This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3301888

Multigrained Hybrid Neural Network for Rotating Machinery Fault Diagnosis Using Joint Local and Global Information

Zhenkun Yang, Bin He, Member, *IEEE*, Gang Li, Member, *IEEE*, Ping Lu, Bin Cheng, Member, *IEEE*, and Pengpeng Zhang

***Abstract*—Deep learning (DL) models such as multilayer perceptrons (MLPs) and convolutional neural networks (CNNs) have strong feature representation and nonlinear mapping capabilities, and their effectiveness has been demonstrated in fault diagnosis. However, fault features usually occur at different scales and are always disturbed by noise, making it difficult for DL-based models to learn local and global information in mechanical vibration signals. To address this issue, a multigrained hybrid neural network named MgHNN is proposed to extract robust features that seamlessly integrate CNN into vision MLP. First, the short-time Fourier transform is performed on original vibration signals to obtain time-frequency images, and each image is then divided into multiple nonoverlapping patches. Second, a novel multigrained feature representation (MFR) block is proposed by constructing hierarchical residual-like connections within one single wave block, which is more suitable for learning hierarchical local and global feature representations among different image patches. Third, we propose a depthwise wave (DWwave) block by integrating depthwise convolution and feature concatenation operations, which can make MFR block better focus on the local information and effectively overcome vanishing gradient problem. Finally, experimental results on two fault diagnosis datasets demonstrate that the MgHNN has improved diagnostic accuracy and reduced model complexity compared to state-of-the-art models. The results of noise interference experiments indicate that the MgHNN exhibits superior robustness against noise.**

***Index Terms*—Convolutional neural networks, deep learning, fault diagnosis, joint local and global information, multilayer perceptrons.**

This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0108905, in part by the National Natural Science Foundation of China under Grant 62273262, Grant 61825303, Grant 62088101, and Grant U1713215, in part by the Shanghai Science and Technology Commission Project under Grant 21ZR1466400, Grant 22QA1408500 and Grant 2021SHZDZX0100, in part by the China University Industry, University and Research Innovation Fund under Grant 2021ZYA03004, in part by the Postdoctoral Innovation Talents Support Program under Grant BX20220235, in part by the Shanghai Postdoctoral Excellence Program under Grant 2022756, in part by the Special Fund for Independent Innovation of Aero Engine Corporation of China under Grant ZZCX-2021-007, and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding authors: Gang Li; Bin Cheng.)*

The authors are with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China, and also with the Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai 200120, China (e-mail: 2111466@tongji.edu.cn; hebin@tongji.edu.cn; lig@tongji.edu.cn; pinglu@tongji.edu.cn; bincheng@tongji.edu.cn; ppzhang@hotmail.com).

I. INTRODUCTION

ROTATING machinery, as a key component of industrial equipment, is the basis for ensuring the continuous and orderly operation of industrial equipment [1]. However, rotating machinery works under various complex external environmental conditions for a long time, such as high temperature, high pressure and high speed, and rotating machinery part failures will inevitably occur during operation [2], [3]. Therefore, it is of great significance to study a fault diagnosis method for rotating machinery to ensure the normal operation of industrial equipment in actual scenarios.

Building accurate and intelligent fault diagnosis methods for rotating machinery has always been the research frontier of scientists and engineers. Traditionally, mechanical fault diagnosis consists of three main stages: 1) sensor signal acquisition and preprocessing; 2) feature extraction and selection; and 3) fault classification. Among the three stages, feature extraction and selection play an important role in fault diagnosis. However, several limitations still need to be considered in these traditional fault diagnosis methods. First, traditional fault diagnosis methods [4]-[6] have a great level of dependence on manually selected features. Therefore, if these manually selected features are not suitable for the given task, the performance of the final fault classification algorithm will be significantly reduced. Second, handcrafted features are tailored to specific tasks, which means that using them to accurately make predictions under certain situations may not be suitable for other scenarios. Third, in the early stage of fault occurrence, the fault characteristics may not be clear, and how to select the appropriate features becomes more difficult. Finally, there may potentially be many unique features hidden within the data themselves that can reveal rotating machinery failures, making it nearly impossible for humans to identify and extract these complex features through manual observation or interpretation.

Unlike traditional fault diagnosis methods, deep learning (DL)-based methods can extract features adaptively instead of manually, which gives them powerful feature learning abilities. In addition, DL has superior generalization and model fitting capabilities. In recent years, a variety of neural network structures have been proposed in the field of DL; these methods, such as deep belief networks (DBNs) [7], multilayer perceptrons (MLPs) [8], autoencoders (AEs) [9], long short-term memory (LSTM) [10], and convolutional neural networks

(CNNs) [11], are widely used in fault diagnosis. Although CNN-based methods have achieved successful applications in fault diagnosis, some issues are still associated with them. A major limitation of CNNs in fault diagnosis is that while they demonstrate an excellent ability to capture local feature representations, they have an inherent disadvantage when capturing the long-range dependencies of time series signals.

