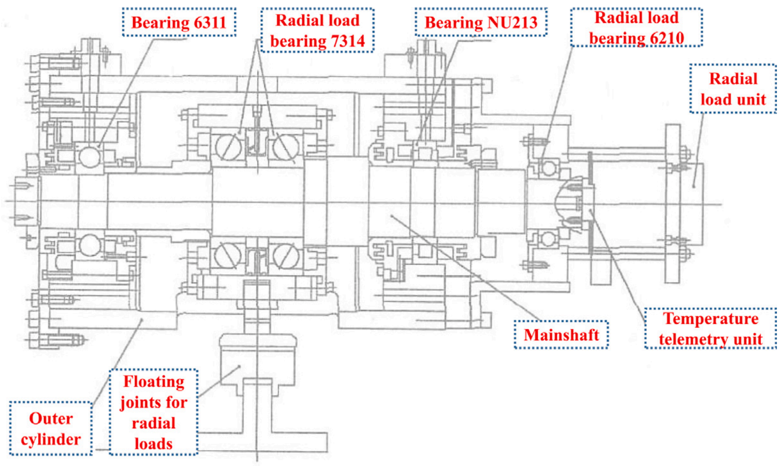


Fig. 1. Schematic diagram of the mainshaft structure of the bearing test bench: the bearing seat of the tested bearing is designed according to the real-world motor structure to simulate the real-world working conditions to the greatest extent possible.

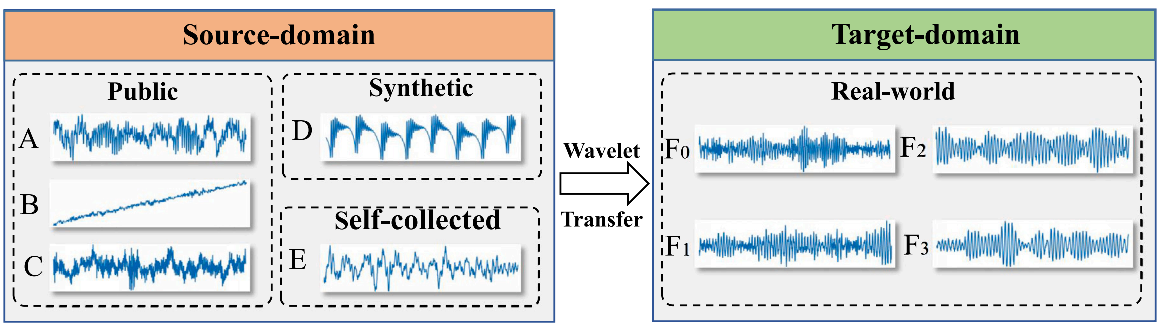


Fig. 2. The description of overall tasks on cross-machine transfer diagnosis. The datasets in source domain are from three public datasets, one self-customized testbed-based dataset, one simulation-based dataset, and target domain is from real-world traction motor platform-based dataset.

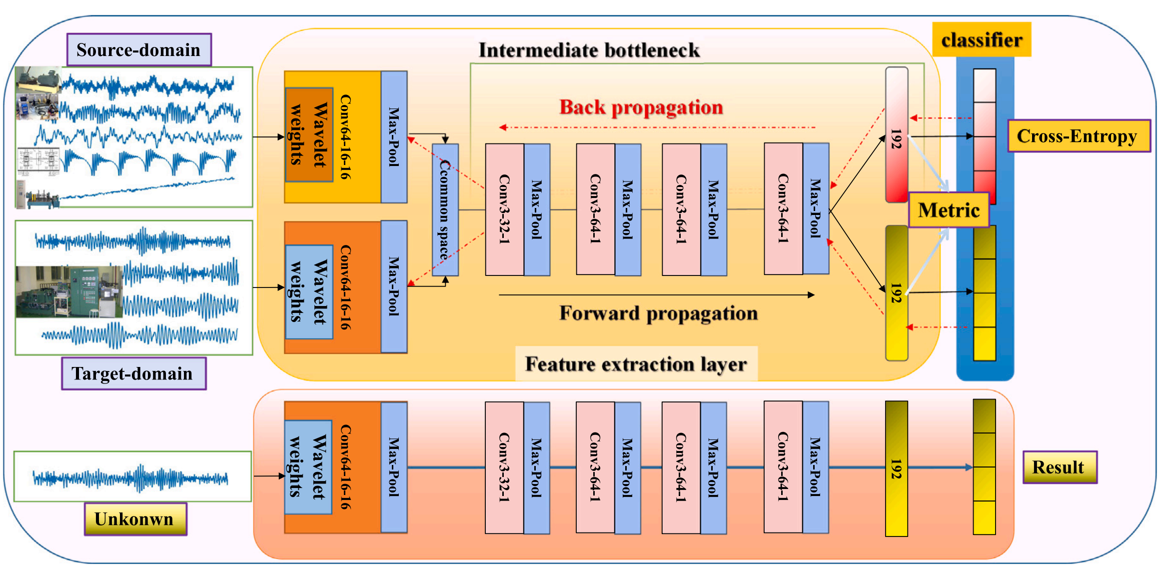
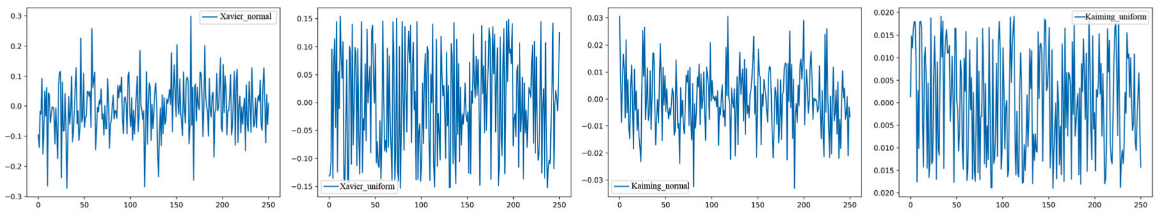


Fig. 3. Framework of WIDAN: ① The source domain and target domain are respectively fed into the feature extraction layer to extract domain-invariant and discriminative fault feature representations. ② The transformation process of the source and target domains into a common feature space is accomplished by utilizing SWK without sharing weights, and meanwhile the domain metric loss can select arbitrary metric function. ③ In testing process, the unknown bearing signals are input into the trained target domain model to produce the final diagnostic precision.



(a) Xavier_normal weights (b) Xavier_uniform weights (c) Kaiming_normal weights (d) Kaiming_uniform weights

Fig. 4. The widely adopted weight initialization. The horizontal axis depicts the convolution kernel size, while the vertical axis denotes the value of weight initialization. Noticeably, these values span all distributed between $(-1, 1)$.

Table 1
Parameter comparison of WCK vs. SWK. In contrast with vanilla WCK, SWK has made notable enhancements in regards to time t, scaling factor s, translation factor u, sampling frequency f, smoothing factor ζ, based on convolution kernel size K and output channels N_k.

Parameter	WCK		SWK
t	$(0, 1)$, step $= K$	$\xrightarrow{revised}$	$(0, K-1)$, step $= K$
s	$(1, 10)$, step $= N_k$	$\xrightarrow{revised}$	$(0, N_k)$, step $= N_k$
u	$(0, 10)$, step $= N_k$	$\xrightarrow{revised}$	$(0, N_k)$, step $= N_k$
f	50 (fixed)	$\xrightarrow{revised}$	Sampling frequency
ζ	w/o	$\xrightarrow{revised}$	$[-0.5, 0.1, 0.2, 0.6, 0.5, 0.9]$
Learnable parameter	s, u	$\xrightarrow{revised}$	$\psi_{u,s}(t)$

sampling frequency across different datasets, but also severs the correlation with the convolution kernel. SWK goes beyond these constraints to achieve a series of ground-breaking improvements.

Consider $N_k = 64$, $K = 300$ as prime examples (Fig. 5). As shown in Figs. 5(a) and 5(d) (originating from WCK), the range of values for the Laplace wavelet is $(-1 \times 10^6, 10^6)$. This fails to satisfy Condition II and to ensure consistency between the order of magnitude of the first-layer value ranges and weights of other layers. Second, the Morlet wavelet takes the values near 0 at the front and back ends. These values do not satisfy Condition I and cannot extract sufficient features.

Firstly, $t \in (0, K-1)$, $step = K$ implies that a time step represents a unit convolution kernel, which establishes a correlation between time t and filter length K, resulting in the production of Figs. 5(b) and 5(e). The above operation imbues the convolution kernel size with the physically interpretable meaning of time. However, Conditions I and II remain unsatisfactory. This is because the Laplace and Morlet values approach zero for a considerable number of time positions and hamper the extraction of the typical feature representation, where the range of values for Laplace wavelets continues to fluctuate wildly. This results in a vanishing-gradient problem.

