Journal of Dynamics, Monitoring and Diagnostics, 2023, 2, 115-132
https://doi.org/10.37965/jdmd.2023.156

ISTP
RESEARCH ARTICLE

RMA-CNN: A Residual Mixed Domain Attention CNN for Bearings Fault Diagnosis and Its Time-Frequency Domain Interpretability

Dandan Peng,[1,2,4] Huan Wang,[3] Wim Desmet,[1,2] and Konstantinos Gryllias[1,2,4]
[1]Department of Mechanical Engineering, Faculty of Engineering Science, KU Leuven, Leuven, Belgium
[2]Flanders Make@KU Leuven, Celestijnenlaan 300, BOX 2420, 3001 Leuven, Belgium
[3]Department of Industrial Engineering, Tsinghua University, Beijing 100084, China
[4]Leuven.AI – KU Leuven Institute for AI, B-3000, Leuven, Belgium

(Received 20 January 2023; Revised 12 April 2023; Accepted 21 April 2023; Published online 21 April 2023)

Abstract: Early fault diagnosis of bearings is crucial for ensuring safe and reliable operations. Convolutional neural networks (CNNs) have achieved significant breakthroughs in machinery fault diagnosis. However, complex and varying working conditions can lead to inter-class similarity and intra-class variability in datasets, making it more challenging for CNNs to learn discriminative features. Furthermore, CNNs are often considered "black boxes" and lack sufficient interpretability in the fault diagnosis field. To address these issues, this paper introduces a residual mixed domain attention CNN method, referred to as RMA-CNN. This method comprises multiple residual mixed domain attention modules (RMAMs), each employing one attention mechanism to emphasize meaningful features in both time and channel domains. This significantly enhances the network's ability to learn fault-related features. Moreover, we conduct an in-depth analysis of the inherent feature learning mechanism of the attention module RMAM to improve the interpretability of CNNs in fault diagnosis applications. Experiments conducted on two datasets—a high-speed aeronautical bearing dataset and a motor bearing dataset—demonstrate that the RMA-CNN achieves remarkable results in diagnostic tasks.

Keywords: attention interpretability; CNN; fault diagnosis; rolling element bearings

I. INTRODUCTION

Rolling element bearings are critical mechanical components extensively used in various machinery systems, providing support and reducing friction [1]. Bearing failure can lead to equipment shutdown or, in severe cases, to equipment damage and casualties [2]. Consequently, ensuring the reliability and safety of bearings has garnered significant attention from industry professionals. Real-time fault diagnosis of bearings is essential for maintaining the normal operation of mechanical systems and facilitating the timely replacement of damaged bearing components.

With advancements in the industrial Internet of Things and sensor equipment, vast amounts of monitoring data can be collected. Machine learning methods, capable of learning fault-related features from historical data, have been extensively researched for bearing fault diagnosis tasks. For example, Kang et al. [3] proposed a fault diagnosis method for rolling bearings based on kernel discriminant feature analysis and support vector machine methods. Baraldi et al. [4] presented an improved method based on the K-nearest neighbor for automatically diagnosing bearing faults under various working conditions. Typically, most methods [5–8] first employ feature extraction techniques (e.g., empirical wavelet transform [5] and empirical mode decomposition [6]) to obtain useful signal features. These features are then input into a machine learning algorithm [4,8] to classify bearing health status. However, these feature extraction techniques often rely on expert knowledge, and their feature extraction capabilities are limited. Additionally, shallow machine learning algorithms struggle to handle the complex non-linear relationships between inputs and outputs.

In recent years, several powerful deep learning algorithms have been proposed [9,10], achieving state-of-the-art performance in fields such as computer vision [9], speech recognition [11], and signal processing [12]. Unlike traditional methods, deep learning can automatically learn and extract features from raw data, allowing for more efficient and accurate diagnostic performance. This is especially advantageous in fault diagnosis, where data can be noisy, complex, and high-dimensional, and thus, the underlying patterns and relationships may be highly non-linear and difficult to capture using traditional methods [13,14]. By leveraging the power of deep learning, researchers and practitioners in the fault diagnosis field can achieve more accurate and reliable diagnoses, leading to improved safety, reduced downtime, and increased efficiency.

Particularly, the convolution operation of convolutional neural networks (CNNs) [15,16] makes a significant breakthrough in machinery fault diagnosis [17–25]. Zhao et al. [26] proposed a novel deep residual shrinkage network that effectively enhanced the network's feature extraction ability. Peng et al. [27] combined multi-scale and multi-branch concepts with CNN to develop an improved CNN model for wheelset bearing fault diagnosis tasks. Zhang et al. [28] introduced a residual learning-based CNN method for diagnosing faults in rotating machinery. Wen et al. [29] first transformed the signal into a two-dimensional (2D) image and then employed 2D convolution to

Corresponding author: Konstantinos Gryllias (e-mail: konstantinos.gryllias@kuleuven.be).

learn the spatiotemporal features of the signal, ultimately outputting diagnostic results. All these methods depend on the powerful feature learning capabilities of CNNs to capture signal features, utilize fully connected layers to encode these features, and finally obtain diagnostic results.

Despite the considerable achievements of CNNs in mechanical fault diagnosis tasks, they still face challenges, including:

- Intra-class variability: Bearing operating conditions, such as loads and speeds, often change in complex ways. Consequently, the characteristic pattern of the same fault type may significantly vary in terms of periodicity and amplitude, making it difficult for the network to learn the inherent features of that fault type.
- Inter-class similarity: Bearings present different types of faults, with some fault features being too weak to be distinguished from normal conditions. Some faults may have similar vibration responses, causing the network to misclassify the fault type easily.
- Weak ability to focus on meaningful features: While CNNs have powerful automatic feature learning abilities, they struggle to focus on meaningful features instead of noise or other masking signals from various sources. This makes it challenging to address the intra-class variability and inter-class similarity problems faced in bearing fault diagnosis tasks.
- Poor interpretability: As a "black box" for researchers, it hinders the development of CNN in machinery health condition monitoring. Interpretability is crucial for both academic research and industrial applications.

To address the aforementioned limitations, attention-based deep learning methods have been developed. Specifically, attention modules, such as channel attention modules or time attention modules, have been integrated into CNN networks to enhance their feature learning ability and improve diagnostic performance. For instance, Wang et al. [30] improved the noise resistance and bearing diagnostic ability of a CNN network by incorporating a channel attention module. Hao et al. [31] introduced a channel attention module into each scale network of a multi-scale network to enhance its feature learning ability. Jia et al. [32] proposed a multi-scale residual attention CNN for bearing fault diagnosis, where a residual attention module was introduced into each scale network as well to improve the model's performance. However, these studies did not provide an interpretability analysis to explain the underlying reasons for the effective performance of attention modules, especially from the perspective of time and frequency domain analysis methods.

Therefore, this paper aims to deeply explore the inherent interpretability of the feature learning mechanism of attention modules. By integrating traditional signal analysis techniques, we provide an in-depth analysis from a time-frequency domain perspective to explain how attention modules contribute to improving bearing diagnostic performance. The proposed residual mixed domain attention module (RMAM) is designed to effectively enhance the network's feature learning ability and strengthen the network's learning of intra-class variability and inter-class similarity features. RMAM constructs attention-based feature learning mechanisms for both time and channel domains. The time domain attention module focuses on extracting signal components related to signal impulses, which are more likely to be associated with fault events. The channel domain attention module focuses on extracting frequency components related to faults, which can help capture relevant information even in the presence of variations due to changes in load and speed. By introducing mixed attention modules, the network can learn to focus on the relevant signal components associated with faults, while ignoring the noise signal components caused by variations due to load and speed. This reduces the impact of intra-class variability and inter-class similarity of signals on the diagnostic performance of the network, ultimately leading to better performance in bearing fault diagnosis tasks. The proposed RMAM has a minor increase in parameters. It can serve as an independent, lightweight network module to form an arbitrary depth network architecture for various fault diagnosis tasks. Additionally, RMAM employs residual connections [9] to optimize the network's gradient transfer, enabling the construction of deeper networks.

Furthermore, an RMAM-based CNN architecture (RMA-CNN) for bearing fault diagnosis is proposed and evaluated on two popular public bearing datasets, a High-Speed Aeronautical (HAS) [1] bearing dataset and a motor bearing dataset [33], achieving competitive performance.

The proposed attention mechanism can automatically learn the time domain and channel domain information most relevant to the task. The relationship between the input and output of RMAM is deeply explored, discussing and analyzing the inherent mechanism of RMAM feature learning. This contributes to the interpretability of CNNs in mechanical fault diagnosis.

The contributions of this paper are summarized as follows:

- This paper designs a novel attention mechanism (RMAM) that can automatically extract fault-related features from noisy signals and enhance the network's discriminative feature learning ability.
- This paper proposes a CNN framework based on residual mixed domain attention for bearing fault diagnosis. This framework is a simple and versatile model that can be flexibly adapted to various health monitoring tasks.
- This paper provides an in-depth analysis of the feature learning mechanism of the attention method from a time-frequency domain perspective.

The paper is organized as follows: Section II describes the proposed RMA-CNN in detail. Section III verifies the effectiveness and the superiority of the RMA-CNN method. Section IV discusses the interpretability of the attention mechanism. Finally, Section V summarizes the conclusions of the paper.

II. METHODOLOGY

A. METHOD OVERVIEW

The proposed framework for bearing fault diagnosis is illustrated in Fig. 1. It primarily comprises a bearing vibration signal acquisition system and an end-to-end condition monitoring model based on deep learning. Figure 1(a) displays the test rig utilized for bearing vibration signal acquisition, which essentially consists of a high-speed spindle driving the rotation of a shaft. An acceleration sensor is

