Multi-scale deep residual shrinkage networks with a hybrid attention mechanism for rolling bearing fault diagnosis

Xinliang Zhang (zxldq@hpu.edu.cn)  
Henan Polytechnic University

Shengqiang Wei  
Henan Polytechnic University

Jianhang Huang  
Henan Polytechnic University

Lijie Jia  
Henan Polytechnic University

Research Article

Keywords: fault diagnosis, deep residual shrinkage network, hybrid attention mechanism, dilated convolution, multi-scale feature fusion

Posted Date: April 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2802750/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Multi-scale deep residual shrinkage networks with a hybrid attention mechanism for rolling bearing fault diagnosis

Xinliang Zhang\textsuperscript{1*}, Shengqiang Wei\textsuperscript{1}, Jianhang Huang\textsuperscript{1}, Lijie Jia\textsuperscript{1}

1. School of Electrical Engineering and Automation, Henan Key Laboratory of Intelligent Detection and Control of Coal Mine Equipment, Henan Polytechnic University, Jiaozuo 454003, China

E-mail: zxldq@hpu.edu.cn

Abstract: The fault diagnosis of rolling bearings based on deep networks is hindered by the unexpected noise involved with accessible vibration signals and global information abatement in deepened networks. To combat the degradation, a multi-scale deep residual shrinkage network with a hybrid-attention-mechanism (MH-DRSN) is proposed in this paper. First, a spatial domain attention mechanism is introduced into the residual shrinkage module to represent the distance dependence of the feature maps. Then, a hybrid attention mechanism considering both the inner-channeled and cross-channeled characteristics is constructed. Through the comprehensive evaluation of the feature map, it provides a soft threshold for the activation function and realizes the feature-map selection adaptively. Second, the dilated convolution with different dilation rates is implemented for multi-scale context information extraction. Through the feature combination of the DRSN and the dilated convolution, the global information of the rolling bearing fault is strengthened and preserved as the fault diagnosis network is deepened. Finally, the performance of the proposed fault-diagnosis model is validated on datasets from Case Western Reserve University (CWRU), Xi’an Jiaotong University and Zhejiang Changxing Sumyoung Technology Co. Ltd (XJTU-SY). Also, the influence of the number of residual shrinkage layers, model optimizers, and different learning rates on the accuracy of the diagnostic model has been discussed. The experimental results show that, compared with common convolution neural networks, the proposed neural diagnosis model provides a higher identification accuracy and better robustness under noise interference.

Keywords: fault diagnosis, deep residual shrinkage network, hybrid attention mechanism, dilated convolution, multi-scale feature fusion

1 Introduction

The effective state monitoring and fault diagnosis of rolling bearings is an important guarantee for the safe and stable operation of mechanical equipment [1]. The emergence of the ‘big-data era’ has provided a new opportunity for the data-driven fault diagnosis of rolling bearings. Based on feature extraction and consequent pattern recognition, machine-learning approaches have shown potential for bearing fault diagnosis due to their self-learning optimization. Examples of such approaches are the support vector machine (SVM)[2], Bayesian networks[3] and deep-learning networks[4]. In particular, deep-learning networks, which employ multi-layered deep convolutional networks to extract semantic features, have been proven experimentally to be an effective method for fault diagnosis[5-8]. However, because of intrinsic structural characteristics, global information suffers an abatement with deepened networks. Consequently, the fault diagnosis performance of deep-learning networks is negatively affected by the insufficient feature extraction of the bearing fault. To supplement the global feature information, Xu et al. developed a hybrid deep-learning model based on the convolutional neural-network (CNN) model and deep forest (gcForest) by using the time-frequency image from vibration signals[9]. Xue et al. proposed a parallel multi-channel structure with both 1D-CNN and 2D-CNN[10]. Then, the in-depth features of the bearings were extracted, and a particle-swarm, optimized SVM was derived with a high accuracy. Zhao et al. proposed a transfer-learning-network diagnosis model for rolling bearings with the introduction of dilated convolution to
capture the multi-scale features of the fault[11].

Internal and external interference inevitably exists in practical-application scenarios, which distorts the features of the original vibration signal and significantly deteriorates the performance of intelligent diagnosis models. Zhang et al. adopted a wide kernel in the first-layer convolution (WDCNN) and the AdaBN algorithm in the network[12]. Through interference pre-training, they yielded a bearing fault diagnosis model with anti-noise ability. Ravikumar et al. introduced a stacked, long short-term memory (LSTM) into a multi-scale deep residual learning network[13]. These networks functioned essentially as a nonlinear moving noise filter and consequently attenuated the signal indiscriminately during the pre-training process. Based on the multi-scale structure, the attention mechanism was gradually applied to the intelligent fault diagnosis networks to combat the environmental noise[13]. Jin et al. proposed a gated recursive neural network with an attention mechanism for bearing fault diagnosis[14]. The noise suppression of the network was improved by combining a data random sampling strategy, gate recurrent unit, and attention mechanism. In consideration of the parallel structure with the residual network, Liang et al. employed a squeeze-and-excitation (SE) attention module to suppress different environmental noise[15], where the dilated convolutions were added into the residual network to construct a multi-level connection[16]. Zhao et al. also embedded an attention block into the soft thresholding function and proposed a deep residual shrinkage network (DRSN) model to solve the problem of fault diagnosis under strong noise interference[17]. The attention mechanism proved to be an alternative applicable approach for noise suppression by preserving the prominent features through the evaluation of the feature map. However, these attention-mechanism-based diagnosis models considered only the inner-channeled relations, i.e., point attention, taking no account of cross-channel relations, i.e., spatial attention[18-19].

The fault diagnosis of rolling bearings based on the deep networks is hindered by the unexpected noise involved with accessible vibration signals and the global information abatement with deepened networks. It is rather necessary and urgent to improve the diagnostic performance from the point of the introduction of the noise suppression block and multi-scale feature extractor into the traditional models. In this paper, a hybrid domain attention module based on the squeeze-and-excitation algorithm is proposed that integrates both the point attention mechanism and spatial attention mechanism. Then, the ordinary residual shrinkage block of the DRSN is replaced by the hybrid attention module to improve fault-diagnosis efficiency. At the same time, the multi-scale context information is extracted by the multi-rated dilated convolutions. Through the combination of the features from both the improved deep residual shrinkage network and dilated convolutions, more complete feature information of the rolling-bearing fault is extracted with the deepening of the network.