To tackle the phenomenon, the sizes of $s \in (1, 10)$ and $u \in (0, 10)$ is inconsistent. In view of the output channels, $s \in (1, 10)$ is thus updated to $s \in (0, N_k)$. In this way, the alignment of the scaling and the translation factors is further associated with the output channels to strengthen the correlation with the convolution kernel. Using the above operation, the convolution channel is given a physically interpretable meaning for s and u.

So that, the various wavelet basic dictionaries $\psi_{u_i,s_i}(t)$ corresponding to $s_i, u_i, (i = 0, 1, 2, \dots, N_k)$ on the same time scaling are captured as:

$$\psi_{u_i,s_i}(t) = \psi(\frac{t-u_i}{s_i}), (i = 0, 1, 2, \dots, N_k), \tag{4}$$

where $s \in (0, N_k), step = K; u \in (0, N_k), step = K$.

However, the phenomenon of gradient vanishing arises because of inappropriate wavelet scaling and excessive amplitude variation in the Laplace wavelet dictionary. Therefore, a smoothing factor (ζ) is introduced to form the smooth-augmented wavelet kernel as depicted in Eq. (5), where the value of ζ is the highly crucial determinant of the feature transferability.

$$\psi_{u_i,s_i}(t) = \psi(\frac{t-u_i}{s_i-\zeta}), (i = 0, 1, 2, \dots, N_k). \tag{5}$$

Simultaneously, the exponential component with base e of the wavelet basis function is modified with the Sigmoid function to restrict it to $(0, 1)$. This in turn makes $\psi_{u,s}(t) \in (-1, 1)$, an operation similar to data normalization to ameliorate the consistency of wavelet weights with the initialization of neural network weights. Through a series of modifications, Conditions I and II are satisfied while possessing wavelet interpretable properties.

Laplace wavelet $\psi^L_{u,s}(t)$ and Morlet wavelet $\psi^M_{u,s}(t)$ as depicted in Figs. 5(c) and 5(f), respectively:

$$\psi^L_{u,s}(t) = Ae^{\{\frac{-\xi}{\sqrt{1-\xi^2}} \times 2\pi f(\frac{t-u}{s-\zeta}) - \tau\}.sigmoid()} \times \sin[2\pi f(\frac{t-u}{s-\zeta} - \tau)], \tag{6}$$

$$\psi^M_{u,s}(t) = Ce^{-\frac{\left(\frac{t-u}{s-\zeta}\right)^2}{2}.sigmoid()} \cos\left(5 * \frac{t-u}{s-\zeta}\right), \tag{7}$$

where ξ is the viscous damping ratio. A is a wavelet normalization function, $A = \frac{1}{e}$. τ means a time parameter. C denotes the normalization constant, $C = \frac{\pi^{\frac{1}{4}}}{\sqrt{s+0.01}}$.

The primary distinctions between SWK and WCK are highlighted in Table 1. The limitations inherent in vanilla WCK are modified to devise SWK that is both physically interpretable and compatible with weights of neural networks, where ζ is a significant parameter that regulates how much the scaling factor is smoothed, which in turn influences the distribution of wavelet energy. ζ will be subjected to empirical and experimental analysis.

Compared with the WCK, the SWK is interpretable under the guidance of wavelets. Additionally, through a series of alterations, Condition s I and II are fulfilled to achieve consistency with the other weights of neural networks and extract more robust feature representations. This is a necessary operation because only the first-layer weights must be modified.

3.3. Initialization of wavelet weights for cross-machine transfer

Based on the insight, in this subsection, a robust initialization of weight method is proposed, attempting to mitigating the inter-domain shift presented in dissimilar mechanical fault data, to a certain extent.

For the first conv layer, a response y is:

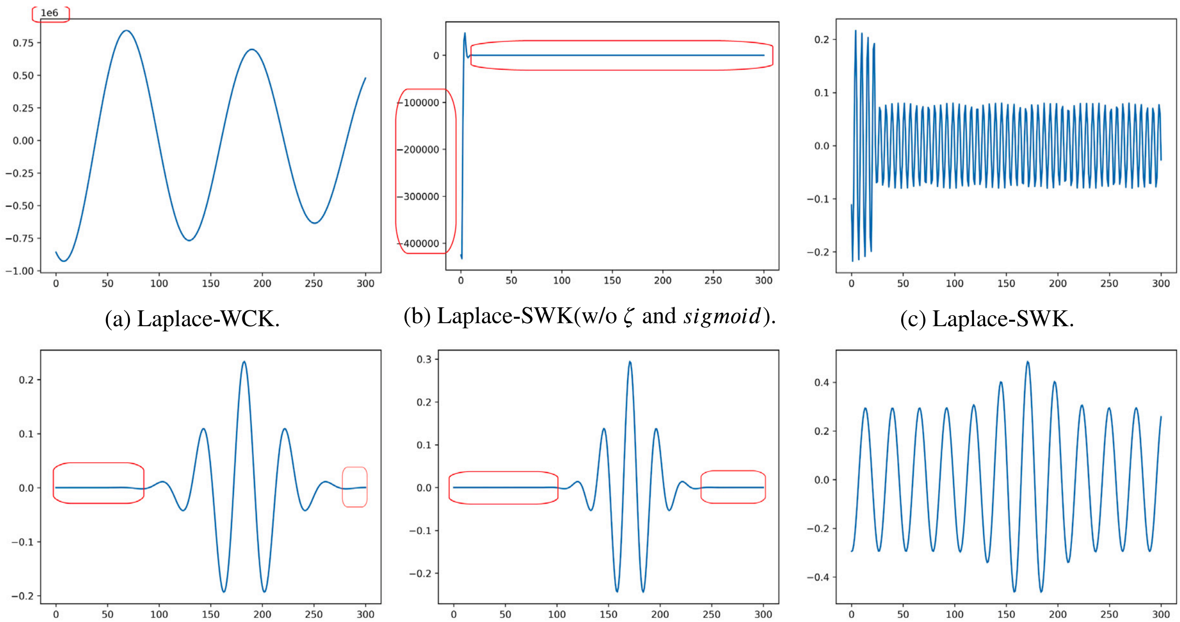
$$y_1 = W_1 x + b_1, \tag{8}$$

where W_1 is weights and b_1 is bias.

The vanilla weights are replaced by the modified wavelet weights: $W_1^s = \psi^L_{u,s}; W_1^t = \psi^M_{u,s}$. The forward propagation mechanism can be expressed as:

$$\begin{cases} y_1^L = \psi^L_{u,s}(t)x + b_1, \\ y_1^M = \psi^M_{u,s}(t)x + b_1, \end{cases} \tag{9}$$

$$\begin{cases} y_l = w_l y^L_{l-1} + b_l (l = 2, \dots, n), \\ y_l = w_l y^M_{l-1} + b_l (l = 2, \dots, n), \end{cases} \tag{10}$$



(a) Laplace-WCK. (b) Laplace-SWK(w/o ζ and $sigmoid$). (c) Laplace-SWK.

(d) Morlet-WCK. (e) Morlet-SWK(w/o ζ and $sigmoid$). (f) Morlet-SWK.

Fig. 5. The weight visualization of WKN and SWK. (a) and (b) represents WCK; (d) and (e) is the intermediate forms of SWK, presenting issues with excessive size or near zero values in weight value distribution ranges; (c) and (f) denotes the proposed SWK.

where l is lth layer. Specifically, $\psi_{u,s}^{1,s} \neq \psi_{u,s}^{1,t} \Leftrightarrow W_1^s \neq W_1^t$.

In addition, as shown by Eqs. (9) and (10), the weights of the deep neural networks are both affected by the first-layer weights, indicating the significance of reasonable the first-layer weight initialization. According to the theory of transfer learning, minimizing a large inter-domain discrepancy facilitates the improvement of the transferability, and we argue that this reasonable first-layer weight initialization contributes to improving the domain transferability.

The backward propagation mechanism can be expressed as:

$$\theta_1 \leftarrow \theta_1 - \alpha(\frac{\partial L_{total}}{\partial \theta_1}), \tag{11}$$

$$\theta_l \leftarrow \theta_l - \alpha(\frac{\partial L_{total}}{\partial \theta_l})(l = 2, \ldots, n), \tag{12}$$

where α represents learning, θ_1 denotes the parameters to be updated in the first layer, and L_{total} is the total loss.

