Electronic Research Archive

ERA, 33(9): 5301–5322.
DOI: 10.3934/era.2025237
Received: 28 June 2025
Revised: 21 August 2025
Accepted: 05 September 2025
Published: 10 September 2025

https://www.aimspress.com/journal/ERA

Research article

Adaptive hybrid attention mechanism deep residual threshold networks for bearing fault diagnosis under noisy environments

Yan Wang[1], Kangwen Sun[1], Yongkao Li[1] and Haoquan Liang[2,*]

[1] School of Aeronautic Science and Engineering, Beihang University, Beijing 100191, China
[2] Institute of Unmanned System, Beihang University, Beijing 100191, China

* **Correspondence:** Email: haoquan.liang@buaa.edu.cn.

Abstract: Intelligent bearing fault diagnosis based on deep learning has immense potential. However, improving the noise immunity, generality, and accuracy of fault diagnosis methods is still challenging. This paper proposes a novel adaptive hybrid attention mechanism deep residual threshold network (AHA-RTN) for bearing fault diagnosis under various noise conditions. First, channel-wise and spatial attention were both integrated into residual blocks to capture multiscale information. Hybrid attention was obtained using the proposed adaptive attention module, which computes adjustment coefficients of each attentional mechanism. Next, a novel noise reduction activation function based on soft thresholding was incorporated to suppress noise. Finally, the method was validated on two distinct bearing datasets under various noise conditions. The results show that the proposed AHA-RTN has better noise immunity and accuracy than the other advanced multiscale convolutional neural networks.

Keywords: deep learning; fault diagnosis; attention mechanism; threshold denoising; deep residual threshold network

1. Introduction

Rotating components are vital parts of mechanical equipment in many industrial sectors, and they are crucial in maintaining operational safety and reliability [1]. The failure of rotating components, such as rolling bearings, can result in substantial economic losses and pose a significant risk to production safety [2]. Their vibration signals exhibit nonlinear and non-stationary characteristics, with typical fault signals characterized primarily by impulsive features [3]. However, the vibration signals

collected from sensors often contain considerable noise, making it difficult to discern the weak fault characteristic signals of rotating components [4]. Therefore, it is necessary to perform intelligent diagnostics on faulty rotating components.

Extracting information from signals that effectively characterize different fault features is crucial for intelligent fault diagnosis. Preprocessing and analyzing collected signals using signal processing methods such as bandwidth-based empirical mode decomposition (BEMD) [5], improved uniform phase empirical modal decomposition (IUPEMD) [6], an adaptive and concise empirical wavelet transform (ACEWT) [7], a ghost module and efficient channel attention module (GE)-improved EfficientNet model (GE-EfficientNet) [8], variational mode decomposition (VMD) [9], and single-valued neutrosophic cross-entropy (SVNCE) [10] can help suppress signal noise. However, these methods require complex algorithms and an understanding of signal processing, which requires significant individual expertise.

Deep learning has been widely applied across various fields and has become a key focus of current academic research [11–13]. Intelligent bearing fault diagnosis using deep learning reduces manual feature extraction, decreases dependence on signal processing knowledge, improves diagnostic accuracy, and simplifies computational complexity. Zhang et al. [14] proposed a data-driven, deep learning approach based on the ConvNeXt architecture (ConvNeXt-TTSP). This approach does not rely on complex physical models or specialized expertise. Yang et al. [15] proposed a twin broad learning system (TBLS) to address deep learning models' limitations of heavy reliance on large datasets and time-consuming training and testing, for rotating machinery fault diagnosis. Hei et al. [16] proposed a novel domain-adaptive Wasserstein conditional generative adversarial network (DA-WGAN) to address the challenges of scarce labeled data and imbalanced samples in early bearing fault diagnosis under various operating conditions. Li et al. [17] proposed a data augmentation method for fault diagnosis based on an adaptive diffusion model integrated with generative adversarial networks (ADGANs), which effectively addresses the data imbalance problem and significantly improves fault diagnosis performance. Zhang et al. [18] proposed a federated transfer learning method featuring prior alignment and feature adaptation schemes, which enables cross-domain shared feature extraction without simultaneous processing of source and target data. Li et al. [19] proposed the multi-branch feature cross-fusion bearing fault diagnosis model (MCFormer), which leverages a multi-branch parallel perception architecture, a feature cross-fusion strategy, and a dynamic classifier module to achieve superior diagnostic performance.

As the depth of neural networks increases, their accuracy becomes saturated and then degrades rapidly. He et al. [20] proposed a residual learning framework to address the problem by introducing shortcut connections. Later, Huang et al. [21] introduced the dense convolutional network (DenseNet), which connects each layer to every other layer in a feed-forward manner, improving information flow and gradient propagation. Deep residual networks are becoming increasingly popular diagnosis tools [22,23]. However, collected industrial data often contain significant noise that traditional deep learning algorithms struggle to discern, hindering the improvement in accuracy [24]. To solve this problem, Zhang et al. [25] configured the first convolutional layer of the convolutional neural network (CNN) with wide convolutional kernels to suppress high-frequency noise. They employed multiple layers of small convolutional kernels to deepen the network, enhancing the CNN's ability to handle noisy data. Furthermore, an adaptive denoising convolutional neural network (ADCNN) that removes noise while preserving fault features was proposed by Wang et al. [26].

The introduction of attention mechanisms provides a new train of thought [27,28]. Snyder et al. [29]

proposed a dual-head ensemble transformer (DHET), which leverages self-attention mechanisms to capture long-range dependencies and global contextual information. Hu et al. [30] proposed a squeeze-and-excitation network (SENet) that enhances the representational power of CNNs by explicitly modeling the interdependencies between channels. Subsequent models, such as the efficient channel attention (ECA) module [31], global second-order pooling network (GSoP-Net) [32], and frequency channel attention networks (FcaNet) [33], have further demonstrated the effectiveness of channel attention mechanisms. Zhao et al. [34] proposed the deep residual shrinkage network with channel-wise thresholds (DRSN-CW), which incorporates a residual shrinkage mechanism into deep neural networks to reduce noise and improve fault diagnosis accuracy. However, methods that only consider channel-wise thresholds struggle to effectively extract features from high-noise data and fail to simultaneously capture the breadth and depth of noise features, leading to biased thresholding and reduced fault diagnosis accuracy.

To address these limitations, we propose that the AHA-RTN set the threshold value by effectively capturing the noise information in the original signal through adaptive hybrid attention, thereby improving accuracy through soft thresholding. The main contributions of this paper are summarized as follows:

1) Incorporating channel and spatial attention mechanisms into residual blocks helps capture multiscale noise information from raw data, thereby suppressing its interference with effective feature information.

2) A denoising activation function based on soft thresholding is proposed in the nonlinear transform layer, which suppresses noise by integrating the channel-wise average and spatial average with attention factors and setting data within the threshold range to zero.

3) The proposed AHA-RTN uses multiscale convolutional kernels to capture cross-channel and spatial information, generating adaptation coefficients to quantitatively blend channel and spatial attention weights; its effectiveness in denoising via spatial attention and adaptability are validated through comparisons with the DRSN-CW, ResNet-18, and the designed DRSN-SW.

2. Proposed methods

2.1. AHA-RTN architecture

This paper proposes the AHA-RTN, which comprises multiscale convolutional layers, channel attention modules, spatial attention modules, and an adaptive hybrid module. Utilizing a residual network architecture addresses network degradation issues that arise with increasing depth through shortcut connections [35]. The continuous learning of residuals facilitates network optimization and enhances generalization as the depth increases. The raw data first undergoes two convolutional layers for initial feature extraction, and then is fed into three branches, whose functions are to obtain channel weight coefficients, spatial weight coefficients, and denoising thresholds, respectively. Figure 1 shows the structure of the AHA-RTN, which can be integrated into network architectures such as ResNets to provide plug-and-play functionality. This enhances the ability of ResNets and similar architectures to handle high-noise data. The overall architecture of the AHA-RTN used for fault diagnosis is shown in Figure 2.