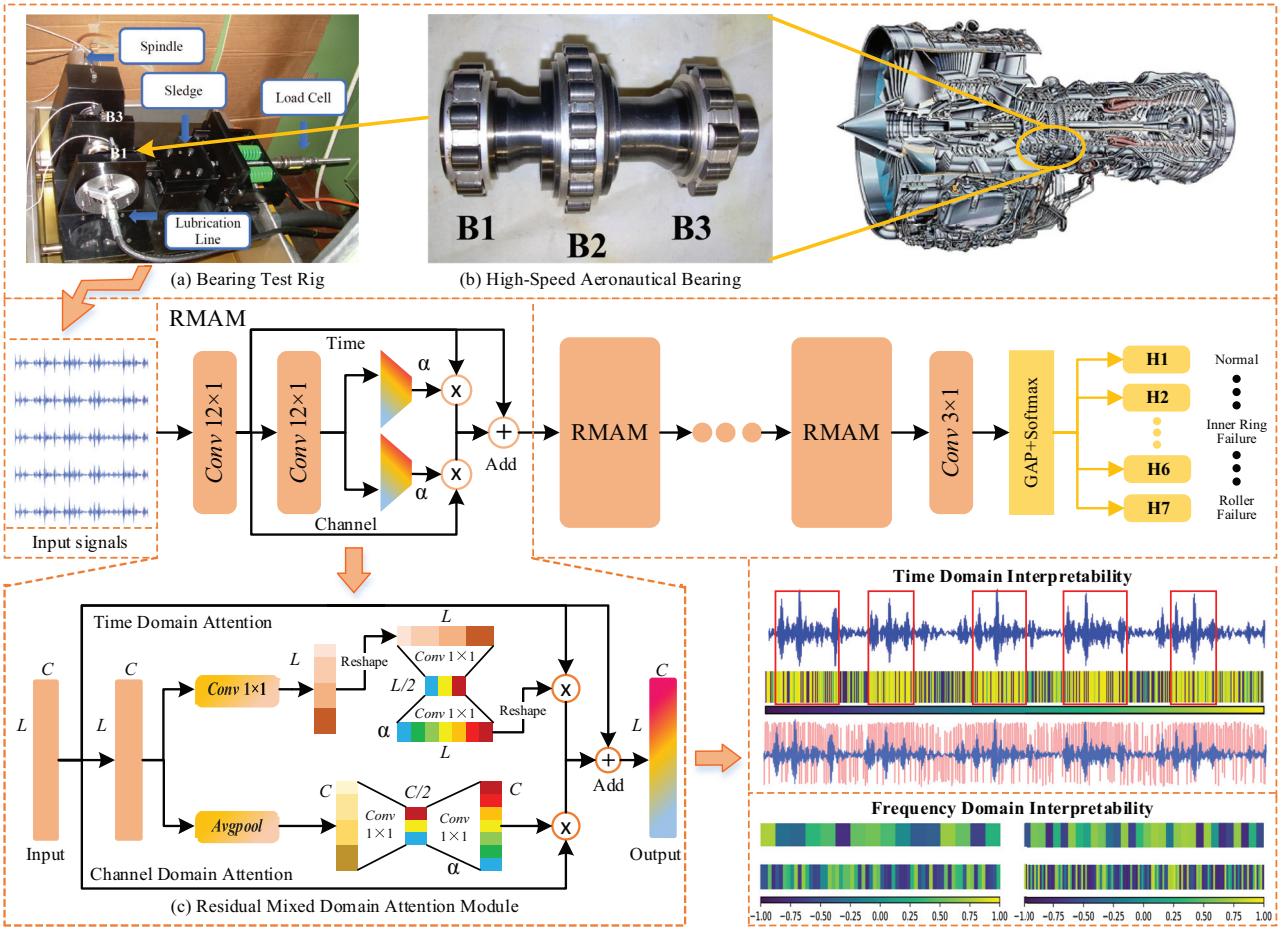


Fig. 1. The detailed network architecture of the proposed RMA-CNN and the RMAM.

mounted on the test rig to monitor the bearing's status information. Figure 1(b) presents the shaft with its three bearings. During the bearing experiment, the fault location is in the B1 bearing.

To achieve precise bearing status monitoring, a novel RMAM-based CNN architecture is presented, boasting exceptional discriminative feature learning capabilities and impressive scalability. As depicted in Fig. 1, the RMA-CNN takes raw vibration signals as input and delivers an end-to-end assessment of the bearing's health condition. Comprised of multiple RMAMs, the architecture allows for flexible adjustments in the number of RMAMs to accommodate various dataset types and sizes.

To address the intra-class variability and inter-class similarity challenges posed by vibration signals, RMAM introduces both channel domain and time domain attention mechanisms, as illustrated in Fig. 1(c). These attention mechanisms work to enhance the network's ability to learn meaningful features. A GAP layer and a classification layer with Softmax are then employed to aggregate the acquired features and produce diagnostic results. Moreover, our approach employs traditional signal analysis methods to examine the feature learning mechanism and the interpretability of the attention mechanisms from both time and frequency domain perspectives. The following parts provide a comprehensive introduction to the proposed RMAM, along with detailed information about the RMA-CNN.

B. RESIDUAL LEARNING

First, consider a plain CNN block consisting of several simple convolution layers. Assume that the function mapping learned by these layers is defined as $H(x)$. Throughout the training process, these layers will directly fit $H(x)$. Here, a basic CNN block can be defined as Eq. (1).

$$y = F(x,W) \tag{1}$$

where x and y denote the input and output of the block, respectively, while $F(\cdot)$ is the function mapping and W represents the parameters learned by these layers.

The fundamental assumption of residual learning [9] is that, compared to having the convolution layers directly learn the complex function $H(x)$, it is more manageable to learn its residual function, which also simplifies the network training. In the residual learning architecture, these convolution layers learn the residual function $H(x) - x$ instead of directly fitting $H(x)$. The definition of residual learning is illustrated in Eq. (2).

$$y = F(x,W) + x \tag{2}$$

where the function $F(\cdot)$ denotes the residual mapping learned by these convolution layers. The residual learning mechanism can be integrated into the network through skip connections. Although the introduction of residual learning allows for a deeper network and, consequently, for enhanced learning capabilities, it is more worthwhile to

explore lighter and more efficient methods to improve the model's discriminative feature learning ability.

C. CHANNEL DOMAIN ATTENTION LEARNING

In 1D-CNN, the convolution layer takes a 1D time step signal $v(t)$ as input and uses a set of convolution kernels to convolve the signal, capturing useful features of the signal, and finally outputs the obtained feature maps. During network training, the CNN model optimizes and updates the parameters of the convolution kernel to improve its feature learning ability, enabling the network to obtain more useful features. From the perspective of signal analysis, the convolution kernel can be defined as a time domain function $c(t)$, and the input signal is defined as $v(t)$. In this way, the feature learning process of CNN is a time domain convolution of $c(t)$ and $v(t)$. Time domain convolution is equivalent to frequency domain multiplication, which means that the convolution kernel is essentially a filter that controls which frequency domain information is retained and which is discarded [34]. Each convolution layer contains multiple filters to capture different frequency domain features from the input signal.

However, each layer of a deep CNN contains a large number of convolutional kernels, and it is usually difficult to collect enough data to effectively optimize all the kernel parameters. Among these kernels, some learned features are useful for diagnosis, while others may be irrelevant. These irrelevant features may affect the network's decision-making. Therefore, it is challenging for CNNs to identify which kernels are more relevant for diagnosis tasks since the model treats all kernels with equal importance weights.

To address this issue, channel domain attention is proposed. As shown in Fig. 1(c), we assume that the output features of the second convolution module (in the second row) of RMAM are $M = [m_1, m_2, \ldots, m_C]$, where $m_i \in \mathfrak{R}^{L\times 1}$ denotes the feature on the ith channel. We first use a GAP layer to aggregate global information in the time domain, obtaining a channel descriptor $z \in \mathfrak{R}^{1\times C}$ for each channel. The core of channel domain attention is to find out which channel feature is more important for the fault diagnosis. Therefore, two non-linear transformation layers are adopted to obtain the relative importance among different channels. The non-linear layers consist of two 1×1 convolution layers, which reduce by half of the original dimension and then restore their original dimensions. The activation function α maps the resulting vector to a fixed weight range and outputs the final channel weight vector $\hat{z} \in \mathfrak{R}^{1\times C}$. The value $\hat{z}$ represents the importance of the corresponding channel feature. Finally, $\hat{z}$ is employed to enhance the meaningful channel features in M, as shown in Eq. (3).

$$N_z = M \otimes \hat{z} = [m_1\hat{z}_1, m_2\hat{z}_2, \cdots, m_C\hat{z}_C] \tag{3}$$

where $\hat{z}_i$ is the ith element of $\hat{z}$, and $\otimes$ denotes the element-wise multiplication of two matrices.

D. TIME DOMAIN ATTENTION LEARNING

Vibration signals of bearings are time domain signals that contain periodic and time-correlated information. There is a strong signal correlation among different time sequences, and much valuable information is hidden in some signal sequences. For example, when a local fault occurs to a bearing, a rolling element passes over the defect each time, and a periodic impulse is generated and excites the natural frequencies of the structure. The fault-impulsive signal components contain more meaningful information than other signal sequences and can more directly reflect the inherent properties of the faulty bearing. Therefore, the goal of the time domain attention learning module is to make the network pay more attention to important signal sequences in the time domain.

As illustrated in Fig. 1(c), we redefine M as $M = [m^1, m^2, \ldots, m^L]$, where $m^j \in \mathfrak{R}^{1\times C}$ represents the feature at the jth point at the time axis. We first use one 1×1 convolution layer to aggregate the global information on the channel domain to generate a time feature vector $q \in \mathfrak{R}^{L\times 1}$. The core of the time domain attention is to find out which signal segment information is more important for the fault diagnosis. In order to facilitate the calculation of convolution, a time feature vector $q \in \mathfrak{R}^{L\times 1}$ is reshaped into $q' \in \mathfrak{R}^{1\times L}$. Similarly, two non-linear layers are used to encode the relative importance among time signal segments. Then, the activation function α is adopted to map the obtained feature vector to a fixed weight range. Through the reshape operation, the final time weight vector $\hat{q} \in \mathfrak{R}^{L\times 1}$ is output. The value of a point in $\hat{q}$ represents the importance of the corresponding time signal segments. Finally, $\hat{q}$ is used to enhance the meaningful signal segment features in M, as shown in Eq. (4).

$$N_t = M \otimes \hat{q} = [m^1\hat{q}_1, m^2\hat{q}_2, \cdots, m^L\hat{q}_L] \tag{4}$$

where $\hat{q}_j$ is the jth element of $\hat{q}$.

E. MIXED DOMAIN ATTENTION

The time domain attention module aims to extract signal components that are more likely to be related to fault events, such as signal impulses. On the other hand, the channel domain attention module focuses on extracting frequency components that are associated with faults and can capture relevant information even in the presence of variations due to changes in load and speed. Therefore, RMAM performs the channel domain attention and the time domain attention in parallel and then combines the optimized features. This mitigates the impact of intra-class variability and inter-class similarity of signals on the diagnostic performance of the network, resulting in improved performance in bearing fault diagnosis tasks.

The structure of RMAM is shown in Fig. 1(c). It shows that, after two convolution modules, the output features are input to two attention branches to enhance the network's learning ability for meaningful features in both the channel and the time domains. Finally, the residual learning idea is introduced to reduce the difficulty of network training. In most attention-based studies [35,36], the Sigmoid function is often used as the activation function α because of its good performance in most cases. However, α is a very important hyperparameter for an attention module, which identifies which kind of weight vectors we can obtain. Different activation functions α will bring different weight vectors, and thus, the network will show different diagnostic performances. Therefore, this paper discusses and demonstrates seven classical activation functions, which are applied to the proposed RMAM to generate different weight vectors. These activation functions are displayed in Table I.

Table I. Seven classical activation functions

Activation function	Equation	Activation function	Equation
Sigmoid	$S(z) = \frac{1}{(1+e^{-z})}$	ReLU	$f(z) = max(0,z)$
Softplus	$\tau(z) = log\ log(\ 1 + e^{z})$	Softsign	$\tau(z) = \frac{z}{(1+}\|z\|)$
Leaky ReLU	$f(z) = \{z, if\ z > 0\ \lambda z, if\ z \leq 0$	ELU	$f(z) = \{z, if\ z > 0\lambda(e^{z} - 1), if\ z \leq 0$
Tanh		$tanh(z) = \frac{(e^{z}-e^{-z})}{(e^{z}+e^{-z})}$	

III. EXPERIMENTAL VERIFICATION

In this section, the parameters of the proposed RMAM-based CNN architecture are elaborated in detail. In addition, the influence of different activation functions on the model performance will be discussed, and the effectiveness of mixed domain attention will be demonstrated on a motor bearing dataset and an HSA bearing dataset. At last, the proposed method will be compared with existing deep-learning-based fault diagnosis methods. Figure 2 presents a detailed flowchart that comprehensively illustrates the diagnosis procedure. The process consists of data normalization and sliding segmentation, training set and test set partitioning using four-fold cross-validation, the RMA-CNN model training on the training set, and subsequently the trained model evaluation on the test set.

A. RMAM-BASED CNN ARCHITECTURE

The proposed RMA-CNN is a universal and flexible end-to-end bearing fault diagnosis architecture. By stacking RMAM, we can easily construct an RMA-CNN architecture with any depth. In this experiment, we use a lightweight version (named RMA-CNN-10, which means that there are only ten learnable layers in the network). For simplicity, in the following description, we use RMA-CNN to refer to RMA-CNN-10. The structure and the parameters of RMA-CNN are shown in Table II.

