IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 February 2023, accepted 8 March 2023, date of publication 14 March 2023, date of current version 22 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3257101

RESEARCH ARTICLE

Identifying Bearing Faults Using Multiscale Residual Attention and Multichannel Neural Network

CHUN-YAO LEE, (Member, IEEE), AND GUANG-LIN ZHUO
Department of Electrical Engineering, Chung Yuan Christian University, Taoyuan 320314, Taiwan

Corresponding author: Chun-Yao Lee (cyl@cycu.edu.tw)

ABSTRACT To solve the problem of the low signal-to-noise ratio and fault features can only be extracted from a single scale of traditional convolutional neural network (CNN) in vibration-based bearing fault diagnosis, this paper proposes a new multi-scale residual attention and multi-channel network (MSCNet), which can effectively reduce noise and fully extract meaningful features from different scales of the signal. The proposed method combines filtering methods to remove redundant parts and noise in the signal, and multiple filtered signals are input into the proposed CNN. The proposed CNN can perform multi-scale feature extraction on the signal and make the network focus on valuable information in the feature through the residual attention mechanism. Therefore, MSCNet achieves better performance. Experimental results on the published bearing datasets at the Paderborn University and the University of Ottawa show that MSCNet achieves 94.28% and 96.6% accuracy in strong noise environments, while outperforming five state-of-the-art (SOTA) networks in terms of accuracy.

INDEX TERMS Convolutional neural network (CNN), bearing fault diagnosis, multi-scale feature extraction, multi-channel network.

I. INTRODUCTION

Bearings are the basic components of rotating machinery. The health condition of bearings has an enormous influence on the stable operation of rotating machinery. Since rotating machinery usually works under harsh conditions of heavy load and strong noise, bearings are prone to failure, which may cause serious losses to operators [1]. Therefore, it is especially important to develop an intelligent diagnosis system that can accurately identify motor bearing faults.

Traditional intelligent bearing fault diagnosis models are usually composed of signal processing methods and machine learning (ML) methods [2]. These models first need to preprocess the acquired signal using signal processing methods, such as envelope analysis [3], wavelet transform [4], Hilbert-Huang transform [5], and nonlinear mode decomposition [6], etc. The features of the processing results are manually extracted [7] and then input to the machine learning method to obtain the diagnosis results. Some classic MLs have been widely used in intelligent bearing fault diagnosis, for example, k-nearest neighbor [8], linear discriminant analysis [9], naive Bayes [10], support vector machine [11], etc. Although these studies have made great contributions to intelligent fault diagnosis, manually extracted features are usually highly dependent on expert knowledge. Furthermore, the performance of MLs strongly depends on the quality of manual features [12]. Finally, shallow network architectures are often unable to effectively learn complex nonlinear relations [13].

In recent years, convolutional neural networks (CNNs) have attracted extensive attention due to their excellent performance in dealing with complex nonlinear features. CNN can remove information irrelevant to faults in vibration signals and learn complex nonlinear relations between features, thereby achieving better performance than MLs. Because of these advantages, CNN-based intelligent bearing fault diagnosis models have been widely proposed. For example, Gao et al. [14] proposed a CNN, the Nesterov momentum is

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

used to improve the capability to find the best solution, and the adaptive learning rate is used to improve the generalization capability of CNN. The results show that the proposed model substantially improves the convergence performance. Wang et al. [15] input the two-dimensional image based on singular value decomposition into CNN to effectively solve the problem of edge information loss. Shen et al. [16] introduced the concept of multi-label classification and successfully completed fault diagnosis in the missing label. Wang et al. [17] proposed a multi-input multi-task CNN based on time-domain signals, frequency-domain signals, and time-frequency graphs, which can effectively improve the accuracy of the model.

Although the above studies successfully prove the superiority of CNNs, these models still have some unresolved technical issues. These issues make them extremely challenging to apply in environments with strong noise and varying speeds and loads. The first technical issue is the multiscale properties of vibration signals. When the bearing fails, periodic impulses are generated due to the impact on the fault point. Then, the impulses are distributed on different time scales due to the large changes in the speed and load, as well as the different types and degrees of faults [18]. However, existing CNN models only use a single receptive field and cannot effectively extract fault features from different time scales. The second technical problem is noise pollution. Sensors collect vibration signals not only on target bearings but also environmental noise and vibration signals from other parts of rotating machinery. Due to the relatively low energy of the impulse excitation generated by the fault, the impulse is heavily polluted by these irrelevant noises [19]. The existing CNN model has a low tolerance to noise, which is easy to cause misdiagnosis.

To overcome the above-mentioned technical problems, researchers have conducted intensive studies. For example, Chen et al. [20] introduced a multi-scale CNN with parallel branches of four kernel sizes for bearing fault diagnosis. But the proposed model lacks an interaction strategy between different branches. Liu et al. [21] design an optimal sparse wavelet decomposition to extract multi-scale signals, and then the extracted signals are used as the input of a parallel CNN. However, the proposed model is highly dependent on the optimal sparse wavelet decomposition and has a large computational burden. Shi et al. [22] proposed a parallel CNN with kernel sizes of geometric series to extract multi-scale features and introduces the introduction of residual connections to prevent the loss of useful information. But the proposed model has the same problem as [20] which lacks the interaction strategies between different branches. Liu et al. [23] proposed a multi-scale parallel CNN consisting of three branches with different kernel sizes. However, the kernel size is too small, and the receptive field is limited, which makes the model easy to fall into the local optimum. In addition, the number of parameters of the proposed model is huge, which makes it difficult to apply to real scenarios. In summary, most of the existing research uses kernels of assorted sizes to complete multi-scale feature extraction. However, features at different scales are completely isolated and lack interaction, which limits the learning capability of the model.

Most of the existing CNN-based fault diagnosis research introduced attention mechanisms to overcome the problem of noise pollution. The attention mechanism can adaptively select the most key features while suppressing useless noise. The weights automatically generated based on learning greatly reduce the computational burden, so attention mechanisms are mostly lightweight and easy to implement. For example, Jin et al. [24] propose an adaptive anti-noise neural network (AAnNet). AAnNet employs a random sampling strategy and a modified attention mechanism for bearing fault diagnosis under heavy noise and varying load conditions. However, AAnNet is done at a single feature scale. Fang et al. [25] proposed a CNN with better anti-noise capability and domain adaptability and studied the splitting of feature maps to greatly reduce the number of parameters. But the proposed model only considers spatial attention. In summary, most of the existing research does not integrate multi-scale and attention at the same time, which limits the richness of features. In addition, we observe limited improvement in the signal-to-noise ratio (SNR) by the attention mechanism. The most fundamental reason is that it is extremely difficult for CNN to learn fault features directly from vibration signals [18]. This result motivates us to introduce filtering techniques to preprocess the vibration signal to reduce the influence of unwanted noise and further improve the performance of the model.