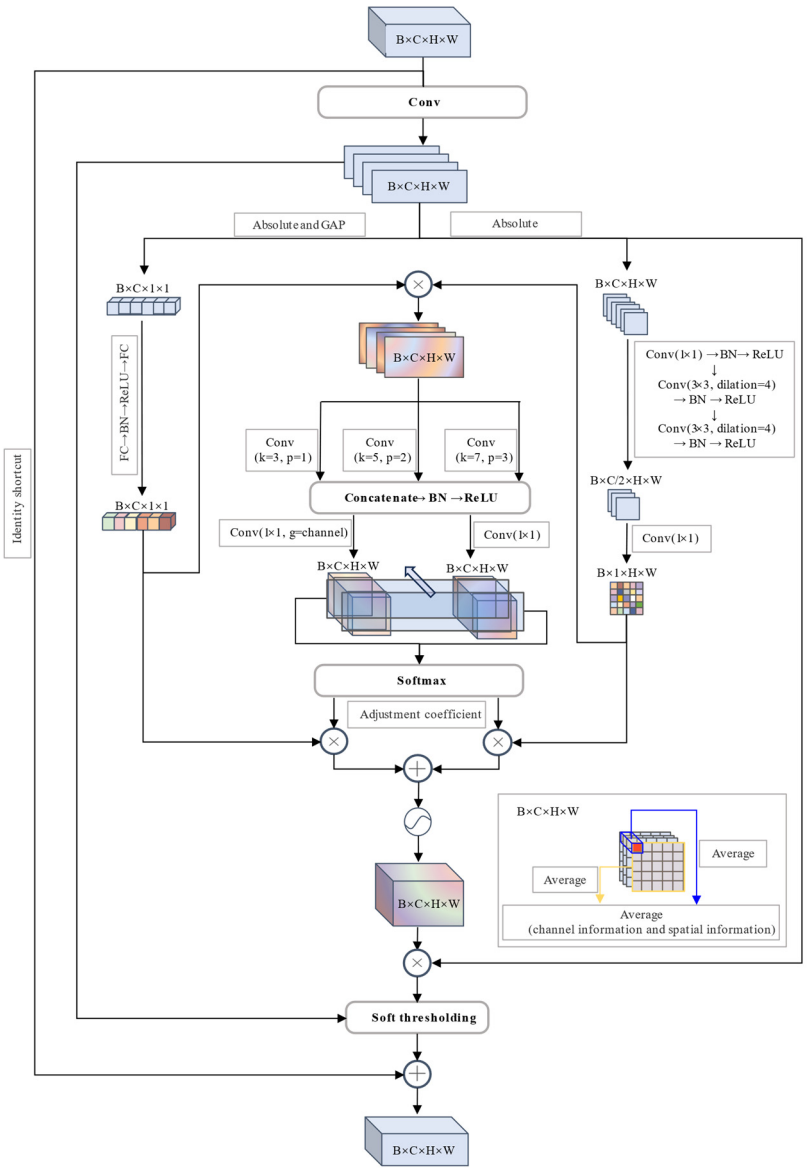


Figure 1. AHA-RTN architecture.

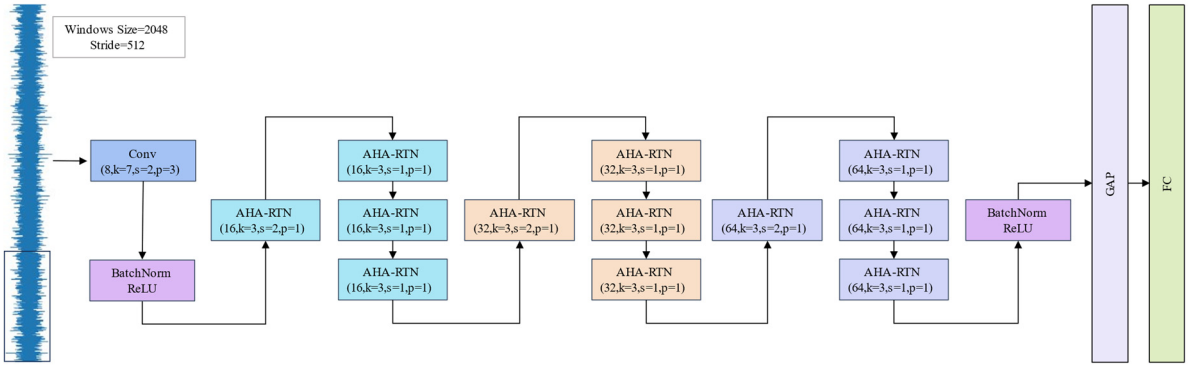


Figure 2. Overall AHA-RTN architecture.

2.2. Attention mechanism

The attention mechanism enables networks to focus selectively on relevant input features, enhancing recognition of complex patterns and critical information while boosting efficiency by reducing irrelevant data processing. Channel attention prioritizes global features in CNNs by weighting feature channels to avoid redundancy, streamlining operations. Complementarily, spatial attention focuses on modeling local features, using spatial coefficient matrices to emphasize locally significant information, with local details enhancing feature representation.

2.2.1. Channel attention branch

In the channel attention branch, data undergoes absolute and global average pooling (GAP), which compresses spatial information into channels and facilitates the acquisition of global information across different feature maps. Subsequently, the data undergoes sequential processing through fully connected (FC) layers, batch normalization layers, and a rectifier linear unit (ReLU) activation function, followed by another pass through FC layers to obtain more comprehensive channel weight coefficients. The subsequent adaptive attention module uses the channel weight coefficients obtained at this stage to determine the corresponding tuning coefficients and is also involved in the hybrid attention fusion. The channel attention weights are obtained as follows:

$$M_{channel} = F_{fc}\left(\gamma\left(F_{fc}(X)\right)\right) \tag{1}$$

where X and γ represent the feature data and the ReLU activation function, respectively.

The above operation facilitates the enhancement of advantageous feature channels while suppressing redundant ones, increasing network adaptability while reducing computational load.

2.2.2. Spatial attention branch

Local detailed information on features is modeled in the spatial attention branch to enable the network to focus on crucial local features. Emphasizing only global features while ignoring local feature information may cause the network to lose sensitivity to subtle changes. The data undergo absolute value operations and then enter the convolutional layer, which are then processed by the batch normalization (BN) layer and the ReLU activation function to obtain spatial coefficient matrices with different descriptive capabilities and relative simplicity. Null convolution is used to extend the perceptual domain to capture the correlation between different spatial location information. The spatial weight coefficients obtained at this stage will be used for subsequent adaptive attention adjustment units to obtain corresponding adjustment coefficients and for the fusion acquisition of hybrid attention. The spatial attention weights are obtained as follows:

$$\begin{aligned} M_s &= \gamma\left(k_{d=4}^{3\times3}\left(\gamma\left(k^{1\times1}(X)\right)\right)\right) \\ M_{spatial} &= k^{1\times1}\left(\gamma\left(k_{d=4}^{3\times3}(M_s)\right)\right) \end{aligned} \tag{2}$$

where k represents a convolutional layer.

The network dynamically assigns weights based on the importance of features in the local space through the above steps, highlighting responses with local detail.

2.3. Adaptive attention module

After obtaining the channel and spatial attention weights, the hybrid attention weights are calculated by multiplying them pointwise. As the useful signal is mixed with noise, employing a simple fused attention mechanism for subsequent denoising tasks may amplify the negative effects associated with channel or spatial attention weights. Therefore, the network needs to incorporate an adaptive attention selection mechanism to attenuate the attention bias induced by noise. By passing the fused hybrid attention weights to the convolutional layers with kernel sizes of 3 × 3, 5 × 5, and 7 × 7, the multiscale convolutional kernels help to expand the receptive field of the network. This allows the initial fused attention features to capture more contextual information about the connections between channels and spatial dimensions. The individual outputs of these convolutional layers are then concatenated along the channel dimension to obtain richer contextual information:

$$\begin{aligned} &M_{simple\_mixture} = M_{channel} \cdot M_{spatial} \\ &M_{cat} = \gamma\left(cat\left(k^{3\times3}\left(M_{simple\_mixture}\right), k^{5\times5}\left(M_{simple\_mixture}\right), k^{7\times7}\left(M_{simple\_mixture}\right)\right)\right) \end{aligned} \tag{3}$$

To obtain mixed matrix guiding attention fusion, measuring the effect of spatial information on channels within the channel domain and that of channels on spatial locations within the spatial domain is necessary. First, the semantically rich channel and spatial contextual information obtained earlier are processed through grouped convolution using 1 × 1 convolutional kernels, with the number of groups set to the corresponding number of channels. This approach allows different convolutional kernels on different feature dimensions to capture the spatial effects within their respective channels. Conversely, point-to-point convolution is performed using 1 × 1 convolutional layers to obtain the effect of different channels on spatial locations. The obtained attention effect information is then processed using the softmax operator, with softmax operations conducted on the same feature dimensions and spatial positions. Finally, two adjustment coefficients are obtained to guide the deep fusion of channel and spatial attention, as follows:

$$\begin{aligned} &M^1 = k^{1\times1}_{g=channel}(M_{cat}) \\ &M^2 = k^{1\times1}(M_{cat}) \\ &M_{guide\_1} = \frac{exp\left(M^1_{i,j,k,l}\right)}{exp\left(M^1_{i,j,k,l}\right)+exp\left(M^2_{i,j,k,l}\right)} \quad (i = B, j = C, k \in H, l \in W) \\ &M_{guide\_2} = \frac{exp\left(M^2_{i,j,k,l}\right)}{exp\left(M^1_{i,j,k,l}\right)+exp\left(M^2_{i,j,k,l}\right)} \quad (i = B, j = C, k \in H, l \in W) \\ &M_{guide\_1} + M_{guide\_2} = 1 \end{aligned} \tag{4}$$

Subsequently, the two adjustment coefficients are multiplied element-wise with the channel and spatial attention weights, followed by addition. The deep fusion of the hybrid attention is obtained as follows:

$$M_{last\_attention} = M_{guide\_1} \cdot M_{channel} + M_{guide\_2} \cdot M_{spatial} \tag{5}$$

This form of deep-fused attention weights verifies the effectiveness of each attention mechanism

for the task, thus determining whether to emphasize global or local information. The mechanism effectively mitigates the negative effects of inadvertently highlighting certain aspects of information due to noise interference, particularly when dealing with noisy data.