The latest DL-based feature extraction model shows that vision MLP [12] can achieve high performance by establishing long-range dependencies. Vision MLP models (such as the residual MLP (Res-MLP) [13], AS-MLP [14] and Wave-MLP [15]) capture long-range dependencies among image patches through projections along different patches implemented by the MLPs. For instance, Tang *et al*. [15] proposed a Wave-MLP architecture for dynamically aggregating tokens; this approach represents each token as a wave function containing both amplitude and phase information and achieves strong performance on different visual recognition tasks. Although vision MLP models have great advantages in capturing global features, much of the current research on vision MLP models focuses on vision tasks such as image classification, object detection, and semantic segmentation. Vision MLP models that are specifically designed to handle vibration signals for rotating machinery fault diagnosis are still lacking. Moreover, due to the lack of local feature representation ability, it is difficult for vision MLP models to simultaneously capture local and global information among different image patches, which limits their ability to represent fine-grained features.

As mentioned above, the advantages of CNN are good at capturing local information, but do not fully account for modeling long-range dependencies; the properties of vision MLP are able to model long-range dependencies, but lack sufficient mechanism to capture local spatial information at different scales. To address this issue, we propose a multigrained hybrid neural network (MgHNN) for fault diagnosis, which takes advantage of vision MLP to learn long-range dependencies, and of CNN to capture local information. The original vibration signals are first transformed by short-time Fourier transform (STFT) [16] that can obtain the time-frequency images. Then, the MgHNN splits an image into multiple nonoverlapping patches. Next, inspired by the exploitation of hierarchical residual-like connections [17], a novel multigrained feature representation (MFR) block is proposed by constructing hierarchical residual-like connections within one single wave block. MFR block can learn hierarchical local and global feature representations and use hierarchical residual-like connections to effectively reduce the model's computational complexity through stronger interactions among different image patches. Finally, a depthwise wave (DWwave) block is developed by integrating depthwise convolution and feature concatenation operations. The feature propagation ability of MFR block is promoted by the feature concatenation operation. The depthwise convolution operation is used to further focus on the local information with negligible extra computational complexity. The main contributions are summarized as follows.

1) In order to solve fault diagnosis problems with multiscale properties and noise disturbance, a multigrained hybrid neural network named MgHNN is proposed to extract robust features that seamlessly integrate CNN into vision MLP with hierarchical residual-like connections, aiming at effectively combining the advantages of CNN and vision MLP.
2) Drawing inspiration from hierarchical residual-like connections, a novel multigrained feature representation (MFR) block is proposed, which could learn hierarchical local and global information within one single wave block, thereby capturing fine-grained fault features and reducing the model's computational complexity.
3) Through the integrate of depthwise convolution and feature concatenation operations, a depthwise wave (DWwave) block is developed, which can make MFR block better focus on the local information and effectively overcome vanishing gradient problem.

The rest of the paper is organized as follows. Section II introduces the related works of the proposed approach, including DL-based fault diagnosis and MLP-based vision models. Section III proposes the MgHNN, which is a core operation of the proposed fault diagnosis method in this paper. In Section IV, the specific implementation of the proposed fault diagnosis method is described in detail. In Section V, experimental studies on two datasets are carried out to verify the effectiveness of the proposed method. Conclusion and future work are presented in Section VI.