The remainder of this paper is organized as follows. Section 2 summarizes the operation of the residual shrinkage network and its soft thresholding block. Section 3 describes the improved residual shrinkage module with a hybrid attention mechanism used for noise suppression. In Section 4, the dilated convolutions with different receptive fields are given in detail, and the deep residual shrinkage network diagnosis model with multi-scale feature extraction and hybrid attention mechanism (MH-DRSN) is derived for rolling bearings. Finally, Section 5 gives the experimental validation results on the benchmark bearing dataset, and Section 6 presents the conclusions and discusses possible future work.

2 Related work and problem statement

DRSN is derived from the residual network (ResNet)[20] and is composed of a number of serially connected residual shrinkage modules for deep feature learning. As shown in Fig. 1, the residual shrinkage module is mainly composed of two, parallel functional blocks. A shortcut connection from the input to the output feature map is used to overcome the problem of gradient
disappearance and degradation during training, while the main feed-forward network consists of
serially connected convolution layers, batch normalization (BN), and activation layers for the
feature extraction.

BN accelerates the convergence speed in network training and prevents gradient
disappearance and explosion[21]. ReLU nonlinear activation functions are used to increase the
nonlinear description ability of the network. The error function for training the residual shrinkage
network is described as a residual function, i.e.,
\[ F(x) = Y(x) - x, \quad (1) \]
where \( x \) is the input, and \( Y(x) \) is the desired output of the neural network model. For the first
convolution layer of the first residual shrinkage network, \( x \) is identical to the gray image derived
from the original time-domain signal. Then, the traditional training algorithm through fitting \( Y(x) \)
is transformed equivalently to fitting the residual function \( F(x) \), i.e.,
\[ F(x) \to 0 \Leftrightarrow Y(x) \to x. \quad (2) \]

\( F(x) \) reaches 0, and \( Y(x) = x \), i.e., \( Y(x) = x \) means the best solution mapping for a set of stacked
network layers. When the network is deepened, it is difficult for the model to fit the actual
mapping \( Y(x) \) directly. The residual structure turns to fit the residual mapping \( F(x) \) by introducing
shortcut connections, and the actual mapping \( Y(x) \) is denoted as \( Y(x) = F(x) + x \). When \( F(x) = 0 \),
an identity mapping \( Y(x) = x \) is obtained. The approximation of the actual mapping is realized by
minimizing the residual function \( F(x) = Y(x) - x \) to solve the performance degradation problem
with network layer stacking. Thus, it gives a more convenient and efficient method compared to
traditional training approaches[20].

At the same time, a SE block is embedded in the main feedforward network of the residual
shrinkage module. Through the evaluation of the effectiveness of the current input feature map, it
realizes an adaptive threshold adjustment of the final activation function for feature compression.

A typical structure of the SE block is composed of a generalized function \( g(.) \), a down-
sampling layer, and a learning block. As shown in Fig. 1, through \( g(.) \) and a down-sampling, the

\[ \text{Conv} \]
\[ \text{BN} \]
\[ \text{ReLU} \]
\[ \downarrow \]
\[ F(x) \]
\[ \text{Conv} \]

Identity shortcut

Soft thresholding

\[ Y(x) = F(x) + x \]

Fig. 1. The structure of the residual shrinkage module.
current input feature map is transformed into a global feature vector $F_P$. Then, the nonlinear relationship between the channels is obtained through $F_P$ by the learning block, which is described as a series of the full connection layers and the following activation functions. Finally, the weight coefficients $A$ are obtained corresponding to the feature maps of every channel. Thus, the SE block gives an evaluation of the features obtained from the convolution layers and brings a feedback impact on representative feature selection. A realistic example of the SE block was constructed by two full connection layers and two activation functions [22]. In its SE block, an absolutization operation of the features was employed alternatively as the generalized function, and global average pooling was used to get the global feature vector. Also, the Relu and Sigmoid functions were adopted in the first and second activation layers, respectively. The Sigmoid activation function at the last layer normalized the feature weight coefficient between channels to a range of $(0, 1)$.

In practical application scenarios, both the internal and ambient interference may distort the fault characteristics in the original vibration signal. Consequently, the DRSN suffers prediction accuracy degradation due to the inefficiency of the extracted features for identifying the fault types. In the DRSN, the adaptive threshold adjustment is realized by the weight coefficients $A$ obtained merely through the channel attention mechanism. It doesn’t take into account the cross-channel spatial constraints among the input feature map and leads to incomplete information for the evaluation. For the bearing fault diagnosis, the original vibration signal is in essence composed of time domain data. Thus, after its transformation to the gray images as the input of the fault diagnosis model, the identifying characteristics and location information of the bearing fault are represented, respectively, by the pixel areas in the image and the features across the channels. Thus, the attention mechanism in the SE block is improved and constructed with the two-sided evaluation of channel and cross-channel dependence. Moreover, due to the inherent characteristic of the deep learning networks, the global information will undergo abatement with the deepened layers. To retain the global feature information, the features should be transmitted backwards to the following extraction layers. By virtue of the parallel structure with ResNet, the parallel multi-level connection realized by the dilated convolution is introduced to residual network layers. Then, the features of preceding layers are incorporated into subsequent layers to supplement the context information.

3 SE block with a hybrid attention mechanism

3.1 Point-spatial attention mechanism

To guarantee the representation of both point domain information and spatial domain information, a hybrid point-spatial attention mechanism is adopted in the squeeze-and-excitation block of the residual shrinkage module. The point attention module focuses on the extraction of the identifying features of specific targets and yields the importance degree on each channel component of the input feature map. Considering that the global average pooling (GAP) operation compresses the feature map of each channel to extract the global feature information, it mainly provides a quantitative reflection of the global central tendency, i.e., the derived coefficients are governed by the total level of each feature-map channel. Therefore, an additional global max pooling (GMP) is introduced into the point attention module to allow for the effect of the extreme data in describing the bearing fault.

The point attention module with mixed pooling operations is shown in Fig. 2. GAP and GMP operations are performed in parallel on the multi-channeled feature map $G$ obtained after the $g(.)$ operation to get the compressed feature vectors $F_p^{avg}$ and $F_p^{Max}$, i.e.,

$$F_p^{avg} = \text{GAP}(G) \text{ and } F_p^{Max} = \text{GMP}(G),$$

(3)
where $G \in \mathbb{R}^{W \times H \times C}$, $F_p^{\text{Avg}} \in \mathbb{R}^{W \times H \times C}$, $F_p^{\text{Max}} \in \mathbb{R}^{W \times H \times C}$, and $W$, $H$, $C$ are the width, height, and channel number of the feature map, respectively. Then the point attention feature weight is derived as,

$$A_p = \sigma \left( f(F_p^{\text{Avg}}) + f(F_p^{\text{Max}}) \right),$$

(4)

where $A_p \in \mathbb{R}^{W \times H \times C}$; $f(.)$ denotes the compound operation upon $F_p^{\text{Avg}}$ and $F_p^{\text{Max}}$, composed of two serial, full connection layers, BN and the ReLU activation function. The activation function is $\sigma(.)$, typically adopted as the Sigmoid function to map the variable value to the range $(0, 1)$.