To overcome the two technical problems mentioned above, this article proposes a bearing fault diagnosis model called multi-scale residual attention and multi-channel network (MSCNet). In the proposed model, the morphological filter [26] and the mean filter [27] are introduced, which allow the network to learn rich features from the impulse signal and the noise-reduced signal. A parallel network of two channels extracts features from the impulse signal and the noise-reduced signal, respectively. Then, the features extracted by the dual-channel network are fused, which can help CNN to identify faults from more diverse perspectives. In the proposed network, a multi-scale feature extraction (MFE) module is adopted to extract multi-scale features and a residual learning feature extraction (RAFE) module is used to focus on key features in different scales to suppress redundant noise. The MFE module introduces wide-kernel convolution [28] and improved Res2Net, allowing CNN to extract diverse multi-scale features under a large receptive field. The hierarchical structure of the improved Res2Net can enhance the interaction between different time scales and enhance the learning capability of the model. At the same time, the improved Res2Net can complete feature extraction with fewer parameters than the original Res2Net [29]. The RAFE module adopts a parallel mechanism of channel attention

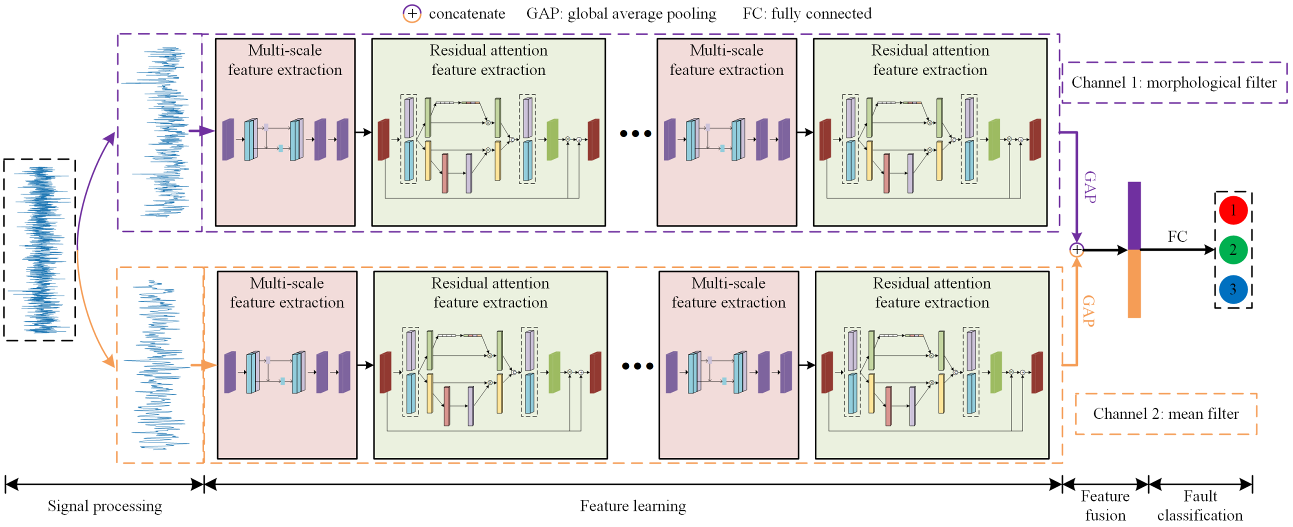


FIGURE 1. **Overall architecture of MSCNet.**

and spatial attention and introduces a residual mechanism to suppress the degradation of the network. Based on published bearing vibration datasets, experimental results show that MSCNet can achieve better performance than five state-of-the-art (SOTA) networks. The main contributions of this article are summarized as follows.

1) In this paper, the morphological filter and mean filter are combined with CNN, and the proposed parallel dual-channel network can extract representative features from the pulse signal and the noise reduction signal.
2) The proposed network consists of full connections of MFE and RAFE. MFE is a combination of wide-kernel convolution and improved Res2Net. This combination successfully solves the problem of traditional multi-scale CNN extracting features in a narrow receptive field and lacking interaction between different scales. RAFE considers parallel channel attention and spatial attention and residual connection. RAFE can make full use of the learned multi-scale features and suppress the network degradation problem, and it is a lightweight solution.
3) Based on the above improvements, a bearing fault diagnosis model based on MSCNet is formed. The proposed model successfully classifies accurately under strong noise and time-varying speed.

The rest of this paper is organized as follows. Section II elaborates on MSCNet. In Section III, MSCNet is validated with five SOTA networks under two bearing datasets. Meanwhile, extensive hyperparameter studies and ablation studies are given in this section. A discussion of the proposed model is conducted in Section IV. Finally, Section V concludes the paper.

II. PROPOSED METHOD

In this section, we introduce the proposed MSCNet in detail. The overall architecture of MSCNet is shown in Fig. 1. MSCNet integrates signal processing, feature learning, feature fusion, and fault classification, thus successfully completing bearing fault diagnosis under strong noise. MSCNet uses a morphological filter [26] and a mean filter [27] as the signal processing stage, which is used to extract the impulse and filter out the noise of the vibration signal, respectively. This stage aims to find complementary features from different input data to achieve more comprehensive feature learning. In the feature learning stage, MSCNet uses a dual-channel network architecture to learn the output signals of the morphological filter and the mean filter, respectively. The filtered signals are sequentially input into the multi-scale feature extraction (MFE) module and the residual attention feature extraction (RAFE) module. Multiple multi-scale feature extraction modules and residual attention feature extraction modules are stacked in MSCNet to construct a deeper network to learn key features. After fusing the complementary features, they are finally input to the fully connected (FC) layer for fault classification.

A. SIGNAL PROCESSING

1) MORPHOLOGICAL FILTER

In this study, the morphological filter is introduced as one of the signal processing layers of MSCNet to extract the impulse features in the bearing signal. The principle of the morphological filter is to extract the impulse of the signal by combining the signal and the structural element (SE) through morphological operations. Several researchers have shown that the shape of SE has insignificant effect on the performance of morphological filters [26]. Therefore, the use of complex SE shapes that affect computational efficiency should be avoided. There are several basic SE shapes, such as flat, triangle, and semicircle, etc. For impulse feature extraction, the shape of the triangle matches the impulse signal better. Therefore, MSCNet adopts triangle SE.

The self-complementary top hat (SCTH) is a morphological operator introduced in this study, which is used to simultaneously extract the positive and negative impulses of the vibration signal. The SCTH is formed by cascading four basic morphological operators of dilation, erosion, and opening and closing, which can be defined as follows:

$$\text{dilation } (x \oplus g) : y = \max\{x + g\} \tag{1}$$

$$\text{erosion } (x \Theta g) : y = \min\{x - g\} \tag{2}$$

$$\text{opening } (x \circ g) : y = x \Theta g \oplus g \tag{3}$$

$$\text{closing } (x \cdot g) : y = x \oplus g \Theta g \tag{4}$$

where x and y are the input signal and output signal, respectively. g indicates the SE. The SCTH is defined as the difference of the closing and opening operators, which is defined as follows:

$$\text{SCTH} = (x \cdot g) - (x \circ g). \tag{5}$$

In addition, choosing a suitable scale for SE is the key to affecting the filtering results. Assuming SE at scale S, the output can be defined as repeating the morphological operation on the input signal S-1 times. Thus, the morphological operators at scale S can be defined as follows:

$$x \oplus Sg = x \oplus \underbrace{(g \oplus g \oplus g)}_{S-1} \tag{6}$$

$$x \Theta Sg = x \Theta \underbrace{(g \Theta g \Theta g)}_{S-1}. \tag{7}$$

To find a suitable SE scale, this study introduces comprehensive metrics to evaluate different SE scales. The comprehensive index is the weighted average of the three statistical values of three commonly used impulse metrics, such as impulse factor, crest factor, and clearance factor, and assigns the same weight to the three metrics. Note that for rotating machinery, the clearance factor has a maximum value for healthy bearings and goes on decreasing for defective bearings. Therefore, in the comprehensive index, the clearance factor will add a negative sign. The comprehensive index defined Q_i as for the SCTH output signal at the i-th scale is defined as follows:

$$Q_i = \frac{\sum_{j=1}^{3} w_j H_{i,j}}{\sum_{j=1}^{3} w_j} \tag{8}$$

where the statistical values including impulse factor, crest factor, and clearance factor are defined as $H_{i,1}$, $H_{i,2}$, and $H_{i,3}$, and w_j represent the weights assigned to the impulse metrics. A comparison study of the best SE scale selection is conducted. In this comparison study, the same bearing dataset as in Section III was used, and four bearing fault signals (1024 data points) were selected from the dataset. The comparison results of Q under different S are shown in Table 1. The results show that the case of $S = 2$ has the highest Q value, indicating that the SCTH output signal contains the most impulse information, which is beneficial to the bearing fault diagnosis of MSCNet.

2) MEAN FILTER

Since rotating machinery often works in harsh environments, fault features are usually hidden by noise, which may cause the CNN model to be misled to learn the wrong features and misdiagnose. Therefore, it is necessary to introduce an effective denoising scheme to improve the performance of MSCNet in a strong noise environment. This study introduces the mean filter [27] as the signal processing layer of MSCNet to improve the anti-noise capability of the model.

The one-dimensional mean filter is defined as follows:

$$f(x) = \frac{1}{m} \sum_{s \in S_x} g(s) \tag{9}$$

where, S_x is the filter window whose center point is at x and the window size is m, $g(s)$ is the input signal.

Then, the window length of the mean filter is set to 3, which is selected from four candidates (3, 5, and 7) through empirical testing on MSCNet. In addition, the mean filter can filter out as much noise as possible by iterative operation. In this study, the number of iterations is determined by calculating the correlation coefficient (CC) between the current filtering result and the previous filtering result. When the CC reaches the threshold, the iteration is stopped, which means that further filtering will not bring about significant performance improvement. The threshold is set to 0.995, which is found by empirical testing of MSCNet.

B. MULTI-SCALE FEATURE EXTRACTION MODULE

The detailed architecture of the proposed MFE module is shown in Fig. 2. The MFE module consists of a wide-kernel Conv1D [28], an improved Res2Net structure [29], a batch normalization (BN) layer [30], and a rectified linear unit (ReLU) activation layer [31]. Based on Fig. 2, an input signal X (batch size: b) is first processed through the wide-kernel Conv1D to obtain a feature map y_w with c channels. Next, the improved Res2Net structure extracts multi-scale features

TABLE 1. Comparison of different SE scales.

Bearing code	S=1	S=2	S=3
KA16	0.82	**0.95**	0.81
KA22	0.71	**0.79**	0.73
KI16	0.93	**0.97**	0.84
KI18	0.72	**0.83**	0.78

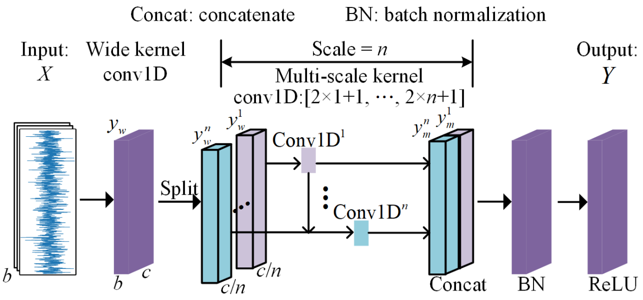


FIGURE 2. Detailed architecture of proposed MFE module.

through hierarchical structure and multi-scale kernels. In the improved Res2Net structure, the feature maps are divided into n subsets, and each subset is a feature scale. Each subset has c/n channels, which means that the Res2Net structure does not change the output size, showing its advantages of easy implementation. To extract features from different receptive fields, convolution kernels of assorted sizes are used in the improved Res2Net structure. The kernel size of the multi-scale kernel Conv1D^i is set to $2 \times i + 1$, where $i \in \{1, 2, \ldots, n\}$ corresponds to the feature subset y_w^i of each scale, and the convolution process is represented by Conv1D$^i(\bullet)$. Assuming the output of Conv1D$^i(\bullet)$ is y_m^i, the hierarchical structure can be expressed as

$$y_m^i = \begin{cases} \mathrm{Conv1D}^i(y_w^i), & i = 1 \\ \mathrm{Conv1D}^i(y_w^i + y_m^{i-1}), & 2 \le i \le n. \end{cases} \tag{10}$$

Finally, the feature maps of each scale are concatenate and then processed by the BN layer and the ReLU activation layer. The final output Y of the proposed MFE module can be expressed as

$$Y = \mathrm{ReLU}(\mathrm{BN}(y_m)). \tag{11}$$

In the proposed MFE module, wide-kernel Conv1D can supply a wider receptive field to help the network receive low-frequency features [28]. Res2Net has been proven for its powerful multi-scale feature learning capability [29]. Res2Net is characterized by fewer parameters and a hierarchical structure that facilitates the fusion of features at different scales. The improved Res2Net has two major improvements compared to the original Res2Net [29]. First, the improved Res2Net uses convolution kernels of assorted sizes and adds a convolution layer at the first scale to strengthen the receptive fields of different scales. Second, the BN layer and the ReLU layer after Conv1D are removed to reduce redundant calculations. In conclusion, the proposed MFE module can effectively extract multi-scale features.

C. RESIDUAL ATTENTION FEATURE EXTRACTION MODULE

In this article, MSCNet introduces the RAFE module to focus on the meaningful information in the features and remove the redundant part. The RAFE module introduces the attention mechanism in [32], which has the advantages of being lightweight and easy to implement. The detailed architecture of the RAFE module is shown in Fig. 3. The RAFE module first splits the number of channels c input feature map Y into g groups, that is, the number of channels for each group is c/g, $Y = [y_1, \ldots, y_g]$. Each group is then divided into two branches, that is, the number of channels for each branch is $c/2g$. One branch uses channel attention to generate channel attention feature, while the other branch uses spatial attention to generate spatial attention feature. In channel attention, the global average pool (GAP) is first performed on the input feature map y_{i1}, where $i \in \{1, 2, \ldots, g\}$, and obtain the channel statistic n through the spatial dimension l shrinking input feature map:

$$n = \mathrm{GAP}(y_{i1}) = \frac{1}{l}\sum_{j=1}^{l} y_{i1}^j. \tag{12}$$

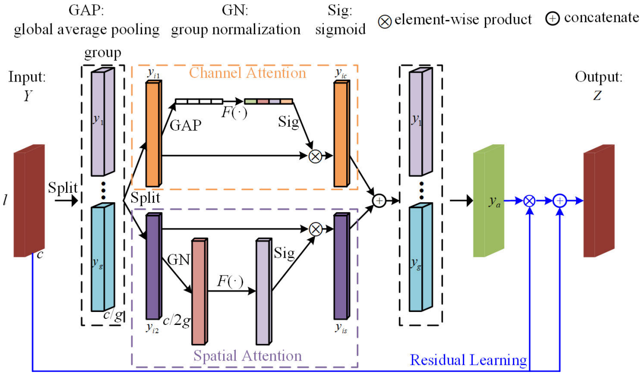


FIGURE 3. Detailed architecture of proposed RAFE module.