2.4. Denoising threshold acquisition branch

Soft thresholding is a pivotal technique in signal denoising, whereby information within the threshold is nullified, while that beyond it is preserved. Unlike hard thresholding, which preserves information entirely, soft thresholding adjusts information based on the threshold, rendering it more gradually. The ReLU activation function sets negative values to zero, but it cannot distinguish between noise and signal data. Similar to the ReLU activation function, soft thresholding can partially alleviate the issue of gradient vanishing. Thus, combining hybrid attention and soft thresholding allows for the denoising activation of data.

It is imperative that thresholds are not set at an excessively elevated level, as this may result in a significant loss of information. Furthermore, it is crucial to ensure that the thresholds are not set at a negative value. In order to circumvent the potential for bias in thresholds resulting from the incorporation of superfluous information from the global average of raw data, the threshold branch employs an absolute value transformation as the initial step. Subsequently, the average values of different channels and spatial positions are computed separately. Finally, the processed data are multiplied pointwise with the hybrid attentional weights to obtain the threshold for the soft thresholding function:

$$\begin{aligned} \beta &= \frac{1}{2}\left(\frac{1}{HW}\sum_{k\in H}\sum_{l\in W} X_{i,j,k,l} + \frac{1}{C}\sum_{j\in C} X_{i,j,k,l}\right) \\ \mu &= M_{last\_attention} \cdot \beta \end{aligned} \tag{6}$$

where μ is the threshold value.

The use of residual networks can mitigate the problem of performance degradation resulting from increased network depth. Therefore, the final output of this layer module can be obtained by adding the data after denoising activation to the data after identity mapping.

2.5. DRSN-SW architecture

The DRSN-CW, which uses the channel attention mechanism, significantly improves the model's resilience to interference by integrating an adaptive threshold learning module within residual blocks and processing input features through residual shrinkage units. To investigate the effectiveness of local information in denoising tasks, we propose deep residual shrinkage networks with spatial-wise thresholds (DRSN-SW) that incorporate spatial thresholds. The core functionality of the network includes the acquisition of spatial attention coefficients and denoising thresholds, as shown in Figure 3.

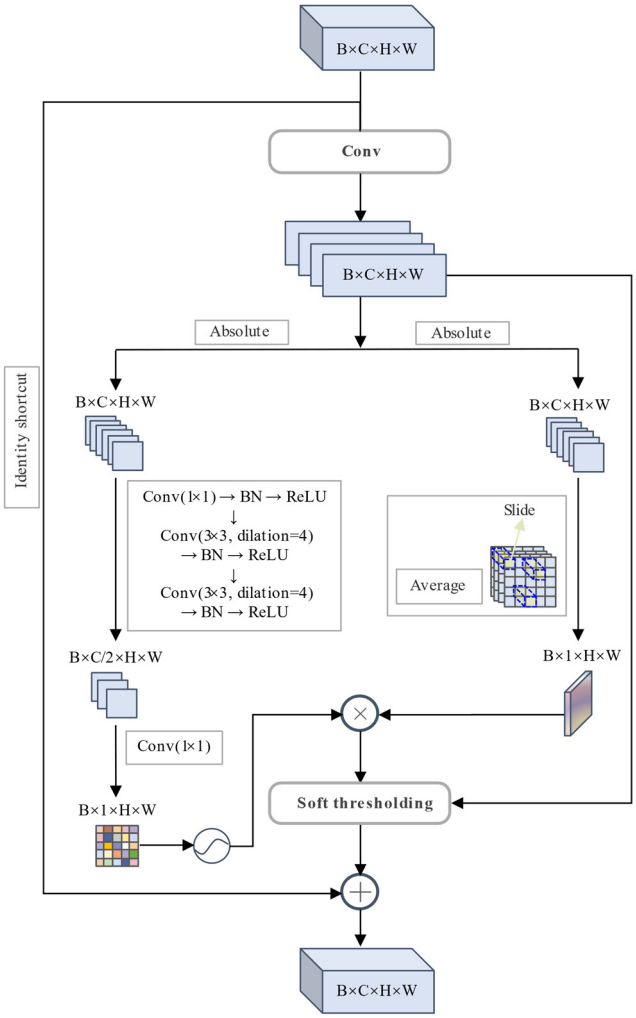


Figure 3. DRSN-SW architecture.

3. Verification and analysis

To evaluate the performance of the proposed AHA-RTN for bearing fault diagnosis under noisy conditions, the Case Western Reserve University (CWRU) and the Southeast University datasets are used for verification and analysis.

3.1. Data preprocessing

In this study, the obtained bearing vibration signals are preprocessed and a sliding window is used for sampling. The window size is set to 2048, with a sliding step of 512. The training and testing data required for the ResNet, DRSN-CW, DRSN-SW, and AHA-RTN all undergo the same operations to ensure that performance differences between different models are not caused by data processing.

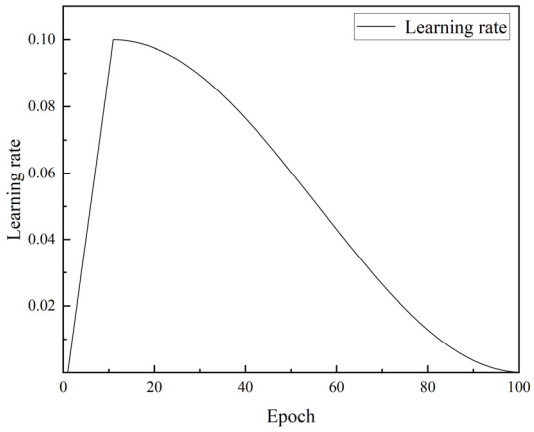


Figure 4. Learning rate curve.

A combination of warm-up and cosine annealing [36] strategies have been used for the learning rate. Specifically, the learning rate varies with the number of epochs, as shown in Figure 4. In the first 10 epochs, the learning rate of the network gradually increases linearly to 0.1. This study used the cross-entropy [37] function to evaluate the loss and optimization objectives during the training, validation, and testing phases. The loss function is defined as follows:

$$Loss(y, ture) = -ln\left(\frac{e^{y[ture]}}{\sum_{i \in class} e^{y[i]}}\right) \tag{7}$$

where y is the final output of the network, $class$ is the total number of categories, and $true$ represents the true category.

The loss function includes a softmax layer and uses regularization techniques to avoid overfitting. Adding a penalty term to the loss function limits the network weights to larger values. The final definition of the loss function is as follows:

$$Loss(y, ture) = -ln\left(\frac{e^{y[ture]}}{\sum_{i \in class} e^{y[i]}}\right) + \lambda \|w\|_2 \tag{8}$$

where λ is the regularization coefficient, set to 0.0001, and w is the weighting coefficient.

Gaussian noise and Laplacian noise are added to the original signal to simulate real conditions with noise interference. Gaussian noise is a type of random noise that conforms to a normal distribution, and it is widely used to simulate random variations and uncertainties in natural phenomena. In contrast, Laplacian noise, characterized by its sharp peaks and long tails, is suitable for simulating real vibration signals with abrupt changes and outliers. These two noise types encompass the two fundamental categories of interference encountered in bearing fault diagnosis, namely continuous and impulsive interference.

$$\begin{aligned} f_{Gaussian}(x|\mu, \sigma) &= \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ f_{Laplacian}(x|\mu, b) &= \frac{1}{2b} exp\left(-\frac{|x-\mu|}{b}\right) \end{aligned} \tag{9}$$

where μ is the mean, σ is the standard deviation, and b is the scale parameter.

The signal-to-noise ratio (SNR) is a key metric for evaluating the quality of acquired signals. The formula for calculating it is as follows:

$$SNR = 10 \cdot lg(P_S/P_N) \tag{10}$$

where P_S is the power of the signal and P_N is the power of the noise.

3.2. *Experimental study on the CWRU dataset*

3.2.1. Data description

The experimental device shown in Figure 5 is widely used in bearing fault diagnosis. The CWRU dataset includes normal baseline data, drive-end bearing fault data sampled at 12 kHz, fan-end bearing fault data, and drive-end bearing fault data sampled at 48 kHz [38]. We evaluated our algorithm using normal baseline data, fan-end bearing fault data, and drive-end bearing fault data sampled at 12 kHz.