Then, the point attention feature map is generated by imposing the weight vector $A_p$ on the input feature map $G$, i.e.,

$$G_p = G \otimes A_p,$$

(5)

where the function `$\otimes$' denotes the element-wise product operation and $G_p \in \mathbb{R}^{W \times H \times C}$. It should be noted that the number of channels in the two FC layers of the module is exactly the number of the original feature map. And so this avoids a dimensionality reduction, which is crucial for the learning capability of the point attention[23].

![Fig. 2. The hybrid point-spatial attention mechanism.](image)

Following the point-attention module, the spatial-attention module supplies the spatial relationship with the location information of the target in the feature map. As shown in Fig. 2, the point-attention feature map $G_p$ is simultaneously fed to the average-pooling and max-pooling operators, and the spatially compressed feature maps are derived as $F_s^{\text{Avg}} \in \mathbb{R}^{W \times H \times 1}$ and $F_s^{\text{Max}} \in \mathbb{R}^{W \times H \times 1}$.

By splicing $F_s^{\text{Avg}}$ and $F_s^{\text{Max}}$ along the channel dimensions as the input of the convolution layer, the spatial attention feature weight $A_s$ is obtained,

$$A_s = \sigma \left( \text{Conv}(\text{Cat}(F_s^{\text{Avg}}, F_s^{\text{Max}})) \right),$$

(6)

where $A_s \in \mathbb{R}^{W \times H \times 1}$. Conv(.) represents the convolution operation. Then, the spatial-attention feature map, i.e., the adjustment coefficient, is derived as,

$$G_H = G_p \otimes A_s,$$

(7)

where $G_H \in \mathbb{R}^{W \times H \times C}$. Considering the description of $G_p$ in Eq. (5), $G_H$ is further revised as,

$$G_H = G \otimes A_p \otimes A_s.$$

(8)

### 3.2 SE block based on the hybrid attention mechanism

The serial-connection mode in the hybrid attention mechanism ensures the effectiveness of
the attention module and yields a comprehensive adjustment coefficient on the thresholds of the bearing image. The structure of the improved residual shrinkage module based on the hybrid attention mechanism is shown in Fig. 3. The H-SE block is embedded into the main feedforward network of the residual shrinkage module. Firstly, the feature map $F$ is absolute valued to obtain the feature map $G$. The feature map $G$ generates a set of feature channel weights $A_P$ through the CA module, and the feature map $G_P$ is obtained by multiplying $A_P$ and the feature map $G$ element-wisely. Secondly, the feature map $G_P$ is used as the input of SA module to obtain a set of feature space weights $A_s$. Finally, the thresholds are obtained by multiplying the feature map $G$ and $A_s$ element by element.

The absolutization on feature map $F$ is employed as the generalized function, and it yields feature map $G$, with a size of $W \times H \times C$, as the input of the mixed attention mechanism, i.e.,

\[
G = \text{abs}(F) \ .
\]  

After the adjustment coefficient is obtained through the attention mechanism according to Eq. (8), a soft, thresholding activation function in DRSN realizes the denoising of the signal by setting the near-zero feature to zero. The function expression of the soft threshold is described as follows,

\[
y = \begin{cases} 
    x - \tau, & x > \tau \\
    0, & -\tau \leq x \leq \tau \\
    x + \tau, & x < -\tau 
\end{cases}
\]  

where $x$ and $y$ are the input feature and the output feature, respectively, and $\tau$ denotes the positive threshold affected by the comprehensive adjustment coefficient, i.e., $\tau = G \otimes A_s$. The output of the soft threshold function brings the values on both sides of the threshold close to zero, retaining the negative values which are absent in the ReLU activation function. Also, the derivative of the soft threshold function is either 0 or 1, which is similar to that of ReLU and avoids the gradient disappearance during the back propagation training of the neural network.

4 The fault diagnosis model for rolling bearings
4.1 MH-DRSN fault diagnosis model

With the introduction of the hybrid point-spatial-attention mechanism into the squeeze-and-excitation block, the DRSN-based fault diagnosis model is constructed with an improvement in feature-extraction robustness and anti-interference ability. Also, to combat the global information degradation caused by the deepening of the network layers, an additional dilated convolution is adopted in parallel with the residual shrinkage modules to supplement the multi-scale context information. The derived fault-diagnosis model MH-DRSN is shown in Fig. 4. It uses the 2-D gray images of vibration signals as input and is mainly composed of the ordinary convolution layer, the multi-scale feature extraction module, the serially connected residual shrinkage layers, the GAP layer, and the FC layer.

![Fig. 4. Network structure of the fault diagnosis model.](image)

The first convolution layer takes the vibration-signal-transformed image as its input. It is composed of the ordinary convolution layer and gives a source feature map for the residual shrinkage layers and the multi-scale feature-extraction module. The residual shrinkage layers assume the network structure of the residual shrinkage module as described in Fig. 3 to acquire the deep features of the input maps. Typically, the input image is of a size of 32×32. Due to the size reduction of the feature map through the down-sampling operation, four serial residual shrinkage layers are adopted in the diagnosis model to provide an appropriately sized feature map for the subsequent full-connection layer. Among them, the preceding three layers realize the multi-scale feature extraction, and the last one generates the prediction feature for the diagnosis output. Every residual shrinkage layer consists of two residual shrinkage modules with a similar network structure to ResNet18[24]. For residual shrinkage Layer 1, in order to utilize more fine-grained feature information, all the convolution layer strides are set to 1. In this case, the overlap area of the receptive field corresponding to each pixel of the feature map is small, which ensures a network with the extraction of more comprehensive detailed features. Considering that the down-sampling operation provides high-level semantic features with a larger receptive field and also that the convolution kernel in the network model is sized as 3×3, the down-sampling operation with a convolution kernel stride of 3 would lead to a gap between the receptive fields of adjacent steps and a consequent information omission in the feature map. Therefore, the strides of the first convolution layers in the remaining three residual shrinkage layers, i.e., Layer 2, Layer 3, and Layer 4 are all set at 2, and the strides of the other three convolution layers remain at 1. Then, the residual shrinkage layers give as their output a feature map with sizes of 32×32, 16×16, 8×8, and 4×4, respectively.
At the same time, a down-sampling operation by the convolution kernel with a stride of 2 in each layer results in a half-size of the current feature map. Therefore, the number of channels in the residual shrinkage layer doubles the former layer along the deepening of the network to prevent a loss of feature information. The channel number of Layer 1 is set to 64 according to the original feature map size of 32×32.