In order to ensure that the input signal sample contains a complete signal period, the input dimension of RMA-CNN-10 is 2048×1. RMA-CNN contains a convolution module and a classification layer, including four RMAMs. Each convolution module consists of a 1D convolution layer, a batch normalization layer, and a ReLU activation function. To obtain features on longer signal segments, we use the wider convolution kernels in the first and the second RMAMs, which are 12×1 and 6×1, respectively. The number of channels gradually increases from 16 to 256. We use Maxpooling technology to reduce feature dimensions while retaining valuable information. In the classification stage, a GAP layer is used, and a fully connected layer with a Softmax function is adopted to give final diagnostic results. In addition, we also constructed a Pure-CNN as a comparison method for our method. Pure-CNN also has ten learnable layers, including nine convolutional layers and a

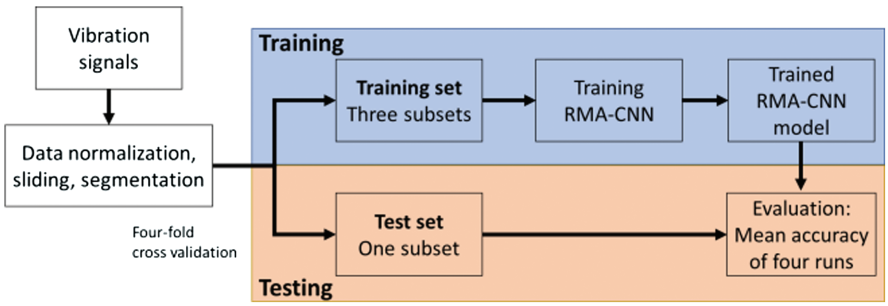


Fig. 2. The diagnosis flowchart.

Table II. Parameters and structure of RMA-CNN

Layer	Type	Kernel/Channel	Stride/Padding	Output
1	RMAM	$12 \times 1/16$	1/yes	2048×16
2	Pooling	–	4/–	512×16
3	RMAM	$6 \times 1/32$	1/yes	512×32
4	Pooling	–	4/–	128×32
5	RMAM	$3 \times 1/64$	1/yes	128×64
6	Pooling	–	2/–	64×64
7	RMAM	$3 \times 1/128$	1/yes	64×128
8	Pooling	–	2/–	32×128
9	Convolution	$3 \times 1/256$	1/yes	32×256
10	Pooling	–	2/–	16×256
9		Global Average Pooling		256
10		Softmax		10

classification layer. The convolution kernel and other parameters are completely consistent with the RMA-CNN.

B. COMPUTATIONAL PARAMETERS

In order to obtain more samples to optimize the parameters of the network, we use the sliding segmentation technique [18] to increase the number of samples, which is a simple and fast method often used. The proposed method is written in Python 3.6 with the deep learning framework Keras and runs on Ubuntu 16.04 with a GTX 2080 GPU. In order to make the training process of the network more stable, we perform a z-score normalization on all samples.

During the training process, the Adam optimizer is used to optimize the network parameters with a learning rate of 0.0001 and a batch size equal to 96. Accuracy is used to evaluate the network's performance. This metric is widely used to evaluate the performance of various classification algorithms and is defined as Eq. (5).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \quad (5)$$

where FP, FN, TN, and TP refer, respectively, to the number of false positive samples, false negative samples, true negative samples, and true positive samples.

To simulate the real operation condition of the bearing and explore the anti-noise performance of our method, we add extra Gaussian white noise to raw signals. The definition of SNR_{dB} is shown as:

$$SNR_{dB} = 10log_{10}(P_{signal}/P_{noise}) \quad (6)$$

where P_{signal} and P_{noise} are, respectively, the power of the signal and the noise.

In this study, we compare the proposed RMA-CNN with the following five deep learning algorithms. First, we compare it with the multi-attention 1D-CNN (MA1DCNN) [35]. In this work, Wang et al. [35] proposed a joint attention module, constructed the MA1DCNN for the fault diagnosis of wheelset bearings, and achieved quite good results. Secondly, we compare RMA-CNN with the deep learning algorithm based on GRU. GRU is an improved version of LSTM with better performance and faster training speed. The GRU architecture has two layers of GRU units. The length of time steps is 64, and the dimension of the input size is 32. In addition, we compare RMA-CNN with three excellent CNN-based fault diagnosis methods. They are named WDCNN [18], ResCNN [28], and Wen-CNN [29]. WDCNN and ResCNN are typical 1D CNNs, which use wide convolution kernels and residual network structures, respectively. Wen-CNN is a 2D convolution network structure, which first converts a 1D signal into a 2D image, and then uses a 2D network to learn fault-related features.

C. CASE 1: MOTOR BEARING FAULT DIAGNOSIS

1) DATA DESCRIPTION. Firstly, the motor bearing dataset of CWRU [2] is adopted to verify the effectiveness of the proposed method. The dataset contains four types of health conditions, which are healthy, outer race fault, inner race fault, and ball fault. Each fault condition contains three levels of fault severity, with faults ranging in diameter from 7 to 21 mil (0.18–0.71 mm) which were seeded on the drive-end bearings. These bearings were then run at a constant speed (approximate motor speeds of 1797–1720 rpm) for motor loads of 0–3 horsepower. We consider the different fault severity degrees as independent health conditions; therefore, this dataset contains 10 types of health conditions in total. The detailed fault information of the experimental bearings is shown in Table III, and the labels are respectively C1, C2, C3, ..., and C10.

Table III. Description of the CWRU bearing dataset information [2]

Fault location	Fault size (mil)	Load (hp)	Label
None	0	0,1,2,3	C1
Ball fault	7	0,1,2,3	C2
Ball fault	14	0,1,2,3	C3
Ball fault	21	0,1,2,3	C4
Inner race fault	7	0,1,2,3	C5
Inner race fault	14	0,1,2,3	C6
Inner race fault	21	0,1,2,3	C7
Outer race fault	7	0,1,2,3	C8
Outer race fault	14	0,1,2,3	C9
Outer race fault	21	0,1,2,3	C10

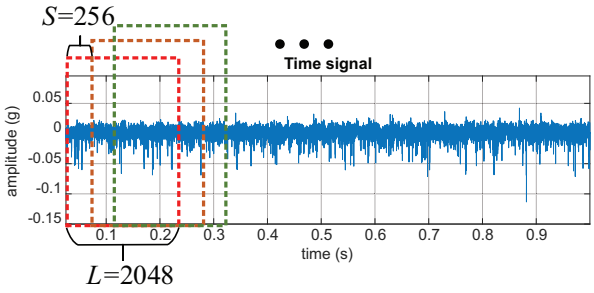


Fig. 3. The signal sliding segmentation description.

To increase the number of samples, a sliding segmentation strategy is employed on the original vibrational signals, as illustrated in Fig. 3. With a stride size of 256, each sample is set to a length of 2048, ensuring that each sample contains at least one complete rotation. This approach results in a total of 106,024 samples.

To evaluate the model's performance more comprehensively and reliably, the four-fold cross-validation method is used. Using multiple folds helps reduce the impact of data variability and noise, improving the accuracy and reliability of the evaluation. All samples are randomly divided into four subsets of equal size. Each of the four subsets is treated as a test set and the remaining three subsets as a training set. The average accuracy across all the test sets was recorded as the final accuracy.

2) DISCUSSION ON THE SELECTION OF ACTIVATION FUNCTIONS. The activation function in RMAM affects the generation of weight vectors, which in turn affects the diagnostic performance of the network. In order to discuss the influence of the activation function on the attention module, we introduce seven common activation functions in RMAM and verify the diagnostic performance of these methods through experiments. These methods are named RMA-CNN-Tanh, RMA-CNN-Sigmoid, RMA-CNN-ReLU, RMA-CNN-Leaky ReLU, RMA-CNN-ELU, RMA-CNN-Softplus, and RMA-CNN-Softsign. In addition,

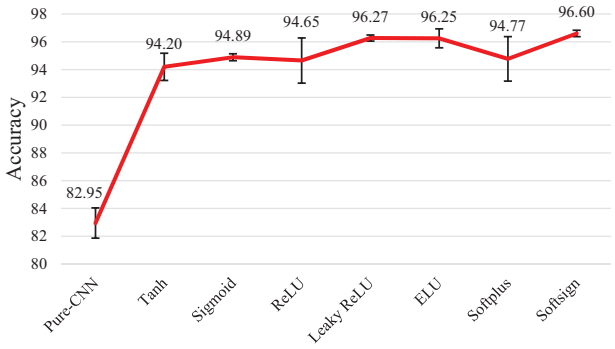


Fig. 4. The experimental results of the network with different activation functions (SNR = −6dB).

in order to more intuitively show the performance improvement brought by the attention mechanism with different activation functions, we also give the experimental results of Pure-CNN. Figure 4 shows the mean accuracy of the four-fold cross-validation for these methods in the case of SNR = −6dB.

It can be observed that, no matter which activation function is used, the network with RMAM is significantly better than the Pure-CNN. For example, Pure-CNN achieves only 82.95% diagnostic accuracy, while RMA-CNN-Sigmoid has a diagnostic accuracy of 94.89%, which is 11.94% higher than the Pure-CNN. This strongly shows that the RMAM can effectively improve the diagnostic performance of the network. This RMAM module can enhance the useful information in the time domain and the channel domain, and suppress useless information, thereby improving the network's ability to learn meaningful features and resist useless information interference (such as noise). In addition, it can be found that different activation functions have an impact on the performance of RMA-CNN. For example, RMA-CNN obtained a diagnostic performance of 96.65% when using Softsign and 94.20% when using Tanh. It is worth noting that Softsign is an improved version of Tanh. Softsign has a flat curve and a slower descending derivative, which can provide a more efficient learning ability than Tanh. In addition, we found that RMA-CNN also has good diagnostic results when using Leaky ReLU and ELU. Due to the excellent performance of Softsign in RMA-CNN, it is adopted in the following experiments.

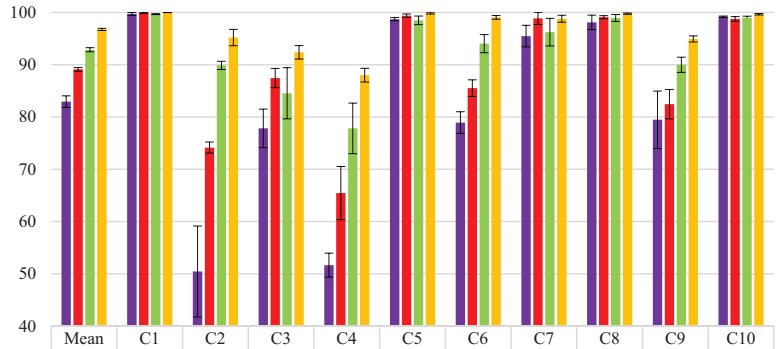
Overall, there are no significant differences in diagnostic performance among the different activation functions in the attention module. This indicates that the choice of activation function may not be the most critical factor influencing the network's performance, and other factors such as model architecture or data representation may have a more significant impact.

However, there is no universally optimal activation function for all applications. Instead, it is important to choose an activation function according to the specific needs of a specific application. The choice of the activation function should be guided by the particular problem and the characteristics of the data being used. For example, ReLU has become a popular choice due to its simplicity and computational efficiency, but it may not be suitable for all applications because of its "dying" tendency when the input is negative. On the other hand, SELU has been proven effective for deep neural networks, but it may require more computational resources and careful initialization [37].

Therefore, the choice of the activation function should be based on the trade-off between computational efficiency and performance, as well as on the specific characteristics of the application and the dataset used. Our research provides some insights into the effectiveness of different activation functions in domain attention modules, but further studies are required to explore their effectiveness for other applications and datasets. Experimenting with different activation functions in the context of a specific problem and selecting the one that provides the best performance is recommended.