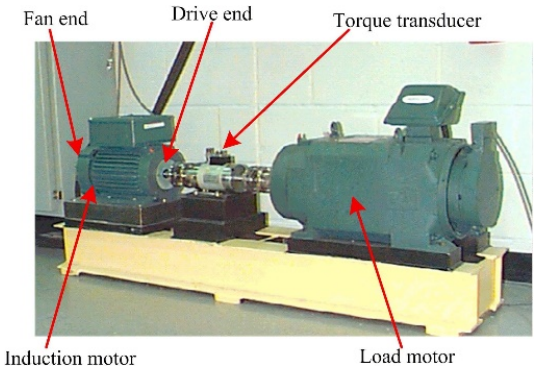


Figure 5. Experimental device of the CWRU dataset [39].

Table 1. CWRU dataset fault type coding

Fault Type	Number	One-Hot Encoding
Ball 7	0	[1,0,0,0,0,0,0,0,0,0,0,0,0]
Inner Race 7	1	[0,1,0,0,0,0,0,0,0,0,0,0,0]
Outer Race 7-3	2	[0,0,1,0,0,0,0,0,0,0,0,0,0]
Outer Race 7-6	3	[0,0,0,1,0,0,0,0,0,0,0,0,0]
Outer Race 7-12	4	[0,0,0,0,1,0,0,0,0,0,0,0,0]
Ball 14	5	[0,0,0,0,0,1,0,0,0,0,0,0,0]
Inner Race 14	6	[0,0,0,0,0,0,1,0,0,0,0,0,0]
Outer Race 14-6	7	[0,0,0,0,0,0,0,1,0,0,0,0,0]
Ball 21	8	[0,0,0,0,0,0,0,0,1,0,0,0,0]
Inner Race 21	9	[0,0,0,0,0,0,0,0,0,1,0,0,0]
Outer Race 21-3	10	[0,0,0,0,0,0,0,0,0,0,1,0,0]
Outer Race 21-6	11	[0,0,0,0,0,0,0,0,0,0,0,1,0]
Normal	12	[0,0,0,0,0,0,0,0,0,0,0,0,1]

The dataset comprises 12 different fault types under inner race faults, outer race faults, and rolling element fault conditions. In the time domain, the dataset was divided into training, validation, and testing sets in the ratio of 6:2:2. One-hot encoding [40] is applied to the fault types, and the specific labels are presented in Table 1. Using one-hot encoding ensures that the distance between any two

fault types is equidistant, thereby eliminating the distance problems commonly associated with simple encoding.

3.2.2. Results and analysis

The loss curve for the AHA-RTN under the above training strategy is shown in Figure 6, which gradually decreases as the number of training epochs increase. The loss value of the final training set remains stables at 0.088, and the loss value of the validation set remains stable at 0.099. The loss curves of the training and validation sets are relatively close.

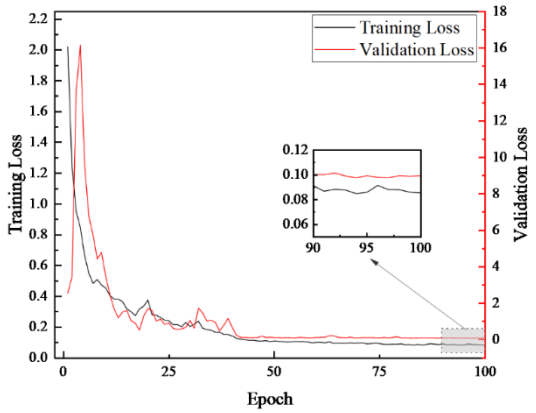


Figure 6. Training and validation loss functions.

a) The CWRU drive-end bearing fault dataset

The performance of the ResNet, DRSN-CW, DRSN-SW, and AHA-RTN under different SNRs of Gaussian noise and Laplacian noise on the CWRU drive-end bearing fault dataset is shown in Figure 7, which also includes the condition with no added noise.

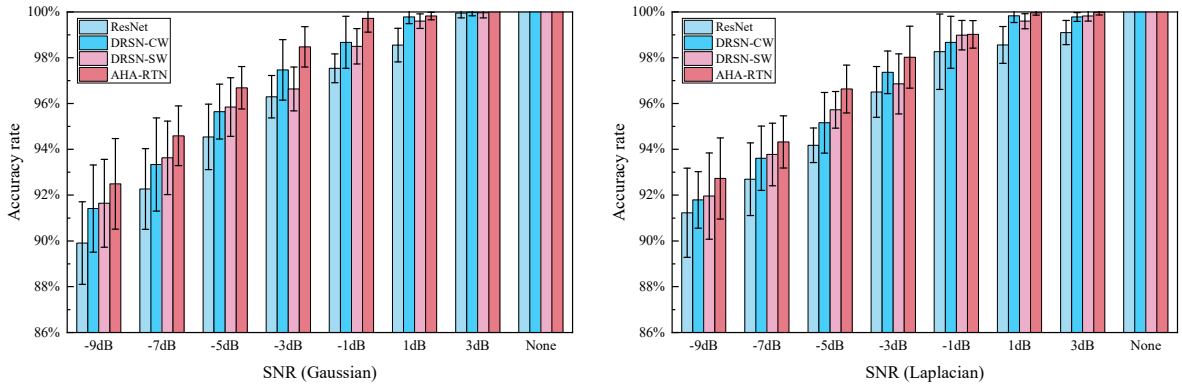


Figure 7. Model performance on the CWRU drive-end bearing dataset.

The results indicate that simply increasing the network depth can degrade performance when dealing with high-noise data, regardless of the type of noise added. This is because, as the depth of the network increases, the features extracted from useful and noisy data become intertwined, making it challenging for the ResNet to distinguish which features are useful for fault diagnosis. The DRSN-CW

can globally capture noise features in each feature space by introducing a channel attention module and setting them to zero through soft thresholding, enhancing the network's robustness against interference. Additionally, the results of the DRSN-SW indicate that local information also contributes positively to improving the network's robustness against interference.

As shown in Table 2 and Table **3**, when SNR > −5 dB, the average accuracy of the AHA-RTN, DRSR-SW, DRSN-CW, and ResNet under Gaussian noise is 99.603%, 98.941%, 99.181%, and 98.468%, respectively. Under Laplacian noise, it is 99.400%, 99.053%, 99.131%, and 98.486%, respectively. The DRSN-CW outperforms the DRSN-SW under both Gaussian and Laplacian noise conditions, and the proposed AHA-RTN outperforms both. This indicates that the channel-wise attention mechanism is effective under these conditions, where the proposed AHA-RTN increases the adjustment coefficient of channel attention and, conversely, increases the adjustment coefficient of spatial attention. When the SNR $\leqslant$ −5 dB, the average accuracy of the AHA-RTN, DRSR-SW, DRSN-CW, and ResNet under Gaussian noise is 94.590%, 93.706%, 93.465%, and 92.237%, respectively. Under Laplacian noise, it is 94.562%, 93.819%, 93.523%, and 92.700%, respectively. The DRSN-SW outperforms the DRSN-CW, indicating that the spatial attention mechanism is also effective under these conditions.

Table 2. Accuracy of each model on the CWRU testing dataset under Gaussian noise.

SNR	ResNet18	DRSN-CW	DRSN-SW	AHA-RTN
None	100(±0)	100(±0)	100(±0)	100(±0)
3 dB	99.951(±0.211)	99.982(±0.152)	99.968(±0.231)	100(±0)
1 dB	98.554(±0.732)	99.783(±0.298)	99.6(±0.315)	99.821(±0.166)
−1 dB	97.538(±0.629)	98.675(±1.135)	98.5(±0.776)	99.717(±0.601)
−3 dB	96.296(±0.927)	97.467(±1.321)	96.638(±0.956)	98.475(±0.879)
−5 dB	94.542(±1.431)	95.646(±1.198)	95.846(±1.276)	96.688(±0.926)
−7 dB	92.267(±1.763)	93.338(±2.032)	93.629(±1.606)	94.592(±1.304)
−9 dB	89.903(±1.802)	91.412(±1.907)	91.644(±1.92)	92.491(±1.979)

Table 3. Accuracy of each model on the CWRU testing dataset under Laplacian noise.

SNR	ResNet18	DRSN-CW	DRSN-SW	AHA-RTN
None	100(±0)	100(±0)	100(±0)	100(±0)
3 dB	99.101(±0.523)	99.782(±0.191)	99.823(±0.221)	99.99(±0.121)
1 dB	98.558(±0.802)	99.833(±0.291)	99.596(±0.328)	99.967(±0.105)
−1 dB	98.263(±1.644)	98.675(±1.135)	98.988(±0.643)	99.017(±0.601)
−3 dB	96.508(±1.11)	97.363(±0.932)	96.858(±1.309)	98.025(±1.354)
−5 dB	94.175(±0.755)	95.163(±1.324)	95.725(±0.802)	96.633(±1.048)
−7 dB	92.696(±1.584)	93.613(±1.403)	93.775(±1.368)	94.325(±1.14)
−9 dB	91.229(±1.943)	91.792(±1.232)	91.958(±1.884)	92.729(±1.771)

As the noise intensity increases, the distribution of the feature mapping across the channels becomes more irregular. The DRSN-CW highlights global information and can set effective denoising thresholds for noise-dominated feature spaces, but it faces challenges in spaces where the noise is weaker than the useful data. In contrast, the DRSN-SW focuses on local information, setting precise

thresholds for each spatial point and enabling effective soft thresholding across noise levels, particularly under high-noise conditions.