The multi-scale feature extraction module is constructed by dilated convolutions with various receptive fields. It is a convolution approach to obtain various ranges of information without losing feature information[25]. As shown in Fig. 5, the dilated convolution provides a larger receptive field by inserting holes into the ordinary convolution kernel. The inserted holes undergo a zero weight during the convolution of the input feature map. Therefore, the dilated convolution can expand the receptive field without changing the amount of calculation. With the introduction of the hyper parameter, i.e., the dilation rate $d$, the size of the convolution kernel of the dilated convolution becomes,

$$k^* = k + (k - 1) \times (d - 1),$$

where $k$ is the size of the ordinary convolution kernel. In the MH-DRSN fault-diagnosis model, various receptive fields are adopted to obtain the multi-scale feature information of the vibration signal by setting different dilation rates to the dilated convolution.

The output features of the dilated convolution are fused with the output features of the residual shrinkage layer 1, 2, and 3, respectively. Governed by the original input image size of 32×32, the dilation rates are set as $d = 2$, $d = 3$, and $d = 4$, respectively. The output feature map of the multi-scale feature extraction module is spliced with the residual shrinkage layer, respectively, with the same size. Thus, for the dilated convolution with $d = 2$, the stride and padding distance are set to 1 and 2, respectively, to be in accordance with the dimensions of the residual shrinkage Layer 1. For $d = 3$, the stride and padding distance are set to 2 and 3, respectively, to match with residual shrinkage Layer 2. The stride and padding distance are set to 4 and 4, respectively, to match with residual shrinkage Layer 3 for $d = 4$. Further, the convolution operation with a kernel with a size of 1×1 adjusts the channel dimension of the spliced features and propagates them to the next residual shrinkage layer in the MH-DRSN model. Finally, through the GAP layer and FC layer, the fault diagnosis result is derived.

### 4.2 Bearing image preparation from the vibration signal

For the rolling bearing fault diagnosis model, the one-dimensional bearing vibration signal is converted into two-dimensional images for the CNN operation. Considering that the commonly used time-frequency distribution images rely on manually extracted features and will cause information loss in the processing of the original signal, this paper adopts a simple and effective data preprocessing method. In this method, a signal interception and matrix transformation-based approach allows the original vibration signal to be converted to a gray image, so as to retain all the
characteristics of the original vibration signal. More importantly, this method does not require any predefined parameters, thus reducing dependence on expert experience as much as possible[26].

As shown in Fig. 6, the dataset sample of the original vibration signal is augmented through overlapping sampling to provide a sufficient number of training samples. It yields the number of obtained samples as,

$$M = \left( \frac{N - L}{S} \right) + 1,$$

where $N$ is the number of data points in the vibration signal, $L$ is the sample length selected when making data samples, and $S$ represents the step size. The overlapping sampling preserves the timing between adjacent data points and can learn more robust features for fault classification.

![Signal-to-image conversion process.](image)

The gray images with sizes of $D \times D$ are obtained by signal interception and matrix transformation according to the input image size requirement. Signal segments with the length of $D^2$ are intercepted from one-dimensional vibration signals and converted into a gray value range from 0 to 255 through a normalization and rounding function, where the length $D^2$ is chosen to permit at least one revolution of the bearing.

Then, the pixel value of the gray image is described as,

$$P(h, i) = \text{Round} \left\{ \frac{L(h-1) \times D + i - \min(L) \times 255}{\max(L) - \min(L)} \right\},$$

where $L(j), j = 1, 2, \ldots, D$, denotes the value of the signal segment. $P(h, i)$ denotes the value of the $h$-th row and the $i$-th column in the two-dimensional gray image with $h = 1, 2, \ldots, D, i = 1, 2, \ldots, D$. The minimal and maximal values of the intercepted signal are represented by $\min(L)$ and $\max(L)$, respectively. $\text{Round}(.)$ is the rounding function.

### 4.3 Loss function formulation

The fault diagnosis of rolling bearings is a multi-classification task. In the proposed MH-DRSN model, softmax is adopted as the activation function in the fully connected output layer to
estimate the probability distribution of samples[27]. It maps the outputs of multiple neurons into the range of (0,1) with a sum of 1.

Then, the loss function of the MH-DRSN fault diagnosis model is defined as the cross-entropy between the estimated output probability distribution and the target probability distribution, i.e.,

\[ H(p, q) = -\sum x p(x) \log q(x), \]  

(14)

where \( p(x) \) is the target distribution, while \( q(x) \) is the estimated distribution. The parameters in the model are trained by minimizing the loss function through optimization algorithms such as the stochastic gradient descent (SGD), root mean square prop (RMSprop), and adaptive momentum estimation (Adam), and others.

5 Experiments and discussion

The effectiveness of the proposed MH-DRSN model for rolling bearing fault diagnosis was validated on the CWRU datasets and XJTU-SY datasets subjected to noise interference. The experimental configuration is shown in Table 1, and the network model was written in the Pytorch framework using Python 3.6.

Table 1. Experimental configuration.

<table>
<thead>
<tr>
<th>Component</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>I5-9400F @2.9GHz</td>
</tr>
<tr>
<td>GPU</td>
<td>NVIDIA GeForce GTX1060 3GB</td>
</tr>
<tr>
<td>GPU acceleration library</td>
<td>CUDA 8.0 cuDNN v5.1</td>
</tr>
<tr>
<td>Deep learning framework</td>
<td>torch 1.0 torchvision 0.2.1</td>
</tr>
</tbody>
</table>

5.1 Experimental validation on the CWRU dataset

A. CWRU dataset description

The CWRU dataset is widely used in rolling bearing fault diagnosis as a benchmark[28]. Its data-acquisition testbed is shown in Fig. 7. The fault bearing was installed in the test motor and vibration data with different loads were recorded by the acceleration sensor at the motor drive end and the fan end. The status of bearings was of four types: normal, inner race fault (IF), ball fault (BF), and outer race fault (OF). Three directional outer race faults were included: 3, 6, and 12 o’clock positions. The fault diameters at 0.007 inches, 0.014 inches, 0.021 inches, and 0.028 inches simulate the various degrees of fault. During the experimental validation, the vibration signals at the driving end were selected for the validation of the proposed fault diagnosis model, which included three types of fault data collected at the sampling frequency of 12 kHz under a load of 0 HP. Ten types of vibration signal data were considered with the fault diameters at 0.007 inches, 0.014 inches, and 0.021 inches, the outer race fault at the 6 o’clock position with normal working conditions and with one sample consisting of 1024 sampling points. The offset step of the vibration signal was set at 240. Then, the input image samples for the MH-DRSN model were constructed by signal-to-image conversion and had a size of 32×32. A total of 5000 samples were adopted with every type containing 400 training samples and 100 test samples, respectively. The data distribution is shown in Table 2.
Table 2. Description of the dataset.