3) EFFECTIVENESS OF MIXED DOMAIN ATTENTION LEARNING

To further explore the effectiveness of the proposed attention module, we construct two architectures, named RMA-CNN-C (only includes the channel domain attention learning) and RMA-CNN-T (only includes the time domain attention learning). Under an SNR equal to −6dB, two networks are compared: Pure-CNN and RMA-CNN. Figure 5 shows the average accuracy of the four-fold cross-validation method for each category and the average accuracy of all fault classes.



	Mean	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Pure-CNN	82.95	99.74	50.43	77.83	51.65	98.72	78.93	95.48	98.11	79.48	99.19
RMA-CNN-C	89.11	99.94	74.15	87.45	65.44	99.38	85.52	98.87	99.09	82.46	98.76
RMA-CNN-T	92.86	99.67	89.88	84.54	77.82	98.49	94.03	96.23	98.96	89.99	98.98
RMA-CNN	96.76	99.98	95.20	92.37	88.01	99.87	99.06	98.81	99.77	94.93	99.66

Fig. 5. The results for each fault category and their average accuracy (SNR = −6 dB).

Both domain attention can effectively improve the fault diagnosis performance of the network. The average accuracy of RMA-CNN-C is 89.11%, which is an increase of 6.16% compared with the Pure-CNN, while that of RMA-CNN-T is 92.86%, which is 9.91% higher than the Pure-CNN. Additionally, RMA-CNN-T has better performance than the RMA-CNN-C, due to the particularity of the vibration signal, for example, the vibration signal of the faulty bearing has a periodic short-time impulse signal segment. The time domain attention mechanism can make the network focus on the learning of the fault impulse signal segment, so as to obtain more fault-related features. In addition, RMA-CNN achieved an average accuracy of 96.76%, which is significantly better than that of the RMA-CNN-C and the RMA-CNN-T. This shows that the time domain attention and the channel domain attention can complement and promote each other, and jointly improve the performance of the network. From the perspective of diagnosis results in each category, RMA-CNN obtains the best results among all methods, with the exception of C7. The diagnosis results of the RMA-CNN-T and the RMA-CNN-C are better than the Pure-CNN in most categories.

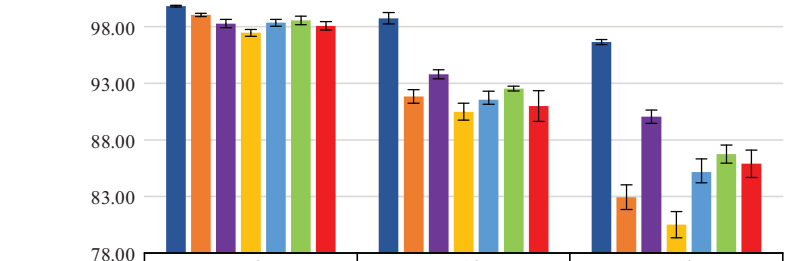
Particularly, our method has the most significant improvement in the diagnosis of the ball fault (i.e., C2, C3, and C4). The accuracy of C2 by the RMA-CNN is 95.20%, which is 44.77% higher than that of the Pure-CNN. These experimental results well illustrate that the proposed method obtains more meaningful features, thereby alleviating the problems of intra-class variability and inter-class similarity.

4) COMPARISON WITH EXISTING DEEP-LEARNING-BASED METHODS. To explore the performance of RMA-CNN under different noise conditions and verify its superiority, we compare the RMA-CNN with five deep learning algorithms and the Pure-CNN under three noise conditions (0 dB, −4 dB, −6 dB). The average accuracy of the four-fold cross-validation of these methods is shown in Fig. 6.

The diagnostic performance of the RMA-CNN is better than the five comparison methods under the three noise conditions. When SNR = −6dB, RMA-CNN-10 obtains a diagnostic accuracy of 96.65%, which is 6.59% higher than that of the MA1DCNN. This shows that our attention method is significantly better than that used in the MA1DCNN. On the other hand, methods with attention mechanisms, such as RMA-CNN and MA1DCNN, are better than other methods. Despite the varying speed and load conditions that result in different vibration signal distributions for the same fault class in this dataset, the proposed model demonstrates promising results, indicating its effectiveness in addressing the limitations of intra-class variability and inter-class similarity of signals caused by varying conditions. As a result, the RMA-CNN is capable of accurately classifying faults even in the presence of different operating conditions. This highlights the applicability and the effectiveness of the attention mechanism in the field of fault diagnosis.

In addition, we can see that as the noise increases, the diagnostic performance of these deep learning models gradually decreases. For example, the diagnostic accuracy of the WDCNN dropped from 98.37% to 85.17%. The diagnostic accuracy of the Wen-CNN dropped from 98.56% to 86.75%. It is worth mentioning that the accuracy of RMA-CNN has only dropped by 3.17%. This shows that the RMA-CNN has good anti-noise performance, and it can extract useful features from noisy signals.

To analyze the recall and the accuracy of the proposed method more clearly, we give the confusion matrix of the RMA-CNN and the Pure-CNN when the SNR is −6 dB. These two confusion matrices are shown in Figs. 7 and 8. The diagonal is the number of accurate diagnoses for each category. The bottom row shows the precision of each category while the rightmost column represents the number of testing samples of each category. It can be seen from Table III that the C2–C4 classes have the same fault location, but the fault severity is different. These three categories have very serious inter-class similarities, resulting in poor classification results. Comparing Figs. 7 and 8, we find that the proposed RMA-CNN alleviates this problem very well and improves the diagnostic performance of each category. To further demonstrate the scalability of the proposed method, we constructed a new network architecture named RMA-CNN-18. The basic architecture of the RMA-CNN-18 is consistent with the RMA-CNN. The difference is that the RMA-CNN-18 has 18 learnable layers. The experimental results of the RMA-CNN and the RMA-CNN-18 are shown in Table IV. We found that



	0dB	−4dB	−6dB
RMA-CNN	**99.82±0.08**	**98.75**±0.50	**96.65±0.23**
Pure-CNN	99.04±0.15	91.84±0.60	82.95±1.09
MA1DCNN	98.28±0.36	93.80±0.41	90.06±0.59
GRU	97.46±0.29	90.49±0.75	80.51±1.15
WDCNN	98.37±0.31	91.55±0.39	85.17±0.96
Wen-CNN	98.56±0.37	92.54±0.22	86.75±0.80
ResNN	98.08±0.36	90.99±1.36	85.89±1.22

Fig. 6. The experimental results of RMA-CNN, MA1DCNN, GRU, WDCNN, Wen-CNN, and ResCNN under three types of noise conditions (0 dB, −4 dB, −6 dB).

True Label \ Predicted Label	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Recall	Test Number
C1	1799	1	0	0	0	0	0	0	0	0	**99.94%**	1800
C2	0	2543	3	115	6	5	4	5	19	0	**94.19%**	2700
C3	0	29	2431	76	0	14	23	0	127	0	**90.04%**	2700
C4	4	99	52	2319	0	30	22	1	173	0	**85.89%**	2700
C5	0	0	0	0	2695	1	4	0	0	0	**99.81%**	2700
C6	0	5	0	2	2	2684	7	0	0	0	**99.41%**	2700
C7	0	0	5	0	1	9	2677	1	1	6	**99.15%**	2700
C8	0	0	3	7	0	16	1	2672	0	1	**98.96%**	2700
C9	0	8	22	74	0	7	2	0	2587	0	**95.81%**	2700
C10	0	0	0	0	0	2	0	6	0	2692	**99.70%**	2700
PRE	**99.78%**	**94.71%**	**96.62%**	**89.43%**	**99.67%**	**96.97%**	**97.70%**	**99.52%**	**88.99%**	**99.74%**	—	26100

Fig. 7. The confusion matrix of the RMA-CNN on the CWRU dataset (SNR = −6dB).

True Label \ Predicted Label	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Recall	Test Number
C1	1796	3	0	0	0	1	0	0	0	0	**99.78%**	1800
C2	8	1400	192	580	20	216	19	16	248	1	**51.85%**	2700
C3	0	69	2206	85	3	129	101	18	81	8	**81.70%**	2700
C4	21	639	144	1334	3	301	72	48	138	0	**49.41%**	2700
C5	0	5	0	0	2678	0	7	0	10	0	**99.19%**	2700
C6	0	68	118	161	17	2175	12	5	98	46	**80.56%**	2700
C7	1	1	11	18	6	22	2622	4	1	14	**97.11%**	2700
C8	0	6	2	46	2	10	14	2619	1	0	**97.00%**	2700
C9	6	244	129	225	12	135	2	2	1941	4	**71.89%**	2700
C10	0	0	0	0	0	0	2	0	0	2698	**99.93%**	2700
PRE	**98.03%**	**57.49%**	**78.73%**	**54.47%**	**97.70%**	**72.77%**	**91.97%**	**96.57%**	**77.08%**	**97.36%**	—	26100

Fig. 8. The confusion matrix of the Pure-CNN on the CWRU dataset (SNR = −6dB).

Table IV. The results of RMA-CNN and RMA-CNN-18 on the CWRU dataset

Method	0dB	−4dB	−6dB
RMA-CNN	99.82 ± 0.08	98.75 ± 0.50	96.65 ± 0.23
RMA-CNN-18	**99.90 ± 0.01**	**98.86 ± 0.21**	**97.01 ± 0.22**

the RMA-CNN-18 has better diagnostic performance than the RMA-CNN, due to the fact that RMA-CNN-18 has stronger learning ability. This shows that the proposed method is a universal and flexible end-to-end bearing fault diagnosis architecture, which can be simply modified to be applied to different situations and applications.

D. CASE 2: HIGH-SPEED AERONAUTICAL BEARINGS FAULT DIAGNOSIS

1) DATA DESCRIPTION. Besides the CWRU dataset, we also considered the Politecnico di Torino rolling bearing dataset to verify the effectiveness of the proposed method [1]. The test rig is shown in Fig. 9(a), which consists of a high-speed spindle, a sledge, a load cell, and a lubrication line. Two accelerometers are located at the two positions of the structure shown in Fig. 9(b).

Bearings with different types and dimensions of damage are mounted in position B1. Local faults on the inner race or on a roller were produced using a Rockwell tool. We consider the different fault severities as independent health conditions; thus, there are seven bearing health conditions H1–H7, as shown in Table V. For every bearing, the

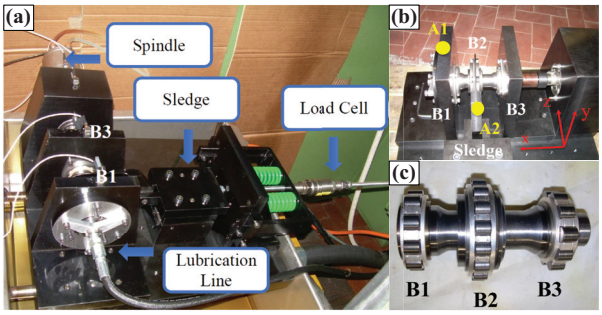


Fig. 9. The high-speed aeronautical bearings test rig [1]: (a) general view of the test rig; (b) positions of the accelerometers and the reference system and (c) the shaft with its three roller bearings.