Next, the final output of channel attention y_{ic} is defined as follows:

$$y_{ic} = \mathrm{Sig}(F(n)) \cdot y_{i1} = \mathrm{Sig}(W_c n + b_c) \cdot y_{i1} \tag{13}$$

where Sig is the sigmoid activation function used to enhance adaptive selection, W_c and b_c are parameters used to scale and shift n.

In spatial attention, the group normalization (GN) is first performed on the input feature map y_{i2} to obtain spatial statistics, where $i \in \{1, 2, \ldots, g\}$. Then, W_s and b_s are parameters adopted to scale and shift the spatial statistics, same as channel attention. Finally, the final output of spatial attention y_{is} is defined as follows:

$$y_{is} = \mathrm{Sig}(W_s \cdot \mathrm{GN}(y_{i2}) + b_s) \cdot y_{i2}. \tag{14}$$

Then, the two branches are concatenated so that the number of channels is the same as the number of inputs. Finally, all groups are aggregated to get the output y_a of the attention module.

Usually, when using CNN to diagnose rotating machinery faults, it is necessary to construct a sufficiently deep network to achieve satisfactory results. However, several studies have shown that too-deep networks may cause degradation problem [23]. Therefore, residual learning [33] is embedded in the proposed RAFE module. The network can avoid the degradation problem by learning residual, and it is easier to perform than traditional feature learning. Assuming Y is the original feature map and y_a is the output of the attention module. The final output of the proposed RAFE module Z is defined as follows:

$$Z = y_a \otimes Y \oplus Y. \tag{15}$$

D. FEATURE FUSION AND FAULT CLASSIFICATION

Before the feature fusion stage, the learned features of the two channels are input to the GAP layer for feature dimension reduction. This mechanism can effectively reduce the number of parameters in the FC layer and prevent overfitting. Then, the pooled feature vectors of the two channels are fused into one vector as the input of the FC layer. In the fault classification stage, the FC layer only uses one layer of network, because the key features have been extracted in the earlier signal processing stage and CNN, so the FC layer can easily complete the fault classification.

III. EXPERIMENTAL RESULTS

In this section, the real bearing damage dataset under constant operating conditions from Paderborn University (PU) [34] and the time-varying speed bearing dataset from the University of Ottawa (UO) [35] are performed for experiments. To evaluate the performance of MSCNet, the proposed model is also compared with other state-of-the-art (SOTA) models.

A. CASE STUDY 1: REAL BEARING DAMAGE DATASET

1) DATASET DESCRIPTION

The dataset for real bearing damage is from the accelerated life test rig at PU. The test rig consists of test bearings, torque-measurement shafts, motor, load motor, and flywheels. As shown in Table 2, there are health bearings and bearings with varying degrees of outer and inner ring faults produced by accelerated life test. A detailed description of real bearing damage can be found in [34]. In addition, each bearing state has four different operating conditions, the operating parameters include rotational speed, load torque, and radial force on the bearing. In this study, data measured at a rotational speed of 1500 rpm, a load torque of 0.7 Nm, and a radial force of 10 N on the bearing were used. An accelerometer was mounted on the top end of the test bearing adapter to record vibration signals. At a sampling rate of 64,000 Hz, each bearing condition was completed with a duration of 4 seconds and independently repeated twenty times. In this study, three times of measurement results were taken to set up a dataset. Therefore, there are 64,000 × 4 × 3 = 768,000 data points per bearing condition. These data points are split into 750 (768,000 / 1024) samples. In total, the dataset has 750 × 15 = 11,250 samples.

2) PARAMETER SETTING AND EVALUATION CRITERION

MSCNet is implemented by PyTorch 1.10.2 platform and runs on a workstation with Windows 10 operating system, Intel Core i7-12700 CPU, and GTX 3060 GPU. During the training process, the loss function is cross-entropy, the optimization is the Adam algorithm with a learning rate of 0.001, the batch size is set to 64, and the model runs for 200 epochs.

The cross-entropy loss function L is expressed as:

$$L = \ell\left(x, y\right) = \left\{l_1, \ldots, l_N\right\}^{\top} \tag{16}$$

$$l_n = -w_{y_n} \log \frac{\exp\left(x_{n,y_n}\right)}{\sum_{c=1}^{C} \exp\left(x_{n,c}\right)} \cdot 1 \tag{17}$$

where x, y, w, C and N refer to the input, target, weight, number of classes, and number of samples.

In this paper, the accuracy defined by (18) is used as the evaluation criterion of the network:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{18}$$

where TP, TN, FP, and FN refer to the number of true positive samples, true negative samples, false positive samples, and false negative samples, respectively. In addition, all experiments in this case study were done under ten-fold cross-validation.

To better simulate the operating environment of the motor in the real world, we consider adding Gaussian noise to the original signal. The signal-to-noise ratio (SNR) is defined as

$$\text{SNR}_{\text{dB}} = 10 \log 10 \left(P_{\text{signal}} / P_{\text{noise}}\right) \tag{19}$$

where Psignal and Pnoise are the powers of the original signal and the added noise, respectively.

3) HYPERPARAMETER SELECTION FOR MSCNET

For CNN, the setting of hyperparameters is especially important. Appropriate hyperparameter settings can improve CNN performance, but it should be noted that we need to balance the number of parameters and accuracy. Table 3 shows the detailed architecture and hyperparameter settings of MSCNet. MSCNet has a dual-channel network composed of two identical CNN architectures (defined here as C_1 and C_2), a layer of concatenate (Concat), and a layer of FC. The CNN architecture consists of three MFEs, three RAFEs, and one GAP layer. The shapes of the output signal of the morphological filter and the mean filter are set to 1 × 1024. The signals are input into respective networks. The parameter setting of MFE involve s the kernel size, stride, and the number of channels in the wide-kernel Conv1D; and the scale in the improved Res2Net. The stride of all MFEs is fixed to 2 so that the output length is halved each time. Since the length of the output is halved, the kernel size is halved. But the number of channels is doubled to extract more valuable features. The scale of all improved Res2Nets is fixed to 2. Only the number of groups g needs to be set in RAFE. However, it was seen in

TABLE 2. Description of the dataset used in this case study.