<table>
<thead>
<tr>
<th>Signal types</th>
<th>Fault diameter/ inch</th>
<th>Training sample</th>
<th>Test sample</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>-</td>
<td>400</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>BF</td>
<td>0.007</td>
<td>400</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>IF</td>
<td>0.007</td>
<td>400</td>
<td>100</td>
<td>2</td>
</tr>
<tr>
<td>OF (6 o’clock)</td>
<td>0.007</td>
<td>400</td>
<td>100</td>
<td>3</td>
</tr>
<tr>
<td>BF</td>
<td>0.014</td>
<td>400</td>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>IF</td>
<td>0.014</td>
<td>400</td>
<td>100</td>
<td>5</td>
</tr>
<tr>
<td>OF (6 o’clock)</td>
<td>0.014</td>
<td>400</td>
<td>100</td>
<td>6</td>
</tr>
<tr>
<td>BF</td>
<td>0.021</td>
<td>400</td>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>IF</td>
<td>0.021</td>
<td>400</td>
<td>100</td>
<td>8</td>
</tr>
<tr>
<td>OF (6 o’clock)</td>
<td>0.021</td>
<td>400</td>
<td>100</td>
<td>9</td>
</tr>
</tbody>
</table>

B. **MH-DRSN validation results on the CWRU dataset**

The MH-DRSN diagnosis model is mainly composed of a residual shrinkage layer, a dilated convolution network, and the feature fusion unit. According to the description in Section 4, its parameter settings are set as shown in Table 3. The parameters \((m_1, m_2, m_3, m_4)\) denote the convolution kernel of the convolution layer with a size of \(m_1 \times m_2\), the stride is governed by \(m_3\), and the number of channels is \(C = m_4\), respectively. The output size parameters \(W \times H \times C\) indicate the length, width, and channel number of the obtained feature map.

Table 3. Parameters of the MH-DRSN model.

<table>
<thead>
<tr>
<th>Layer type</th>
<th>Parameter</th>
<th>Padding</th>
<th>Dilation rate</th>
<th>Output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input layer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>32 \times 32 \times 1</td>
</tr>
<tr>
<td>Convolution</td>
<td>(3,3,1,64)</td>
<td>1</td>
<td>-</td>
<td>32 \times 32 \times 64</td>
</tr>
<tr>
<td>Dilated convolution 1</td>
<td>(3,3,1,64)</td>
<td>2</td>
<td>2</td>
<td>32 \times 32 \times 64</td>
</tr>
<tr>
<td>Residual shrinkage layer 1</td>
<td>(3,3,1,64)</td>
<td>1</td>
<td>-</td>
<td>32 \times 32 \times 64</td>
</tr>
<tr>
<td>Feature fusion unit 1</td>
<td>(1,1,1,64)</td>
<td>-</td>
<td>-</td>
<td>32 \times 32 \times 64</td>
</tr>
<tr>
<td>Dilated convolution 2</td>
<td>(3,3,2,128)</td>
<td>3</td>
<td>3</td>
<td>16 \times 16 \times 128</td>
</tr>
<tr>
<td>Residual shrinkage layer 2</td>
<td>(3,3,1,128)</td>
<td>1</td>
<td>-</td>
<td>16 \times 16 \times 128</td>
</tr>
<tr>
<td>Feature fusion unit 2</td>
<td>(1,1,1,128)</td>
<td>-</td>
<td>-</td>
<td>16 \times 16 \times 128</td>
</tr>
<tr>
<td>Dilated convolution 3</td>
<td>(3,3,4,256)</td>
<td>4</td>
<td>4</td>
<td>8 \times 8 \times 256</td>
</tr>
<tr>
<td>Residual shrinkage layer 3</td>
<td>(3,3,1,256)</td>
<td>1</td>
<td>-</td>
<td>8 \times 8 \times 256</td>
</tr>
<tr>
<td>Feature fusion unit 3</td>
<td>(1,1,1,256)</td>
<td>-</td>
<td>-</td>
<td>8 \times 8 \times 256</td>
</tr>
</tbody>
</table>
For comparison, six representative deep-learning networks – namely, LeNet5[29], CNN, VGG16[30], ResNet18[24], SENet18[16], and DenseNet121[31] were also adopted for bearing fault diagnosis on the CWRU dataset. They are representative in deep-learning approaches and classical convolutional neural networks for bearing diagnosis. The LeNet5 model adopts a classical structure with its convolution kernel size of 5×5 and a pooling layer stride of 2. The CNN model consists of four convolution layers followed by a BN, one pooling layer, and three fully connected layers. The size of the convolution kernel was set to 3×3, and the pooling layer stride was set to 2. The VGG16 diagnosis model was composed of 13 convolution layers with a convolution kernel size of 3×3, three full-connection layers, and one pooling layer with a stride of 2. The ResNet18 model consists of one convolution layer, eight ordinary residual blocks, one GAP layer, one FC layer, and the convolution kernels with a size of 3×3. The SENet18 model is based on the ResNet18 diagnosis model, and the SE block is embedded into the residual block to form a residual block with an attention mechanism. It is composed of one convolution layer, eight residual blocks with the attention mechanism, one GAP layer, one FC layer, and convolution kernels with a size of 3×3. The DenseNet121 model is composed of one convolution layer with 3×3 convolution kernels, 58 dense blocks, three transition layers, one GAP layer, and one FC layer.

Further, for the comparison of the MH-DRSN model with other representative methods, the parameters of LeNet5, CNN, VGG16, ResNet18, SENet18 and DenseNet121 were chosen as equivalent overall complexity according to the number of the computation parameters as shown in Table 4. The learning rate, optimizer, momentum and weight decay were chosen according to the models in the original literatures and modified to achieve their best diagnostic performance during the experiments. For the MH-DRSN model, the SGD algorithm with a momentum was used to train the network with a momentum of 0.9 and a weight decay of 0.0005. A stepwise-attenuated learning rate was adopted to improve the learning ability of the model. During the 20 training epochs with a batch size of 32, a damping coefficient was used to attenuate the learning rate to a smaller value. The initial learning rate was 0.01. The learning rate was 0.001 for 6 to 10 epochs, and the subsequent learning rate was set to 0.0001 for the remaining epochs.