Table V. Description of the HSA bearing dataset information

Defect	Dimension	Load	Speed	Label
No defect	–	0–1800N	100–500 Hz	H1
Diameter of an indentation on the inner ring	450 μm	0–1800N	100–500 Hz	H2
Diameter of an indentation on the inner ring	250 μm	0–1800N	100–500 Hz	H3
Diameter of an indentation on the inner ring	150 μm	0–1800N	100–500 Hz	H4
Diameter of an indentation on a roller	450 μm	0–1800N	100–500 Hz	H5
Diameter of an indentation on a roller	250 μm	0–1800N	100–500 Hz	H6
Diameter of an indentation on a roller	150 μm	0–1800N	100–500 Hz	H7

Table VI. List of the tested load and speed cases

Nominal load (N)	Nominal speed (Hz)				
0	100	200	300	400	500
1000	100	200	300	400	500
1400	100	200	300	400	/
1800	100	200	300	/	/

operating load changes from 0 N to 1800 N (0, 1000, 1400, and 1800 N), and the operating speed also changes from 100 Hz to 500 Hz (100, 200, 300, 400, 500 Hz), as shown in Table VI. Similarly, the sliding segmentation strategy is applied with a stride of 256 and a sample length of 2048, resulting in a total of 118,286 samples. We then used the four-fold cross-validation method to evaluate the performance of our model. The data were randomly divided into four subsets of equal size, with each subset serving as a test set while the remaining three were used as training sets. We conducted four runs and recorded the average accuracy as the final evaluation metric.

2) COMPARISON WITH EXISTING DEEP-LEARNING-BASED METHODS. In order to explore the performance of the RMA-CNN under the high-speed aviation bearing dataset, we conducted experiments on the dataset with three different noise levels. As the fault diagnosis of high-speed aviation bearings is rather difficult, we only consider the cases when the SNR is 0 dB, 4 dB, and 6 dB. In addition, we also report the experimental results of five comparison methods and of the Pure-CNN. The experimental results of these methods are shown in Fig. 10. Consistent with the previous experimental results, the RMA-CNN seems to achieve better performance compared to the other comparison methods in all the three noise environments.

First, the diagnostic accuracy of the RMA-CNN is 7.65 %, 3.34 %, and 2.87 % higher than that of the Pure-CNN under the three noise conditions, respectively. This demonstrates the effectiveness of the proposed RMAM. Secondly, the performance of the RMA-CNN is also better than the MA1DCNN, indicating that the proposed attention method has a stronger feature optimization ability. Thirdly, methods with the attention mechanism have better performance than other methods, which once again illustrates the effectiveness of the attention mechanism in the bearing fault diagnosis task. Similarly, as the noise increases, the performance of these deep learning models gradually decreases. RMA-CNN shows better noise immunity, which also shows that it can extract useful fault-related information from noisy (aeronautical bearing) data.

In order to further analyze the performance of the proposed method in solving the problem of intra-class variability and inter-class similarity, two confusion matrices are shown. Figure 11 shows the confusion matrix of the RMA-CNN on the HSA bearings dataset (SNR = 6 dB). Figure 12 shows the confusion matrix of the Pure-CNN on

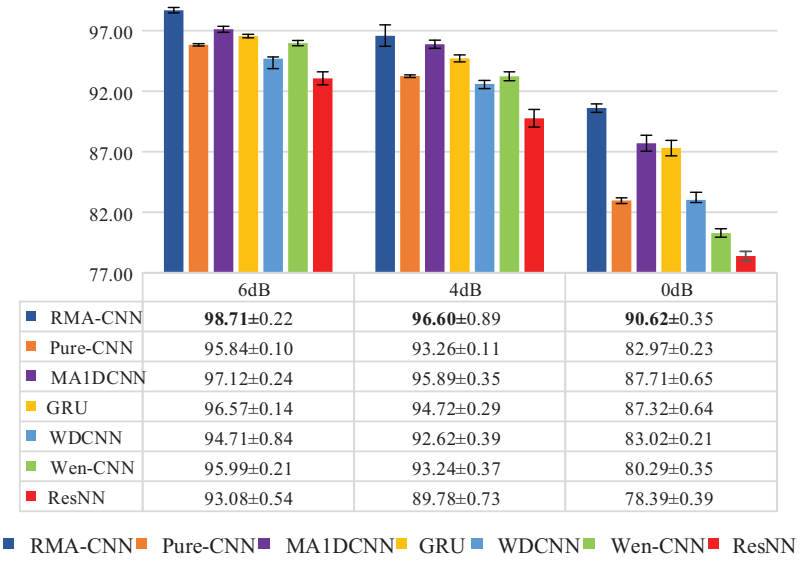


Fig. 10. The experimental results of RMA-CNN, MA1DCNN, GRU, WDCNN, Wen-CNN, and ResCNN under three SNRs.

True Label \ Predicted Label	H1	H2	H3	H4	H5	H6	H7	Recall	Test Number
H1	4141	7	6	42	0	3	0	**98.62%**	4199
H2	27	4141	0	23	0	6	2	**98.62%**	4199
H3	1	2	4151	14	0	29	2	**98.86%**	4199
H4	46	27	3	4069	0	53	1	**96.90%**	4199
H5	0	5	22	16	4155	1	0	**98.95%**	4199
H6	2	1	72	62	0	4061	1	**96.71%**	4199
H7	2	1	5	9	0	2	4180	**99.55%**	4199
PRE	**98.15%**	**98.97%**	**97.46%**	**96.08%**	**100%**	**97.74%**	**99.86%**	—	29393

Fig. 11. The confusion matrix of the RMA-CNN on the HSA bearings dataset (SNR = 6 dB).

True Label \ Predicted Label	H1	H2	H3	H4	H5	H6	H7	Recall	Test Number
H1	3888	55	9	124	0	57	66	**92.59%**	4199
H2	5	4013	1	35	31	13	101	**95.57%**	4199
H3	7	0	4112	11	0	59	10	**97.93%**	4199
H4	69	94	8	3845	4	95	84	**91.57%**	4199
H5	0	1	0	2	4195	0	1	**99.90%**	4199
H6	17	8	33	103	6	4004	28	**95.36%**	4199
H7	13	125	9	24	0	29	3999	**95.24%**	4199
PRE	**97.22%**	**93.41%**	**98.56%**	**92.79%**	**99.03%**	**94.06%**	**93.24%**	—	29393

Fig. 12. The confusion matrix of Pure-CNN on HSA bearings dataset (SNR = 6 dB).

Table VII. The results of RMA-CNN and RMA-CNN-18 on HSA bearings dataset

Method	0 dB	4 dB	6 dB
RMA-CNN	90.62 ± 0.35	96.60 ± 0.89	98.71 ± 0.22
RMA-CNN-18	**91.91 ± 0.48**	**97.54 ± 0.28**	**98.88 ± 0.10**

the HSA bearings dataset (SNR = 6 dB). The diagonal is the number of the accurate diagnoses for every category. The bottom row shows the precision of every category. The rightmost column represents the number of testing samples for every category. From Table V, H2–H4 classes have the same fault location, but the fault severity is different. The same stands also for H5–H7. As shown in Fig. 12, Pure-CNN cannot clearly solve the problem of intra-class variability and inter-class similarity. More specifically, there are many mispredictions between categories H2–H4. In addition, there are many mispredictions between H2, H4, and H6, H7. As shown in Fig. 11, the proposed RMA-CNN alleviates this problem and effectively improves the performance of each category. Similarly, we explored the performance of the RMA-CNN-18 on the HSA bearings dataset. The results are shown in Table VII. The RMA-CNN-18 also has better performance than the RMA-CNN, which once again illustrates the flexibility and scalability of the proposed method.

IV. INTERPRETABILITY OF ATTENTION MECHANISM

In this section, we explore the interpretability of the attention mechanism on the CWRU dataset. We creatively introduce time-frequency analysis technology to deeply analyze the feature learning mechanism of attention and discuss the interpretability of CNNs in the field of mechanical fault diagnosis.

A. INTERPRETABILITY OF TIME DOMAIN ATTENTION

To deeply understand the learning mechanism of time domain attention, we conduct a detailed visualization and analysis of the weight vector of the time domain attention. Figure 13 shows the visualization results of four signal samples (A, B, C, and D). These four samples belong to two categories: C7 (A, B) and C10 (C, D). The vibration signal of the bearing is a time domain signal with periodicity and time correlation information. When a local fault occurs to a bearing, each time a rolling element passes over the defect, a periodic impulse is generated exciting natural frequencies of the structure. The fault-impulsive segment of the vibration signal contains rich fault-related features. The design goal of the time domain attention mechanism is to make the network focus on learning meaningful signal segment

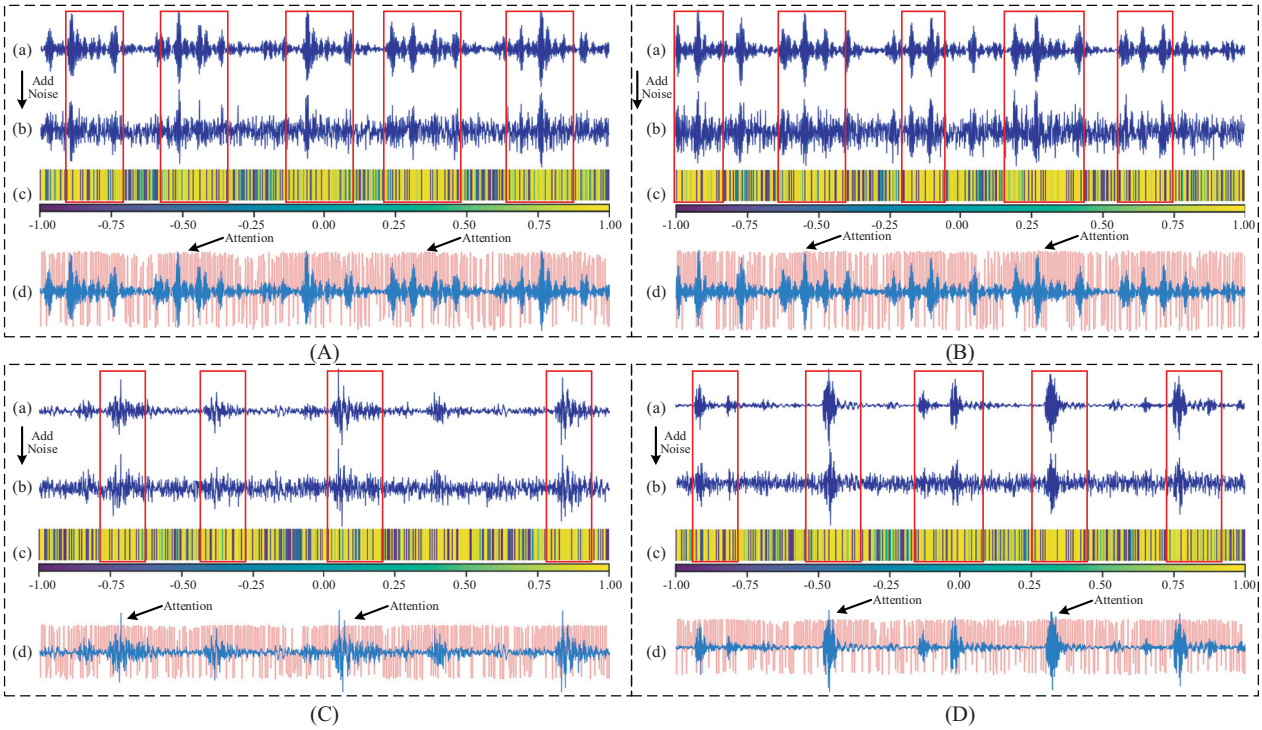


Fig. 13. The visualization result at the time domain for four vibration signal samples (A–D). (a) The vibration signal (no noise added); (b) the noisy vibration signal (SNR = 0 dB); (c) the visualization result of the time weight vector, where each small grid represents a weight and the color represents the weight value; (d) the time weight vector values (pink curve) and the signal without noise (blue curve).

features. If we make the network pay more attention to those periodic impulsive signal segments, it will be helpful for the network to obtain more meaningful information. A similar phenomenon is observed in Fig. 13. The time weight vector shows a higher weight in the area of the impulse signal segmentation and a lower weight in the non-impulse signal segments. It is worth noting that this phenomenon is observed when the SNR is 0 dB. The energy of the noise is the same as that of the original signal, and it can be seen that the noise has basically overwhelmed the normal waveform of the original signal. However, the proposed attention module has good anti-noise performance. It is also found that the attention mechanism tries to filter meaningful information more precisely. For example, there are also a few cases with lower weights in the impulse signal segment. The experimental results from Fig. 5 show that the performance of the RMA-CNN-T with time domain attention has been greatly improved compared with the Pure-CNN, which strongly proves the effectiveness of this learning mechanism. In addition, Fig. 6 demonstrates that the proposed method has very good anti-noise performance, which is consistent with our visualization results. More examples of time domain attention visualization, including the HAS bearings data case, can be found in the Appendix.

