Journal of Manufacturing Systems 70 (2023) 579–592

Contents lists available at ScienceDirect

Journal of Manufacturing Systems

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

Physics-informed interpretable wavelet weight initialization and balanced dynamic adaptive threshold for intelligent fault diagnosis of rolling bearings

Chao He [a,b], Hongmei Shi [a,b,*], Jin Si [c], Jianbo Li [a,b]

[a] School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China
[b] Collaborative Innovation Center of Railway Traffic Safety, Beijing 100044, China
[c] Key Laboratory of Information System and Technology, Beijing Institute of Control and Electronic Technology, Beijing 100038, China

ARTICLE INFO

Keywords:
Convolutional neural network
Wavelet weight initialization
Continuous wavelet transform
Balanced adaptive threshold
Normalization activation mapping
Interpretable fault diagnosis

ABSTRACT

Intelligent fault diagnosis of rolling bearings using deep learning-based methods has made unprecedented progress. However, there is still little research on weight initialization and the threshold setting for noise reduction. An innovative deep triple-stream network called EWSNet is proposed, which presents a wavelet weight initialization method and a balanced dynamic adaptive threshold algorithm. Initially, an enhanced wavelet basis function is designed, in which a scale smoothing factor is defined to acquire more rational wavelet scales. Next, a plug-and-play wavelet weight initialization for deep neural networks is proposed, which utilizes physics-informed wavelet prior knowledge and showcases stronger applicability. Furthermore, a balanced dynamic adaptive threshold is established to enhance the noise-resistant robustness of the model. Finally, normalization activation mapping is devised to reveal the effectiveness of Z-score from a visual perspective rather than experimental results. The validity and reliability of EWSNet are demonstrated through four data sets under the conditions of constant and fluctuating speeds. **Source code is available at:** https://github.com/liguge/EWSNet.

1. Introduction

Rolling bearing is one of the essential components in rotating machinery, the failure of which can potentially result in significant accidents. Accordingly, fault diagnosis of rolling bearings has been widely concerned. Thanks to the advances in artificial intelligence and increasing computational power, intelligent fault diagnosis (IFD) with deep learning technology has achieved unprecedented success. However, complex high-dimensional linear deep models with the inherent "black box" nature are difficult to comprehend, which resort to a substantial amount of samples to achieve the ideal performance. Recently, a bunch of researchers pay attention to introduce physical information and device mechanism into deep neural networks, which can make the model interpretable [1,2]. According to Ref. [3], there are two distinct interpretable perspectives, specifically intrinsic interpretability based on physical prior knowledge, and post-hoc explanations grounded in activation mapping.

In the realm of intrinsic interpretability, Ref. [4] conducted a survey on informed machine learning, which discussed knowledge embedding is an intrinsic interpretability paradigm inserting domain knowledge into data-driven models. Signal processing technology embedding into data-driven models is one of the main directions of knowledge embedding [5,6].

WaveletKernelNet [7,8] is one of the splendid works in knowledge embedding fault diagnosis. WaveletKernelNet brings the continuous wavelet transform into the convolution kernel to construct the wavelet convolution. Tai et al. [9] applied WaveletKernelNet to acoustic emission signals for fault detection. Liao et al. [10] integrated Daubechies wavelet with convolution to assess its effectiveness on deep transfer IFD. Liu et al. [11] introduced the nonparametric wavelet convolution as the backbone and suggested a time-scattering CNN for small sample cross-domain fault diagnosis. Li et al. [12] introduced a graph wavelet denoising network to denoise noisy graph signals. Li et al. [13] proposed an interpretable wavelet packet constrained convolutional network, which integrates the feature extraction capability of wavelet bases with the learning capability of convolutional kernels. Besides, WaveletKernelNet has also been used for the reliability evaluation of rolling bearings [14].

Signal noise has a great impact on deep learning-based fault diagnosis, so it is a current research direction to embed noise-resistant knowledge into network models. Threshold algorithms have been extensively used in IFD to decrease the noise interference. Deep residual

* Corresponding author at: School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China.
E-mail addresses: chaohe@bjtu.edu.cn (C. He), hmshi@bjtu.edu.cn (H. Shi), sijin_sj@126.com (J. Si), jbli@bjtu.edu.cn (J. Li).

https://doi.org/10.1016/j.jmsy.2023.08.014
Received 23 May 2023; Received in revised form 15 July 2023; Accepted 23 August 2023
Available online 4 September 2023

shrinkage networks (DRSN) [15] incorporates soft threshold into residual networks (ResNet) with Squeeze-and-Excitation. Pei et al. [16] incorporated the residual shrinkage block with soft threshold into densely connected convolutional networks. Refs. [17,18] respectively applied soft threshold to cross-domain or imbalanced fault diagnosis. Zhang et al. [19] included the overall self-adaptive slope strategy into soft threshold, utilizing the attention mechanism to search for the most appropriate slope. Salimy et al. [20] replaced soft threshold with firm threshold or garrote threshold, and investigated the application in fault detection of electromagnetic interference data. Chen et al. [21] came up with an end-to-end dual-path mixed domain residual threshold network that combined convolutional block attention module, soft threshold with dilated convolution, performing well under noise circumstances. Shang et al. [22] put forward an aero-engine fault diagnosis method by combining WaveletKernelNet, soft threshold, and energy-to-entropy ratio weights.

Post-hoc explanations provide interpretability after a model has been constructed. There are some studies on post-hoc explanation methods, such as visualizing weights and gradients [23]. Refs. [24,25] captured the weights of the attention mechanism and displayed them with the thermal maps. Grezmak et al. [26] applied Layer-wise Relevance Propagation (LRP) to interpret the time–frequency maps, while Gradient-weighted Class Activation Mapping (Grad-CAM) [27] was used. Grad-CAM [28], Grad-CAM++[29] and Multilayer Grad-CAM [30] are respectively applied to raw signals for further vital information. Kim et al. put forward the frequency activation mapping (FAM) [31] and the multi-variable data-based frequency-domain Grad-CAM [32] to display frequency-domain information.

The inspiring results of prior gains have brought great contributions to the field. There are still some issues that require attention:

(1) The wavelet kernel is utilized as the feature extractor with an emphasis on applications, lacking targeted analysis and modification. Since some wavelet basis functions contain exponential components, Laplace wavelet leads to a large distribution gap of wavelet values, which is detrimental to training. How to narrow the gap and bring the factors into consistency is an issue that needs to be addressed.
(2) The commonly used initialization methods such as Kaiming and Xavier are more suitable for image data. To the best of our knowledge, it is particularly noteworthy that the IFD domain lacks a tailored weight initialization method.
(3) Soft and hard thresholds have their own limitations, where the hard threshold function is discontinuous and the soft threshold function is prone to discarding high-frequency information. It is therefore essential to balance them without increasing too much model complexity.
(4) As for the post-hoc explanations, FAM still has shortcomings, because they are merely appropriate for the meticulously designed models. Concurrently, these post-hoc explanations are backward inference of prediction results to improve model interpretability. However, few studies focus on the explanations of pre-processed signals, which is forward inference.

In order to tackle the above challenges, a novel and general deep triple-stream network called EWSNet is proposed in this paper, which presents an approach to initialize weights of the first convolution kernel without additional network layer added. Specifically, the signal processing-based modified wavelet dictionary is treated as the prior knowledge. WDCNN indicates that the first convolution layer is critical to the robustness of model. Balanced Dynamic Adaptive Threshold (BDAT) is devised to reduce the noise interference. Besides, FAM is extended to the visual interpretation of data pre-processing called Normalization Activation Mapping (NAM). Time-series Gradient Activation Mapping (TGAM) explains the interpretability of EWSNet with thermal maps.

The main contributions of the research are as follows:

(1) A novel triple-stream network for bearing fault diagnosis is proposed, which is embedded the wavelet signal processing technique and threshold noise reduction algorithm to improve the robustness of EWSNet under the operating scenario on constant speed and variable speed.
(2) A modified wavelet basis function is designed, and a scale smoothing factor is defined to make the scale factor and translation factor consistent in wavelet transform, enabling it better interpretability and trainability.
(3) A plug-and-play and interpretable weight initialization method is devised, and the above-mentioned enhanced wavelet basis function is utilized to initialize the weights of the first convolutional layer, which makes the model better extrapolation capability and more suitable for application in real-world industrial scenarios than the vanilla initialization methods such as Kaiming or Xavier initialization.
(4) A differentiable threshold correction coefficient for noise reduction is designed. In the process of training, the automatic trade-off of hard and soft thresholds is realized by adjusting this parameter adaptively, so that the model has better generalization while alleviating the interference of hand-crafted hyper-parameters.
(5) An interpretable mechanism termed Normalized Activation Mapping is presented to reveal the validity of Z-score preprocessing. Through NAM, the signals processed by Z-score are represented as a visual form to prove that it can enhance the frequency-domain information of raw signals.

Four bearing data sets, which are separately from Jiangnan University (JNU), Xi'an Jiaotong University (SQV), Case Western Reserve University (CWRU) and Beijing Jiaotong University (BJTU), are utilized to verify the effectiveness of EWSNet. The remaining sections of this paper are as follows.

Second 2 is mostly concerned about prior works. In Section 3, EWSNet is discussed in detail. Section 4 includes several experiments and analyses to demonstrate the superior performance of EWSNet. Section 5 will discuss interpretability of the model. A conclusion and recommendation for further research are formed in Section 6.

2. Preliminaries

2.1. Wavelet convolution layer and residual shrinkage block

Wavelet convolution (WConv) is formed based on the vanilla convolution, and the output of vanilla convolution y is indicated:

$$y = \sum_{w=1}^{W} k_w x + b_w \tag{1}$$

where k_w and b_w mean weights and bias, respectively. x is raw signals.

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

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

where $\psi(\cdot)$ denotes wavelet basis function, t is time, s is scale factor, u is translation factor, s and u are dynamic adaptive parameters and totally differentiable. The wavelet convolution h is constructed as follows:

$$h = \psi_{u,s}(t) * x \tag{3}$$

In Fig. 1(a), residual shrinkage block employs SE to acquire adaptive threshold η, which is used to set soft threshold to denoise, as stated in Eq. (4). The correction coefficient α is inserted into Eq. (4) shown in Fig. 1(b).

$$y = \begin{cases} \operatorname{sgn}(x)(|x| - \eta) & |x| > \eta \\ 0 & |x| \le \eta \end{cases} \tag{4}$$

where y denotes the wavelet coefficient.