Table 4. Hyperparameters of the network model.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>MH-DRSN</th>
<th>LeNet5</th>
<th>CNN</th>
<th>VGG16</th>
<th>ResNet18</th>
<th>SENet18</th>
<th>DenseNet121</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch size</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Learning rate (Epoch)</td>
<td>6~10</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>11~20</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>Optimizer</td>
<td>SGD</td>
<td>SGD</td>
<td>SGD</td>
<td>SGD</td>
<td>SGD</td>
<td>SGD</td>
<td>SGD</td>
</tr>
<tr>
<td>Momentum</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Weight decay</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0004</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Fig. 8 shows the changing curves of the training accuracy and test accuracy with the training times. We can see that the proposed, MH-DRSN diagnosis model yields a somewhat higher convergence speed than the other six models. Although there are small fluctuations, its training accuracy gradually tends to be stable after 280 iterations and arrives at a training accuracy of 100%. At the same time, the test accuracy of the MH-DRSN model reaches 99.5% at the first...
epoch and more for the subsequent epochs. As a result, it gives a 100% accuracy when it stabilizes. In comparison, the LeNet5, CNN, VGG16, ResNet18, SENet18, and DenseNet121 models encounter obvious fluctuations during both the training and test processes. Particularly, the test accuracy of the VGG16 model encounters an obvious fluctuation, and the VGG16 model undergoes an unexpected fluctuation at the seventh epoch.

Moreover, as shown in Fig. 9, in the training-loss change tendency for training iterations, the MH-DRSN model also presents a faster and more stable convergence ability compared to the other six deep-learning diagnosis models.

![Fig. 8. The training and test accuracies of the diagnosis models on the CWRU dataset.](image)

![Fig. 9. The change tendency of training loss.](image)

### C. Diagnosis Performance on CWRU dataset with Noise Interference

In order to evaluate the fault-diagnosis performance of the MH-DRSN model subjected to noise interference, Gaussian noise was added to the original vibration signal, and the signal-to-noise ratio (SNR) was used to measure the noise level. The SNR is defined as:

\[
\text{SNR} = 10 \log_{10} \left( \frac{P_{\text{signal}}}{P_{\text{noise}}} \right),
\]

(15)
where $P_{signal}$ and $P_{noise}$ represent the powers of signal and noise, respectively. $SNR$ is inversely proportional to noise intensity. During the experimental validation, all the diagnosis models were trained by samples without noise, while they were tested by samples with the added Gaussian white noise. Five groups of test samples were generated by superposing the noise on the original signals with $SNR = 2$ dB, 4 dB, 6 dB, 8 dB, and 10 dB, respectively.

Fig. 10 shows diagnosis accuracies of the models under noise interference, where the accuracy is obtained by the average of three repeated experiments. As shown in Fig. 10, with a declining $SNR$ from 10 dB to 2 dB, i.e., the gradually rising effect of the noise, the MH-DRSN model yields a 12.5% accuracy drop from 100% to 87.5%, while the other six network models show a drop of more than 20%. For the test vibration signal involved with the Gaussian white noise of 10 dB, LeNet5 yields an 85.13% diagnosis accuracy because of its simple structure and the absence of a BN layer. Comparatively, VGG16, ResNet18, SENet18, and DenseNet121 still provide a satisfactory diagnosis result with an accuracy of more than 92%. In particular, SENet18 and DenseNet121 have diagnostic accuracies of 99.53% and 99.17%, respectively. Noticeably, the MH-DRSN model still retains a consistently high level of fault diagnosis at 10 dB and 8 dB, providing an accuracy of 100% and 99.9%, respectively - roughly the same as that in noise-absent conditions. On the other hand, for the considerable noise interference condition of $SNR=2$ dB, the MH-DRSN model provides an 87.5% accuracy for fault diagnosis, which is 26.77%, 18.9%, 15.53%, 15.8%, 9.73%, and 15.03% higher than LeNet5, CNN, VGG16, ResNet18, SENet18, and DenseNet121, respectively. The MH-DRSN model obtains the context global information of the vibration signal image through the multi-scale feature extraction module, and denoises through the improved residual shrinkage module with a hybrid attention mechanism, and so it has a strong ability to suppress noise interference.

Fig. 10. The gray image and its time-frequency diagram after adding noise.
The ablation experiments were also carried out with and without the introduction of the hybrid attention mechanism and the multi-scale feature extraction module into the diagnosis MH-DRSN model. As shown in Fig. 11, the experiment was repeated three times to acquire the average value for an ablation results demonstration. DRSN is the original residual shrinkage network - exactly the same as in Ref.[15]. M-DRSN and H-DRSN are the modified DRSN models with the introduction of merely the multi-scale feature extraction modules and merely the hybrid attention mechanism. We can see that the hybrid attention mechanism used to select and soft-threshold the interference features effectively improves the model robustness. At the same time, the multi-scale feature extraction module can extract the global feature information in the noise data to its maximal extent and contribute to the anti-noise capability of the model. The MH-DRSN model combines the advantages of the two modules and achieves the best fault diagnosis accuracy.

Fig. 10. Comparison of fault diagnosis accuracy under different noise environments.

Fig. 11. Ablation experiment results of MH-DRSN on the CWRU dataset.

Fig. 12(a) plots the confusion matrix of the MH-DRSN model when classifying different fault types from the test samples with a noise interference of 8 dB, where 10 types of bearing fault
and 100 test samples for each type are investigated. The diagnosis accuracy of the model, the misjudgment number, and the misjudgment fault type are visually demonstrated simultaneously. We can see that out of the 1000 test samples, only one sample was predicted incorrectly with a diagnosis accuracy of 99.9%. One fault labelled as 1 is misclassified as label 3. All the misjudgments occur between different fault types while the diagnosis accuracy remains 100% for the existence of the fault state. Similar experimental results were derived for the test samples with a 2 dB noise interference, though there was a 12.5% drop in diagnosis accuracy as shown in Fig. 12(b).