In Table 1, the another key distinction between SWK and WCK is that WCK has two learnable parameters s, u, while the proposed method has learnable parameters $\psi_{u,s}(t)$ as weights.

3.4. Optimization objective

The source-domain datasets are derived from laboratory or simulation scenarios with access to abundant annotated data. Cross-entropy loss is employed for directing parameter updating.

$$L_{cls} = -\sum_{k}^{K} \log(p(k))q(k), \tag{13}$$

where $p(k)$ is predicted distribution, $q(k)$ is real distribution.

The crux of this work is not to design more precise statistical metrics; rather, typical arbitrary typical statistical metrics, such as Maximum Mean Discrepancy (MMD), Sliced Wasserstein Distance (SWD), and CORAL, can be utilized. The final loss function is formulated as follows.

$$L_{total} = L_{cls} + \lambda L_{metric}, \tag{14}$$

where λ is a trade-off coefficient that adjusts the proportion of the corresponding loss term during back propagation, which is obtained as follows [44]:

$$\lambda = \frac{-4}{\sqrt{\frac{epoch}{\max\_epoch} + 1}} + 4. \tag{15}$$

The specific training procedure is displayed in Algorithm 1.

Algorithm 1 The training process of WIDAN.

```
Input: W_1^s = ψ_{u,s}^{1,s}; W_1^t = ψ_{u,s}^{1,t}; max_epoch
Output: Parameters of WIDAN: θ
  while not converged do
    if epoch < max_epoch then
      Forward based on Eq. (9) and Eq. (10)
      Update λ based on Eq. (15)
      Calculate L_total based on Eq. (14)
      Update θ based on Eq. (11) and Eq. (12)
    end if
  end while
```

4. Case analysis and discussion

4.1. Description of experimental set-up and datasets

The detailed description of datasets can be seen in Table 2.

4.1.1. Public datasets

The CWRU [45] dataset widely serves as a benchmark for fault diagnosis and includes raw signals from four states: normal (NC), inner ring fault (IR), outer ring fault (OR), and ball fault (BF). The Jiangnan University (JNU) [46] dataset faces greater difficulty in identifying faults by preliminary experimental analysis. The DDS [47] bearing data is acquired using a Drivetrain Dynamics Simulator (DDS) from the Southeast University without illustrating the sampling frequency.

Fig. 6. DDB: rotor-gear comprehensive fault experiment platform, which comprises three failure state: inner ring, outer ring, ball.

Table 2
The detailed description of datasets.

Name	Data source	Speed	Sampling frequency	Health status
A	CWRU	1797 r/min	48 kHz	
B	DDS	20		
C	JNU	600 r/min	50 kHz	
D	Simulation		10 kHz	
E	DDB	900 r/min	20 kHz	N,IF,OF,BF
F_0		200 km/h		
F_1	NTN	250 km/h	100 kHz	
F_2		300 km/h		
F_3		350 km/h		

4.1.2. Synthetic dataset

A simulation dataset [48] is derived from a rotor-bearing system simulation model with signals sampled at 10 kHz. The details of the public and synthetic datasets can be obtained from the corresponding literature.

4.1.3. Self-collected dataset

As shown in Fig. 6, a double-span dual-rotor fault test bench, comprises a control cabinet, variable-frequency motor, drive motor, loading device, single-span rolling-bearing rotor system, and sensors. Vibration acceleration signals are gleaned using a triaxial integrated circuits piezoelectric (ICP) sensor with sampling frequency 20 kHz, placed on the bearing housing of a deep-groove ball bearing (NSK-6308).

4.1.4. Traction motor bearing dataset

A dataset is obtained from the NTN traction motor bearing bench, which is a specialized equipment with an advanced international level and designed for high-speed train traction motor bearing experiments, as shown in Fig. 7. A vibration acceleration sensor is positioned on the outer side of the bearing. The bearing model used for the experiment is HRB NU214 EM 32214H, with the same dimensions as the actual bearings. The experiment is conducted with four types of mechanical states, including three types of failures (IF, OF, and BF), and a healthy bearing (NC), at a sampling frequency of 100 kHz. The bearing rotational speeds are equivalent to the speeds of high-speed trains (200, 250, 300 and 350 km/h).

4.2. Implementation details

The purpose of this research is to solve the diagnosis problem of traction motors in high-speed trains; therefore, only F is used as the target domain. To comprehensively evaluate the effectiveness of the initialization wavelet weights for cross-machine diagnosis, three datasets and a real-world dataset are collected. Each dataset is split into samples by employing sliding sampling with a length of 1024. The total number of samples is 480 K, where 240 K samples constitute the source and target domains, while the remaining 240 K samples formed the test

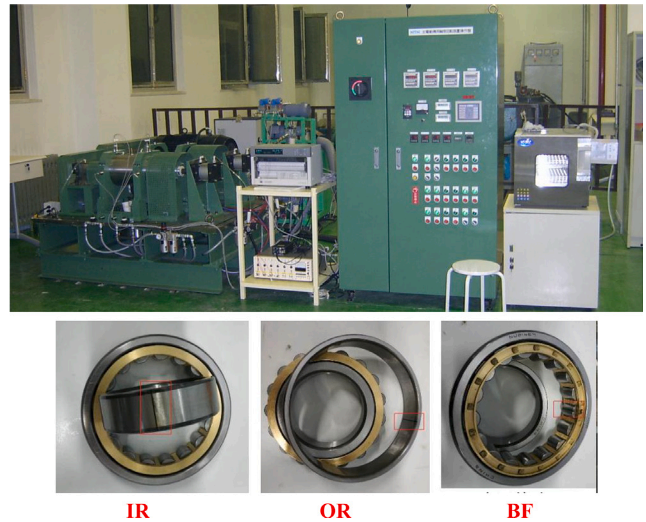


Fig. 7. NTN: real-world high-speed train traction motor bearing platform and failure material objects.

Table 3
Basic hyper-parameters of WIDAN.

Metric loss	Learning_rate	Decay	Epoch	Batch_size	Early stopping	ζ
MMD	0.001	150	200	32	150	0.5

Table 4
The training time of WIDAN on F_2 (NTN at 300 km/h).

Task	$A \rightarrow F_2$	$B \rightarrow F_2$	$C \rightarrow F_2$	$D \rightarrow F_2$	$E \rightarrow F_2$
Training time (s)	235.28	271.19	274.72	268.86	283.79

set, drawn from an additional target domain (where K is the number of categories). Multiple transfer tasks are constructed from A, B, C, D, E to F_0, F_1, F_2, F_3. $A \rightarrow F_0$ denotes from A as the source-domain and F_0 as the target-domain. Also, the models are consistently implemented on PyTorch 1.12.0 with the GPU of NVIDIA Tesla V100 32 GB. The parameter count of WIDAN is 48324.

The SOTA WIDAN is evaluated with respect to several typical benchmark approaches. These benchmarks encompassed: BN-WDCNN without transfer strategy; conventional methods based on statistical metrics (such as DDC, DCORAL, DJDA, and DSAN); adversarial-based methods (such as DANN); SOTA methods for cross-machine diagnosis (CK-CNN [49], DDTLN [50]); and interpretable models applied for this scenario (including NTScatNet [36], WaveletkernelNet [38], DFAWNet [39], SincNet [41], GTFENet [51], GaborNet [52]. MWPN [37]) incorporates the WCK architecture originally proposed by WaveletKernelNet. Code is available at https://github.com/liguge/WIDAN.

For a fair comparison, all the algorithms are rigorously configured based on the parameters specified in corresponding studies. The other parameter settings based on the experience are listed in Table 3. All experiments are performed five times to obtain the mean and standard deviation values. The sampling frequency of B has not been publicly disclosed. Therefore, when B is identified as either the source or target domain, Morlet wavelet weights are employed. ζ is a key parameter that has been elaborated on.

4.3. Discussion and analysis

4.3.1. Generalized experimental analysis

BN-CNN denotes the absence of a dual-stream module, metric loss, and wavelet weight initialization. Table 4 lists the training times of WIDAN. The empirical conclusions for multiple tasks are shown in

Table 5
Performance comparison with baseline methods on dataset F_0 (NTN at 200 km/h). 'Average' is the average performance of a method across various dedicated tasks.