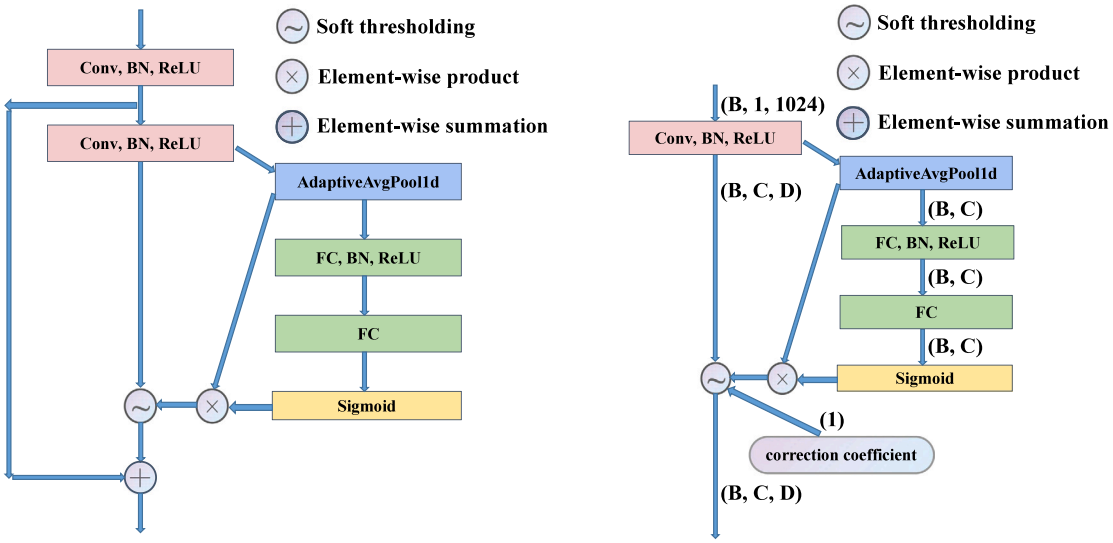


Fig. 1. Shrinkage block and semi-shrinkage block.

2.2. FAM and TGAM

FAM visualizes frequency-domain information based on the relationship between extracted features and classification weights. The frequency map γ_k is defined as:

$$\gamma_k = S_{F_k}(w) \tag{5}$$

where $S_x(w)$ denotes the power spectrum, and F_k represents the output of the last convolution layer, k is channels.

FAM is expressed as the weighted summation of the frequency map γ^c_{FAM}:

$$\gamma^c_{FAM} = \sum_{k=1}^{K} max(0, W_{k,c}) \times \gamma_k \tag{6}$$

where c is sample class, $W_{k,c}$ represents the weights of the full connection layer.

TGAM [29] means one-dimensional Grad-CAM++. Grad-CAM++ is designed for image data, not suitable for one-dimensional time series signals. TGAM calculates the importance of feature maps w^c_k in the time dimension.

$$w^c_k = \sum_l \alpha^{kc}_l relu(\frac{\partial Y^c}{\partial A^k_l}) \tag{7}$$

where K is represented as the number of channels, c is represented as the target category, l is the time position, Y^c is the value of the SoftMax function, A^k_l is denoted as the value of the feature layer activation graph, and α^{kc}_l is the weighted weights of the gradients.

The resulting class gradient activation value is as shown in Eq. (8)

$$L^c_{TGAM} = \text{interp}(\sum_k w^c_k A^k_l) \tag{8}$$

where interp () is expressed as the linear interpolation.

3. Enhanced Semi-shrinkage Wavelet weight initialization Network (EWSNet)

3.1. Pipeline of EWSNet fault diagnosis method

The workflow of fault diagnosis is shown in Fig. 2, where the solid line is denoted as the training process, and the dotted line is represented as the testing process. Firstly, raw signals are gathered with the acquisition equipment, and the samples are normalized using Z-score. Then, the training set, validation set and test set are split. Afterwards, EWSNet is constructed, in which the first convolutional weights are initialized to the wavelet weights. After the training, the trained model parameters are save. Finally, the performance of the models is assessed under limited samples.

3.2. Proposed wavelet weight initialization

3.2.1. Enhanced wavelet convolution layer

The input channels are denoted as N_i. The output channels are denoted as N_k, and convolution kernel size is K.

Dissecting the theory of WaveletKernelNet, it can be noticed WConv additionally needs to follow several criteria in addition to what are mentioned in Ref. [7]:

(1) $u \in (0, N_k)$, $s \in (1, N_k)$. s, u are arithmetic progression of size N_k, which indicates that the channel transform is related to s and u, and a channel for a convolution kernel can represent a wavelet basis function.
(2) $t \in (0, K-1)$, an arithmetic progression of size K, indicates that a unit of the convolution kernel represents a time step.
(3) The symmetry of wavelet basis function needs to be judged to construct the time scale t.

Taking $N_k = 64$, $K = 250$, Laplace wavelet as a prime example, we construct enhanced wavelet convolution (EWconv). Since raw signals belong to real signals, real Laplace wavelet basis function [33] is adopted as follows:

$$\psi(t) = Ae^{\frac{-\xi}{\sqrt{1-\xi^2}} \times 2\pi f(t-\tau)} \times \sin[2\pi f(t-\tau)] \tag{9}$$

where f is the sampling frequency and ξ is the viscous damping ratio. A is a wavelet normalization function. τ means a time parameter.

The real Laplace wavelet dictionary (LWdic) $\psi^L_{u,s}(t)$ can be calculated by Eq. (2) and Eq. (9):

$$\psi^L_{u,s}(t) = Ae^{\frac{-\xi}{\sqrt{1-\xi^2}} \times 2\pi f(\frac{t-u}{s}-\tau)} \times \sin[2\pi f(\frac{t-u}{s}-\tau)] \tag{10}$$

It is noticed that the ranges of s and u are not consistent. Therefore, $s \in (1, N_k)$ is replaced by $s \in (0, N_k)$, and the scale smoothing factor ζ is introduced as follows:

$$\psi^L_{u,s}(t) = Ae^{\frac{-\xi}{\sqrt{1-\xi^2}} \times 2\pi f(\frac{t-u}{s-\zeta}-\tau)} \times \sin[2\pi f(\frac{t-u}{s-\zeta}-\tau)] \tag{11}$$

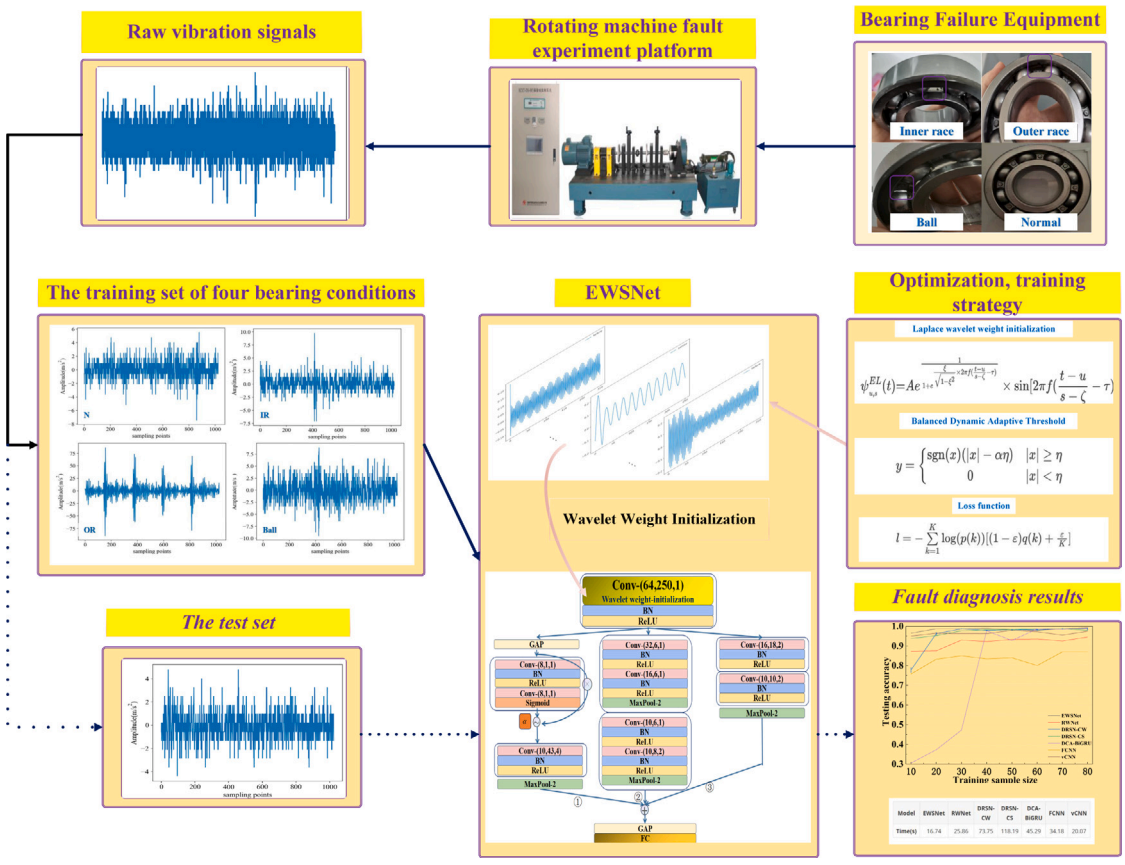


Fig. 2. Pipeline of EWSNet Fault diagnosis method.

As illustrated in Figs. 3(a) and 3(d), the range of the wavelet values with a large fluctuation is adverse to optimization, so the exponential component needs to be handled with four activation functions as follows: Sigmoid (F_1), Tanh (F_2), $\frac{2}{\pi}\arctan x$ (F_3), Softmax (F_4).

$$F_1 = \frac{1}{1+e^{-x}} \in (0,1) \tag{12}$$

$$F_2 = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \in (-1,1) \tag{13}$$

$$F_3 = \frac{2}{\pi}\arctan x \in (-1,1) \tag{14}$$

$$F_4 = \frac{e^{-x_i}}{\sum_{i=1}^{C} e^{-x_i}} \in (-1,1) \tag{15}$$

For example, substituting Eq. (12) into Eq. (11), the enhanced Laplace wavelet can be expressed as follows:

$$\psi_{u,s}^{EL}(t) = Ae^{\frac{1}{1+e^{\frac{\xi}{\sqrt{1-\xi^2}}\times 2\pi f\left(\frac{t-u}{s-\zeta}-\tau\right)}}} \times \sin[2\pi f(\frac{t-u}{s-\zeta}-\tau)] \tag{16}$$

Fig. 3(a) shows the LWdic constructed by the rules in Paragraph 2, Section 3.2.1. In Fig. 3(a), the unit convolution kernel represents the unit time, yet $\psi_{u,s}^{L}(t)$ is of little regularity. Hence, LWdic limits $t \in (0,1)$, but the unit convolution kernel loses the physical meaning because the unit convolution kernel no longer represents the unit time (Fig. 3(b)). From Figs. 3(a) to 3(e), when $t \in (0,K)$, adopting $s \in (0,N_k)$ and Sigmoid can satisfy regularity and sparsity better (Figs. 3(e)), and 3(f) adds a scale smoothing factor compared to Fig. 3(e). Furthermore, Figs. 3(f)~3(i) are enhanced LWdic constructed by the activation function $F_1 \sim F_4$, respectively.