Using soft thresholding, both channel and spatial attention mechanisms can enhance the network's ability to resist interference. Using only the channel attention mechanism cannot flexibly set thresholds for some feature spaces, whereas using only the spatial attention mechanism cannot flexibly set thresholds for places outside the local attention area. However, the proposed AHA-RTN can adaptively select attention mechanisms based on tasks to minimize the negative effects of using attention mechanisms alone. The results indicate that the AHA-RTN outperforms the other three networks under both high and low noise conditions. The confusion matrix of the AHA-RTN at SNR = −5 dB is shown in Figure 8.

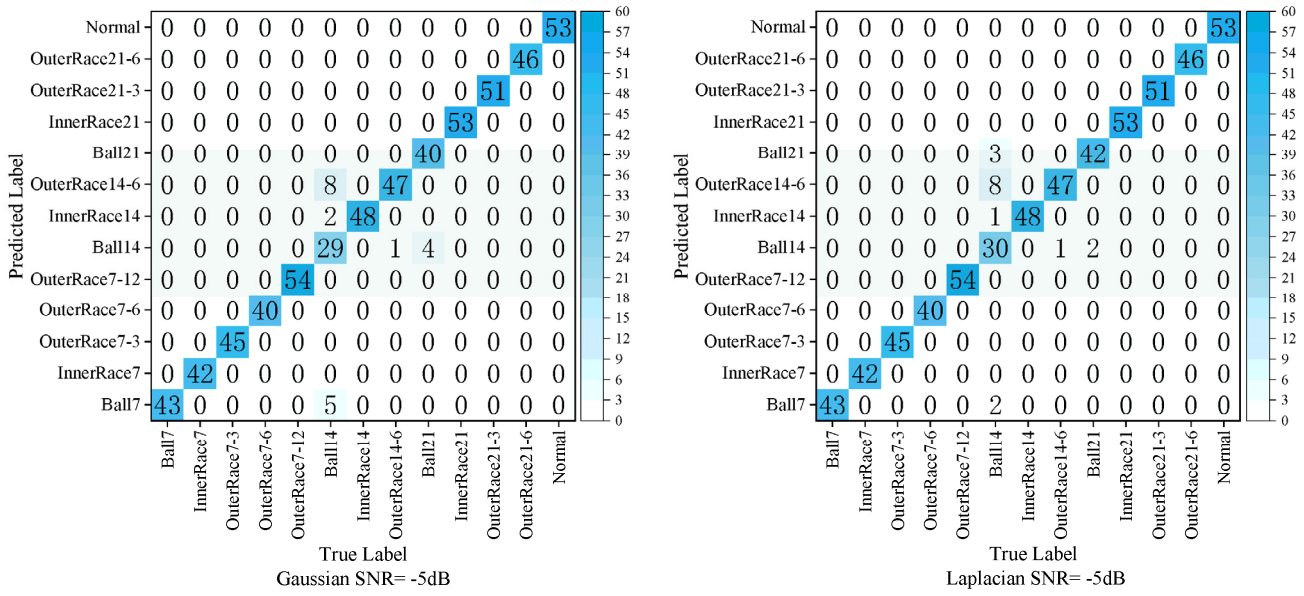
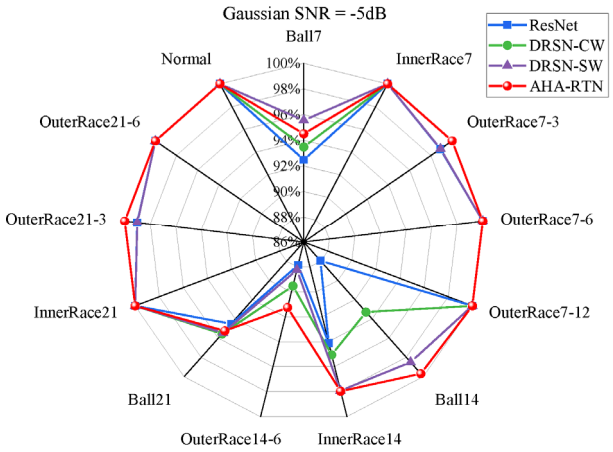


Figure 8. Confusion matrix results with SNR = −5 dB.

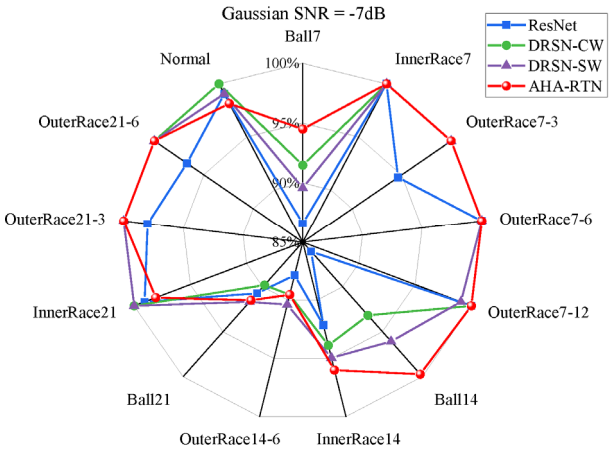
The F1 score was used to evaluate the performance of each network under Gaussian and Laplacian noise with SNR = −5 dB, −7 dB, and −9 dB. The formula for calculating the F1 score is as follows:

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{11}$$

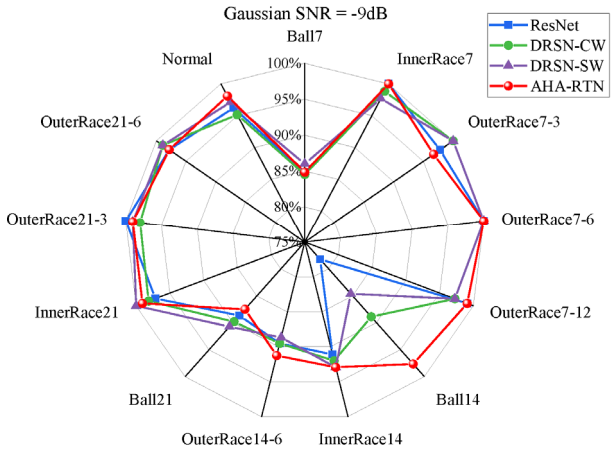
As shown in Figure 9 and Figure **10**, the coverage area of the proposed AHA-RTN essentially encompasses the radar coverage area of the other networks. Under Gaussian noise with SNR = −5 dB, the average F1 score of the AHA-RTN increases by 2.869%, 1.471%, and 0.497% compared to that of the ResNet, DRSN-CW, and DRSN-SW, respectively. Similarly, under Laplacian noise with SNR = −5 dB, the average F1 score of the AHA-RTN increased by 2.825%, 1.649%, and 1.018% compared to that of the ResNet, DRSN-CW, and DRSN-SW, respectively. The average F1 scores under Gaussian and Laplacian noise are shown in Table 4 and 5.



(a) Gaussian noise with SNR = −5 dB



(b) Gaussian noise with SNR = −7 dB



(c) Gaussian noise with SNR = −9 dB

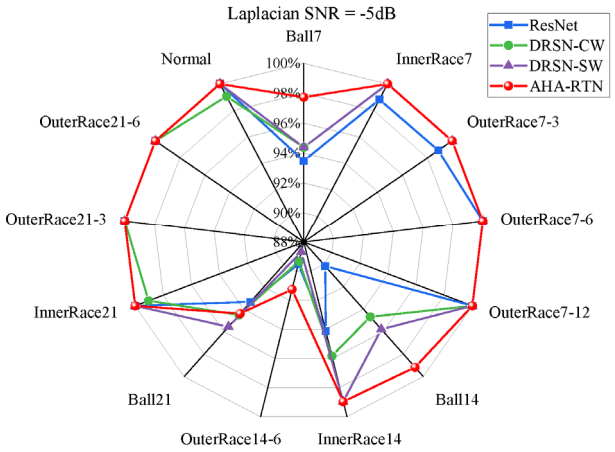
Figure 9. Scores for each model under Gaussian noise.

Table 4. Average F1 scores under Gaussian noise.

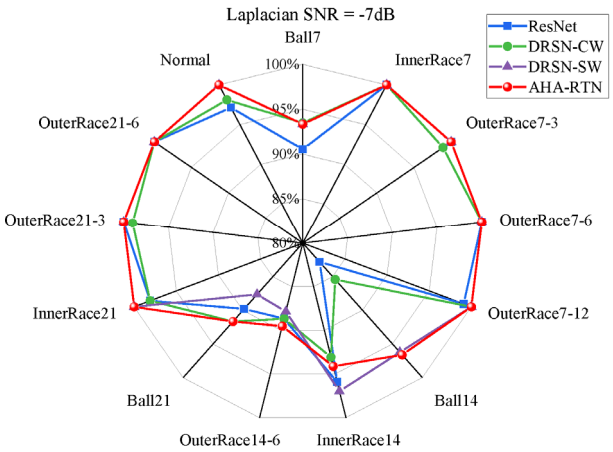
Gaussian	ResNet18	DRSN-CW	DRSN-SW	AHA-RTN
−5 dB	93.541	94.939	95.913	96.41
−7 dB	90.171	92.981	93.505	94.394
−9 dB	89.931	90.421	90.698	91.146

Table 5. Average F1 scores under Laplacian noise.