II. Related Works

A. DL-based Fault Diagnosis

In recent years, a variety of neural network structures have been proposed in the field of DL, and these approaches are widely used in fault diagnosis. For example, to improve the reliability of DBN-based fault diagnosis, Chen *et al*. [18] fused the time domain and frequency domain feature information extracted from different sensor signals in a stacked autoencoder (SAE) neural network and used the fused feature vector to train a DBN. Moreover, Pei *et al*. [19] proposed a novel deep neural network based on transfer learning for rotating machinery fault diagnosis, which combines the advantages of a transformer [20] and a CNN. Zhao *et al*. [21] presented a local feature-based gated recurrent unit (LFGRU), which incorporates local feature information from time series signals for machine health monitoring. Unlike purely black-box DL models, some studies focus on how to improve the interpretability of DL-based fault diagnosis models. For instance, Zhu *et al*. [11] proposed a new CNN-based diagnostic model (DEFT-CNN), which achieves a good balance between diagnosis accuracy and model interpretability by combining feature information and temporal information. In [22], a continuous wavelet convolutional (CWConv) layer was designed for interpretable intelligent diagnosis problems. The designed CWConv can better match the shock responses caused by faults. Therefore, the essential fault features with physical meaning can be extracted from raw mechanical signal. To solve the problem of fault diagnosis of variable working conditions, Zhao *et al*. [23] proposed a novel fault diagnosis method (AIICNN) based on a CNN. In the AIICNN, the intraclass and interclass constraints are designed to optimize the distribution differences among samples. To enrich the fault information acquired from different views and better discriminate the features learned from a diagnosis network, Sun *et al*. [24] proposed a bearing fault diagnosis method based on multidomain information fusion and an improved residual dense network. In [25], a benchmark study was provided for nine DL-based fault diagnosis models. Experimental results on nine fault diagnosis datasets have demonstrated that some DL models can achieve good performance in different fault diagnosis tasks. Inspired by the capsule network [26], Chen *et al*. [27] proposed a new network architecture (ICN) for rolling bearing fault diagnosis. This method used a capsule network to learn abundant spatial information between features, and further used a CNN to improve the generalization ability of the capsule network. Different from CNN-based models, graph neural network (GNN)-based models learn latent feature representations between time series signals from the perspective of graph topology. For instance, a rolling bearing fault diagnosis method (WHVG-GIN+) was presented by Li *et al*. [28], where an improved horizontal visibility graph (HVG) was applied to obtain the geometric transformation of time series. Meanwhile, this method used the graph isomorphism network (GIN) to learn graph representations. Li *et al*. [29] provided a systematic and practical guide to building a GNN-based framework for intelligent fault diagnosis, and further verified the effectiveness of some GNNs, such as graph convolutional network (GCN), simplifying GCN (SGCN), and graph sample and aggregate (GraphSage).

Fault features usually occur at different scales and are always disturbed by noise, which requires the utilized fault diagnosis method to be able to capture fine-grained signal features. However, due to the lack of multigrained feature representation ability, it is difficult for DL-based fault diagnosis methods to simultaneously capture local and global information, which limits their ability to represent fine-grained fault features. To address this issue, in this work, we focus on how to effectively represent local and global information at a granular level.

B. MLP-based Vision Models

Recently, vision architectures based entirely on MLPs have received much attention. Inspired by the vision transformer (ViT) [30], Touvron *et al*. [12] also proposed a purely MLP-based model, called MLP-Mixer, that replaces the self-attention sublayer with a linear layer. MLP-Mixer utilizes two types of MLP layers, i.e., channel-mixing MLP and token-mixing MLP layers. The features of each patch are extracted by the channel-mixing MLP, while the spatial information among different image patches is captured by the token-mixing MLP. By using the same training scheme as that of CaiT [31], MLP-Mixer performs more robustly than the ViT, eliminating the need to use cross-channel or batch-specific normalization. Moreover, Rao *et al*. [32] presented a global filter network (GFNet), which employs a 2D Fourier transform to learn long-range spatial dependencies with log-linear complexity and provides favorable accuracy/complexity trade-offs. gMLP [33] contained novel MLP layers with gating operations to enhance the dependencies between spatial locations, and this approach achieves higher accuracy than MLP-Mixer. Yu *et al*. [34] rethought the design of the token-mixing MLP and proposed a circulant channel-specific token-mixing MLP. This method requires fewer parameters but achieves higher classification accuracy. Moreover, Chen *et al*. [35] proposed an MLP-like architecture, CycleMLP, which introduces a cyclic fully connected layer to capture spatial information. Extensive experiments conducted on various visual tasks, such as image classification, object detection and image segmentation, demonstrated that CycleMLP achieves competitive results.

Many MLP-based vision models further optimize MLP-Mixer based on its advantages concerning long-range spatial dependencies. Moreover, much of the current research on vision MLP focuses on vision tasks such as image classification, object detection, and semantic segmentation. However, vision MLP that is specifically designed to handle vibration signals for rotating machinery fault diagnosis are still lacking. In this work, we combine the properties of CNN and vision MLP, so that the proposed MgHNN not only maintains long-range spatial dependencies but also learns local information.

III. Multigrained Hybrid Neural Network

In this section, we discuss the proposed MgHNN detailly. After introducing the MLP-Mixer and Wave-MLP models, we present the MFR block, which split image patches into several groups and aggregate them by a set of DWwave blocks. At last, we describe the DWwave block in MgHNN.