3.2.2. Wavelet weight initialization

The primary purpose of WConv is to build a physical prior wavelet kernel, which adaptively adjusts the parameters by back-propagation to infer the optimal wavelet kernel.

Fine-tuning retrains pre-training parameters with fewer samples to achieve better results than training directly in downstream tasks, which is an idea of constructing the prior embedded convolution weights.

In this paper, by converting the enhanced WConv layer to wavelet weight initialization, vanilla weights are embedded in prior physical knowledge, which is beneficial to boosting training speed and better cooperation with models. The prior knowledge called signal processing-based modified wavelet dictionary is treated as weight initialization for neural networks.

In the paper, five wavelet basis functions are constructed into the proposed wavelet weight initialization, which are Laplace, Morlet, Mexican Hat, Gaussian, and Shannon shown in Eqs. (9), (17)~(20).

$$\psi(t) = Ce^{-\frac{t^2}{2}}\cos(5t) \tag{17}$$

$$\psi(t) = \frac{2}{\sqrt{3\sqrt{\pi}}}(1-t^2)e^{-\frac{t^2}{2}} \tag{18}$$

$$\psi(t) = \frac{-1}{\sqrt{2\pi}}te^{-\frac{t^2}{2}} \tag{19}$$

$$\psi(t) = \frac{\sin\left(2\pi\left(t-{}^{1}/_{2}\right)\right)-\sin\left(\pi\left(t-{}^{1}/_{2}\right)\right)}{\pi\left(t-{}^{1}/_{2}\right)} \tag{20}$$

In Eq. (2), the initialization methods for k_w are generally random initialization, Kaiming initialization, and Xavier initialization, which are primarily designed for image-related fields. However, the current

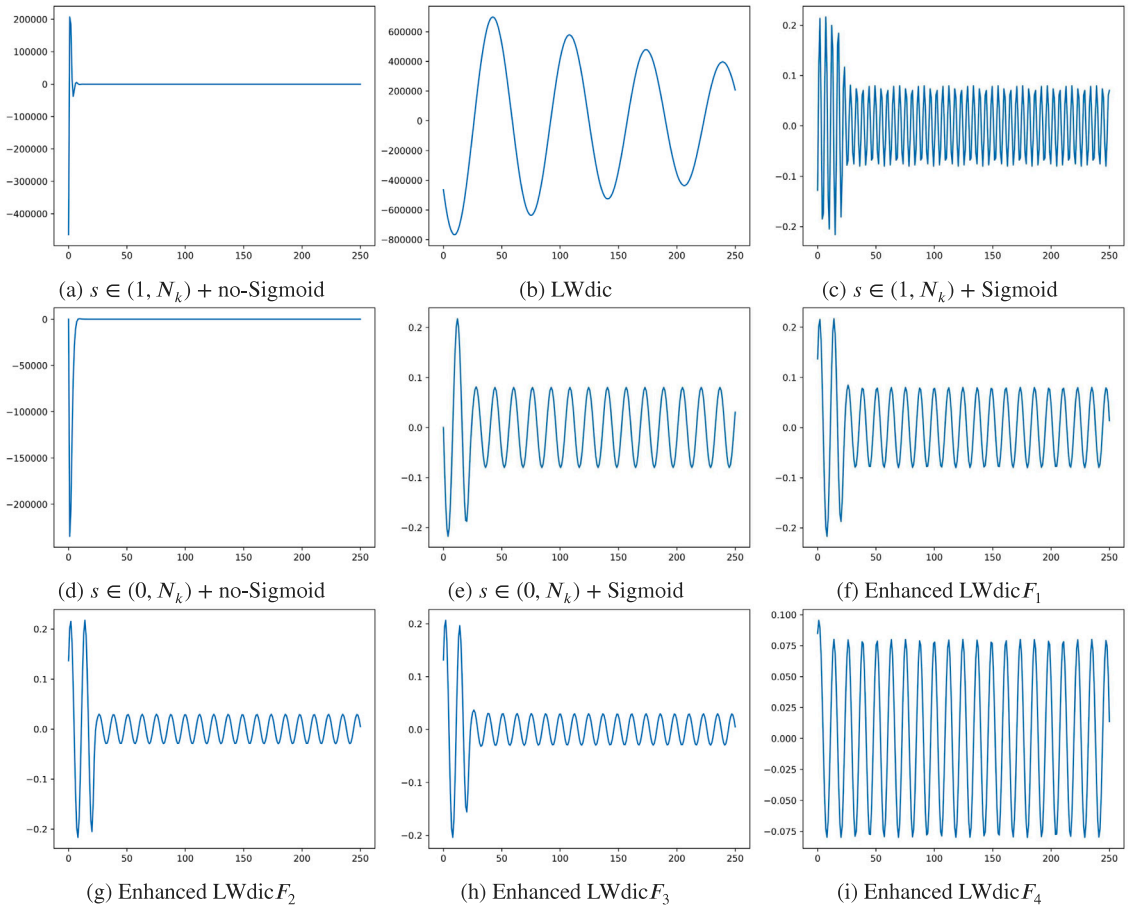


Fig. 3. Laplace Wavelet dictionary.

research lacks initialization methods specifically tailored for faulty signals. Obviously, the wavelet initialization proposed in this paper will fill this gap. In particular, $k_w = \psi_{u,s}(t)$. The wavelet initialization method performs particularly well.

Algorithm 1 provides the pseudo-code of the enhanced wavelet weight initialization for CNN weights.

Algorithm 1 Wavelet weight initialization

Input: N_i: input channels; N_k: out channels; K: convolution kernel size; ζ: scale smoothing factor; s: scale factor; u: translation factor; t: time.

Output: $\psi_{u,s}^{EL}(t)$: wavelet weights.

1: Constructing Wavelet dictionary by Eq. (9), Eq. (17) ∼ Eq. (20) and Eq. (2)
2: Introducing scale smoothing factor by $\psi(\frac{t-u}{s-\zeta})$
3: Calculating wavelet weights by Eq. (12) ∼ Eq. (15) and enhanced Wavelet dictionary
4: Convolution kernel weights are initialized as wavelet weights: $k_w = \psi_{u,s}(t)$

3.2.3. Multi-type EWConv and multi-channel EWConv

The input channel of Wconv is generally limited to one channel, with a single wavelet basis function.

In order to acquire more information, we put forward multi-type and multi-channel EWConv, as can be seen from Fig. 4. At the first step, multi-channel WConv is improved from WConv, where $[N_k, 1, K]$ repeats N_i times to $[N_k, N_i, K]$. Then, Multi-type EWConv combines various wavelet kernels, where $[N_k, N_i, 2K]$ denotes that each channel contains two kinds of kernels, and $[2N_k, N_i, K]$ means that the forward N_k channels are one wavelet kernel, and the backward N_k channels are the other.

3.3. Balanced Dynamic Adaptive Threshold (BDAT)

As shown in Fig. 1(b), especially considering adaptive threshold correction coefficient, the correction coefficient α is inserted into vanilla soft threshold Eq. (4) as follows:

$$y = \begin{cases} \operatorname{sgn}(x)(|x| - \alpha\eta) & |x| \geq \eta \\ 0 & |x| < \eta \end{cases} \tag{21}$$

where α and η are differentiable ($\alpha \in (0,1), \alpha \neq 0,1$). When $\alpha = 0$ or 1, Eq. (21) respectively degenerates into hard threshold and soft threshold, and thus we can adjust α appropriately to make y closer to the genuine wavelet coefficient.

Based on the above BDAT and Adam, the parameters can be updated as follows:

$$\begin{cases} \alpha \leftarrow \alpha - \mu\left(\frac{\partial y}{\partial \alpha}\right) \\ \eta \leftarrow \eta - \mu\left(\frac{\partial y}{\partial \eta}\right) \end{cases} \tag{22}$$

where μ is the learning rate.

We do not seek to tackle hard threshold discontinuities, but tackle soft threshold coefficients. When $\alpha = 0$ or 1, it will be discussed separately.

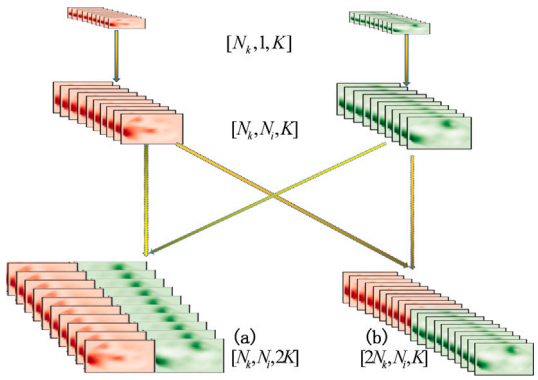


Fig. 4. Multi-type EWConv is on the left and Multi-channel EWConv is on the right.

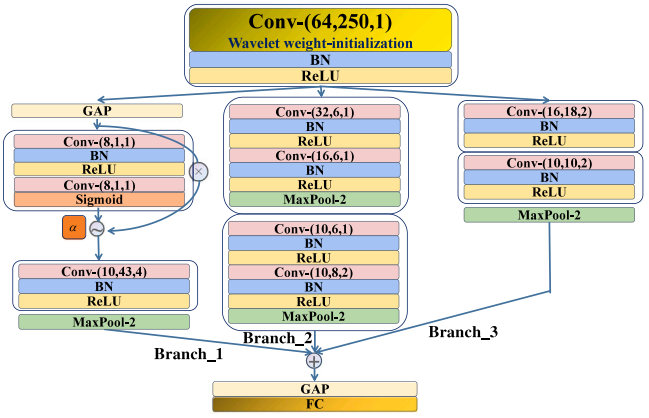


Fig. 5. The specific structure of triple-stream EWSNet.

3.4. Triple-stream EWSNet

It shows the specific structure of EWSNet in Fig. 5, taking $N_k = 64$, $N_i = 1$ and $K = 250$ as an illustrative example.

WDCNN indicates that the first convolution layer is critical to the robustness of model [34]. Therefore, the first convolution layer weights are restricted to wavelet weights. Branch_1 performs wavelet BDAT denoising; Branch_2 integrates four nonlinear activation layers to extract high-frequency features; Branch_3 extracts low-frequency features. In a nutshell, multi-stream gathers multi-scale features to promote the model to possess richer features.