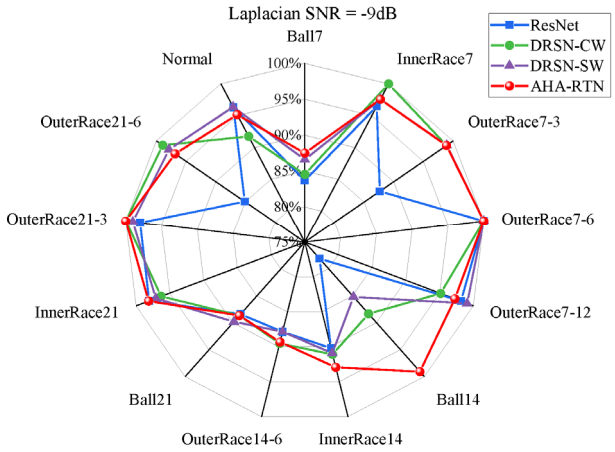
Gaussian	ResNet18	DRSN-CW	DRSN-SW	AHA-RTN
−5 dB	94.118	95.294	95.925	96.943
−7 dB	92.413	92.7	93.69	93.931
−9 dB	87.151	89.856	90.093	91.254



(a) Laplacian noise with SNR = −5 dB



(b) Laplacian noise with SNR = −7 dB



(c) Laplacian noise with SNR = −9 dB

Figure 10. Scores for each model under Laplacian noise.

As demonstrated in Figure 9, the accuracy of categories such as Ball21 experiences a slight decrease under noisy conditions. To better observe the feature distributions extracted by each network, we employed the t-distributed stochastic neighbor embedding (t-SNE) algorithm [41], a nonlinear unsupervised dimensionality reduction method. The results under Gaussian noise at SNR = −5 dB are shown in Figure 11. In the context of high-noise environments, the feature embeddings of fault categories such as Ball21 exhibit significant overlap in the high-dimensional space. Therefore, the performance degradation of the model is caused by the ambiguity of the features themselves. Overall, we found that the AHA-RTN has a high recognition accuracy for different fault types.

b) The CWRU fan-end bearing fault dataset

The performance of the ResNet18, DRSN-CW, DRSN-SW, and AHA-RTN under different SNRs of Gaussian and Laplacian noise on the CWRU fan-end bearing fault dataset is shown in Figure 12. The average accuracy of the AHA-RTN, DRSR-SW, DRSN-CW, and ResNet18 under Gaussian noise is 89.823%, 89.253%, 89.106%, and 87.127%, respectively, while the average accuracy under Laplacian noise is 90.418%, 89.445%, 89.539%, and 87.659%, respectively. Therefore, the results also show that the proposed AHA-RTN outperforms the other networks.

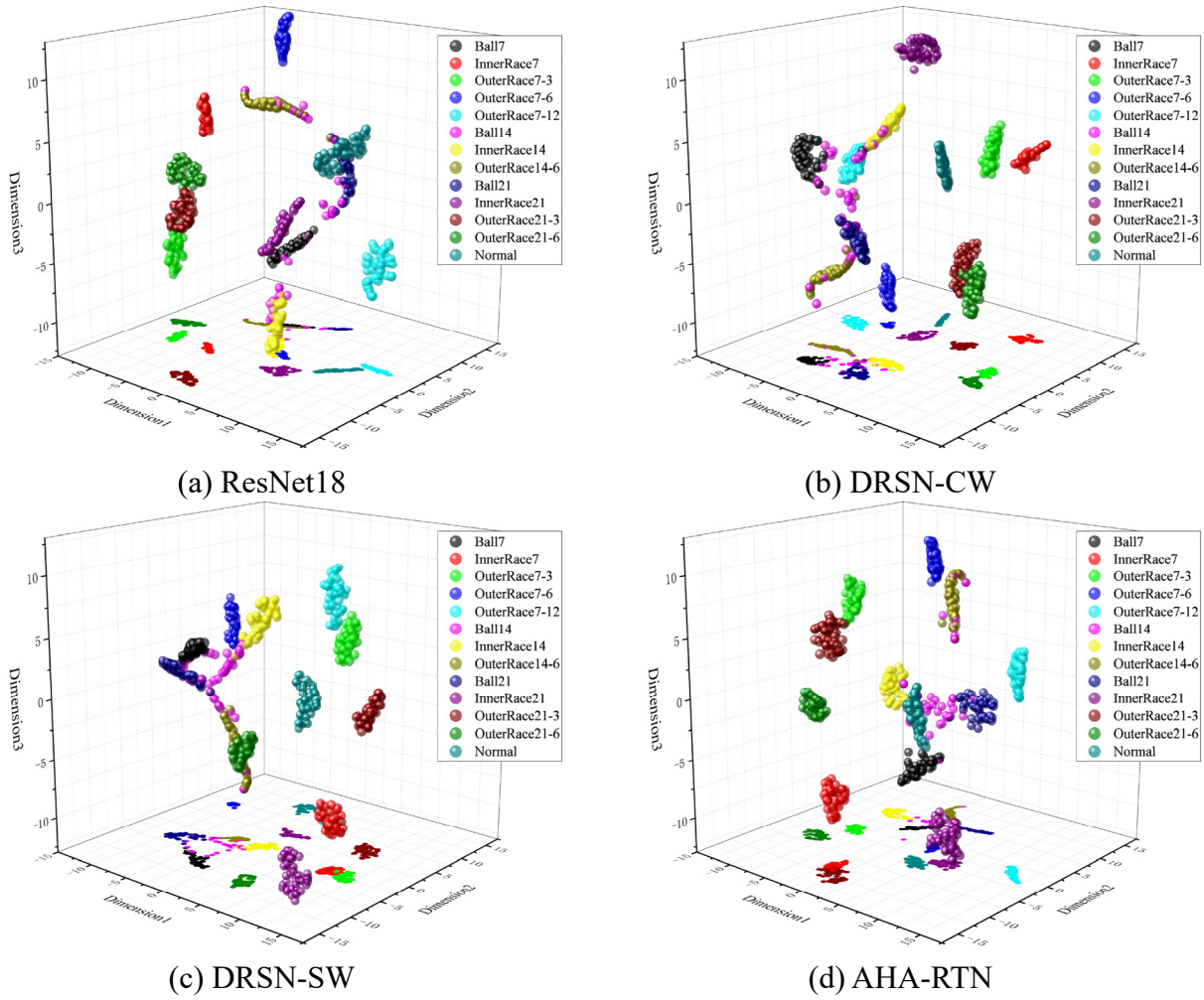


Figure 11. 3D visualization of high-dimensional features.

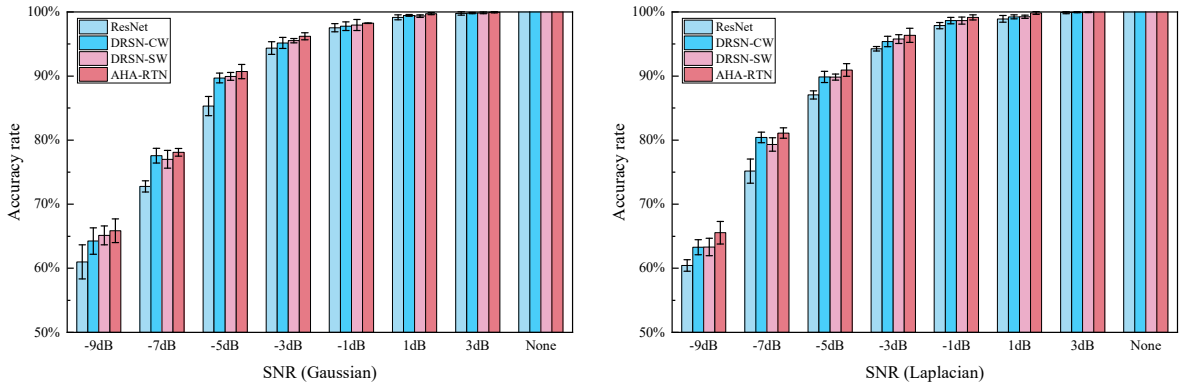


Figure 12. Model performance on the CWRU fan-end bearing dataset.

3.3. Experimental study on the Southeast University dataset

3.3.1. Data description

The experimental device of the Southeast University gear dataset is shown in Figure 13. This study tested the proposed AHA-RTN using the bearing data and operating conditions at 1800rpm with a load of 2 V. The bearing data contain the outer ring, inner ring, ball, and combination faults as well as one health state. The coding method for fault types also adopts the one-hot coding method.