Bearing code	Fault type	Label
K001	healthy	1
K002		
K003		
K004		
K005		
KA04	outer ring	2
KA15		
KA16		
KA22		
KA30		
KI04	inner ring	3
KI14		
KI16		
KI18		
KI21		

TABLE 3. Architecture and hyperparameter settings of MSCNET.

		Channel 1, 2 (C_1, C_2)				Output shape
		MFE				
		Wide-kernel Conv1D			Improved Res2Net	
Layer	type	Kernel	Stride	c	Scale	(c, l)
	Input: (C_1, C_2)	-	-	-	-	(1, 1024)
1	MFE	96	2	16	2	(16, 512)
2	RAFE	-	-	-	-	(16, 512)
3	MFE	48	2	32	2	(32, 256)
4	RAFE	-	-	-	-	(32, 256)
5	MFE	24	2	64	2	(64, 128)
6	RAFE	-	-	-	-	(64, 128)
7	GAP	-	-	-	-	(64, 1)
8	Concat	-	-	-	-	(128, 1)
9	FC	-	-	-	-	Class number

the case study that g does not affect the model performance, so g is fixed at 8, which is set to match the division of the number of model channels.

To find the best hyperparameter settings for MSCNet, a hyperparameter study was conducted on the PU dataset under Gaussian noise with SNR $= -6$ dB. As shown in Table 4, the number of model parameters, floating-point operations (FLOPs), and average precision of ten-fold cross-validation are used as evaluation metrics. In the hyperparameter study, the kernel size, stride, and channel number of MFE; and the scale of the improved Res2Net are compared and analyzed. The "Selected" row indicates the final hyperparameter setting of MSCNet, and the blanks in Table 4 indicate that the parameter uses the same setting as "Selected". In the kernel size experiment, all MFEs with the same kernel size and different kernel sizes (halved sequentially) were evaluated. In the case that all MFEs have the same kernel size, the model achieves lower accuracy with similar or more parameters and FLOPs, and the larger size of the kernel does not improve the accuracy. In the case of different kernel sizes (halved sequentially), when the kernel size increases from 32, the accuracy increases. The accuracy reaches 94.28% at a kernel size of 96 but drops to 94.2% at a kernel size of 128. The stride is related to dimensionality reduction. When the stride is higher, the length of the output feature will be shorter, so the FLOPs will be smaller. However, if the stride is too large, plenty of information will be lost. These results can be seen when the model has a stride size of 3 and 4 with accuracies of 93.2% and 92.13%, respectively. When the model has two times the number of channels than "Selected", the parameters and FLOPs are close to four times that of "Selected", but the accuracy is still at 94.27%. When the number of channels of all layers is 16, the accuracy is 93.65%.

Scales in improved Res2Net are evaluated under 3 and 4. When the scale is 3, the number of channels in the model is adjusted to "24, 48, 96" to split the channels evenly within the model. It can be noticed that the parameter and FLOPs become more than doubled, but the accuracy is only 0.05% higher than "Selected", reaching 94.33%. Meanwhile, the accuracy of the model at scale 4 is 93.76%. Therefore, the "Selected" model in Table 4 is the best hyperparameter setting for MSCNet.

4) ABLATION STUDY: THE FILTERED SIGNAL

To confirm the contribution of each filter to fault diagnosis, an ablation experiment is conducted by using the morphological filter and the mean filter separately on the PU dataset. It is noted that since a single filtered signal is used, the models of the ablation experiment use a single-channel MSCNet. In this experiment, C_1 is the model using the morphological filter alone, and C_2 is the model using the mean filter alone. In addition, the original signal is included in the ablation experiment. Table 5 shows the number of parameters, FLOPs, and classification results under different Gaussian noises for all models. C_2 contributes significantly to fault diagnosis, especially in the case of strong noise. At SNR $= -6$ dB, C_2 brings a 10.72% improvement compared to the original signal. The accuracy of C_1 fluctuates under all noise conditions, which shows that the model cannot learn the impulse features alone. The classification results of the proposed model show that there are complementary features between the two filtered signals, which improves the robustness of the model.

5) ABLATION STUDY: FEATURE EXTRACTION MODULE

To verify the effectiveness of the feature extraction module, three ablation models are designed. The first model is to remove the wide-kernel Conv1D in the MFE module, the second model is to remove the improved Res2Net in the MFE module, and the third model is to remove the RAFE module. These three ablation models are named Model 1, Model 2 and Model 3 in this paper, respectively. Note that Model 1 is replaced by a convolutional layer with a kernel size of 1. Table 6 shows the number of parameters, FLOPs, and classification results for all models under different Gaussian noises. The wide-kernel Conv1D in the MFE module contributes greatly to fault diagnosis, especially in the case of strong noise. At SNR $= -4$ dB and -6 dB, the wide-kernel Conv1D brings a 3.5% accuracy improvement. As mentioned in the earlier section, the advantage of the wide-kernel Conv1D is to provide the model with a wide receptive field to extract global and local features. It can be clearly saw that Model 1 falls into the local optimum at SNR $= -6$ dB. The improved Res2Net accounts for about 13% of the number of parameters, bringing a 0.5% performance improvement to the model in all noise cases. RAFE is one of the valuable modules in MSCNet, and the contribution of RAFE becomes increasingly significant as the noise power increases. At SNR $= -6$ dB, Model 3 degrades the accuracy by 1.63%. It is worth noting that RAFE does not need additional parameters and FLOPs only account for 50,000, indicating that RAFE is lightweight and very suitable for actual environments.

6) COMPARISON WITH OTHER METHODS

To verify the superiority of MSCNet, MSCNet is compared with five SOTA models published in the field of rotating

TABLE 4. Hyperparameter study of MSCNET.

	Wide-kernel Conv1D			Improved Res2Net			
	Kernel	Stride	*c*	Scale	Parameter	Flops	Accuracy (%)
Selected	96, 48, 24	2	16, 32, 64	2	173,091	61.29M	94.28 ± 0.7
	32, 32, 32				187,427	59.2M	93.19 ± 0.7
	64, 64, 64				352,291	110.58M	93.21 ± 1.16
	32, 16, 8				72,739	25.64M	93.15 ± 1.29
	64, 32, 16				122,915	43.47M	93.61 ± 0.78
	128, 64, 32				223,267	79.12M	94.20 ± 0.8
		3			173,091	23.87M	93.2 ± 0.89
		4			173,091	12.72M	92.13 ± 1.12
			32, 64, 128		684,099	237.93M	94.27 ± 0.74
			16, 16, 16		43,395	20.98M	93.65 ± 1.43
			24, 48, 96	3	378,291	132.45M	94.33 ± 0.64
			16, 32, 64	4	167,715	59.46M	93.76 ± 0.76

TABLE 5. Ablation experiment results of filtered signal.

Models	Parameter	Flops	−6dB	−4dB	−2dB	0dB
MSCNet	**173,091**	**61.29M**	**94.28 ± 0.7**	**96.43 ± 0.84**	**97.57 ± 0.52**	**98.41 ± 0.42**
C_1	86,547	30.65M	72.04 ± 1.78	71.46 ± 2.63	74.37 ± 1.45	74.65 ± 3.0
C_2	86,547	30.65M	92.33 ± 1.12	91.11 ± 1.83	94.32 ± 0.82	96.41 ± 0.51
Original signal	86,547	30.65M	81.61 ± 1.65	85.81 ± 2.14	88.36 ± 3.35	91.87 ± 1.14

TABLE 6. Ablation experiment results of feature extraction module.