In the process of the forward propagation, the features extracted by the three branches are fused by linear summation:

$$y_{123} = y_1 + y_2 + y_3 \tag{23}$$

where y_1, y_2, y_3 denote the features extracted by the three branches Branch_1, Branch_2, Branch_3, respectively. y_{123} is the fused feature.

Ultimately, the back-propagation process utilizing the Adam optimizer can be written:

$$\begin{cases} \theta_1 \leftarrow \theta_1 - \mu(\frac{\partial L}{\partial \theta_1}) \\ \theta_2 \leftarrow \theta_2 - \mu(\frac{\partial L}{\partial \theta_2}) \\ \theta_3 \leftarrow \theta_3 - \mu(\frac{\partial L}{\partial \theta_3}) \end{cases} \tag{24}$$

where $\theta_1, \theta_2, \theta_3$ denote the set of parameters for each of the three branches, and L denotes the cross-entropy loss.

In terms of interpretability, the first convolution kernel is constrained to wavelet weights, where Branch_1 constitutes wavelet BDAT denoising module. TGAM visually explains EWSNet.

3.5. The interpretability of Normalization Activation Mapping (NAM)

Data normalization can accelerate the process of convergence. Also, Z-score makes CNN get better accuracy [35]. Unlike experimental methods, we notice that Z-score enhances frequency-domain information of signals so that CNN can learn these features better.

FAM illustrates the frequency-domain information by utilizing the weights of the classification layer and extracted features, but it cannot reveal the influence of normalization methods. Therefore in NAM, the weight of the correct label is 1.0, and the features are signals processed by the normalization methods and it can visualize which normalization method possesses more frequency-domain knowledge. NAM is defined by Eq. (25) as:

$$\gamma^c_{NAM} = \begin{cases} S_x(w) & l_{real} = l_{target} \\ 0 & otherwise \end{cases} \tag{25}$$

where l_{real} is the real label and l_{target} is the tested label.

With JNU(800 r/min) data set and SQV data set, Figs. 6 and 7 show the effects of four normalization methods, which respectively are g_1: [0,1] Normalization; g_2: Maximum absolute Value Normalization; g_3: [−1,1] Normalization and g_4: Z-score. The vertical coordinate is the fault type, and the horizontal coordinate is the frequency corresponding to the fault type. The color mapping represents the magnitude of the energy. Whether the speed is constant or sharply variable, Z-score can enhance the intensity of frequency-domain information (Figs. 6(e) and 7(e)), compared with unprocessed samples (Figs. 6(a) and 7(a)).

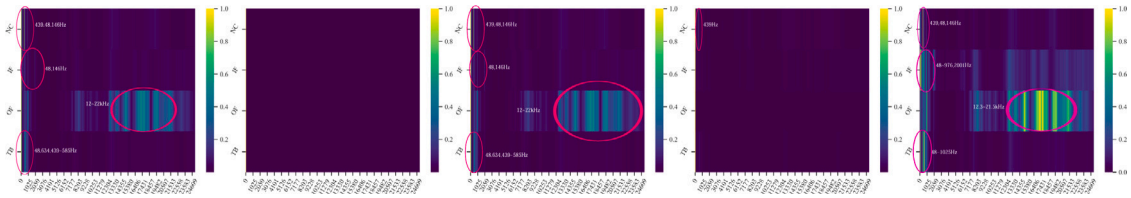
As to other normalization methods about speed fluctuation (Figs. 7(b)~7(d)), Z-score shows the most abundant features. For SQV, the frequency features of raw signals are apparent under conditions of the mild and moderate inner race faults. But after using Z-score, normal condition (12,25,62 Hz), mild inner race fault (8.6~8.7 kHz), moderate inner race fault (12.3~12.5 kHz), acute inner race fault (12 Hz), mild outer race fault (12 Hz,8 kHz), moderate outer race fault (4 kHz,7.9~8.1 kHz) and acute outer race fault (12,25,37 Hz), the features all show more obvious. Nevertheless, the frequencies of the normal condition and the acute fault are relatively close, possibly leading to incorrect fault classification.

4. Case analysis and discussion

4.1. Case analysis

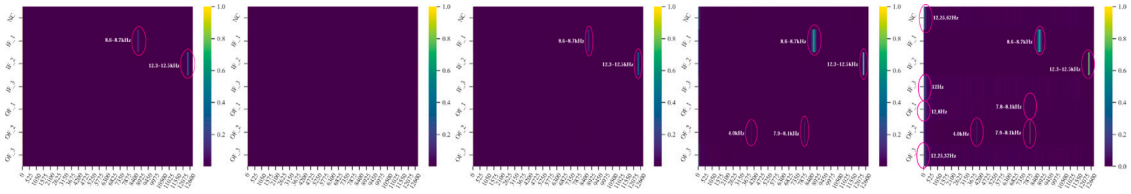
We benchmark EWSNet on four data sets: JNU data set [36], CWRU data set [37], SQV data set [38], and BJTU data set [39,40]. The experiments are implemented in PyTorch 1.10.0, Python 3.8.5, running on Intel(R) Core i7-6700HQ CPU @3.40 GHz (8G RAM), GTX970M GPU.

WaveletKernelNet(RWNet) [7], DRSN-CW [15], DRSN-CS [15], DCA-BiGRU [29], FAM-CNN (FCNN) [31], vanilla CNN(vCNN) are utilized for comparison, the structures of which are the same as that mentioned in the published papers, except for a little adjusting to input different sample size. In addition to the first convolution layer in EWSNet is initialized by wavelet weights, and vCNN has the same as EWSNet. Both WaveletKernelNet and EWSNet employ Laplace wavelet. The loss function adopts label smoothing regularization (LSR). The parameters are defined: $\mu = 0.001$, $\tau = 0.1$ [7], batch_size = 16. Early stopping is applied to alleviate over-fitting. All methods are experimentalized five times to ensure the stability of algorithms.



(a) Non (b) g_1 (c) g_2 (d) g_3 (e) g_4

Fig. 6. Normalized visualization in JNU. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Non (b) g_1 (c) g_2 (d) g_3 (e) g_4

Fig. 7. Normalized visualization in SQV. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Comparing different initialization methods (%)

Model	Wavelet weights	Random	Kaiming	Xavier
LeNet	**60.25**	50.50	56.25	58.25
AlexNet	**88.00**	76.00	76.50	83.00
Inception	**94.25**	86.75	92.50	93.50
ResNet	**99.00**	89.75	86.50	91.75

4.1.1. Constant speed condition

The sampling frequency of JNU is 50 kHz at three different speeds (600, 800, and 1000 r/min). The states of bearings are divided into normal conditions (NC), inner race fault (IF), outer race fault (OF), and ball fault (BF), respectively corresponding to Label 0,1,2,3.

The case of 600 r/min (Data set A) and 800 r/min (Data set B) are examined. The length of sample is 1024 using a sliding window and then normalized with Z-score. There are 200 samples in each state, while training samples are from 10 to 80 with a step 10. The parameter of ζ is -0.3.

Compared with Data set B, Data set A contains stronger noises, as shown in Figs. 8(a) and 8(b), the accuracy degrades on various models. DCA-BiGRU is the worst because it mainly uses small convolution kernels, so it cannot cope with strong noises. But wavelet weight initialization is used in the dual-stream branches, which is improved by 6%. In Data set A, FCNN gets the accuracy between 60% ∼ 70%, indicating its poor robustness, and the accuracy of vCNN without wavelet weight initialization is inferior to EWSNet under limited samples. Owing to BDAT denoising module and wavelet weight initialization, EWSNet has the highest accuracy. Even with a training sample size of 10 per class, it achieves 90.39%, which improves about 10% above vCNN. ESWNet can enhance by roughly 4.5% for the cases where the signal features are salient in Data set B. Additionally, method based on soft threshold (DRSN) outperforms that based on the wavelet convolution layer (RWNet).

The experimental steps are the same as described above, with 80 for training and 100 for testing. The experimental results compared with other initialization methods are shown in Table 1. Wavelet weight initialization has achieved best in classical models, especially in some simple classical models(LeNet and AlexNet).

4.1.2. Sharply variable speed condition

The test bench of SQV [38] (Data set C), as depicted in Fig. 9(a), consists of three major components: motor, rotor and load. The data set is acquired from the motor bearing model NSK6203 at a sampling frequency of 25.6 kHz. Also, SQV data set is acquired during the continuous change of speed, which contains a complete acceleration and deceleration process from the stationary state to 3000 rpm gradually, then remains stable, and finally is gradually decelerated to 0. As shown in Fig. 9(b), in addition to the normal state (N), six fault states are simulated, which are mild inner race fault (IF_1), moderate inner race fault (IF_2), acute inner race fault (IF_3), mild outer race fault (OF_1), moderate outer race fault (OF_2) and acute outer race fault (OF_3). The frequency-domain information following data pre-processing is shown in Fig. 7, and the sample length is 2048, $\zeta = 0.1$.

As shown in Fig. 8(c), FCNN performs the worst. Besides, when the sample size is 10∼30, the accuracy of EWSNet decreases by about 1% than vCNN, but EWSNet improves by about 2% with a sample size of 40∼60, and it consistently beats DRSN.