Method	$A \rightarrow F_0$	$B \rightarrow F_0$	$C \rightarrow F_0$	$D \rightarrow F_0$	$E \rightarrow F_0$	Average
BN-CNN	29.77	25.17	48.09	27.95	37.17	33.63
DDC	62.87 ± 0.19	53.59 ± 2.23	68.36 ± 1.86	93.56 ± 5.54	21.78 ± 5.50	60.03
DCORAL	31.21 ± 2.65	30.38 ± 1.01	48.50 ± 5.24	33.62 ± 4.66	48.61 ± 14.15	38.64
DANN	47.86 ± 2.72	36.28 ± 6.16	39.75 ± 1.30	29.17 ± 4.26	44.39 ± 2.08	39.49
DDTLN [50]	38.02 ± 0.94	16.25 ± 2.56	64.63 ± 0.13	42.81 ± 0.01	51.92 ± 0.02	42.73
MWPN [37]	26.56 ± 0.78	41.45 ± 2.60	37.87 ± 6.36	31.25 ± 0.26	16.67 ± 3.91	30.76
CK-CNN [49]	40.80 ± 6.55	35.42 ± 5.18	76.50 ± 1.07	42.88 ± 5.35	52.00 ± 11.81	49.52
DJDA	41.93 ± 3.52	34.72 ± 4.86	64.24 ± 1.29	51.33 ± 8.06	49.31 ± 0.70	48.31
DSAN	43.84 ± 4.07	45.49 ± 6.68	70.25 ± 10.19	30.15 ± 11.63	45.14 ± 9.29	46.97
WIDAN	**99.63** ± **0.37**	**98.29** ± **0.30**	**99.80** ± **0.07**	**99.14** ± **0.78**	**94.30** ± **4.98**	**98.23**

Table 6
Performance comparison with baseline methods on dataset F_1 (NTN at 250 km/h).

Method	$A \rightarrow F_1$	$B \rightarrow F_1$	$C \rightarrow F_1$	$D \rightarrow F_1$	$E \rightarrow F_1$	Average
BN-CNN	25.17	25.52	25.00	37.67	27.95	28.26
DDC	95.76 ± 4.24	**99.93** ± **0.07**	42.37 ± 7.92	64.23 ± 0.23	69.95 ± 0.81	74.45
DCORAL	27.43 ± 1.65	35.59 ± 12.07	26.04 ± 5.64	48.78 ± 2.14	25.43 ± 6.29	32.65
DANN	21.99 ± 6.19	28.65 ± 1.05	25.62 ± 0.70	28.12 ± 1.22	54.91 ± 0.03	31.86
DDTLN [50]	44.10 ± 0.89	27.03 ± 5.06	52.04 ± 0.19	38.64 ± 0.02	47.29 ± 0.02	41.82
MWPN [37]	23.70 ± 1.30	40.62 ± 5.99	28.65 ± 1.71	57.81 ± 3.91	23.44 ± 4.17	34.84
CK-CNN [49]	53.99 ± 2.08	24.88 ± 10.68	37.84 ± 5.50	85.24 ± 0.35	29.51 ± 10.85	46.29
DJDA	52.46 ± 1.93	41.84 ± 19.53	73.73 ± 2.58	50.00 ± 0.26	34.38 ± 16.15	50.48
DSAN	43.68 ± 9.11	24.88 ± 8.33	25.35 ± 12.56	46.88 ± 0.28	27.78 ± 13.63	27.78
WIDAN	**98.55** ± **1.08**	99.78 ± 0.07	**99.40** ± **0.07**	**97.85** ± **0.59**	**99.64** ± **0.29**	**99.04**

Table 7
Performance comparison with baseline methods on dataset F_2 (NTN at 300 km/h).

Method	$A \rightarrow F_2$	$B \rightarrow F_2$	$C \rightarrow F_2$	$D \rightarrow F_2$	$E \rightarrow F_2$	Average
DDC	32.00 ± 0.75	**99.77** ± **0.16**	71.13 ± 1.19	73.76 ± 0.03	97.76 ± 0.10	74.88
DANN	45.16 ± 4.58	25.14 ± 9.00	27.31 ± 11.81	44.97 ± 4.40	28.30 ± 11.72	34.18
DCORAL	32.81 ± 3.07	28.07 ± 1.91	28.99 ± 1.31	50.00 ± 3.13	56.60 ± 0.17	39.29
DDTLN [50]	41.25 ± 4.99	32.02 ± 0.78	45.05 ± 2.40	74.27 ± 0.01	48.64 ± 0.01	48.25
MWPN [37]	26.56 ± 1.82	23.44 ± 1.56	28.72 ± 1.75	44.27 ± 4.69	56.25 ± 2.08	35.85
CK-CNN [49]	42.41 ± 2.38	23.21 ± 8.25	50.49 ± 5.27	**99.86** ± **0.14**	48.75 ± 11.20	52.94
DJDA	98.88 ± 0.07	47.69 ± 4.75	35.82 ± 2.55	98.75 ± 0.38	43.32 ± 5.63	64.89
DSAN	47.10 ± 7.51	28.94 ± 12.79	36.52 ± 11.0	97.31 ± 2.69	32.20 ± 14.06	48.41
WIDAN	**99.97** ± **0.03**	99.02 ± 0.81	**99.18** ± **0.60**	99.68 ± 0.03	**99.48** ± **0.26**	**99.47**

Table 8
Performance comparison with baseline methods on dataset F_3 (NTN at 350 km/h).

Method	$A \rightarrow F_3$	$B \rightarrow F_3$	$C \rightarrow F_3$	$D \rightarrow F_3$	$E \rightarrow F_3$	Average
DDC	48.85 ± 10.08	99.71 ± 0.23	73.44 ± 2.83	95.52 ± 0.90	58.37 ± 2.51	75.18
DANN	39.06 ± 5.30	47.05 ± 8.16	29.11 ± 2.05	47.78 ± 1.36	24.87 ± 12.08	37.57
DCORAL	65.91 ± 10.04	49.47 ± 3.64	37.33 ± 1.31	69.20 ± 7.81	65.00 ± 2.17	57.38
DDTLN [50]	15.15 ± 0.50	16.09 ± 2.24	37.76 ± 1.38	46.30 ± 0.02	38.54 ± 0.01	30.77
MWPN [37]	31.77 ± 2.56	42.19 ± 3.91	27.08 ± 6.25	50.00 ± 0.26	11.46 ± 5.26	32.50
CK-CNN [49]	31.47 ± 0.67	99.28 ± 0.72	36.67 ± 7.52	60.53 ± 10.50	30.73 ± 3.53	51.74
DJDA	34.81 ± 9.81	72.47 ± 2.01	75.58 ± 4.37	**99.22** ± **0.78**	55.00 ± 14.97	67.42
DSAN	30.80 ± 2.98	23.50 ± 12.56	35.35 ± 12.56	34.32 ± 16.96	26.37 ± 4.98	30.07
WIDAN	**99.89** ± **0.11**	**99.93** ± **0.03**	**99.81** ± **0.19**	99.13 ± 0.12	**99.97** ± **0.03**	**99.75**

Tables 5–8. Monitoring the service status of the traction motor bearings from multiple data sources cannot be accomplished without transfer learning. However, the initialization of wavelets can significantly improve the accuracy to over 99%, indicating that the initialization of wavelets and dual-stream transform layer can significantly diminish the domain discrepancy, thereby facilitating the training stage.

Compared with other statistical measures, owing to the significant distribution shift between the datasets from different machines, relying solely on statistical metrics can result in a negative transfer and poor performance compared to BN-CNN. The efficacy of purely data-driven approaches varies depending on the task at hand, with certain tasks proving to be robust and others demonstrating subpar performances. Wavelet initialization can be considered as a flexible form of data preprocessing. Distinct from general data preprocessing, the initialization of wavelets has been integrated into an end-to-end WIDAN.

Purely data-driven SOTA methods have high benchmarks for the quality and quantity of data, which are not always feasible in real-world scenarios owing to low quality, limited data, or low sample length. However, the wavelet weights can be deemed as a valid method to alleviate this dependency. In summary, the initialization of wavelets can improve data quality.

The impact on several well-known metrics is evaluated as shown in Fig. 8 (Task: $A \rightarrow F_0$). As per this, the impact of statistics on performance is decisive; however, for poorer metrics, WIDAN can significantly boost performance. It can upgrade by at least 10% or more. This confirms that the wavelet weight initialization indeed diminishes the inter-domain discrepancy.

The initialization of wavelet weights serves the dual purpose of maintaining interpretability and the consistency of the weight distribution of neural networks. The experiments in Table 9 (Task: $A \rightarrow F_2$) showcase that WIDAN, an interpretable and consistent model,

Table 9
Performance comparison with signal mechanism-guided backbones on F_2 (NTN at 300 km/h).