B. INTERPRETABILITY OF CHANNEL DOMAIN ATTENTION

As explained in Section 2.3, a convolutional layer comprises multiple filters. These filters learn various signal features, but not all of the learned features are useful. The purpose of channel domain attention is to enable the network to concentrate on acquiring meaningful channel features. To thoroughly understand the learning mechanism of channel domain attention, we first carry out an analysis of the channel weight vectors.

Figure 14 presents the visualization of four channel weight vectors (W1–W4) within the RMA-CNN, where each small grid represents a weight. It is observed that the method assigns different weights to distinct channels, indicating its attempt to discern which features are more valuable. Notably, W3 and W4 contain numerous channels with negative weights. This suggests that a considerable amount of irrelevant information exists within the network.

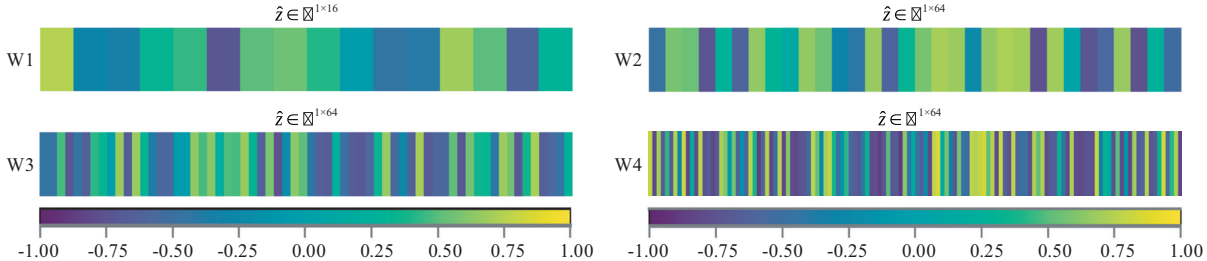


Fig. 14. The visualization of the channel weight vectors of the RMA-CNN, where each small grid represents a weight value.

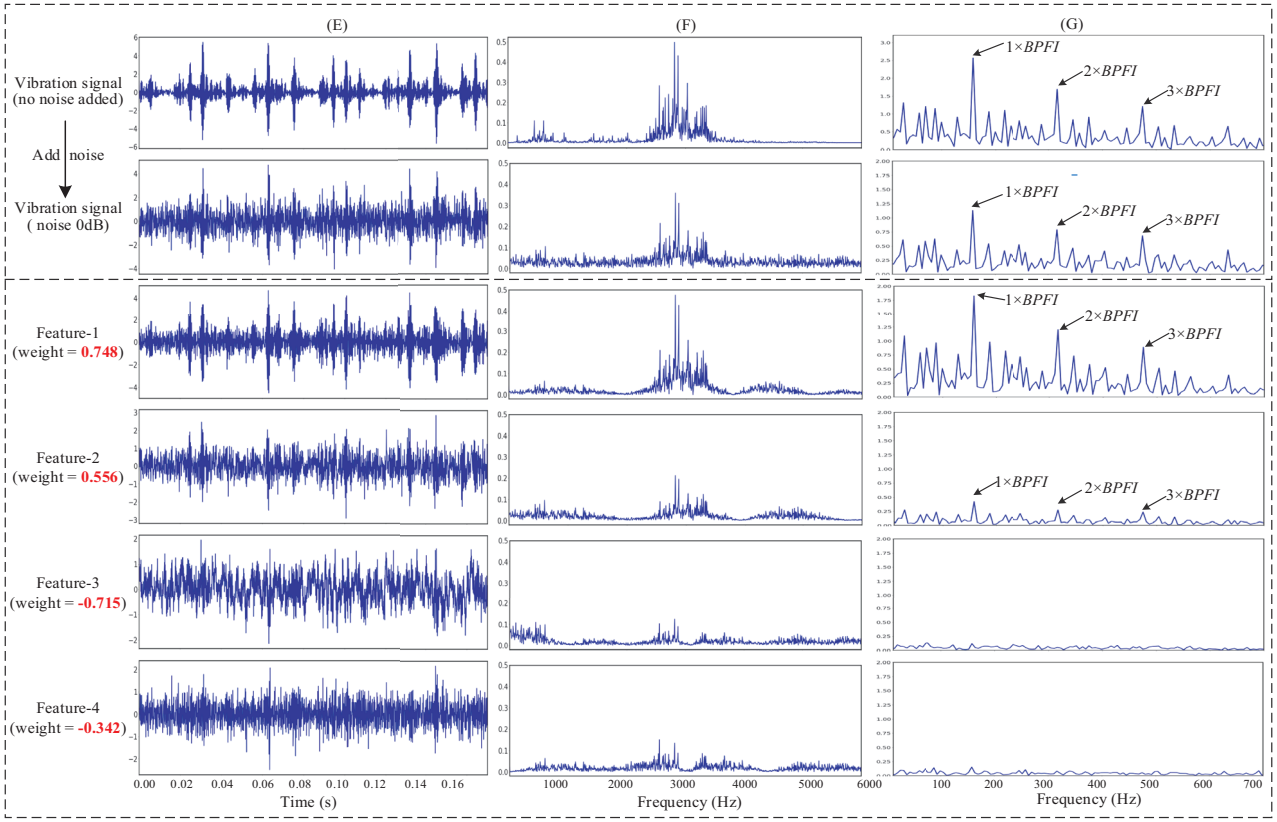


Fig. 15. The visualization results of the vibration signal (E), its Fourier Spectrum (F) and its square envelope spectrum (G) and the corresponding features (Feature-1 to Feature-4) learned by the convolution layer.

Secondly, we focus on analyzing the first RMAM of CNNs, as it is responsible for extracting basic features from the input data. These features are crucial for building higher-level representations in subsequent layers. Furthermore, CNN typically includes pooling layers to reduce the dimensionality of parameters and avoid overfitting. As the network deepens and multiple pooling layers are introduced, the sampling rate of features decreases, posing difficulties for the attention module's learning process from a time-frequency analysis perspective. Therefore, we further analyze the features learned by different convolutional kernels of the first RMAM to understand the learning mechanism of the attention module.

Figure 15 consists of two parts: the top and the bottom sections. The top section's first row displays the original vibration signal (E), the Fourier spectrum of the original vibration signal (F), and the squared envelope spectrum (G) of (E). The second row shows the vibration signal after adding white Gaussian noise with SNR = 0 dB, along with its Fourier spectrum and the corresponding squared envelope spectrum. The bottom section illustrates the feature signals learned by some convolutional kernels when the signal with SNR = 0 dB is used as network input, for example, Feature-1, Feature-2, Feature-3, and Feature-4. The channel attention module assigns different weight values to these four convolutional kernels, for example, 0.748 to Feature-1, 0.556 to Feature-2, −0.715 to Feature-3, and −0.342 to Feature-4. Therefore, among these four convolutional kernel feature maps, Feature-1 is the most relevant feature selected by the channel attention module.

Let us then observe the spectrum of Feature-1 (denoted as A) and the spectrum of the network input signal (denoted as B). Both A and B have the same y-axis, with an amplitude range [0, 0.5]. It can be seen that A not only retains the 3000 Hz high-frequency impulse signal and its surroundings from B but also has a higher amplitude than B, thus enhancing this impulse signal. Meanwhile, in the low-frequency range (0–2000 Hz), A has a lower amplitude than B, achieving some noise suppression. The attention module assigns a higher weight value (0.748) to this convolutional kernel feature, thus highlighting the fault-related impulse signal. In the square envelope spectrum, BPFI is clearly visible for feature-1, which shows that while removing some noise, the fault-related features are well preserved. Therefore, our method gives higher weight to these features.

Secondly, as shown in the spectra of Feature-3 and Feature-4, the high-frequency impulse signals are completely submerged in the noise, making it impossible to clearly identify the fault characteristic frequencies BPFI in the squared envelope spectrum. This indicates that the features learned by these two convolutional kernels are mainly noise signals unrelated to the fault. The attention module assigns lower weight values to these two convolutional kernels (such as −0.715 and −0.342), thereby suppressing the unrelated noise signals learned by these kernels.

Therefore, the attention module can focus on fault-related convolutional kernel features while suppressing convolutional kernel features related to noise signals. Moreover, the performance of the RMA-CNN-C in Fig. 5 is improved compared to the Pure-CNN, which also proves the effectiveness of this learning mechanism. Furthermore, Fig. 6 shows that the proposed method has good anti-noise performance, which is consistent with our visualization

results. In addition, all channel feature signals and channel attention weights of the first RMAM are displayed in the additional material, and it can be seen that the attention mechanism tries to select meaningful features more precisely. On the other hand, it gives high weight to a few channels with unobvious fault characteristic frequency.

C. ADVANTAGES OF INTERPRETABILITY WITH AN ATTENTION MECHANISM

It is possible to analyze the weights of the kernels of the last CNN layer and visualize the resulting signals for the highest weight kernels, as done in the Gradient-weighted Class Activation Mapping (Grad-CAM) method [38]. However, the interpretability provided by the attention mechanism differs from other explainable artificial intelligence (XAI) methods like Grad-CAM in several aspects:

- Direct focus on important signal components: Attention mechanisms help identify important signal components in the input data by assigning higher weights to them, which provides a more straightforward way to interpret which parts of the data are considered most important by the network. In contrast, Grad-CAM analyzes the gradients of the output with respect to the feature maps, which is a more indirect way of understanding the network's decision-making process.
- Easier visualization and interpretation: The attention mechanism can be easier to visualize and interpret compared to Grad-CAM, as it directly highlights the significant signal components without relying on back-propagation-based explanations.
- Comprehensive understanding: Attention mechanisms can provide insights into both time-sequence and channel-wise importance, giving a more comprehensive understanding of the network's focus, while Grad-CAM primarily focuses on the time-sequence importance.
- Integrated interpretability: Attention mechanisms allow for interpretability to be integrated directly into the network architecture, rather than relying on post hoc analysis, such as Grad-CAM. This can lead to a better understanding and trust in the model's decisions during the training process and help improve the network's performance by guiding it to focus on the relevant signal components of the data. Grad-CAM and other post hoc analysis methods are applied after the model is trained, which may not provide the same level of understanding during the training process.

Therefore, compared to other XAI methods like Grad-CAM, the proposed methodology enhances interpretability by incorporating an attention mechanism into the network architecture, which provides clearer visualizations and more comprehensive insights into the important signal components of the input data. On the other hand at the same time, attention mechanisms add complexity to the network architecture, which may increase the computational requirements and training time.

V. CONCLUSIONS

This paper explores a solution to the problems of intra-class variability and inter-class similarity in the field of mechanical fault diagnosis and designs an attention module RMAM. It constructs the attention feature optimization mechanism from the time domain and the channel domain, and can highlight fault-related features from noisy signals. Based on the RMAM, a flexible and universal framework named RMA-CNN is proposed. Experiments on two datasets, the HAS bearing dataset and the CWRU dataset, show that the RMA-CNN has a very competitive fault diagnosis performance, which is significantly better than the performance of five comparison algorithms. In addition, the proposed channel domain attention and time domain attention can effectively improve the feature learning ability of the network to obtain better results. More importantly, we analyzed in detail the internal mechanism of feature learning of the proposed attention mechanism focusing on the interpretability of the CNN network, applied in the fault diagnosis field. In the future, we will continue exploring the theory of the attention mechanism focusing on the interpretability of the network.