Models	Parameter	Flops	−6dB	−4dB	−2dB	0dB
MSCNet	**173,091**	**61.29M**	**94.28 ± 0.7**	**96.43 ± 0.84**	**97.57 ± 0.52**	**98.41 ± 0.42**
w/o wide-kernel Conv1D	27,715	11.53M	90.78 ± 1.1	92.93 ± 1.1	95.13 ± 0.91	96.51 ± 0.6
w/o improved Res2Net	151,587	53.95M	93.62 ± 1.05	95.94 ± 0.8	97.07 ± 1.9	97.96 ± 0.45
w/o RAFE	173,091	61.24M	92.65 ± 0.7	94.92 ± 1.12	96.85 ± 0.75	97.61 ± 0.66

TABLE 7. Comparison results using PU dataset.

Models	Parameter	Flops	−6dB	−4dB	−2dB	0dB
MSCNet	**173,091**	**61.29M**	**94.28 ±0.7**	**96.43 ±0.84**	**97.57 ±0.52**	**98.41 ±0.42**
MC-CNN [36]	4,486	1.86M	83.86 ±4.0	88.49 ±2.78	90.93 ±2.28	93.1 ±1.36
MK-ResCNN [23]	2.15M	167.37M	89.76 ±1.84	92.97 ±1.42	95.26 ±1.0	95.65 ±1.56
MBSCNN [18]	587,291	31.7M	86.8 ±2.72	91.64 ±1.68	95.14 ±1.59	97.8 ±0.49
MSCNN [37]	15.68M	229.33M	92.35 ±1.57	95.08 ±1.04	96.28 ±1.22	97.89 ±0.4
MRA-CNN [38]	598,536	117M	93.43 ±0.79	95.64 ±0.9	96.99 ±0.64	97.67 ±0.66

machinery fault diagnosis in recent years. These five methods include multi-scale [23], [36], [37], multi-scale residual attention learning [38], and multi-branch and multi-scale method [18]. These methods and MSCNet use the same parameter settings to complete comparison experiments.

Table 7 shows the number of parameters, FLOPs, and classification results for all models under different Gaussian noises. MSCNet achieves the best accuracy at all noise powers. In strong noise environments (SNR = −6 dB, −4 dB, and −2 dB), MRA-CNN [38] performs best among the five methods. MRA-CNN is based on the multi-scale mechanism to improve feature learning ability and residual attention learning to improve anti-noise capability, but it lacks to find complementary features from different signals. MC-CNN [36] only achieves 83.86% accuracy at SNR = −6 dB. MC-CNN [36] does not fully consider the noise problem, so the performance of the model degrades severely in a stronger noise environment. MK-ResCNN [23] fails to use a sufficiently wide kernel to provide global and local information in the first convolutional layer, resulting in the subsequent multi-channel multi-scale kernels that can only capture local features. MBSCNN [18] uses filtered signals and multi-channel multi-scale methods to enhance feature learning. But MBSCNN [18] lacks an attention mechanism to focus on learning key features, resulting in its weak denoising capability. MSCNN [37] uses coarse-grained filtering to obtain multi-scale signals. However, this filtering method cannot change the contour of the signal, resulting in the limited performance of the model. At the same time, it can be observed that MSCNet only uses 29% of the number of parameters and 52% of FLOPs of MRA-CNN [38], which is a relatively lightweight solution.

In addition, the average training loss curves of each method at SNR = −6 dB are shown in Fig. 4. The training loss of each method gradually stabilizes in the first 40 epochs. MSCNet, MRA-CNN, and MK-ResCNN have the fastest convergence speed, and the loss can converge to 0. However, it can be found that MC-CNN, MK-ResCNN, and MBSCNN fall into the local optimum according to the average validation accuracy given in Table 7. Then, the convergence process of MSCNN is unstable. Based on the above analysis results, MSCNet successfully proved its superiority in bearing fault diagnosis.

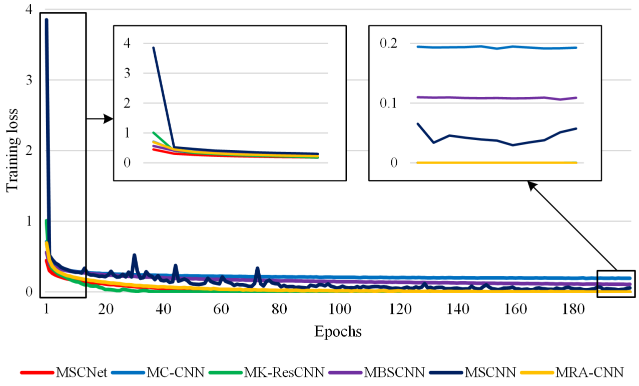


FIGURE 4. Average training loss curves for each method.

B. CASE STUDY 2: TIME-VARYING SPEED BEARING DATASET

Rotating machinery often runs under different operating conditions. Fault diagnosis under different operating conditions is a challenging task because fault features may be blurred under varying operating conditions [35]. Therefore, in this case study, experiments are performed using the time-varying speed bearing dataset from UO.

1) DATASET DESCRIPTION

The time-varying speed bearing dataset is from UO's machinery fault simulator (MFS-PK5M). The simulator consists of a motor, AC drive, shaft, and two ER16K ball bearings (one healthy and the other experimental). Mount an accelerometer (ICP accelerometer, Model 623C01) on the bearing housing of the experimental bearing to record vibration signals. The experimental data were collected by the NI data acquisition board (NI USB-6212 BNC) at a sampling rate of 200,000 Hz for a duration of 10 seconds. In addition, an incremental encoder (EPC model 775) was installed to measure the shaft speed.

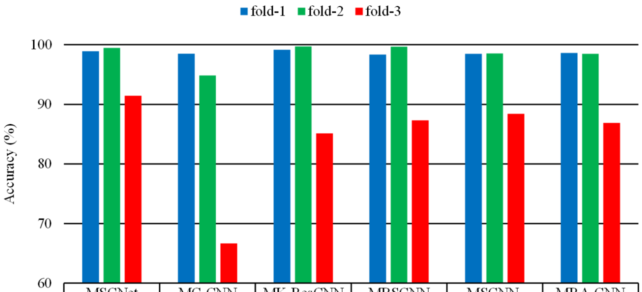
The UO dataset has twelve experimental setups consisting of three bearing health conditions and four varying speed conditions. Bearing health conditions include health, inner race faults, and outer race faults. Varying speed conditions include increasing speed, decreasing speed, increasing and then decreasing speed, and decreasing and increasing speed. Then three trials were completed for each experimental setup, resulting in a total of thirty-six datasets. This case study uses a varying speed condition of decreasing and then increasing velocity, so a total of nine datasets are used. Table 8 shows the detailed operating conditions for the nine datasets used in this case study. Each dataset has 200,000 × 10 = 2,000,000 data points. These data points are divided into 1953 (1,999,872 / 1024) samples, and the part that does not meet 1,024 data points is removed. The dataset has a total of 1953 × 9 = 17,577 samples. This case study uses three-fold cross-validation for training and testing, with two trials for each bearing health condition for training and one for testing, resulting in three combinations.