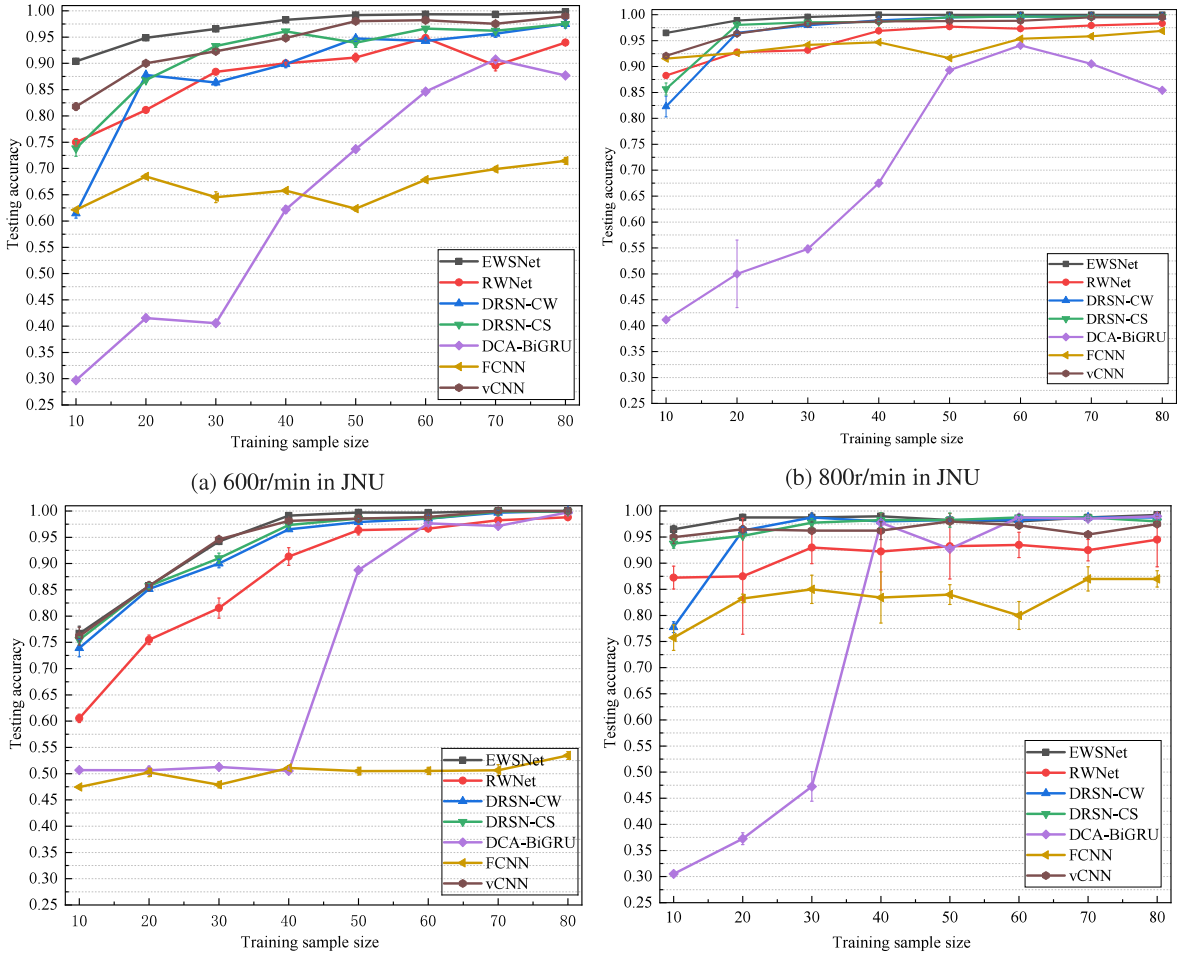
4.1.3. BJTU data set

As shown in Fig. 2, BJTU data set comes from the double-span and double-rotor fault experimental platform, with a model named NSK-6308 deep groove ball bearing. The sampling frequency is 10 kHz, and the speed is 900rmp/min. Four states, inner ring failure, outer ring failure, rolling element failure and normal, are simulated (shown in Fig. 2) in Data set D, $\zeta = 0.5$.

From Fig. 8(d), for one thing, EWSNet achieves the greatest 96.5% when there are 10 samples per class, an improvement of 1.5% over the cases where wavelet weight initialization is not used. And the performance of soft threshold is marginally better than the wavelet convolution layer using ResNet. The combination of wavelet weight initialization and BDAT outperforms the SOTA models that use only one strategy.

For another thing, the performance gaps between models become smaller as the sample size increases. In Table 2, EWSNet can vastly reduce training time, roughly 3∼10 times faster than more sophisticated models like DRSN. The training speed is about 3.5 s faster than vCNN, which indicates that wavelet weight initialization is more likely to drive neural networks to converge faster, making it more suitable for processing raw signals.

In Eq. (16), different scale smoothing factors ζ have essential impacts on the model. According to Table 3, we investigate the effects of different ζ when the training set is 20. 99.75% is attained when $\zeta = 0.5$ because a proper scale smoothing factor can construct more appropriate wavelet scales to build the weight initialization.



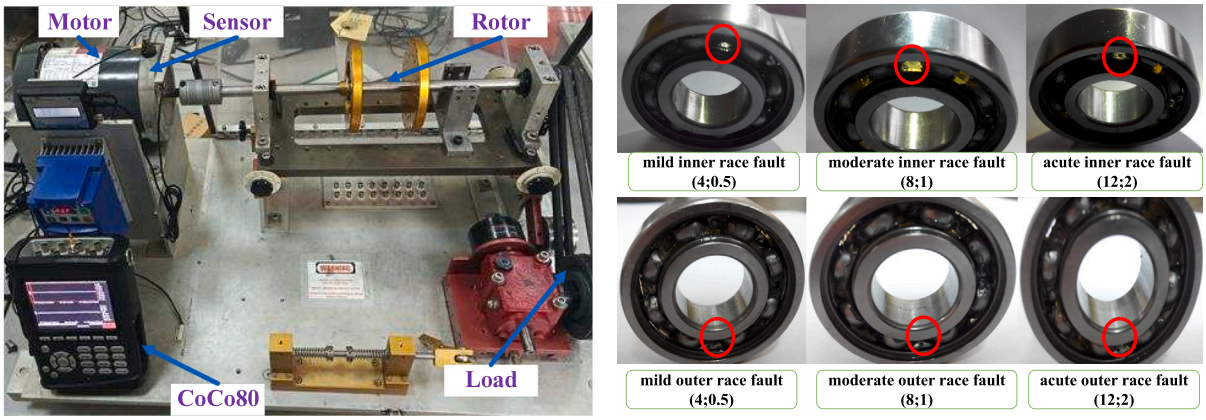
(a) 600r/min in JNU

(b) 800r/min in JNU

(c) sharply variable speed in SQV

(d) constant speed in BJU

Fig. 8. Performance of different models.



(a) Test bench for SQV

(b) Six simulated fault types (Area(mm²); Depth(mm))

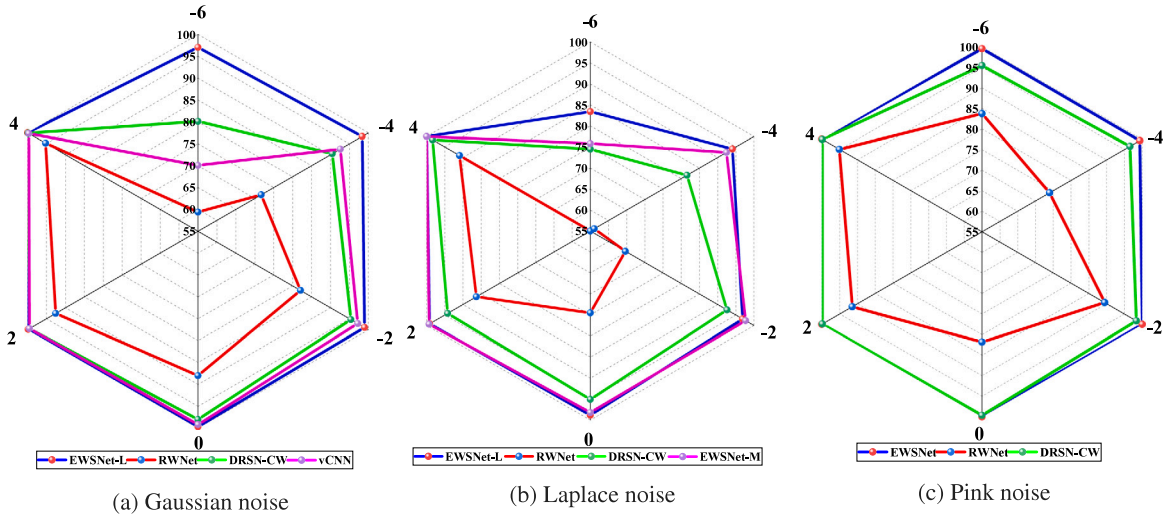
Fig. 9. Data acquisition platform and simulated fault bearing.

Table 2
Training time of models when sample amount is 80.

Model	EWSNet	RWNet [7]	DRSN-CW [15]	DRSN-CS [15]	DCA-BiGRU [29]	FCNN [31]	vCNN
Time (s)	**16.74**	25.86	73.75	118.19	45.29	34.18	20.07

Table 3
The accuracy under different ζ.

ζ	−0.5	−0.4	−0.3	−0.2	−0.1	0.1	0.2	0.3	0.4	0.5
Acc	97.00%	97.50%	99.00%	99.50%	97.00%	99.25%	98.50%	94.25%	95.00%	99.75%



(a) Gaussian noise (b) Laplace noise (c) Pink noise

Fig. 10. Performance of different models under noisy conditions.

4.1.4. Anti-noise robustness

CWRU data set has been widely applied to verify various algorithms [37]. Ten states are collected with the sampling frequency of 12khz. The proportion of training set and test set is 7:3. The noisy immunity of EWSNet, RWNet, and DRSN-CW, are respectively performed under Gaussian noise, Laplace noise, and Pink noise with the signal-to-noise ratio(SNR) from −6 dB to 4 dB:

$$\mathrm{SNR}_{\mathrm{dB}} = 10\lg(P_{signal}/P_{nosie}) \tag{26}$$

where P_{signal} is the original signal power and P_{noise} is the noise power.

According to Fig. 10, RWNet performs poorly because the extracted time–frequency information contains more noises. EWSNet adopts the wavelet weight initialization and BDAT, which can filter noises effectively. In addition, the low SNR Laplace noise causes more disruption to models. However, by Fig. 10(b), Laplace wavelet weight initialization (EWSNet-L) does not focus too much on Laplace noise and may even be more progressive than Morlet wavelet weight initialization (EWSNet-M) with SNR = −6.

4.2. Discussion

4.2.1. Wavelet weight initialization

In Table 4, A denotes Data set A, and C denotes Data set C. Five types of wavelet weight initialization are established. Laplace and Morlet wavelet weights are superior at extracting the features in the constant speed data set, where they are less sensitive to the sample size. When the sample size is 10, the other three wavelet basis functions. The accuracy of methods is all below 86%, and unable to reserve signal features. As for SQV, the sample size has the most significant impacts on models. In conjunction with Fig. 8(c), when the sample size is larger, Gaussian wavelet can extract features better. However, for the convolution kernel of the first layer, with fewer samples, wavelet weight initialization is of little improvement relative to random weight initialization.

4.2.2. Balanced Dynamic Adaptive Threshold

Selecting Data set A with 40 training samples and 160 test samples for each class, the performance of five threshold methods is shown in Table 5. (Without special instructions, training and test samples are kept consistent with the data division above.)

By inserting the correction parameter η, BDAT can achieve a better balance between hard and soft thresholds, making the estimated y closer to the real y. In addition, the parameter of x, which is introduced adaptive slope, does not effectively eliminate noises, the performance of which is worse than soft threshold. The adaptive correction factor α is only added to η.

4.2.3. Normalization strategy

Table 6 shows the performance of various wavelet functions under various normalization methods, where f_{no} indicates no normalization.