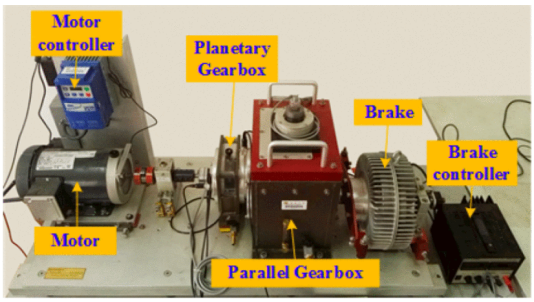


Figure 13. Experimental setup for the Southeast University dataset [42].

3.3.2. Results and analysis

The performance of the ResNet, DRSN-CW, DRSN-SW, and AHA-RTN under different SNRs of Gaussian and Laplacian noise is shown in Figure 14. The accuracy of all networks decreases because of the combination fault. However, the proposed AHA-RTN still outperforms the other networks. The average accuracy of the AHA-RTN, DRSR-SW, DRSN-CW, and ResNet under Gaussian noise is 88.706%, 87.514%, 87.641%, and 83.978%, respectively, and under Laplacian noise is 89.558%, 88.416%, 88.501%, and 85.526%, respectively.

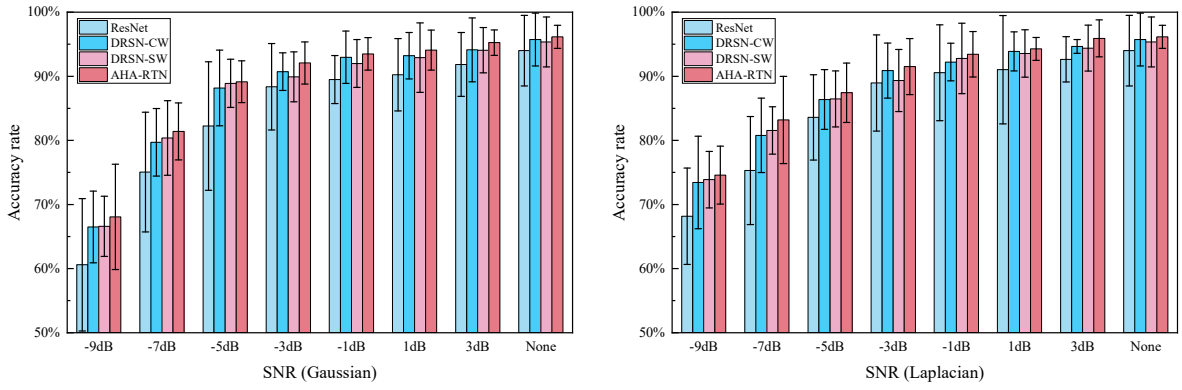


Figure 14. Model performance differences on the Southeast University dataset.

3.4. Ablation experiment

The ablation experiment includes four models: a baseline model, two comparative models, and the proposed model.

Baseline (ResNet): As the fundamental architecture for the subsequent comparative and proposed models, its original structure for image processing is not suitable for the current task. A rapid increase in the number of channels will lead to the neglect of a large number of abstract features at different levels and also increase the risk of overfitting in the network. Therefore, adjustments are made to structures such as the convolution kernel size and maximum pooling layer in the original architecture of the ResNet to adapt to the current fault diagnosis task.

Model 1 (DRSN-CW): It introduces the channel attention mechanism into the ResNet, designs an adaptive threshold calculation module, suppresses noise by means of soft thresholding, and improves the anti-interference ability of the model through denoising activation.

Model 2 (DRSN-SW): It introduces the spatial attention mechanism into the ResNet and also has an adaptive threshold calculation module.

Proposed (AHA-RTN): It incorporates both the channel attention mechanism and spatial attention mechanism into the ResNet, designs an adaptive hybrid module, and is also equipped with an adaptive threshold calculation module.

The comparative results are shown in Table 2 and Table **3**. The proposed method has demonstrated excellent accuracy in all experiments, which indicates that each module contributes to enhancing the feature recognition capability of the AHA-RTN under different operating conditions. The DRSN-SW allows for testing the rationality of the adaptive attention selection module design in the AHA-RTN. Since the AHA-RTN can adaptively select between channel attention and spatial attention, its performance is comparable to that of the DRSN-CW when spatial attention negatively impacts the denoising algorithm. This serves as a validation of the reasonableness of the adaptive design in the AHA-RTN.

To observe the effect of different initial fusion methods of channel-wise and spatial attention mechanisms on the performance of the AHA-RTN, we validate four fusion methods: addition, multiplication, minimum, and maximum. All other configuration and data processing parameters remain the same except the initial fusion method of the network, which was varied. The results of SNR = –5 dB are shown in Table 6.

Table 6. Influence of fusion methods on network performance.

Fusion Method	Gaussian	Laplacian
addition	96.574(±0.721)	96.264(±1.086)
multiplication	96.688(±0.926)	96.633(±1.048)
minimum	95.949(±0.859)	95.974(±1.055)
maximum	96.338(±0.899)	96.070(±0.759)

The initial fusion process uses the multiplication method, which can enhance feature characterization. This approach enables features to display high responsiveness in spatial and channel-wise domains, helping models capture local and global data characteristics better. This process reduces the impact of unrelated features on generating subsequent adjustment coefficients.

The threshold is calculated by multiplying $M_{last\_attention}$ and β. The computation of β significantly impacts subsequent threshold settings. To determine the impact of the calculation methods

for β, such as not preprocessing the initial data, using global averaging, maximum selection, and channel-wise spatial alignment averaging, on the final results, all other configuration parameters and data processing remain unchanged. The results of SNR = –5 dB are presented in Table 7.

Table 7. Influence of β calculation methods on network performance.

β Calculation Method	Gaussian	Laplacian
no preprocessing	94.579(±2.908)	94.997(±2.856)
global averaging	95.702(±1.346)	95.109(±1.221)
maximum selection	94.934(±2.727)	94.784(±2.362)
channel-spatial alignment averaging	96.688(±0.926)	96.633(±1.048)

Unreasonable thresholds will set valid features to zero. Methods such as no preprocessing, global averaging, and maximum selection all lead to β containing redundant information from other channel-wise or spatial dimensions, making it difficult to extract specific information needed for denoising. However, channel-spatial alignment averaging effectively captures noise information across specific noise reduction ranges across channels and spatial dimensions, thus eliminating interference with thresholds set for other channels and spatial dimensions.

4. Conclusions

This paper proposes a novel adaptive hybrid attention residual thresholding network that integrates concurrent channel-wise and spatial attention processes. Additionally, the adaptive attention module integrates these two processes quantitatively in response to the needs of noise reduction, thereby evaluating the extent and depth of noise signal features. We compared the independent contributions and synergistic effects of the two attention mechanisms, and analyzed the impact of their different fusion methods on the performance of the AHA-RTN. It was found that using the multiplication method in the initial fusion process and channel-spatial alignment averaging in threshold acquisition can enhance network performance. Finally, evaluating two bearing cases demonstrates that the proposed AHA-RTN performs well in fault diagnosis under different noise environments.

Although the AHA-RTN has good feature extraction capabilities and high fault diagnosis accuracy, it increases computational complexity and parameter scale. In the future, we will further explore lightweight optimization schemes to achieve a better balance between accuracy and real-time performance. We will also explore fault diagnosis methods tailored for imbalanced data to improve the robustness of the model in scenarios with small sample sizes and extreme faults. Additionally, future research will focus on model interpretability and deployment efficiency to enhance trust in model decisions in real-world industrial scenarios.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No.