TABLE 8. Detailed operating conditions of UO dataset.

Health condition	Trial	Rotational speed (Hz)
Healthy	1	From 24.2 to 14.8, then from 14.8 to 20.6.
	2	From 24.6 to 14.0, then from 14.0 to 18.6.
	3	From 26.0 to 16.9, then from 16.9 to 23.2.
Inner race fault	1	From 25.3 to 14.8, then from 14.8 to 19.4.
	2	From 25.3 to 15.1, then from 15.1 to 19.8.
	3	From 23.1 to 15.7, then from 15.7 to 23.6.
Outer race fault	1	From 26.0 to 18.9, then from 18.9 to 24.5.
	2	From 25.2 to 14.9, then from 14.9 to 19.5.
	3	From 25.5 to 15.0, then from 15.0 to 19.6.

2) COMPARISON WITH OTHER METHODS

Same as case study 1, MSCNet is compared with five SOTA models under Gaussian noise with SNR = -6 dB. The results of the comparison are shown in Fig. 5. MSCNet achieves the best average accuracy among the six models. From the accuracy trend of each fold, it can be clearly observed that all models have a significant drop in accuracy in the third fold (Trial 2 and Trial 3 for training, and Trial 1 for testing). Except for MC-CNN [37] which suffers the most performance drop with about a 29% decrease in accuracy. This result indicates that for the UO dataset, the main fault features are distributed in the working conditions of Trial 1, so the features learned from Trial 2 and Trial 3 are blurred in Trial 1. But MSCNet is still the most accurate among the six models. MK-ResCNN [23], MBSCNN [18], MSCNN [37], and MRA-CNN [38] all achieve satisfactory performance, and MC-CNN [37] has the worst average accuracy. It is shown that multi-scale mechanisms are crucial for fault diagnosis with time-varying speeds. Because MC-CNN [37] only uses convolutional layers with three kinds of kernel sizes to achieve multi-scale. In conclusion, the above analysis results prove that MSCNet can handle fault diagnosis with time-varying speed.



	MSCNet	MC-CNN	MK-ResCNN	MBSCNN	MSCNN	MRA-CNN
🟦 fold-1	98.89	98.50	99.18	98.34	98.48	98.60
🟩 fold-2	99.47	94.85	99.71	99.66	98.55	98.46
🟥 fold-3	91.43	66.67	85.13	87.34	88.41	86.89
Average	96.60 ± 4.48	86.67 ±17.42	94.67 ± 8.27	95.11 ± 6.77	95.15 ± 5.83	94.65 ± 6.72

FIGURE 5. Comparison results using uo dataset.

IV. DISCUSSION

The MSCNet proposed in this study has two key concepts. One is a multi-channel mechanism combining various filters.

The other is a CNN network constructed by MFE and RAFE. Two key concepts are discussed in detail next.

For the multi-channel mechanism, MSCNet filters the original signal from two aspects to obtain richer fault features, namely impulse extraction, and noise reduction. According to Table 5, from the original signal to C_2 (MSCNet with mean filter alone), the accuracy of CNN at SNR $= -6$ dB is improved from 81.61% to 92.33%. On the one hand, it proves the noise reduction capability of the mean filter, and on the other hand, it shows the superiority of the proposed CNN network. Compared with C_2, the accuracy of MSCNet at SNR $= -6$ dB is improved from 92.33% to 94.28%. This indicates that the learned impulse features in C_1 can provide additional fault information and thus improve the model performance. However, the accuracy of MSCNet using the mean filter alone fluctuates under different noise powers. This shows that although the fault pulse of the bearing is extracted, the different fault types cannot be well distinguished. This result motivates us to explore other advanced impulse filtering methods to further improve MSCNet in the future.

For the CNN network, as shown in Table 6, the wide-kernel Conv1D in MFE is the part that mainly affects the performance. The role of wide-kernel Conv1D is to provide CNN with a wide receptive field to capture local and global fault information, so as to achieve better performance. It is worth noting that the difference in kernel size will affect the performance of the model, but this setting may only be applicable to the currently used dataset. But in this paper, for a fair comparison with other SOTA models, the kernel size is not adjusted for different datasets. According to [18], the size of the convolution kernel should become larger as the sampling frequency increases. On the other hand, the improved Res2Net can find rich features in different time scales, and then the RAFE module guides the model to focus on meaningful parts of the features.

V. CONCLUSION

This paper proposed an MSCNet model for bearing fault diagnosis. The main contribution of MSCNet is the combination of filtering techniques, multi-channel mechanism, multi-scale mechanism, and residual attention mechanism. MSCNet can effectively remove redundant information in the signal, learn from the filtered signal and focus on complementary features of different scales, to achieve reliable bearing fault diagnosis. In this study, a hyperparameter selection experiment was performed to optimize model performance. At the same time, the ablation experiment was carried out to verify the effectiveness of the filters and CNN network in MSCNet. MSCNet is evaluated against five SOTA networks on a fixed working condition PU dataset and a time-varying rotational speed UO dataset to demonstrate its superiority. For the PU dataset, MSCNet achieves 94.28% accuracy at SNR $= -6$ dB while outperforming the other five SOTA networks. For the UO dataset, MSCNet achieves 96.60% accuracy at SNR $= -6$ dB while outperforming the other five SOTA networks. These results prove that MSCNet can keep reliable performance under strong noise and time-varying rotational speed. Therefore, this study proposes an effective intelligent diagnosis model for bearing faults, which can help to improve the reliability of rotating machinery operations.

REFERENCES

[1] B. Ma, W. Cai, Y. Han, and G. Yu, "A novel probability confidence CNN model and its application in mechanical fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[2] C.-Y. Lee and T.-A. Le, "An enhanced binary particle swarm optimization for optimal feature selection in bearing fault diagnosis of electrical machines," *IEEE Access*, vol. 9, pp. 102671–102686, 2021.

[3] E. Bechhoefer, M. Kingsley, and P. Menon, "Bearing envelope analysis window selection using spectral kurtosis techniques," in *Proc. IEEE Conf. Prognostics Health Manage.*, Jun. 2011, pp. 1–6.

[4] Y. Xin, S. Li, Z. Zhang, Z. An, and J. Wang, "Adaptive reinforced empirical Morlet wavelet transform and its application in fault diagnosis of rotating machinery," *IEEE Access*, vol. 7, pp. 65150–65162, 2019.

[5] N. E. Huang, *Hilbert–Huang Transform and Its Applications*, vol. 16. Singapore: World Science, 2014.

[6] D. Iatsenko, P. V. E. McClintock, and A. Stefanovska, "Nonlinear mode decomposition: A noise-robust, adaptive decomposition method," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 92, no. 3, Sep. 2015, Art. no. 032916.