Acknowledgments

The authors would like to acknowledge the support of the China Scholarship Council, the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" Program and the Research Foundation–Flanders (FWO) under the ROBUSTIFY research grant no. S006119N.

CONFLICT OF INTEREST STATEMENT

Konstantinos Gryllias is an associate editor for the *Journal of Dynamics, Monitoring and Diagnostics*, and he was not involved in the editorial review or the decision to publish this article. The authors declare that they have no conflict of interest.

References

[1] A.P. Daga, A. Fasana, S. Marchesiello, and L. Garibaldi, "The Politecnico di Torino rolling bearing test rig: description and analysis of open access data," *Mech. Syst. Signal Process.*, vol. 120, pp. 252–273, 2019.

[2] Y. Lei, J. Lin, Z. He, and M.J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mech. Syst. Signal Process.*, vol. 35, no. 1, pp. 108–126, 2013.

[3] M. Kang, J. Kim, J. Kim, A.C.C. Tan, E.Y. Kim, and B. Choi, "Reliable fault diagnosis for low-speed bearings using individually trained support vector machines with Kernel discriminative feature analysis," *IEEE Trans. Power Electron.*, vol. 30, no. 5, pp. 2786–2797, 2015.

[4] P. Baraldi, F. Cannarile, F. Di Maio, and E. Zio, "Hierarchical k-nearest neighbours classification and binary differential evolution for fault diagnostics of automotive bearings operating under variable conditions," *Eng. Appl. Artif. Intell.*, vol. 56, pp. 1–13, 2016.

[5] D. Wang, Y. Zhao, C. Yi, K. Tsui, and J. Lin, "Sparsity guided empirical wavelet transform for fault diagnosis of rolling element bearings," *Mech. Syst. Signal Process.*, vol. 101, pp. 292–308, 2018.

[6] Y. Lv, R. Yuan and G. Song, "Multivariate empirical mode decomposition and its application to fault diagnosis of rolling bearing," *Mech. Syst. Signal Process.*, vol. 81, pp. 219–234, 2016.

[7] B. Xu, F. Zhou, H. Li, B. Yan, and Y. Liu, "Early fault feature extraction of bearings based on Teager energy operator and optimal VMD," *ISA Trans.*, vol. 86, pp. 249–265, 2019.

[8] X. Yan and M. Jia, "A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing," *Neurocomputing*, vol. 313, pp. 47–64, 2018.

[9] K. He, X. Zhang, S. Ren, and S. Jian, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[10] G. Huang, Z. Liu, V.D.M. Laurens, and K.Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, 2017, pp. 2261–2269.

[11] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[12] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification," *J. Syst. Eng. Electron.*, vol. 28, no. 1, pp. 162–169, 2017.

[13] T. de Bruin, K. Verbert, and R. Babuška, "Railway track circuit fault diagnosis using recurrent neural networks," *IEEE Trans. Neural Netw. Learn.*, vol. 28, no. 3, pp. 523–533, 2017.

[14] Y. Xu, X. Li, D. Chen, and H. Li, "Learning rates of regularized regression with multiple Gaussian Kernels for multi-task learning," *IEEE Trans. Neural Netw. Learn.*, vol. 29, no. 11, pp. 5408–5418, 2018.

[15] N. Qin, K. Liang, D. Huang, L. Ma, and A.H. Kemp, "Multiple convolutional recurrent neural networks for fault identification and performance degradation evaluation of high-speed train Bogie," *IEEE Trans. Neural Netw. Learn.*, vol. 31, pp. 1–14, 2020.

[16] H. Wang, Z. Liu, D. Peng, M. Yang, and Y. Qin, "Feature-level attention-guided multitask CNN for fault diagnosis and working conditions identification of rolling bearing," *IEEE Trans. Neural Netw. Learn.*, vol. 33, pp. 1–13, 2021.

[17] D. Peng, Z. Liu, H. Wang, Y. Qin, and L. Jia, "A Novel deeper one-dimensional CNN with residual learning for fault diagnosis of wheelset bearings in high-speed trains," *IEEE Access*, vol. 7, pp. 10278–10293, 2018.

[18] Z. Wei, P. Gaoliang, L. Chuanhao, C. Yuanhang, and Z. Zhujun, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors-Basel*, vol. 17, no. 2, p. 425.

[19] J. Jiao, M. Zhao, J. Lin, and C. Ding, "Deep coupled dense convolutional network with complementary data for intelligent fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9858–9867, 2019.

[20] L. Su, L. Ma, N. Qin, D. Huang, and A.H. Kemp, "Fault diagnosis of high-speed train Bogie by residual-squeeze net," *IEEE Trans. Ind. Inform.*, vol. 15, no. 7, pp. 3856–3863, 2019.

[21] R. Liu, F. Wang, B. Yang, and S.J. Qin, "Multi-scale Kernel based residual convolutional neural network for motor fault diagnosis under non-stationary conditions," *IEEE Trans. Ind. Inform.*, p. 1, 2019.

[22] M. Xia, T. Li, L. Xu, L. Liu, and C.W. de Silva, "Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks," *IEEE/ASME Trans. Mechatron.*, vol. 23, no. 1, pp. 101–110, 2018.

[23] Z. Chen, A. Mauricio, W. Li, and K. Gryllias, "A deep learning method for bearing fault diagnosis based on cyclic spectral coherence and convolutional neural networks," *Mech. Syst. Signal Process.*, vol. 140, p. 106683, 2020.

[24] Z. Chen, K. Gryllias, and W. Li, "Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network," *IEEE Trans. Ind. Inform.*, vol. 16, no. 1, pp. 339–349, 2020.

[25] Z. Chen, K. Gryllias, and W. Li, "Mechanical fault diagnosis using convolutional neural networks and extreme learning machine," *Mech. Syst. Signal Process.*, vol. 133, p. 106272, 2019.

[26] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, "Deep residual shrinkage networks for fault diagnosis," *IEEE Trans. Ind. Inform.*, vol. 16, p. 1, 2019.

[27] D. Peng, H. Wang, Z. Liu, W. Zhang, M.J. Zuo, and J. Chen, "Multi-branch and multi-scale CNN for fault diagnosis of wheelset bearings under strong noise and variable load condition," *IEEE Trans. Ind. Inform.*, vol. 16, pp. 4949–4960, 2020.

[28] W. Zhang, X. Li, and Q. Ding, "Deep residual learning-based fault diagnosis method for rotating machinery," *ISA Trans.*, vol. 95, pp. 295–305, 2018.

[29] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, 2018.

[30] H. Wang, Z. Liu, D. Peng, and Z. Cheng. "Attention-guided joint learning CNN with noise robustness for bearing fault diagnosis and vibration signal denoising," *ISA Trans.*, vol. 128, pp. 470–484, 2022.

[31] Y. Hao, H. Wang, Z. Liu, and H. Han, "Multi-scale CNN based on attention mechanism for rolling bearing fault diagnosis," in *2020 Asia-Pac. Int. Symp. Adv. Reliab. Maint. Model. (APARM)*, IEEE, pp. 1–5, 2020.

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

[33] Case Western Reserve University Bearing Data Center Website. Available: https://engineering.case.edu/bearingdatacenter

[34] F. Jia, Y. Lei, N. Lu, and S. Xing, "Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization," *Mech. Syst. Signal Process.*, vol. 110, pp. 349–367, 2018.

[35] H. Wang, Z. Liu, D. Peng, and Y. Qin, "Understanding and learning discriminant features based on multi-attention 1DCNN for wheelset bearing fault diagnosis," *IEEE Trans. Ind. Inform.*, vol. 16, pp. 5735–5745, 2019.

[36] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal.*, pp. 1, vol. 42, pp. 2011–2023, 2019.

[37] C. Nwankpa, I. Winifred, G. Anthony, and M. Stephen, "Activation functions: comparison of trends in practice and research for deep learning," *arXiv preprint arXiv:1811.03378*, 2018.

[38] R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. "Grad-CAM: why did you say that?," *arXiv preprint arXiv:1611.07450*, 2016.

APPENDIX

INTERPRETABILITY ANALYSIS

Here, we delve into a more detailed visualization and interpretability analysis. Figure 16 displays the visualization results for six vibration signal samples (A–F) taken from the CWRU bearing dataset. Samples A and B represent signals from bearings with rolling element faults, while C and D represent signals with inner race faults, and E and F correspond to signals with outer race faults. Figure 17 exhibits the visualization results for four vibration signal samples (A–D) associated with inner race faults in the HSA bearing dataset. For each sample's visualization results, (a) signifies the vibration signal without added noise; (b) denotes the noisy vibration signal; and (c) illustrates

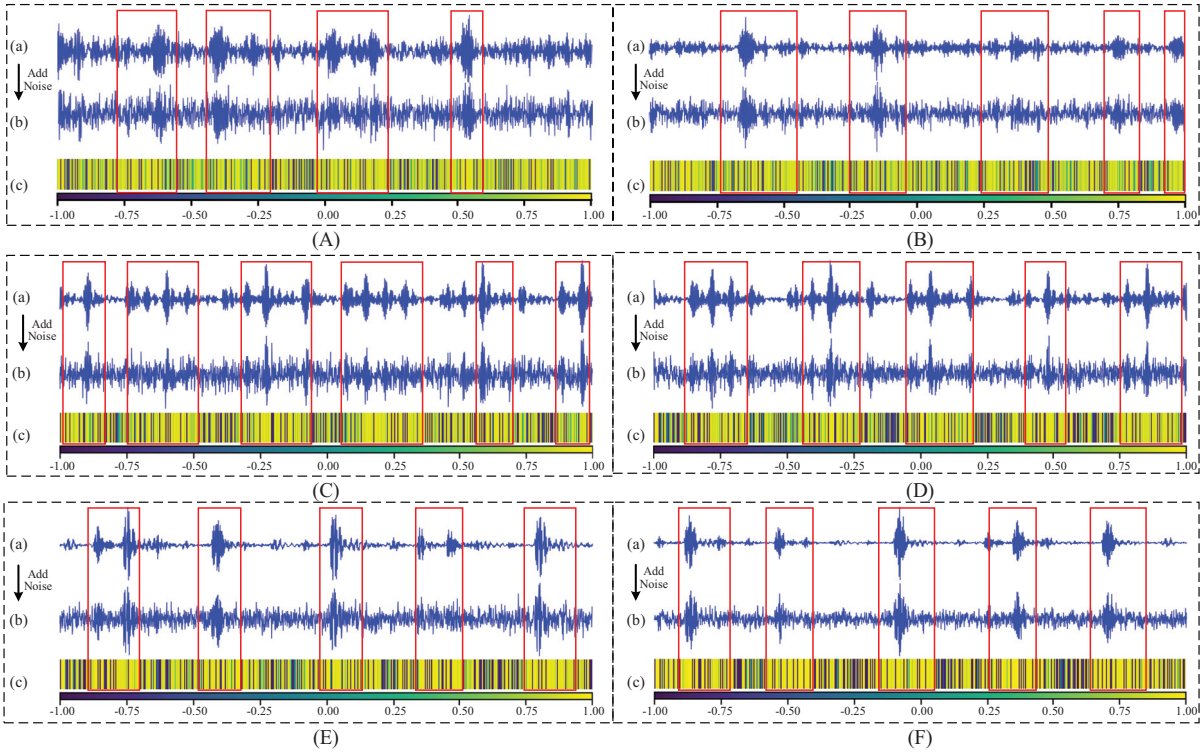


Fig. 16. CWRU bearing dataset: The visualization result at the time domain for six vibration signal samples (A–F). (a) The vibration signal (no noise added); (b) the noisy vibration signal (SNR = 0 dB); (c) the visualization result of the time weight vector. Each small grid represents a weight, and the color represents the weight value.