Method	$A \rightarrow F_2$	$B \rightarrow F_2$	$C \rightarrow F_2$	$D \rightarrow F_2$	$E \rightarrow F_2$	Average
NTScatNet [36]	26.56 ± 1.82	23.44 ± 1.56	28.72 ± 1.75	44.27 ± 4.69	56.25 ± 2.08	35.85
WaveletkernelNet [38]	50.00 ± 6.57	44.56 ± 12.21	42.71 ± 5.10	**100.00 ± 0.00**	35.13 ± 10.61	54.48
DFAWNet [39]	54.79 ± 0.08	34.53 ± 1.04	53.91 ± 1.06	68.39 ± 0.94	96.00 ± 0.89	61.53
SincNet [41]	66.93 ± 3.19	**99.22 ± 7.06**	73.91 ± 0.51	99.39 ± 0.57	99.44 ± 0.19	87.78
GTFENet [51]	49.38 ± 6.93	44.95 ± 4.60	46.67 ± 10.60	51.25 ± 8.28	46.51 ± 10.55	47.75
GaborNet [52]	25.00 ± 1.47	38.49 ± 4.29	29.22 ± 2.99	50.68 ± 4.48	25.00 ± 5.32	33.68
$\text{WIDAN}_{\text{(ours)}}$	**99.97 ± 0.03**	99.02 ± 0.81	**99.18 ± 0.60**	99.68 ± 0.03	**99.48 ± 0.26**	**99.47**

significantly improves the performance and exhibits lower sensitivity to various datasets compared to models that solely prioritize interpretability. Although WIDAN is inferior to certain models for certain tasks, it exhibits exceptional stability and robustness, with an average accuracy improvement of 11.69% compared to SincNet, and on average, WIDAN improves the accuracy by 37.94% over DFAWNet.

Exactly, from Tables 7 and 8, we can see that the baselines such as DDC, CK-MMD, and DJDA show better performance in certain tasks. Some tasks can yield commendable results utilizing baselines, which is related to the smaller distribution discrepancy between the source domain and target domain in certain tasks such as $B \rightarrow F_2$, $D \rightarrow F_2$, $D \rightarrow F_3$, $B \rightarrow F_3$. These methods can effectively alleviate distribution shifts between datasets, but this advantage is limited to the aforementioned tasks. Conversely, its performance in other tasks is poor, which indicates that the purely data-driven methods rely on the quality of the source and target domain datasets. The difference between these baselines and WIDAN lies in the former's reliance on random initialization while the latter employs SWK initialization. utilizing physics-informed wavelet weights allows the method to demonstrate better robustness, stability, and the best average accuracy in multiple tasks. WIDAN marginally underperforms certain baselines, possibly due to the influence of the smoothing factor ζ in SWK. As shown in Table 11, the reasonableness of the smoothing factor directly affects the final accuracy, which also indicates the importance of reasonable first-layer weight initialization on cross-machine transfer diagnosis. Indeed, owing to the acceptable accuracy gap, ζ need not to adjust meticulously.

4.3.2. Ablation experiments

From Table 10, B and C respectively employ Morlet and Laplace wavelets with $\zeta = -0.5$. a_{21} and b_{21} additionally add Sigmoid on the basis of a_2 (Fig. 5(b)) and b_2 (Fig. 5(e)). (w/o) denotes random initialization exploited, (0.3) denotes the fraction of the test set, and a_1 and b_1 denote a purely interpretable wavelet. The accuracy of the identification reflects the need for improvement at each step.

'NAN' denotes the gradient vanishing occurring in the process of training, leading to training disruption. The ultimate output is 'NAN'. As illustrated in a_2 (Fig. 5(b)) and b_2 (Fig. 5(e)), the weight of a_2 is predominantly situated within $(-500000, 0)$, whereas the weight range of generic neural networks spans between $(-1, 1)$, which represents a considerable order of magnitude difference between a_2 and the weight of generic neural networks. Therefore, the gradient is incapable of a smooth propagation, a phenomenon known as gradient vanishing. Furthermore, the partial weights of a_2 and b_2 are approximately zero, a condition that can induce the gradient to become excessively small, culminating in gradient vanishing and ultimately resulting in an output loss of the 'NAN'. This, in turn, disrupts the training process.

As can be seen from a_3 vs. $a_{(w/o)}$ and b_3 vs. $b_{(w/o)}$, wavelet weight initialization can be improved by more than 30% over random initialization. A purely physically interpretable model can compromise the performance or even lead to training failure owing to gradient update issue. To address this challenge, the interpretable component devised should not only exhibit basic physical interpretability but also reflect consistency with neural networks. However, SWK is more time-consuming than WCK, but it is less time-consuming than random initialization. Furthermore, it can achieve a trade-off in terms of time and accuracy.

Table 10
Ablation experimental performance of SWK, and (0.3) denotes the fraction of test set. 'NAN' denotes training disruption.

	$C \rightarrow F_3(0.3)$			$B \rightarrow F_3(0.7)$	
	Accuracy (%)	Training time (s)		Accuracy (%)	Training time (s)
a_1	21.28	123.71	b_1	29.99	73.43
a_2	NAN		b_2	NAN	
a_{21}	4.17	114.37	b_{21}	61.31	97.70
$a_{(w/o)}$	61.11	170.46	$b_{(w/o)}$	69.97	99.51
a_3	100.00	149.57	b_3	100.00	97.87

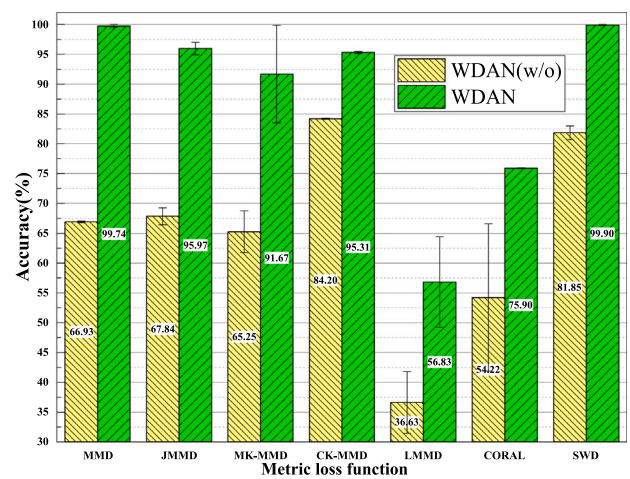


Fig. 8. Performance of wavelet weights for typical metrics on Task $C \rightarrow F_3$.

4.3.3. Parameter sensitivity analysis about ζ

The technical smoothing factor ζ assists the scaling factor s in controlling the degree of energy concentration of the wavelets. However, the estimation of ζ currently lacks an effective solution. Hence, we evaluate the performances of $\zeta \in [-0.5, 0.5]$ listed in Table 11 by fixing the other hyperparameter to the optimal one. For Task $C \rightarrow F_3$, the robustness can be obtained when $a = -0.3, 0.1, 0.5$. Choosing an appropriate ζ is crucial. An inappropriate value leads to a negative transfer and destroys the similarity distribution of the data. Additionally, different tasks correspond to different smooth-assisted factors ζ. The precise evaluation of these factors will be the focus of future research. Apart from it, other parameters are fixed and have not been discussed specifically for analysis.

4.4. Domain transferability analysis

The $\mathscr{A}$-distance is extensively adopted to evaluate the similarity between two domains and is defined as $\mathscr{A}(\mathscr{D}_s, \mathscr{D}_t) = 2(1 - 2err(h))$, where $err(h)$ is the generalization error of the binary classifier in the source-and target-domain. With regard to accuracy, Fig. 8 demonstrates initialization of wavelet weights has an enhancement for various statistical metrics. And Fig. 9 quantitatively showcases it decreases inter-domain discrepancy and promotes similarity across domains.

Table 11
The accuracy under different smoothing factor ζ for Task $C \rightarrow F_3$.

ζ	−0.5	−0.4	−0.3	−0.2	−0.1	0.1	0.2	0.3	0.4	0.5
Acc (%)	34.72 ± 3.30	88.37 ± 11.63	98.61 ± 1.39	47.31 ± 9.46	70.75 ± 0.43	**99.83** ± **0.17**	33.07 ± 2.52	91.67 ± 8.33	37.07 ± 0.26	99.74 ± 0.26

Table 12
Performance comparison with other methods on public dataset A, B, C.