52302511, and the Fundamental Research Funds for the Central Universities under Grant No. 501JCGG2024129005.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. W. Huang, Z. Li, X. Ding, D. He, Q. Wu, J. Liu, Digital-analog driven multi-scale transfer for smart bearing fault diagnosis, *Eng. Appl. Artif. Intell.*, **137** (2024), 109186. https://doi.org/10.1016/j.engappai.2024.109186
2. J. Tong, S. Tang, Y. Wu, H. Pan, J. Zheng, A fault diagnosis method of rolling bearing based on improved deep residual shrinkage networks, *Measurement*, **206** (2023), 112282. https://doi.org/10.1016/j.measurement.2022.112282
3. B. Pang, M. Nazari, G. Tang, Recursive variational mode extraction and its application in rolling bearing fault diagnosis, *Mech. Syst. Signal Process.*, **165** (2022), 108321. https://doi.org/10.1016/j.ymssp.2021.108321
4. F. Li, L. Wang, D. Wang, J. Wu, H. Zhao, An adaptive multiscale fully convolutional network for bearing fault diagnosis under noisy environments, *Measurement*, **216** (2023), 112993. https://doi.org/10.1016/j.measurement.2023.112993
5. Y. Li, M. Xu, W. Huang, M. J. Zuo, L. Liu, An improved EMD method for fault diagnosis of rolling bearing, in *2016 Prognostics and System Health Management Conference (PHM-Chengdu)*, (2016), 1–5. https://doi.org/10.1109/PHM.2016.7819842
6. J. Zheng, M. Su, W. Ying, J. Tong, Z. Pan, Improved uniform phase empirical mode decomposition and its application in machinery fault diagnosis, *Measurement*, **179** (2021), 109425. https://doi.org/10.1016/j.measurement.2021.109425
7. K. Zhang, C. Ma, Y. Xu, P. Chen, J. Du, Feature extraction method based on adaptive and concise empirical wavelet transform and its applications in bearing fault diagnosis, *Measurement*, **172** (2021), 108976. https://doi.org/10.1016/j.measurement.2021.108976
8. H. Hu, Y. Lv, R. Yuan, S. Xu, W. Zhu, A novel vibro-acoustic fault diagnosis approach of planetary gearbox using intrinsic wavelet integrated GE-EfficientNet, *Meas. Sci. Technol.*, **35** (2024), 025131. https://doi.org/10.1088/1361-6501/ad0afe
9. K. Dragomiretskiy, D. Zosso, Variational mode decomposition, *IEEE Trans. Signal Process.*, **62** (2014), 531–544. https://doi.org/10.1109/TSP.2013.2288675
10. S. Chauhan, G. Vashishtha, R. Kumar, R. Zimroz, M. K. Gupta, P. Kundu, An adaptive feature mode decomposition based on a novel health indicator for bearing fault diagnosis, *Measurement*, **226** (2024), 114191. https://doi.org/10.1016/j.measurement.2024.114191
11. B. Zheng, M. Zhu, X. Guo, J. Ou, J. Yuan, Path planning of stratospheric airship in dynamic wind field based on deep reinforcement learning, *Aerosp. Sci. Technol.*, **150** (2024), 109173. https://doi.org/10.1016/j.ast.2024.109173
12. H. Hu, B. Tang, X. Gong, W. Wei, H. Wang, Intelligent fault diagnosis of the high-speed train with big data based on deep neural networks, *IEEE Trans. Ind. Inf.*, **13** (2017), 2106–2116. https://doi.org/10.1109/TII.2017.2683528

13. C. Liu, X. Li, X. Chen, S. Khan, Neuromorphic computing-enabled generalized machine fault diagnosis with dynamic vision, *Adv. Eng. Inf.*, **65** (2025), 103300. https://doi.org/10.1016/j.aei.2025.103300
14. W. Zhang, M. Xu, H. Yang, X. Wang, S. Zheng, X. Li, Data-driven deep learning approach for thrust prediction of solid rocket motors, *Measurement*, **225** (2024), 114051. https://doi.org/10.1016/j.measurement.2023.114051
15. L. Yang, Z. Yang, S. Song, F. Li, C.P. Chen, Twin broad learning system for fault diagnosis of rotating machinery, *IEEE Trans. Instrum. Meas.*, **72** (2023), 1–12. https://doi.org/10.1109/TIM.2023.3259022
16. Z. Hei, W. Sun, H. Yang, M. Zhong, Y. Li, A. Kumar, et al., Novel domain-adaptive Wasserstein generative adversarial networks for early bearing fault diagnosis under various conditions, *Reliab. Eng. Syst. Saf.*, **257** (2025), 110847. https://doi.org/10.1016/j.ress.2025.110847
17. X. Li, X. Wu, T. Wang, Y. Xie, F. Chu, Fault diagnosis method for imbalanced data based on adaptive diffusion models and generative adversarial networks, *Eng. Appl. Artif. Intell.*, **147** (2025), 110410. https://doi.org/10.1016/j.engappai.2025.110410
18. W. Zhang, N. Jiang, S. Yang, X. Li, Federated transfer learning for remaining useful life prediction in prognostics with data privacy, *Meas. Sci. Technol.*, **36** (2025), 076107. https://doi.org/10.1088/1361-6501/ade552
19. X. Li, S. Xiao, Q. Li, L. Zhu, T. Wang, F. Chu, The bearing multi-sensor fault diagnosis method based on a multi-branch parallel perception network and feature fusion strategy, *Reliab. Eng. Syst. Saf.*, **261** (2025), 111122. https://doi.org/10.1016/j.ress.2025.111122
20. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
21. G. Huang, Z. Liu, L.V.D. Maaten, K.Q. Weinberger, Densely connected convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 2261–2269. https://doi.org/10.1109/CVPR.2017.243
22. W. Zhang, X. Li, Q. Ding, Deep residual learning-based fault diagnosis method for rotating machinery, *ISA Trans.*, **95** (2019), 295–305. https://doi.org/10.1016/j.isatra.2018.12.025
23. M. Zhao, B. Tang, L. Deng, M. Pecht, Multiple wavelet regularized deep residual networks for fault diagnosis, *Measurement*, **152** (2020), 107331. https://doi.org/10.1016/j.measurement.2019.107331
24. X. Zheng, P. Yang, K. Yan, Y. He, Q. Yu, M. Li, Rolling bearing fault diagnosis based on multiple wavelet coefficient dimensionality reduction and improved residual network, *Eng. Appl. Artif. Intell.*, **133** (2024), 108087. https://doi.org/10.1016/j.engappai.2024.108087
25. W. Zhang, G. Peng, C. Li, Y. Chen, Z. Zhang, A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals, *Sensors*, **17** (2017). https://doi.org/10.3390/s17020425
26. Q. Wang, F. Xu, A novel rolling bearing fault diagnosis method based on adaptive denoising convolutional neural network under noise background, *Measurement*, **218** (2023) 113209. https://doi.org/10.1016/j.measurement.2023.113209
27. Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, *Neurocomputing*, **452** (2021) 48–62. https://doi.org/10.1016/j.neucom.2021.03.091
28. K. Zhang, B. Tang, L. Deng, X. Liu, A hybrid attention improved ResNet based fault diagnosis method of wind turbines gearbox, *Measurement*, **179** (2021), 109491. https://doi.org/10.1016/j.measurement.2021.109491

29. Q. Snyder, Q. Jiang, E. Tripp, Integrating self-attention mechanisms in deep learning: A novel dual-head ensemble transformer with its application to bearing fault diagnosis, *Signal Process.*, **227** (2025), 109683. https://doi.org/10.1016/j.sigpro.2024.109683
30. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *2018 IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 7132–7141. https://doi.org/10.48550/arXiv.1709.01507
31. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 11531–11539. https://doi.org/10.48550/arXiv.1910.03151
32. Z. Gao, J. Xie, Q. Wang, P. Li, Global second-order pooling convolutional networks, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 3019–3028. https://doi.org/10.1109/CVPR.2019.00314
33. Z. Qin, P. Zhang, F. Wu, X. Li, FcaNet: Frequency channel attention networks, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 763–772. https://doi.org/10.1109/ICCV48922.2021.00082
34. M. Zhao, S. Zhong, X. Fu, B. Tang, M. Pecht, Deep residual shrinkage networks for fault diagnosis, *IEEE Trans. Ind. Inf.*, **16** (2020), 4681–4690. https://doi.org/10.1109/TII.2019.2943898
35. M. Zhao, S. Zhong, X. Fu, B. Tang, S. Dong, M. Pecht, Deep residual networks with adaptively parametric rectifier linear units for fault diagnosis, *IEEE Trans. Ind. Electron.*, **68** (2021), 2587–2597. https://doi.org/10.1109/TIE.2020.2972458
36. M. Nagubandi, R. Walia, A. Karanath, G.N. Pillai, Electric load forecasting using dual-stage attention network with cosine annealed warm restart schedule, in *2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, (2022), 141–146. https://doi.org/10.1109/ICETCI55171.2022.9921361
37. P. T. De Boer, D. P. Kroese, S. Mannor, R.Y. Rubinstein, A tutorial on the cross-entropy method, *Ann. Oper. Res.*, **134** (2005), 19–67. https://doi.org/10.1007/s10479-005-5724-z
38. W. A. Smith, R. B. Randall, Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study, *Mech. Syst. Signal Process.*, **64** (2015), 100–131. https://doi.org/10.1016/j.ymssp.2015.04.021
39. Y. Wang, P. W. Tse, B. Tang, Y. Qin, L. Deng, T. Huang, Kurtogram manifold learning and its application to rolling bearing weak signal detection, *Measurement*, **127** (2018), 533–545. https://doi.org/10.1016/j.measurement.2018.06.026
40. P. Rodríguez, M. A. Bautista, J. Gonzàlez, S. Escalera, Beyond one-hot encoding: Lower dimensional target embedding, *Image Vision Comput.*, **75** (2018), 21–31. https://doi.org/10.1016/j.imavis.2018.04.004
41. L. Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.*, **9** (2008), 2579–2605.
42. S. Shao, S. McAleer, R. Yan, P. Baldi, Highly accurate machine fault diagnosis using deep transfer learning, *IEEE Trans. Ind. Inf.*, **15** (2018), 2446–2455. https://doi.org/10.1109/TII.2018.2864759

AIMS Press