[7] Y. Li, W. Dai, and W. Zhang, "Bearing fault feature selection method based on weighted multidimensional feature fusion," *IEEE Access*, vol. 8, pp. 19008–19025, 2020.

[8] Q. Wang, S. Wang, B. Wei, W. Chen, and Y. Zhang, "Weighted K-NN classification method of bearings fault diagnosis with multi-dimensional sensitive features," *IEEE Access*, vol. 9, pp. 45428–45440, 2021.

[9] X. Yu, F. Dong, E. J. Ding, S. P. Wu, and C. Y. Fan, "Rolling bearing fault diagnosis using modified LFDA and EMD with sensitive feature selection," *IEEE Access*, vol. 6, pp. 3715–3730, 2018.

[10] B. R. Nayana and P. Geethanjali, "Analysis of statistical time-domain features effectiveness in identification of bearing faults from vibration signal," *IEEE Sensors J.*, vol. 17, no. 17, pp. 5618–5625, Sep. 2017.

[11] Y. Cheng, H. Zhu, K. Hu, J. Wu, X. Shao, and Y. Wang, "Health degradation monitoring of rolling element bearing by growing self-organizing mapping and clustered support vector machine," *IEEE Access*, vol. 7, pp. 135322–135331, 2019.

[12] G. Xu, M. Liu, Z. Jiang, W. Shen, and C. Huang, "Online fault diagnosis method based on transfer convolutional neural networks," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 2, pp. 509–520, Feb. 2020.

[13] H. Tang, Z. Liao, P. Chen, D. Zuo, and S. Yi, "A novel convolutional neural network for low-speed structural fault diagnosis under different operating condition and its understanding via visualization," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[14] S. Gao, Z. Pei, Y. Zhang, and T. Li, "Bearing fault diagnosis based on adaptive convolutional neural network with Nesterov momentum," *IEEE Sensors J.*, vol. 21, no. 7, pp. 9268–9276, Apr. 2021.

[15] F. Wang, G. Deng, L. Ma, X. Liu, and H. Li, "Convolutional neural network based on spiral arrangement of features and its application in bearing fault diagnosis," *IEEE Access*, vol. 7, pp. 64092–64100, 2019.

[16] J. Shen, S. Li, F. Jia, H. Zuo, and J. Ma, "A deep multi-label learning framework for the intelligent fault diagnosis of machines," *IEEE Access*, vol. 8, pp. 113557–113566, 2020.

[17] Y. Wang, M. Yang, Y. Li, Z. Xu, J. Wang, and X. Fang, "A multi-input and multi-task convolutional neural network for fault diagnosis based on bearing vibration signal," *IEEE Sensors J.*, vol. 21, no. 9, pp. 10946–10956, May 2021.

[18] D. Peng, H. Wang, Z. Liu, W. Zhang, M. J. Zuo, and J. Chen, "Multibranch and multiscale CNN for fault diagnosis of wheelset bearings under strong noise and variable load condition," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4949–4960, Jul. 2020.

[19] H. Tang, Z. Liao, P. Chen, D. Zuo, and S. Yi, "A novel convolutional neural network for low-speed structural fault diagnosis under different operating condition and its understanding via visualization," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[20] J. Chen, R. Huang, K. Zhao, W. Wang, L. Liu, and W. Li, "Multiscale convolutional neural network with feature alignment for bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.

[21] X. Liu, J. Centeno, J. Alvarado, and L. Tan, "One dimensional convolutional neural networks using sparse wavelet decomposition for bearing fault diagnosis," *IEEE Access*, vol. 10, pp. 86998–87007, 2022.

[22] Y. Shi, A. Deng, M. Deng, J. Zhu, Y. Liu, and Q. Cheng, "Enhanced lightweight multiscale convolutional neural network for rolling bearing fault diagnosis," *IEEE Access*, vol. 8, pp. 217723–217734, 2020.

[23] R. Liu, F. Wang, B. Yang, and S. J. Qin, "Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 3797–3806, Jun. 2020.

[24] G. Jin, T. Zhu, M. W. Akram, Y. Jin, and C. Zhu, "An adaptive anti-noise neural network for bearing fault diagnosis under noise and varying load conditions," *IEEE Access*, vol. 8, pp. 74793–74807, 2020.

[25] H. Fang, J. Deng, B. Zhao, Y. Shi, J. Zhou, and S. Shao, "LEFE-Net: A lightweight efficient feature extraction network with strong robustness for bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[26] J. Serra, "Introduction to mathematical morphology," *Comput. Vis., Graph., Image Process.*, vol. 35, no. 3, pp. 283–305, 1986.

[27] P. Zhang and F. Li, "A new adaptive weighted mean filter for removing salt-and-pepper noise," *IEEE Signal Process. Lett.*, vol. 21, no. 10, pp. 1280–1283, Oct. 2014.

[28] H. Qiao, T. Wang, P. Wang, L. Zhang, and M. Xu, "An adaptive weighted multiscale convolutional neural network for rotating machinery fault diagnosis under variable operating conditions," *IEEE Access*, vol. 7, pp. 118954–118964, 2019.

[29] S. H. Gao, M. M. Cheng, and K. Zhao, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37. Lille, France, 2015, pp. 448–456.

[31] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 1–8.

[32] Q. L. Zhang and Y. B. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," in *Proc. ICASSP*, Jun. 2021, pp. 2235–2239.

[33] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[34] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, "Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification," in *Proc. Eur. Conf. Prognostics Health Manage. Soc.*, vol. 2016, pp. 5–8.

[35] H. Huang and B. Natalie, "Bearing vibration data collected under time-varying rotational speed conditions," *Data Brief*, vol. 21, pp. 1745–1749, Dec. 2018.

[36] W. Huang, J. Cheng, Y. Yang, and G. Guo, "An improved deep convolutional neural network with multi-scale information for bearing fault diagnosis," *Neurocomputing*, vol. 359, pp. 77–92, Sep. 2019.

[37] G. Jiang, H. He, J. Yan, and P. Xie, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 3196–3207, Apr. 2019.

[38] L. Jia, T. W. S. Chow, Y. Wang, and Y. Yuan, "Multiscale residual attention convolutional neural network for bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

CHUN-YAO LEE (Member, IEEE) received the Ph.D. degree from the Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, in 2007. From 1998 to 2008, he was a Distribution System Designer with the Engineering Division, Taipei City Government, and CECI Engineering Consultants Inc., Taiwan. From 2004 to 2006, he was a Visiting Scholar with the Department of Electrical Engineering, University of Washington, Seattle, WA, USA, sponsored by the Ministry of Science and Technology, Taiwan. Since 2008, he has been a Faculty Member of electrical engineering with Chung Yuan Christian University, Taoyuan, Taiwan, where he is currently a Full Professor. His research interests include power distribution, optimization algorithms, and damage diagnosis. He received the National Excellent Teachers Award, Taiwan, in 2020.

GUANG-LIN ZHUO received the B.Eng. degree in electrical engineering from Chung Yuan Christian University, Taoyuan, Taiwan, in 2018, where he is currently pursuing the Ph.D. degree with the Electrical Engineering Department. His current research interests include fault diagnosis and deep learning.