The precision, recall, specificity, and overall achievement of the MH-DRSN model with noise of 8 dB and 2 dB noise are respectively illustrated in Table 5 and Table 6. They are used to evaluate the overall diagnosis performance of the MH-DRSN model under the noise interference and are defined as follows,

\[ P = \frac{TP}{TP + FP} \]  
\[ R = \frac{TP}{TP + FN} \]  
\[ S = \frac{TN}{TN + FP} \]  
\[ O = \frac{2TP}{2TP + FP + FN} \]

where \(TP, FP, FN,\) and \(TN\) indicate the number of positive samples predicted correctly as positive, the negative samples predicted incorrectly as positive, the positive samples predicted incorrectly as negative, and the negative samples predicted correctly as negative, respectively. We can see that for the test samples with a noise of 8 dB, the precision, recall, specificity, and overall achievement of the MH-DRSN model for all ten fault types are higher than 0.99 and the model provides excellent diagnostic performance. For the case of noise of 2 dB, a considerable, incorrect prediction occurs to the fault types labeled by 1. It yields a recall rate and an overall performance value of 0.14 and 0.246, respectively. Similar results occur to the fault types labeled 2, 3, and 6, while the other six fault types still give a relatively high diagnosis performance.
Table 5. Evaluation of the diagnosis results of the MH-DRSN model (SNR = 8 dB).

<table>
<thead>
<tr>
<th>Label</th>
<th>P</th>
<th>R</th>
<th>S</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.99</td>
<td>1</td>
<td>0.995</td>
</tr>
<tr>
<td>2</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0.999</td>
<td>0.995</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6. Evaluation of the diagnosis results of the MH-DRSN model (SNR = 2 dB).

<table>
<thead>
<tr>
<th>Label</th>
<th>P</th>
<th>R</th>
<th>S</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0.01</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.14</td>
<td>1</td>
<td>0.246</td>
</tr>
<tr>
<td>2</td>
<td>0.588</td>
<td>1</td>
<td>0.922</td>
<td>0.741</td>
</tr>
<tr>
<td>3</td>
<td>0.866</td>
<td>0.71</td>
<td>0.988</td>
<td>0.78</td>
</tr>
<tr>
<td>4</td>
<td>0.99</td>
<td>1</td>
<td>0.999</td>
<td>0.995</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0.725</td>
<td>1</td>
<td>0.958</td>
<td>0.84</td>
</tr>
<tr>
<td>7</td>
<td>0.99</td>
<td>0.97</td>
<td>0.999</td>
<td>0.98</td>
</tr>
<tr>
<td>8</td>
<td>0.962</td>
<td>1</td>
<td>0.996</td>
<td>0.98</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.93</td>
<td>1</td>
<td>0.964</td>
</tr>
</tbody>
</table>

5.2 Experimental validation on the XJTU-SY dataset

A. XJTU-SY dataset description

The XJTU-SY dataset contains the vibration signals of 15 rolling bearings in three working conditions. Three types of faults - the inner race, the outer race, and cage defect - are included for every five bearings. The vibration signals were obtained by two accelerometers installed on the bearing horizontally and vertically with a sampling frequency of 25.6 kHz and a sampling period of 1.28 seconds. The rolling bearing testbed is shown in Fig. 13[33]. In this experiment, five types of fault data under the condition of 2100 rpm with a load of 12 kN were selected for the validation of the diagnosis models as shown in Table 7. Since the XJTU-SY dataset has sufficient data, the overlapping sampling method was not used for data augmentation, and each sample was composed of 1024 data points. Five types of faults were considered within the generated 8320 samples, where each type corresponded to 1330 training samples and 334 test samples.

Fig. 13. The data acquisition testbed of the XJTU-SY dataset.
### Table 7. Detailed description of the XJTU-SY dataset.

<table>
<thead>
<tr>
<th>Signal types</th>
<th>Fault location</th>
<th>Training sample</th>
<th>Test sample</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bearing 1_1</td>
<td>Outer race</td>
<td>1330</td>
<td>334</td>
<td>0</td>
</tr>
<tr>
<td>Bearing 1_2</td>
<td>Outer race</td>
<td>1330</td>
<td>334</td>
<td>1</td>
</tr>
<tr>
<td>Bearing 1_3</td>
<td>Outer race</td>
<td>1330</td>
<td>334</td>
<td>2</td>
</tr>
<tr>
<td>Bearing 1_4</td>
<td>Cage</td>
<td>1330</td>
<td>334</td>
<td>3</td>
</tr>
<tr>
<td>Bearing 1_5</td>
<td>Inner race and outer race</td>
<td>1330</td>
<td>334</td>
<td>4</td>
</tr>
</tbody>
</table>

### B. MH-DRSN validation results on the XJTU-SY dataset

In the experiment, the MH-DRSN model took the same structure and configuration as in Section 5.1. The deep-learning methods, that is, LeNet5, CNN, VGG16, ResNet18, SENet18, and DenseNet121 were also employed under the same conditions as listed in Table 4 for comparison. The curves of the training accuracy and testing accuracy of the seven diagnosis models on the XJTU-SY dataset against the training epochs are shown in Fig. 14. We can see that the MH-DRSN model shows a faster convergence rate than the other six models. After about 210 iterations, it arrived at a training accuracy of 100% and remained stable with minor fluctuations. In comparison with the MH-DRSN model, the convergence speeds of the other six models are relatively slow. In particular, the LeNet5 model reaches a stable value of 99.5% after 1050 iterations, giving it the slowest convergence speed and obvious fluctuations due to its limitation of a simple structure. At the same time, the MH-DRSN model shows satisfactory test accuracy on diagnosing the involved faults. It gives a test accuracy of 99.4% at the first epoch and 100% at the ninth epoch and later. In comparison, LeNet5, CNN, VGG16, and SENet18 encounter considerable fluctuations with their test accuracies. Especially LeNet5 yields 93.89% at the 4th epoch, the VGG16 model gives 58.86% and 83.59% at the 4th and 18th epochs, respectively, and the SENet18 model provides 94.85% at the tenth epoch. Although the test accuracies of the ResNet18 model and the DenseNet121 model show no significant fluctuations, they are significantly lower than those of the MH-DRSN model with accuracies of 85.581% and 93.05% in the first epoch, respectively. Similar behaviors can be observed in Fig. 15, where the changing curves of the training loss for the MH-DRSN model and the other six models are demonstrated. The MH-DRSN model provides a rapid decrease in training loss with minimum fluctuation.
C. Diagnosis Performance on XJTU-SY dataset with Noise Interference

In the experiment, the fault diagnosis performance of the MH-DRSN model under ambient noise was validated on the XJTU-SY dataset with the addition of Gaussian white noise. The noises with SNRs of 0 dB, 2 dB, 4 dB, 6 dB, 8 dB, and 10 dB were superposed on the original vibration signal, and six groups of test samples were obtained to get the input gray images for the diagnosis models.