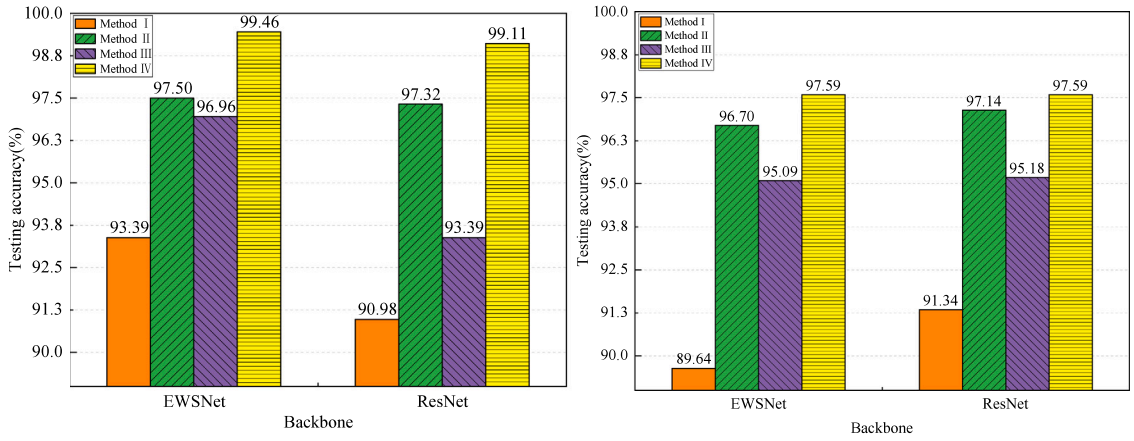
Overall, Laplace wavelet, Morlet wavelet, and Shannon wavelet possess high diagnostic efficiency without the substantial performance gap. For constant speed, the normalized Laplace wavelet weight initialization is more helpful in extracting effective signal features, with about 4.5% upgrade and faster convergence compared to f_{no}, but Mexican Hat wavelet and Gaussian wavelet maybe not generate the weights in favor of EWSNet. Regarding SQV, EWSNet with the f_3 normalized Gaussian wavelet weight initialization reaches 98.04% but requires more training time.

4.2.4. Ablation experiments

Ablation experiments are conducted for Laplace wavelet weight initialization.

From Table 7, each modification can improve the performance of EWSNet. For constant speed, EWSNet shines when the unit time t represents the single convolutional kernel, indicating that the single convolutional kernel is more effective in preserving more information from raw signals. For the condition of sharply various speed, the scale factor s, which changes to $(0, N_k)$, can improve by 2.59%. Sigmoid improves by 3.39%, and using (d) strategy improves by 2.23%. Based on the results above, when the constructed convolutional kernel is physically meaningful, it is more capable of extracting effective features and reducing the training difficulty of neural networks. At the same time, combining optimizing strategies can upgrade convergence speed (e.g., normalization methods).

In Fig. 11, the performance of several methods is compared on EWSNet and ResNet: Method I uses Laplace wavelet convolution layer,



(a) 600r/min in JNU (b) SQV

Fig. 11. Performance of EWSNet and ResNet.

Table 4
Performance of different wavelet basis functions with different samples (%).

Sample amount		10	20	30	40	50	60	70	80
A	Laplace	90.39 ± 0.48	**94.86 ± 0.21**	95.56 ± 0.18	**98.21 ± 0.19**	**99.17 ± 0.21**	**99.35 ± 0.08**	99.31 ± 0.24	**99.79 ± 0.01**
	Morlet	**91.15 ± 0.26**	92.92 ± 0.41	**96.37 ± 0.00**	97.81 ± 0.15	99.01 ± 0.21	99.11 ± 0.15	**99.42 ± 0.54**	99.58 ± 0.17
	Mexican Hat	78.52 ± 0.52	87.22 ± 0.82	91.13 ± 0.24	93.59 ± 0.29	97.53 ± 0.47	98.04 ± 0.30	98.75 ± 0.49	99.58 ± 0.59
	Gaussian	77.60 ± 0.33	83.89 ± 0.17	87.50 ± 0.29	92.50 ± 0.52	95.56 ± 0.00	95.54 ± 0.08	96.59 ± 0.18	98.52 ± 0.09
	Shannon	85.55 ± 0.00	91.39 ± 0.30	93.61 ± 0.25	97.03 ± 0.07	99.01 ± 0.36	99.29 ± 0.08	99.24 ± 0.00	99.79 ± 0.09
C	Laplace	76.84 ± 1.42	**87.54 ± 0.42**	**94.20 ± 0.46**	97.68 ± 0.40	**99.71 ± 0.19**	**99.69 ± 0.01**	99.79 ± 0.20	**100 ± 0.00**
	Morlet	73.31 ± 0.27	83.97 ± 0.04	90.92 ± 0.58	95.62 ± 0.13	98.09 ± 0.47	99.39 ± 0.17	99.79 ± 0.18	**100 ± 0.00**
	Mexican Hat	75.11 ± 0.21	86.43 ± 0.39	90.84 ± 0.43	**98.04 ± 0.26**	98.57 ± 0.09	99.18 ± 0.09	99.23 ± 0.10	**100 ± 0.00**
	Gaussian	**77.22 ± 0.48**	85.39 ± 0.09	93.03 ± 0.14	96.79 ± 0.28	98.67 ± 0.25	99.59 ± 0.04	**100 ± 0.00**	**100 ± 0.00**
	Shannon	74.14 ± 0.59	84.60 ± 0.33	93.53 ± 0.71	97.68 ± 0.32	98.48 ± 0.48	99.29 ± 0.08	99.56 ± 0.41	**100 ± 0.00**

Table 5
Performance of different threshold strategies.

Method	Acc. (%)	α
Ref. [15]	99.06 ± 0.14	1.0000
Ref. [19]	98.72 ± 0.30	0.1140
Ref. [41]	97.96 ± 0.60	0.0942
Hard threshold	98.53 ± 0.59	0.0000
$BDAT_{(ours)}$	**99.25 ± 0.23**	0.7326

Method II turns Laplace wavelet convolution layer into Laplace wavelet weight initialization, Method III introduces enhanced Laplace wavelet convolution layer, and Method IV is the enhanced Laplace wavelet weight initialization proposed in the paper.

Both for EWSNet and ResNet, wavelet weight initialization maintains the physical prior knowledge, is more appropriate for extracting the essential information and boosting the accuracy compared to the wavelet convolution. Method I and II show that replacing the wavelet convolution layer with wavelet weight initialization can substantially improve performance. Compared with Method I and Method III, in the case of using the wavelet convolution layer, the enhanced Laplace wavelet convolution is more capable of extracting sufficient information. In conclusion, the proposed wavelet weight initialization method is universally adaptable.

4.2.5. Analysis of multi-type EWConv and multi-channel EWConv

We also has studied multi-type and multi-channel EWConv. Table 8 shows the performance of different schemes. Scheme I indicates that enhanced Laplace wavelet weight initialization is adopted only in the first layer; Scheme II is that enhanced Laplace wavelet weight initialization is adopted for all convolution layers, which resembles Algorithm Unrolling [6] to a large extent; Scheme III and Scheme IV use Morlet wavelet; Scheme V means that a kernel with spliced Laplace and Morlet wavelet is adopted for the first convolution layer, as shown in Fig. 4(a); Scheme VI denotes that a kernel with concatenating Laplace and Morlet channels is adopted, as shown in Fig. 4(b). Additionally, two feature fusion approaches are established, wherein Scheme VII denotes features extracted by Laplace and Morlet wavelet weights are fused, and Scheme VIII denotes features extracted by Laplace and Morlet wavelet convolutional layers are fused. Except for Scheme II and Scheme IV, others only deal with the first convolutional layer.

Scheme VI can achieve the best accuracy, which proves that hybrid wavelets have better feature extraction capability than individual wavelet. For the condition of constant speed, wavelet weights are utilized in all convolution kernels, which instead is not conducive to extracting useful information. Notably, the accuracy of Morlet wavelet decreases by 6.09% (Scheme IV), indicating that the wavelet weights may not be excel at mining deep information.

For SQV data set (Data set C), most convolutions, which are initialized with wavelet weights, gain outstanding performance. It is because data has tight and complex pulses due to continuous speed changes. Wavelet weights can grasp the feature information from shallow to deep better. Comparing Scheme VII and VIII with Scheme VI, the features extracted by Scheme VI are more effective than Scheme VII.

With poor performance, Scheme V indicates that each channel contains two kinds of kernels such as Laplace and Morlet that destroys the physical principle that only one type of wavelet is involved in the continuous wavelet transform, and due to the existence of stride, some signal points can only be convolved with one type of wavelet, and some signal points can be convolved with two types of wavelets, resulting in feature confusion and thus causing a decrease in recognition accuracy. In contrast, Scheme VI with high performance called the multi-channel EWconv without facing this issue, performs a Laplace wavelet convolution on the front N_k channels and a Morlet wavelet convolution on the back N_k channels, which extracts richer representation features and contributes to the improvement of the recognition accuracy.

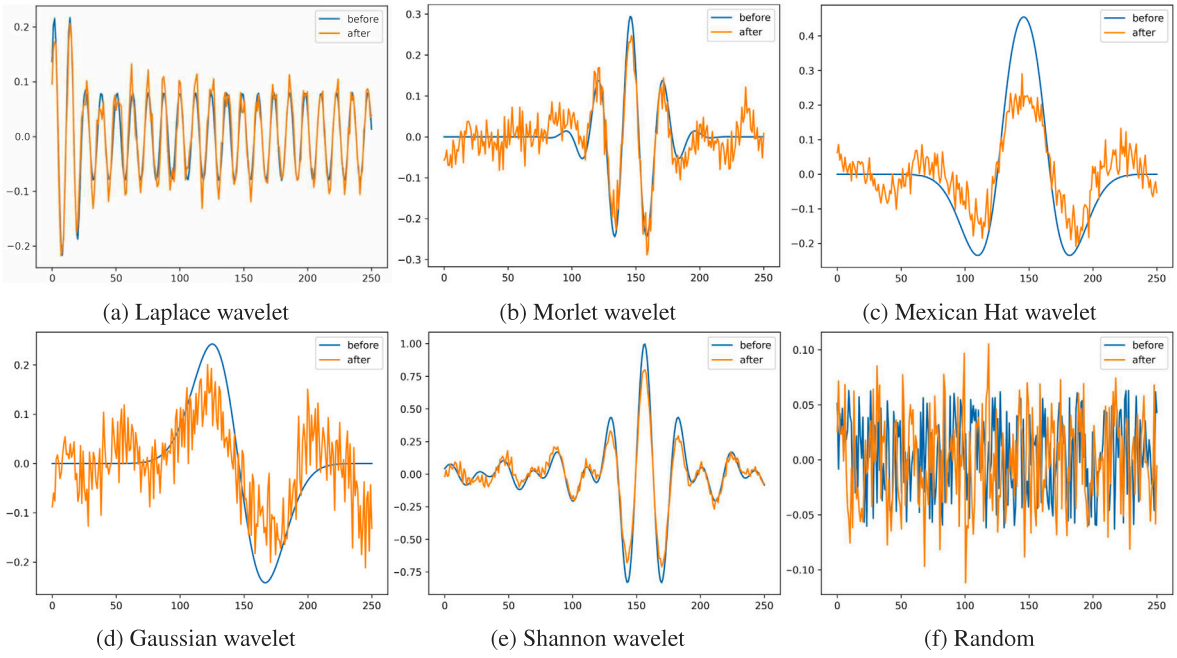


Fig. 12. Wavelet weight initialization of EWSNet before and after training. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Effect of different mapping functions on different wavelet basis functions.