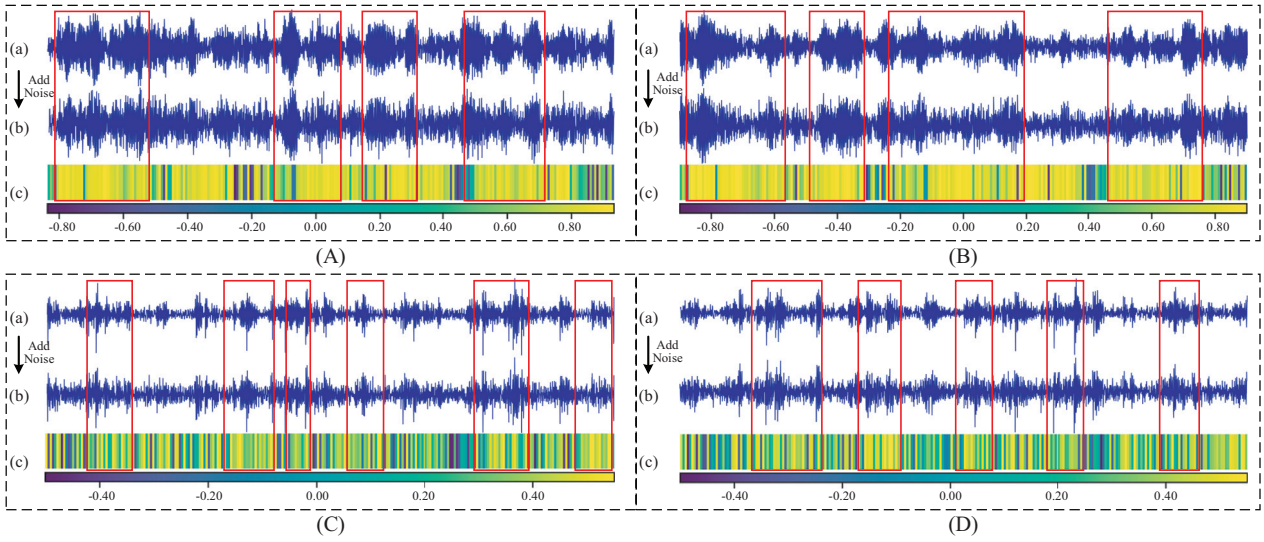


Fig. 17. HSA bearing dataset: The visualization result at the time domain for four vibration signal samples (A–D). (a) The vibration signal (no noise added); (b) the noisy vibration signal (SNR = 6 dB); (c) the visualization result of the time weight vector. Each small grid represents a weight, and the color represents the weight value.

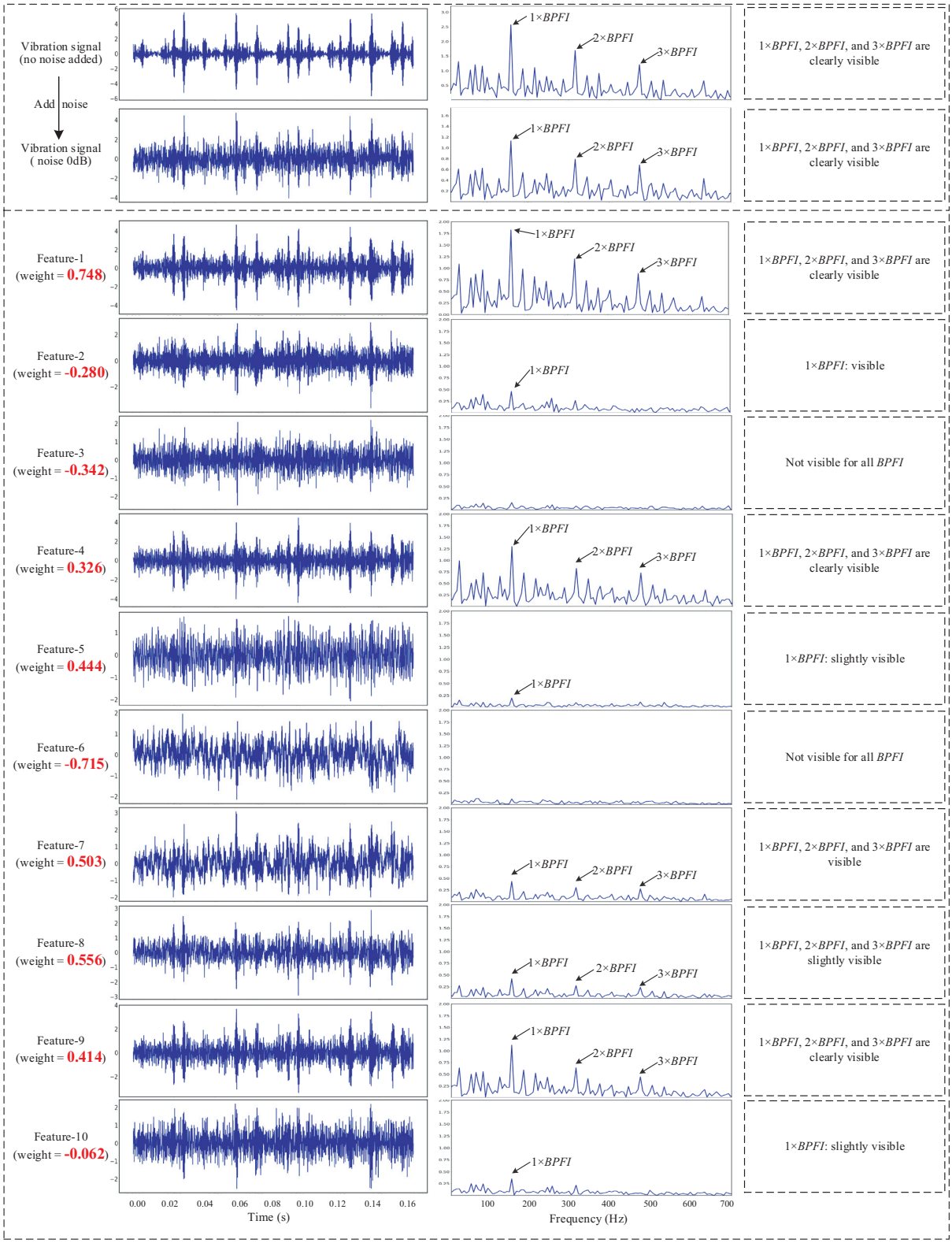


Fig. 18. The visualization of all channel feature signals (Feature-1 to Feature-16) and channel attention weights of the first RMAM. The time domain and the squared envelope spectrum of these signals are visualized.

the time weight vector visualization, where each small grid symbolizes a weight and the color represents the weight value.

As seen in Fig. 16, the CWRU bearing signals are relatively clean, with clearly visible impulse signal components. Despite the addition of strong noise, which overwhelms the original vibration signal waveform, the proposed method remains focused on impulsive signal component, irrespective of the fault type. This observation aligns with the paper's conclusions. Figure 17 reveals that the HSA bearing dataset signals are complex, exhibiting no discernible trends. Nonetheless, the proposed method

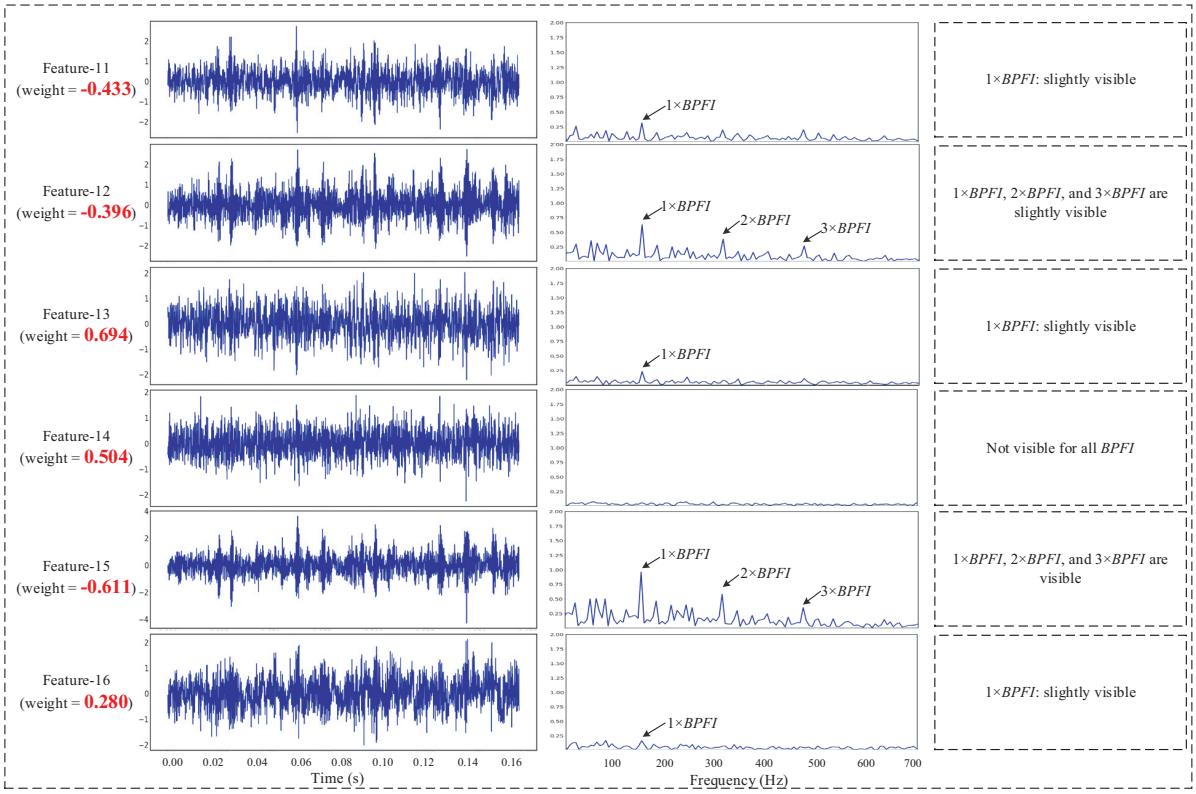


Fig. 18. (*Continued*)

strives to identify useful features for fault diagnosis, demonstrating a particular focus on impulsive signals.

As a supplementary analysis, we dive deeper into the feature learning mechanism of channel domain attention within the CWRU bearing dataset and explore the interpretability of the first RMAM module. Figure 18 visualizes all channel feature signals (from Feature-1 to Feature-16) and channel attention weights. The frequency spectrum and the square envelope spectrum of these signals are also presented. BPFI represents the inner race fault characteristic frequency. It can be seen that the proposed method effectively differentiates between useful and irrelevant features. For instance, Feature-1, Feature-4, Feature-7, Feature-8, and Feature-9 contain substantial fault-related information, resulting in the assignment of greater weight. Conversely, Feature-2, Feature-3, Feature-6, and Feature-10 possess minimal fault information or lack identifiable fault characteristic frequencies, leading to lower assigned weights. However, certain exceptions exist, such as Feature-14 receiving a larger weight and Feature-15 receiving a smaller weight. Two possible explanations can account for this situation: (1) the attention mechanism may not accurately discern useful features 100% of the time, leading to occasional weight assignment errors; and (2) the attention mechanism aims to more precisely select meaningful features, implying that fault characteristic frequencies do not fully represent the information required by the network. Consequently, these signals might contain other features needed by the network. In summary, while the proposed method can generally distinguish between useful and irrelevant features, some weight assignment errors occur. Further exploration and analysis in future work are necessary to address these discrepancies.