Method	$A \rightarrow B$	$A \rightarrow C$	$B \rightarrow A$	$B \rightarrow C$	$C \rightarrow A$	$C \rightarrow B$	Average
DDC	40.39 ± 1.01	42.30 ± 0.43	25.64 ± 1.10	66.38 ± 3.56	25.69 ± 0.26	27.20 ± 1.77	37.93
DCORAL	39.93 ± 0.14	36.28 ± 6.22	26.39 ± 1.04	67.94 ± 1.48	25.58 ± 9.78	30.09 ± 3.18	37.70
DANN	47.74 ± 0.03	47.51 ± 0.86	47.74 ± 0.41	57.17 ± 0.93	44.50 ± 0.95	52.26 ± 0.69	49.49
DDTLN	46.35 ± 0.01	51.51 ± 0.36	30.00 ± 0.01	45.10 ± 0.00	70.83 ± 0.03	35.36 ± 0.01	46.53
CK-CNN	42.13 ± 0.67	42.94 ± 4.83	27.20 ± 4.63	65.97 ± 7.06	26.56 ± 1.10	30.61 ± 3.85	39.24
DJDA	39.41 ± 2.31	45.83 ± 1.91	33.33 ± 8.19	60.42 ± 1.56	26.22 ± 2.26	27.84 ± 2.60	38.84
DSAN	40.97 ± 5.32	43.75 ± 2.23	31.25 ± 7.09	62.91 ± 5.03	26.10 ± 5.24	29.05 ± 3.56	39.01
WIDAN	**91.75** ± **8.19**	**82.99** ± **0.06**	**82.23** ± **6.05**	**100** ± **0.00**	**99.02** ± **0.98**	**98.26** ± **0.75**	**92.38**

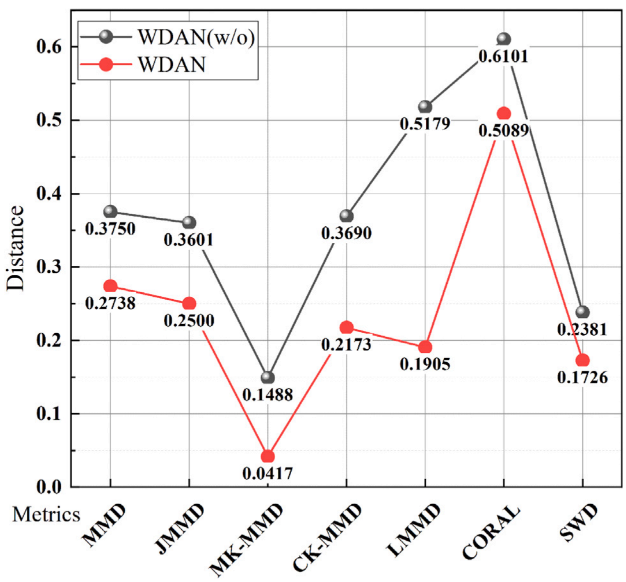


Fig. 9. $\mathscr{A}$-distance for typical metrics on Task $C \rightarrow F_3$.

4.5. Experimental analysis for reproducibility

To ensure the reproducibility of the reported results, the performances of the three public datasets are testified. As shown in Table 12, WIDAN demonstrates better generalization. It exhibits better robustness for different datasets. The results are less effective compared to the previous ones on $A, B, C, D, E \rightarrow F_0, F_1, F_2, F_3$. This indicates that it has requirements on the quality of the dataset. Additionally, high-quality source-domain data can enhance transfer learning accuracy.

These well-known baselines exhibit competitive performance in cross-domain tasks such as cross-speed, cross-sensor, and cross-load. However, the data distribution discrepancy of cross-machine tasks is greater, rendering the challenge greater. These metric-based and data-driven domain matching methods are incapable of addressing the aforementioned problems due to the dependence on data quality, resulting in negative transfer. There is a greater distribution shift in transfer diagnosis between public datasets, such as CWRU, JNU, and DDS, the tasks of which are more difficult, so the average accuracy is relatively not good. The said approaches underperform notably in scenarios on relatively limited data. In the experiment, there are 240 data in each class. When the data amount is small, the performance of these methods will degrade.

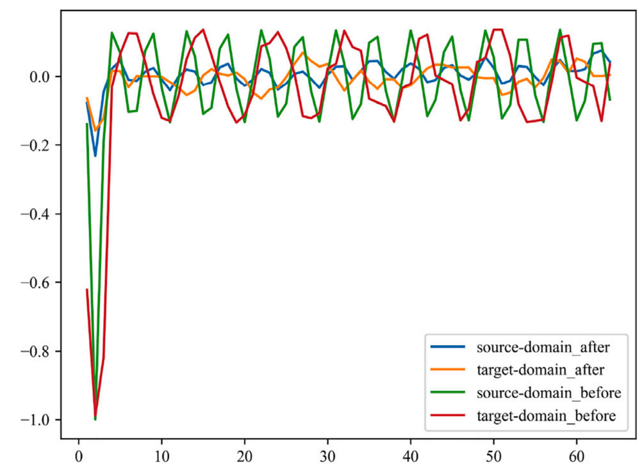


Fig. 10. SWK initialization before and after training. It indicates that SWK adheres to the physical principles, where the convolutional kernel size is 64.

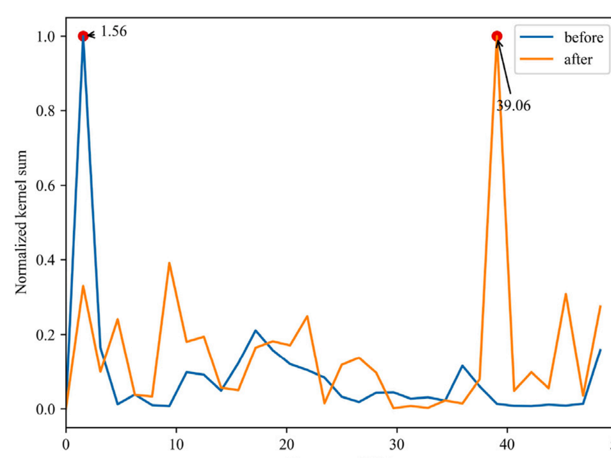


Fig. 11. Cumulative frequency band of the SWK initialization. The cumulative frequency band has transited from 1.56 kHz (before training) up to 39.06 kHz (after training), located near the four fault frequency bands. Per this observation, it indicates that SWK adepts at to extract superior fault frequency bands.

4.6. The interpretability analysis of SWK

As depicted in Fig. 10, WIDAN establishes dual-stream modules with distinct weights to facilitate the transformation of different data into a common pattern space. The proposed SWK initialization assists in guiding the search for optimal weights, thereby improving domain

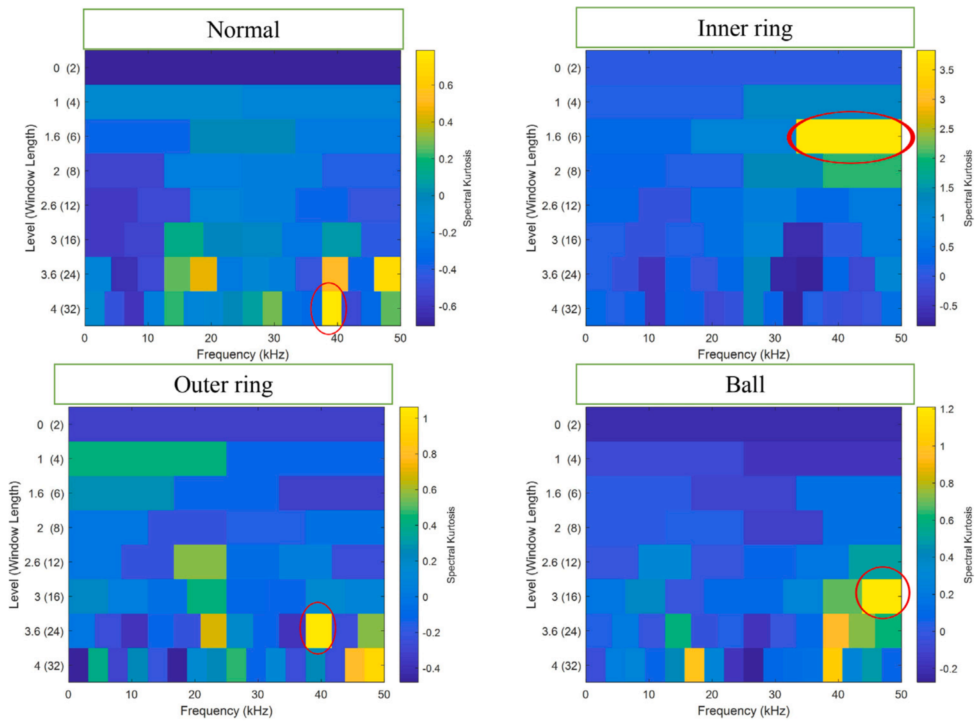


Fig. 12. The paving spectral kurtosis values and their associated frequency bands for signals in different states: normal state (the optimal central frequency is 39.06 kHz and the bandwidth is 3.13 kHz); inner race fault (the optimal central frequency is 41.67 kHz and the bandwidth is 16.67 kHz); outer race fault (the optimal central frequency is 39.58 kHz and the bandwidth is 4.17 kHz); ball fault (the optimal central frequency is 46.88 kHz and the bandwidth is 6.25 kHz).

transferability. The variation trends of the weights before and after training remain largely similar.