		Laplace		Morlet		Mexican Hat		Gaussian		Shannon	
		Acc. (%)	Time (s)	Acc. (%)	Time (s)	Acc. (%)	Time (s)	Acc. (%)	Time (s)	Acc. (%)	Time (s)
A	F_{no}	94.06	25.62	97.81	27.65	**93.59**	32.22	**92.50**	33.93	96.88	24.76
	F_1	**98.91**	19.51	97.81	23.06	70.16	51.22	79.53	62.21		
	F_2	98.75	15.87	97.19	37.95	68.75	28.15	76.88	32.56		
	F_3	98.59	20.16	97.50	30.68	72.03	40.44	80.16	37.64		
	F_4	98.44	18.29	**97.97**	23.83	53.28	21.15	80.78	49.24		
C	F_{no}	96.88	90.08	97.68	93.69	**97.68**	124.52	96.79	101.28	97.68	85.93
	F_1	97.14	70.84	95.63	75.39	69.46	195.86	90.36	222.92		
	F_2	**97.86**	85.99	**97.95**	71.22	94.11	269.19	97.86	187.47		
	F_3	97.41	68.53	97.59	75.53	93.21	287.28	**98.04**	184.77		
	F_4	96.96	57.55	96.25	73.73	58.53	99.07	94.82	228.42		

Table 7
Ablation experiments of various improved components.

Case	$s \in (0, N_k)$	ζ	Sigmoid	$t \in (0, K-1)$	A acc. (%)	C acc. (%)
EWSNet-L	✓	✓	✓	✓	99.06	99.02
(a)	✗	✓	✓	✓	98.44(−0.62)	96.43(−2.59)
(b)	✗	✗	✓	✓	97.81(−1.25)	98.13(−0.89)
(c)	✓	✓	✗	✓	98.19(−0.87)	95.63(−3.39)
(d)	✓	✓	✓	✗	93.59(−5.47)	96.79(−2.23)

Table 8
Performance comparison of different wavelet kernel design schemes.

Schemes	Accuracy (%)	
	A	C
Scheme I	$98.75_{(\text{baseline})}$	$97.14_{(\text{baseline})}$
Scheme II	97.66(−1.09)	98.48(+1.34)
Scheme III	97.19(−1.56)	97.23(+0.09)
Scheme IV	92.66(−6.09)	98.04(+0.90)
Scheme V	93.13(−5.62)	96.52(−0.62)
Scheme VI	**99.38(+0.63)**	**98.21(+1.07)**
Scheme VII	98.59 ± 0.41	97.64 ± 0.49
Scheme VIII	97.31 ± 0.33	96.05 ± 0.29

5. Model interpretability analysis

5.1. Intrinsic interpretability of wavelet weight initialization

Wavelet weight initialization and random weight initialization are shown in Fig. 12. Each figure contains the wavelet weights before training (blue) and after training (orange). The convolution kernels initialized by wavelet weights transform signals into multiscale frequency bands and adaptively learn more relevant features to the faults.

After training, the Laplace wavelet weights adjust minimally, indicating that Laplace wavelet weights are closer to the optimal wavelet weights (Fig. 12(a)). From the consistency of wavelet weights before and after training, Laplace wavelet, Morlet wavelet, and Shannon wavelet have better consistency and need not make extensive adjustments. But Mexican Hat wavelet and Gaussian wavelet require more significant adjustments.

5.2. Post-hoc explanations of EWSNet

Figs. 13 and 14 display the results of Data set A and Data set C visualized by TGAM, respectively. The blue color represents the low attention of EWSNet. In contrast, the red color indicates intense attention, where they show the concentration in different regions of raw signals.

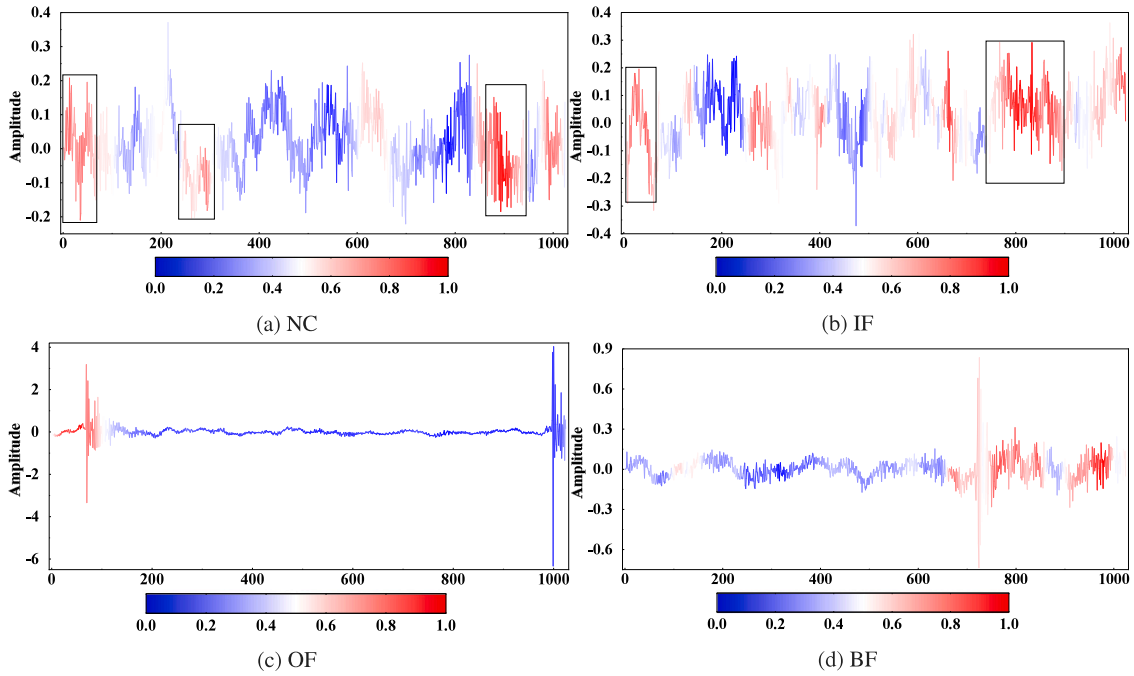


Fig. 13. Visualization of classification weights for different faults in JNU. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

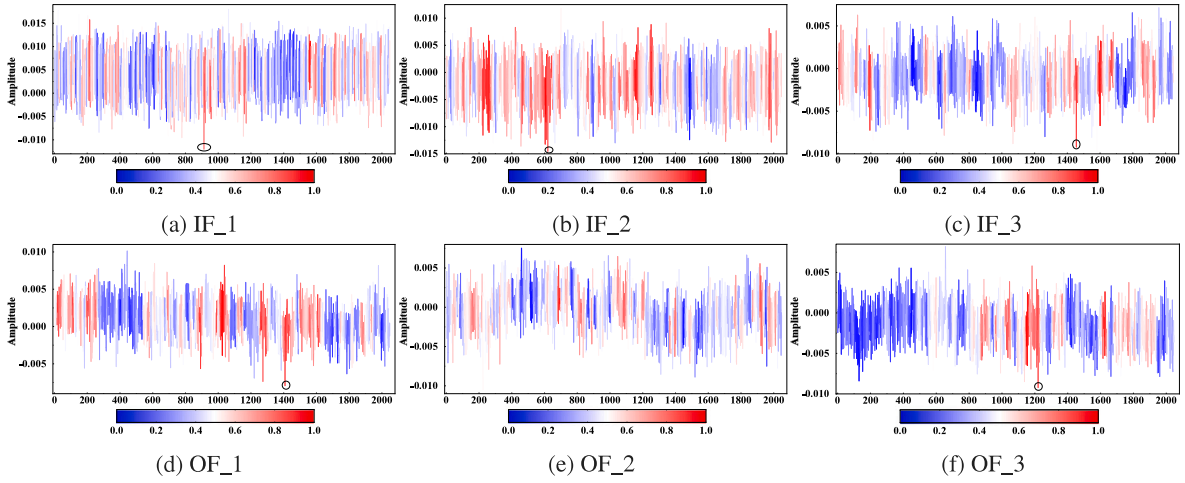


Fig. 14. Visualization of classification weights for different faults in SQV. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

From Data set A, the main focus for NC is on 'peak-concave' pulses, whereas the primary focus for IR is on 'peak-convex' pulses. For OR, the primary emphasis is on the first shorter pulse; for BF, the main focus is on the area after the pulse.

When it comes to Data set C, the vibration amplitude of the IR fault is more stable, while the amplitude of the OR fault has an upward and downward trend. In addition, the acute fault contains more concentrated blue areas. EWSNet mainly focuses on the peak threshold in the negative direction in addition to the moderate outer ring fault (OF_2). For example, IF_1 whose peak in the adverse order is mapped to red, demonstrates that EWSNet pays attention to this part, which is one of the bases for determining fault classification.

6. Conclusion and future work

In the paper, we propose the prior embedded wavelet weight initialization with generics and physical interpretability, which can be widely utilized in weight initialization of the first convolution layer based on CNN. For rolling bearings, under conditions of constant or sharply various speed, the performance of various wavelet weight initialization strategies is investigated by different data sets. In addition, a balanced dynamic adaptive threshold algorithm is presented, where superior outcomes are achieved in experiments.

In summary, this study showcases that in the future backbone networks for fault diagnosis, not only wide convolution kernel is needed, but also the blessing of wavelet weight initialization proposed in this paper.

The following questions merit further consideration. Firstly, wavelet weight initialization is sensitive to ζ, where an inappropriate ζ may lead to erratic results, and how to obtain the appropriate parameters by means of calculation remains a challenge. In addition, the paper mainly concerns bearings, and future research can be extended to other rotating machinery such as gearboxes or shafts. Lastly, the promotion

effect of wavelet weight initialization on transfer learning will be studied in the future.