As shown in Fig. 16 concerning the average accuracy of the three repeated experiments of the models, the MH-DRSN model shows 100% diagnosis accuracy for 10 dB and 8 dB noise interference, and an accuracy decline with a growing effect of the noise up to 95.45% for the 2 dB noise. Even for the equivalent power of white noise addition on the test samples, i.e., \( SNR = 0 \text{ dB} \), it still provides an accuracy of 88.98%. In comparison, the LeNet5, CNN, VGG16, ResNet18, SENet18, and DenseNet121 models undergo an obvious drop in test accuracy with the increase of noise content. For the case of the 0 dB noise, the diagnosis accuracy of the MH-DRSN model is 20.58%, 20.04%, 27.84%, 15.01%, 17.9%, and 18.38% higher than LeNet5, CNN, VGG16, ResNet18, SENet18, and DenseNet121, respectively. The MH-DRSN model benefits from its multi-scale feature extraction module and an improved residual shrinkage module with a hybrid attention mechanism and derives a rather satisfactory diagnosis performance in the noise condition.

Further, ablation experiments on the MH-DRSN model were carried out to evaluate the effect of the hybrid attention mechanism and the multi-scale feature extraction module on noise suppression. Fig. 17 shows the test results on the XJTU-SY dataset superposed by white noise of 10 dB, 8 dB, 6 dB, 4 dB, 2 dB, and 0 dB, respectively, where the experiment assumes the same configuration as in Section 5.1 C, and the averages of the three experiments are used as the final diagnostic accuracy. We can see that all four DRSN-based models show a satisfactory diagnosis performance for the small degree of noise involvement. The models provide an accuracy of nearly 100% for noise of 10 dB, 8 dB, and 6 dB. With the increasing power of the noise from 4 dB to 0 dB, the accuracy of the DRSN decreases considerably from 95.09% to 78.2%, while the improved models, - H-DRSN introducing only the hybrid attention mechanism, M-DRSN with only the multi-scale feature extraction module, and MH-DRSN - present a slow declining tendency. In particular, the MH-DRSN model achieves the best diagnosis performance under different noise environments because of the advantages of its use of the two modules.
Fig. 16. Comparison of fault diagnosis accuracy under different noise environments.

The confusion matrix is used to visualize the classification results of the MH-DRSN model on different fault types under noise of 6 dB and 0 dB. As shown in Fig. 18(a), in the 1670 test samples, all the fault samples were successfully detected though one fault sample labelled as 1 was misjudged as label 2, and one fault sample labelled as 4 was misjudged as label 1. As shown in Fig. 18(b), similar results have been derived for test samples with noise of 0 dB, where all the fault samples were successfully detected, while 62 fault samples labelled as 0, 29 fault samples labelled as 1, 1 fault sample labelled as 2, 36 fault samples labelled as 3, and 56 fault samples labelled as 4 are misidentified as other fault types.
Fig. 18. Confusion matrix of the MH-DRSN model on the XJTU-SY dataset with 6 dB and 0 dB noise.

The precision, recall, specificity, and overall achievement of the MH-DRSN model with 6 dB and 0 dB noise are illustrated in Table 8 and Table 9, respectively.

Table 8. Evaluation of the diagnosis results of MH-DRSN model (SNR = 6 dB).

<table>
<thead>
<tr>
<th>Label</th>
<th>P</th>
<th>R</th>
<th>S</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.997</td>
<td>0.997</td>
<td>0.999</td>
<td>0.997</td>
</tr>
<tr>
<td>2</td>
<td>0.997</td>
<td>1</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.997</td>
<td>1</td>
<td>0.999</td>
</tr>
</tbody>
</table>

Table 9. Evaluation of the diagnosis results of the MH-DRSN model (SNR = 0 dB).

<table>
<thead>
<tr>
<th>Label</th>
<th>P</th>
<th>R</th>
<th>S</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.996</td>
<td>0.814</td>
<td>0.999</td>
<td>0.896</td>
</tr>
<tr>
<td>1</td>
<td>0.847</td>
<td>0.913</td>
<td>0.959</td>
<td>0.879</td>
</tr>
<tr>
<td>2</td>
<td>0.74</td>
<td>0.997</td>
<td>0.912</td>
<td>0.849</td>
</tr>
<tr>
<td>3</td>
<td>0.974</td>
<td>0.892</td>
<td>0.994</td>
<td>0.931</td>
</tr>
<tr>
<td>4</td>
<td>0.989</td>
<td>0.832</td>
<td>0.998</td>
<td>0.904</td>
</tr>
</tbody>
</table>

6 Conclusions

To improve the anti-noise performance and global information extraction of deep-learning networks in diagnosing rolling bearing faults, an improved DRSN model is proposed with the introduction of a hybrid attention mechanism and multi-scale extraction networks. The hybrid attention mechanism considers both the point relation in the feature map and the spatial dependence across channels simultaneously. Consequently, a comprehensive weight vector is derived for the soft thresholding function in the DRSN to select the robust and discriminative fusion features and reduce the impact of noise interference. At the same time, the dilated convolutional extraction module is adopted to obtain the multi-scale fault-sensitive information and fuse it with the feature information extracted by the residual shrinkage network, which solves the problem of global information loss during network deepening. Experiments results on the bearing fault datasets have shown the derived MH-DRSN model has shown great merits in convergence ability and diagnostic performance compared to representative models. It can
effectively discriminate the fault samples against the normal ones with a satisfactory anti-noise ability.

The application of the attention mechanism and the multi-scale feature extraction module in this paper is still insufficient for fault diagnosis subjected to various working loads. The load changes of rolling bearings in the actual environment cause different data distributions between the source domain and the target domain. Traditional noise suppression approaches cannot be used directly to deal with such distribution mismatch. Transfer learning can pre-train a model in one working condition (source domain) and adapts it to fault diagnosis in another working conditions (target domain) [34]. It shows great potential due to its simple feature extraction and good model migration performance. Subsequent research will focus on transfer learning methods based on MH-DRSN to realize the fault diagnosis of rolling bearings under variable working loads.

**Declarations**

**Ethical Approval**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

**Authors' contributions**

Xinliang Zhang proposed the idea and conceptualization of the approach, and supervised the work. Shengqiang Wei carried out the experiments and wrote the original main manuscript text. Jianhang Huang did formal analysis, investigation, and the data visualization about the model validation experiments. Lijie Jia collected and processed the data, participated in the results investigation during the experiments. All authors reviewed the manuscript.

**Funding**

This work is supported by Scientific and Technological Research Projects in Henan Province(222102210274, 232102210092), and the Fundamental Research Funds for the Universities of Henan Province (NSRF210305).

**Availability of data and materials**


**References**


[8] Li X et al., “Gear pitting fault diagnosis with mixed operating conditions based on adaptive 1D separable convolution with residual connection” Mechanical Systems and Signal Processing 142: 106740 (2020).