The weight initialization method proposed in this study is clarified based on the analysis of interpretability in Ref. [8]. In transfer learning analysis, the target domain accuracy is most pivotal. Thus, focusing on the target domain involves the calculation of the cumulative frequency band of the SWK initialization. This calculation encompasses all SWKs, with the center frequency ranging from 1.56 to 39.06 kHz (see Fig. 11). As shown in Fig. 12, the calculated center frequency is within the normal, inner, and outer ring bandwidths. Furthermore, it is also close to the bandwidth of the rolling element. Consequently, the optimal frequency band is identified under the guidance of the reasonable initialized weights.

5. Conclusion

A physics-informed cross-machine adaptation network (WIDAN) that creatively regards smooth-assisted wavelets as weights, is devised. This approach maintains a certain degree of physically intrinsic interpretability, while simultaneously ensuring consistency with optimization processes. Notably, the framework is a valid solution for cross-machine diagnosis, as confirmed by public, synthetic, and self-collected datasets from three different sources, along with real-world high-speed train data. In conclusion, this study builds a potential guideline indicating that the reasonable first-layer weight initialization is extremely valid for cross-machine diagnosis. Through the initialization of the physics-informed weights, the following empirical conclusions can be drawn:

1. Knowledge embedded or physics-informed interpretable backbone can facilitate cross-machine transfer.
2. The novel bottleneck should not only prioritize intrinsic interpretability, but also consider its cooperation with the optimization process of neural networks, taking into account factors such as gradient, activation function, and normalization.

On one hand, it is sensitive to the parameter ζ, which is a challenging task to evaluate precisely and quantitatively. On the other hand, the approach based on constant rotational speed performs well, but the scenario on speed sharp fluctuations is warranted to be explored in the future, especially to devise a specialized initialization method of weights.

As we all know, there is no general method to solve the Analytical solution of the nonlinear partial differential equation. The symbol calculation method based on neural networks proposed by Zhang et al. [53–56] open up a general symbolic computing path for the analytic solution of Nonlinear partial differential equation, and lays the foundation for the universal method of symbolic calculation of Analytical expression. The problems studied in this paper can be solved by using this method in the future work.

CRediT authorship contribution statement

Chao He: Writing – original draft, Software, Validation, Visualization, Methodology, Investigation, Proofreading. **Hongmei Shi:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Xiaorong Liu:** Writing – review & editing. **Jianbo Li:** Supervision.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors are grateful for the supports of the National Natural Science Foundation of China (No. 52272429), and State Key Laboratory of Advanced Rail Autonomous Operation (Contract No. RAO2023ZZ003).

References

[1] Z. Zhao, Q. Zhang, X. Yu, C. Sun, S. Wang, R. Yan, X. Chen, Applications of unsupervised deep transfer learning to intelligent fault diagnosis: A survey and comparative study, IEEE Trans. Instrum. Meas. 70 (2021) 1–28, http://dx.doi.org/10.1109/TIM.2021.3116309.

[2] W. Li, R. Huang, J. Li, Y. Liao, Z. Chen, G. He, R. Yan, K. Gryllias, A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges, Mech. Syst. Sig. Process. 167 (2022) 108487.

[3] I. Misbah, C. LEE, K. KEUNG, Fault diagnosis in rotating machines based on transfer learning: literature review, Knowl.-Based Syst. 283 (2024) 111158.

[4] S. Luo, X. Huang, Y. Wang, R. Luo, Q. Zhou, Transfer learning based on improved stacked autoencoder for bearing fault diagnosis, Knowl.-Based Syst. 256 (2022) 109846.

[5] Y. Zhang, S. Li, A. Zhang, C. Li, L. Qiu, A novel bearing fault diagnosis method based on few-shot transfer learning across different datasets, Entropy 24 (9) (2022) 1295.

[6] Y. Li, W. Jiang, G. Zhang, L. Shu, Wind turbine fault diagnosis based on transfer learning and convolutional autoencoder with small-scale data, Renew. Energ. 171 (2021) 103–115.

[7] X. Liu, J. Chen, K. Zhang, S. Liu, S. He, Z. Zhou, Cross-domain intelligent bearing fault diagnosis under class imbalanced samples via transfer residual network augmented with explicit weight self-assignment strategy based on meta data, Knowl.-Based Syst. 251 (2022) 109272.

[8] T. Han, C. Liu, R. Wu, D. Jiang, Deep transfer learning with limited data for machinery fault diagnosis, Appl. Soft Comput. 103 (2021) 107150.

[9] Q. Qian, J. Zhou, Y. Qin, Relationship transfer domain generalization network for rotating machinery fault diagnosis under different working conditions, IEEE Trans. Ind. Inform. 19 (9) (2023) 9898–9908.

[10] Y. Shen, B. Chen, F. Guo, W. Meng, L. Yu, A modified deep convolutional subdomain adaptive network method for fault diagnosis of wind turbine systems, IEEE Trans. Instrum. Meas. 71 (2022) 1–10, http://dx.doi.org/10.1109/TIM.2021.3128708.

[11] Y. Xiao, H. Shao, S. Han, Z. Huo, J. Wan, Novel joint transfer network for unsupervised bearing fault diagnosis from simulation domain to experimental domain, IEEE/ASME Trans. Mech. 27 (6) (2022) 5254–5263.

[12] Q. Qian, Y. Wang, T. Zhang, Y. Qin, Maximum mean square discrepancy: a new discrepancy representation metric for mechanical fault transfer diagnosis, Knowl.-Based Syst. 276 (2023) 110748.

[13] J. Lin, H. Shao, X. Zhou, B. Cai, B. Liu, Generalized MAML for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals, Expert Syst. Appl (2023) 120696.

[14] H. Tao, L. Cheng, J. Qiu, V. Stojanovic, Few shot cross equipment fault diagnosis method based on parameter optimization and feature mertic, Meas. Sci. Technol. 33 (11) (2022) 115005.

[15] Y. Xiao, J. Wang, Z. He, H. Zhou, H. Zhu, Deep transfer learning with metric structure for fault diagnosis, Knowl.-Based Syst. 256 (2022) 109826.

[16] Y. Zhang, Z. Ren, S. Zhou, K. Feng, K. Yu, Z. Liu, Supervised contrastive learning-based domain adaptation network for intelligent unsupervised fault diagnosis of rolling bearing, IEEE/ASME Trans. Mech. 27 (6) (2022) 5371–5380.

[17] J. Li, Y. Wang, Y. Zi, Z. Zhang, Whitening-Net: A generalized network to diagnose the faults among different machines and conditions, IEEE Trans. Neural Netw. Learn. Syst. 33 (10) (2022) 5845–5858.

[18] F. Lu, Q. Tong, Z. Feng, Q. Wan, Unbalanced bearing fault diagnosis under various speeds based on spectrum alignment and deep transfer convolution neural network, IEEE Trans. Ind. Inform. 19 (7) (2023) 8295–8306.

[19] B. Liu, C. Yan, Y. Liu, Z. Wang, Y. Huang, L. Wu, Multiscale residual antinoise network via interpretable dynamic recalibration mechanism for rolling bearing fault diagnosis with few samples, IEEE Sens. J. 23 (24) (2023) 31425–31439.

[20] C. Li, S. Li, H. Wang, F. Gu, A.D. Ball, Attention-based deep meta-transfer learning for few-shot fine-grained fault diagnosis, Knowl.-Based Syst. 264 (2023) 110345.

[21] Y. Feng, J. Chen, S. He, T. Pan, Z. Zhou, Globally localized multisource domain adaptation for cross-domain fault diagnosis with category shift, IEEE Trans. Neural Netw. Learn. Syst. 34 (6) (2023) 3082–3096.

[22] G.-B. Jang, S.-B. Cho, Cross-domain adaptation using domain interpolation for rotating machinery fault diagnosis, IEEE Trans. Instrum. Meas. 71 (2022) 1–17, http://dx.doi.org/10.1109/TIM.2022.3204093.