CRediT authorship contribution statement

Chao He: Writing – original draft, Software, Validation, Visualization, Methodology, Investigation, Proof-reading. **Hongmei Shi:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Jin Si:** Investigation, Writing – review & editing, Data curation, Resources. **Jianbo Li:** Investigation, Writing – review & editing, Data curation, Resources.

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.

Acknowledgments

The authors are grateful for the supports of the Fundamental Research Funds for the Central Universities (Science and Technology Leading Talent Team Project) (2022JBXT005), and the National Natural Science Foundation of China (No. 52272429).

References

[1] Zhang J, Gao RX. Deep learning-driven data curation and model interpretation for smart manufacturing. Chin J Mech Eng 2021;34(03):65–85. http://dx.doi.org/10.1186/s10033-021-00587-y.

[2] Wang J, Li Y, Gao RX, Zhang F. Hybrid physics-based and data-driven models for smart manufacturing: Modelling, simulation, and explainability. J Manuf Syst 2022;63:381–91. http://dx.doi.org/10.1016/j.jmsy.2022.04.004.

[3] Vollert S, Atzmueller M, Theissler A. Interpretable machine learning: A brief survey from the predictive maintenance perspective. In: 26th IEEE international conference on emerging technologies and factory automation, ETFA 2021, Vasteras, Sweden, September 7-10, 2021. IEEE; 2021, p. 1–8. http://dx.doi.org/10.1109/ETFA45728.2021.9613467.

[4] von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, Kirsch B, Pfrommer J, Pick A, Ramamurthy R, Walczak M, Garcke J, Bauckhage C, Schuecker J. Informed machine learning – A taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans Knowl Data Eng 2023;35(1):614–33. http://dx.doi.org/10.1109/TKDE.2021.3079836.

[5] Lu H, Nemani VP, Barzegar V, Allen C, Hu C, Laflamme S, Sarkar S, Zimmerman AT. A physics-informed feature weighting method for bearing fault diagnostics. Mech Syst Signal Process 2023;191:110171. http://dx.doi.org/10.1016/j.ymssp.2023.110171.

[6] Monga V, Li Y, Eldar YC. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. IEEE Signal Process Mag 2021;38(2):18–44. http://dx.doi.org/10.1109/MSP.2020.3016905.

[7] Li T, Zhao Z, Sun C, Cheng L, Chen X, Yan R, Gao RX. WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis. IEEE Trans Syst Man Cybern Syst 2022;52(4):2302–12. http://dx.doi.org/10.1109/TSMC.2020.3048950.

[8] Yan R, Shang Z, Xu H, Wen J, Zhao Z, Chen X, Gao RX. Wavelet transform for rotary machine fault diagnosis: 10 years revisited. Mech Syst Signal Process 2023;200:110545. http://dx.doi.org/10.1016/j.ymssp.2023.110545.

[9] Tai J, Liu C, Wu X, Yang J. Bearing fault diagnosis based on wavelet sparse convolutional network and acoustic emission compression signals. Math Biosci Eng 2022;19(8):8057–80. http://dx.doi.org/10.3934/mbe.2022377.

[10] Liao M, Liu C, Wang C, Yang J. Research on a rolling bearing fault detection method with wavelet convolution deep transfer learning. IEEE Access 2021;9:45175–88. http://dx.doi.org/10.1109/ACCESS.2021.3067152.

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

[12] Li T, Sun C, Li S, Wang Z, Chen X, Yan R. Explainable graph wavelet denoising network for intelligent fault diagnosis. IEEE Trans Neural Netw Learn Syst 2022;1–14. http://dx.doi.org/10.1109/TNNLS.2022.3230458.

[13] Li S, Li T, Sun C, Chen X, Yan R. WPConvNet: An interpretable wavelet packet kernel-constrained convolutional network for noise-robust fault diagnosis. IEEE Trans Neural Netw Learn Syst 2023;1–15. http://dx.doi.org/10.1109/TNNLS.2023.3282599.

[14] Dai L, Guo J, Wan J-L, Wang J, Zan X. A reliability evaluation model of rolling bearings based on WKN-BiGRU and Wiener process. Reliab Eng Syst Saf 2022;225:108646. http://dx.doi.org/10.1016/j.ress.2022.108646.

[15] Zhao M, Zhong S, Fu X, Tang B, Pecht M. Deep residual shrinkage networks for fault diagnosis. IEEE Trans Ind Inform 2020;16(7):4681–90. http://dx.doi.org/10.1109/TII.2019.2943898.

[16] Pei X, Dong S, Tang B, Pan X. Bearing running state recognition method based on feature-to-noise energy ratio and improved deep residual shrinkage network. IEEE/ASME Trans Mechatron 2022;27(5):3660–71. http://dx.doi.org/10.1109/TMECH.2021.3120755.

[17] Yang X, Chi F, Shao S, Zhang Q. Bearing fault diagnosis under variable working conditions based on deep residual shrinkage networks and transfer learning. J Sensors 2021;2021. http://dx.doi.org/10.1155/2021/5714240.

[18] Yu Y, Guo L, Gao H, Liu Y, Feng T. Pareto-optimal adaptive loss residual shrinkage network for imbalanced fault diagnostics of machines. IEEE Trans Ind Inf 2022;18(4):2233–43. http://dx.doi.org/10.1109/TII.2021.3094186.

[19] Zhang Z, Li H, Chen L. Deep residual shrinkage networks with self-adaptive slope thresholding for fault diagnosis. In: 2021 7th international conference on condition monitoring of machinery in non-stationary operations. CMMNO, 2021, p. 236–9. http://dx.doi.org/10.1109/CMMNO53328.2021.9467549.

[20] Salimy A, Mitiche I, Boreham P, Nesbitt A, Morison G. Robust deep residual shrinkage networks for online fault classification. In: 2021 29th European signal processing conference. EUSIPCO, 2021, p. 1691–5. http://dx.doi.org/10.23919/EUSIPCO54536.2021.9616148.

[21] Chen Y, Zhang D, Zhang H, Wang Q-G. Dual-path mixed domain residual threshold networks for bearing fault diagnosis. IEEE Trans Ind Electron 2022;69(12):13462–72. http://dx.doi.org/10.1109/TIE.2022.3144572.

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

[23] Mey O, Neufeld D. Explainable AI algorithms for vibration data-based fault detection: Use case-adadpted methods and critical evaluation. Sensors 2022;22(23):9037. http://dx.doi.org/10.3390/s22239037.

[24] Li X, Zhang W, Ding Q. Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. Signal Process 2019;161:136–54. http://dx.doi.org/10.1016/j.sigpro.2019.03.019.

[25] Li Y, Zhou Z, Sun C, Chen X, Yan R. Variational attention-based interpretable transformer network for rotary machine fault diagnosis. IEEE Trans Neural Netw Learn Syst 2022;1–14. http://dx.doi.org/10.1109/TNNLS.2022.3202234.

[26] Grezmak J, Zhang J, Wang P, Loparo KA, Gao RX. Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis. IEEE Sens J 2020;20(6):3172–81. http://dx.doi.org/10.1109/JSEN.2019.2958787.

[27] Xu Y, Yan X, Sun B, Liu Z. Global contextual residual convolutional neural networks for motor fault diagnosis under variable-speed conditions. Reliab Eng Syst Saf 2022;225:108618. http://dx.doi.org/10.1016/j.ress.2022.108618.

[28] Chen B, Liu T, He C, Liu Z, Zhang L. Fault diagnosis for limited annotation signals and strong noise based on interpretable attention mechanism. IEEE Sens J 2022;22(12):11865–80. http://dx.doi.org/10.1109/JSEN.2022.3169341.

[29] Zhang X, He C, Lu Y, Chen B, Zhu L, Zhang L. Fault diagnosis for small samples based on attention mechanism. Measurement 2022;187:110242. http://dx.doi.org/10.1016/j.measurement.2021.110242.

[30] Li S, Li T, Sun C, Yan R, Chen X. Multilayer Grad-CAM: An effective tool towards explainable deep neural networks for intelligent fault diagnosis. J Manuf Syst 2023;69:20–30. http://dx.doi.org/10.1016/j.jmsy.2023.05.027.

[31] Kim MS, Yun JP, Park P. An explainable neural network for fault diagnosis with a frequency activation map. IEEE Access 2021;9:98962–72. http://dx.doi.org/10.1109/ACCESS.2021.3095565.

[32] Kim MS, Yun JP, Park P. Deep learning-based explainable fault diagnosis model with an individually grouped 1D convolution for 3-axis vibration signals. IEEE Trans Ind Inf 2022;18(12):8807–17. http://dx.doi.org/10.1109/TII.2022.3147828.

[33] Feng K, Jiang Z, He W, Qin Q. Rolling element bearing fault detection based on optimal antisymmetric real Laplace wavelet. Measurement 2011;44(9):1582–91. http://dx.doi.org/10.1016/j.measurement.2011.06.011.

[34] Zhang W, Peng G, Li C, Chen Y, Zhang Z. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors 2017;17(2):425. http://dx.doi.org/10.3390/s17020425.

[35] Zhao Z, Li T, Wu J, Sun C, Wang S, Yan R, Chen X. Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study. ISA Trans 2020;107:224–55. http://dx.doi.org/10.1016/j.isatra.2020.08.010.

[36] Li K. School of Mechanical Engineering, Jiangnan University. 2020, http://mad-net.org:8765/explore.html?t=0.5831516555847212.

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

[38] Shi Z, Chen J, Zi Y, Zhou Z. A novel multitask adversarial network via redundant lifting for multicomponent intelligent fault detection under sharp speed variation. IEEE Trans Instrum Meas 2021;70:1–10. http://dx.doi.org/10.1109/TIM.2021.3055821.

[39] Si J, Shi H, Han T, Chen J, Zheng C. Learn generalized features via multi-source domain adaptation: Intelligent diagnosis under variable/constant machine conditions. IEEE Sens J 2022;22(1):510–9. http://dx.doi.org/10.1109/JSEN.2021.3126864.

[40] Si J, Shi H, Chen J, Zheng C. Unsupervised deep transfer learning with moment matching: A new intelligent fault diagnosis approach for bearings. Measurement 2021;172:108827. http://dx.doi.org/10.1016/j.measurement.2020.108827.

[41] Zhang Z, Chen L, Zhang C, Shi H, Li H. GMA-DRSNs: a novel fault diagnosis method with global multi-attention deep residual shrinkage networks. Measurement 2022;196:111203. http://dx.doi.org/10.1016/j.measurement.2022.111203.