[23] S. Wan, J. Liu, X. Li, Y. Zhang, K. Yan, J. Hong, Transfer-learning-based bearing fault diagnosis between different machines: A multi-level adaptation network based on layered decoding and attention mechanism, Measurement 203 (2022) 111996.

[24] R.-F. Zhang, M.-C. Li, Bilinear residual network method for solving the exactly explicit solutions of nonlinear evolution equations, Nonlinear Dynam. 108 (1) (2022) 521–531.

[25] R.-F. Zhang, S. Bilige, T. Chaolu, Fractal solitons, arbitrary function solutions, exact periodic wave and breathers for a nonlinear partial differential equation by using bilinear neural network method, J. Syst. Sci. Complex 34 (2021) 122–139.

[26] R.-F. Zhang, S. Bilige, Bilinear neural network method to obtain the exact analytical solutions of nonlinear partial differential equations and its application to p-gBKP equation, Nonlinear Dynam. 95 (2019) 3041–3048.

[27] R.-F. Zhang, M.-C. Li, H.-M. Yin, Rogue wave solutions and the bright and dark solitons of the (3+ 1)-dimensional Jimbo–Miwa equation, Nonlinear Dynam. 103 (2021) 1071–1079.

[28] T. Kim, J. Chai, Pre-processing method to improve cross-domain fault diagnosis for bearing, Sensors 21 (15) (2021) 4970.

[29] H. Zheng, Y. Yang, J. Yin, Y. Li, R. Wang, M. Xu, Deep domain generalization combining a priori diagnosis knowledge toward cross-domain fault diagnosis of rolling bearing, IEEE Trans. Instrum. Meas. 70 (2021) 1–11, http://dx.doi.org/10.1109/TIM.2020.3016068.

[30] X. Yu, Z. Liang, Y. Wang, H. Yin, X. Liu, W. Yu, Y. Huang, A wavelet packet transform-based deep feature transfer learning method for bearing fault diagnosis under different working conditions, Measurement 201 (2022) 111597.

[31] I. Kim, S. Wook Kim, J. Kim, H. Huh, I. Jeong, T. Choi, J. Kim, S. Lee, Single domain generalizable and physically interpretable bearing fault diagnosis for unseen working conditions, Expert Syst. Appl. 241 (2023) 122455.

[32] C. He, H. Shi, J. Li, IDSN: A one-stage interpretable and differentiable STFT domain adaptation network for traction motor of high-speed trains cross-machine diagnosis, Mech. Syst. Sig. Process. 205 (2023) 110846.

[33] C. He, H. Shi, J. Si, J. Li, Physics-informed interpretable wavelet weight initialization and balanced dynamic adaptive threshold for intelligent fault diagnosis of rolling bearings, J. Manuf. Syst. 70 (2023) 579–592.

[34] G. Jiang, J. Wang, L. Wang, P. Xie, Y. Li, X. Li, An interpretable convolutional neural network with multi-wavelet kernel fusion for intelligent fault diagnosis, J. Manuf. Syst. 70 (2023) 18–30.

[35] C. Liu, C. Qin, X. Shi, Z. Wang, G. Zhang, Y. Han, TScatNet: An interpretable cross-domain intelligent diagnosis model with antinoise and few-shot learning capability, IEEE Trans. Instrum. Meas. 70 (2021) 1–10, http://dx.doi.org/10.1109/TIM.2020.3041905.

[36] C. Liu, X. Ma, T. Han, X. Shi, C. Qin, S. Hu, NTScatNet: An interpretable convolutional neural network for domain generalization diagnosis across different transmission paths, Measurement 204 (2022) 112041.

[37] K. Yue, J. Li, J. Chen, R. Huang, W. Li, Multiscale wavelet prototypical network for cross-component few-shot intelligent fault diagnosis, IEEE Trans. Instrum. Meas. 72 (2023) 1–11, http://dx.doi.org/10.1109/TIM.2022.3230480.

[38] T. Li, Z. Zhao, C. Sun, L. Cheng, X. Chen, R. Yan, R.X. Gao, WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis, IEEE Trans. Syst. Man Cybern. Syst. 52 (4) (2022) 2302–2312.

[39] Z. Shang, Z. Zhao, R. Yan, Denoising fault-aware wavelet network: A signal processing informed neural network for fault diagnosis, Chin. J. Mech. Eng-En. 36 (1) (2023) 9, http://dx.doi.org/10.1186/s10033-023-00838-0.

[40] S. Ramasinghe, L.E. MacDonald, M. Farazi, H. Saratchandran, S. Lucey, How much does initialization affect generalization? in: International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, in: Proceedings of Machine Learning Research, vol. 202, PMLR, 2023, pp. 28637–28655, https://proceedings.mlr.press/v202/ramasinghe23a.html.

[41] M. Ravanelli, Y. Bengio, Speaker recognition from raw waveform with Sinc-Net, in: 2018 IEEE Spoken Language Technology Workshop, SLT, 2018, pp. 1021–1028.

[42] D. Wang, B. Hou, T. Yan, C. Shen, Z. Peng, New statistical learning perspective for design of a physically interpretable prototypical neural network for machine condition monitoring, Mech. Syst. Sig. Process. 188 (2023) 110041.

[43] K. Zhong, J. Wang, S. Xu, C. Cheng, H. Chen, Overview of fault prognosis for traction systems in high-speed trains: A deep learning perspective, Eng. Appl. Artif. Intell. 126 (2023) 106845.

[44] H. Cao, H. Shao, X. Zhong, Q. Deng, X. Yang, J. Xuan, Unsupervised domain-share CNN for machine fault transfer diagnosis from steady speeds to time-varying speeds, J. Manuf. Syst. 62 (2022) 186–198.

[45] Case western reserve university (CWRU) bearing data center, 2019, [Online]. Available: https://engineering.case.edu/bearingdatacenter/download-data-file.

[46] K. Li, X. Ping, H. Wang, P. Chen, Y. Cao, Sequential fuzzy diagnosis method for motor roller bearing in variable operating conditions based on vibration analysis, Sensors 13 (2013) 8013–8041.

[47] S. Shao, S. McAleer, R. Yan, P. Baldi, Highly accurate machine fault diagnosis using deep transfer learning, IEEE Trans. Ind. Inform. 15 (4) (2019) 2446–2455.

[48] K. Yu, Q. Fu, H. Ma, T.R. Lin, X. Li, Simulation data driven weakly supervised adversarial domain adaptation approach for intelligent cross-machine fault diagnosis, Struct. Health Monit. 20 (4) (2021) 2182–2198.

[49] J. Luo, H. Shao, H. Cao, X. Chen, B. Cai, B. Liu, Modified DSAN for unsupervised cross-domain fault diagnosis of bearing under speed fluctuation, J. Manuf. Syst. 65 (2022) 180–191.

[50] Q. Qian, Y. Qin, J. Luo, Y. Wang, F. Wu, Deep discriminative transfer learning network for cross-machine fault diagnosis, Mech. Syst. Sig. Process. 186 (2023) 109884.
[51] L. Jia, T.W. Chow, Y. Yuan, GTFE-Net: A gramian time frequency enhancement CNN for bearing fault diagnosis, Eng. Appl. Artif. Intell. 119 (2023) 105794.
[52] A. Alekseev, A. Bobe, GaborNet: Gabor filters with learnable parameters in deep convolutional neural network, in: 2019 International Conference on Engineering and Telecommunication (EnT), 2019, pp. 1–4, http://dx.doi.org/10.1109/EnT47717.2019.9030571.
[53] R.-F. Zhang, M.-C. Li, J.-Y. Gan, Q. Li, Z.-Z. Lan, Novel trial functions and rogue waves of generalized breaking soliton equation via bilinear neural network method, Chaos Solitons Fractals 154 (2022) 111692.
[54] R.-F. Zhang, M.-C. Li, M. Albishari, F.-C. Zheng, Z.-Z. Lan, Generalized lump solutions, classical lump solutions and rogue waves of the (2+ 1)-dimensional Caudrey-Dodd-Gibbon-Kotera-Sawada-like equation, J. Comput. Appl. Math. 403 (2021) 126201.
[55] R.-F. Zhang, M.-C. Li, A. Cherraf, S.R. Vadyala, The interference wave and the bright and dark soliton for two integro-differential equation by using BNNM, Nonlinear Dynam. 111 (9) (2023) 8637–8646.
[56] R.-F. Zhang, S. Bilige, J.-G. Liu, M. Li, Bright-dark solitons and interaction phenomenon for p-gBKP equation by using bilinear neural network method, Phys. Scr. 96 (2) (2020) 025224.