A. Preliminaries

The MLP-Mixer model stacks N blocks, each with two components, i.e., a channel-mixing MLP and a token-mixing MLP. The features of each patch are extracted by the channel-mixing MLP, while the spatial information among different image patches is captured by the token-mixing MLP. A patch feature is denoted by $z_j \in \mathbb{R}^d$, where d is the number of channels. The n patch features of an image are denoted as $Z = [z_1, z_2 \cdots, z_n]$. The channel-mixing MLP can be formulated as:

$$\text{Channel-mixer MLP}(z_j, W^c) = W^c z_j, j = 1, 2, \cdots, n, \quad (1)$$

where W^c represents the channel-mixing MLP weight.

To aggregate the information acquired from different tokens, the token-mixing MLP is calculated as follows:

$$\text{Token-mixer MLP}(Z, W^t)_j = \sum_k W^t_{jk} \otimes z_k, j = 1, 2, \cdots, n, \quad (2)$$

where W^t denotes the token-mixing MLP weight. ⊗ denotes

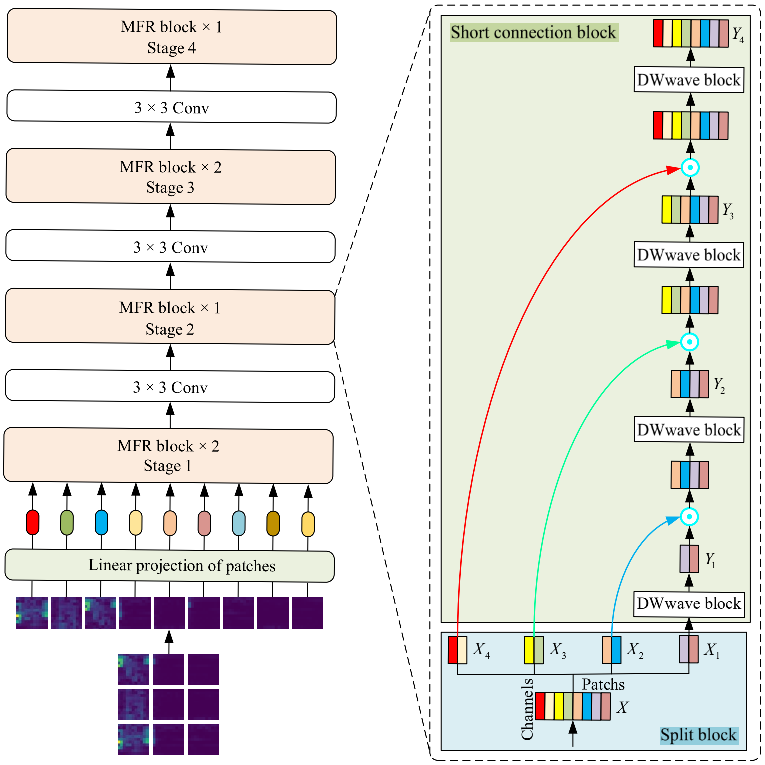


Fig. 1. The overall architecture of the MgHNN.

the elementwise multiplication operation and the subscript *j* denotes the *j*-th token among all output tokens.

Such a simple token-mixing MLP with fixed weights neglects the fine-grained feature information contained in different input image patches, which easily causes vision MLP models to suffer from vanishing gradient and model degradation problems during training. However, unlike the simple token-mixing MLP with fixed weights in MLP-Mixer, Wave-MLP aggregates token information by dynamically modulating the relationships between fixed weights and tokens. Specifically, each token is treated as a wave $\tilde{z}_j$ with magnitude and phase information, i.e.,

$$\tilde{z}_j = |z_i| \otimes e^{i\theta_j}, j = 1, 2, \cdots, n, \qquad (3)$$

where i indicates an imaginary unit such that $i^2 = -1$. $|\cdot|$ represents the absolute value operator. $|z_i|$ denotes the amplitude, which is a real-valued feature representing the content of each token. The phase θ_j denotes the location of token within a wave period. $e^{i\theta_j}$ is a periodic function whose elements have the unit norm.

A patch feature is denoted by $v_j \in \mathbb{R}^c$. The *n* patch features of an image are denoted as $V = [v_1, v_2, \cdots, v_n]$. A token's amplitude z_j can be obtained by a plain channel-mixing MLP operation, i.e.,

$$z_j = \text{Channel-mixer MLP}(v_j, W^c), j = 1, 2, \cdots, n. \qquad (4)$$

Wave-MLP employs a plain channel-mixing MLP operation to capture the specific properties of each input, which in turn generates the phase information θ_j using the input features x_j, where θ_j can be calculated as:

$$\theta_j = \text{Channel-mixer MLP}(x_j, W^\theta), \qquad (5)$$

where W^θ denotes the channel-mixing MLP weight.

In Wave-MLP, each wave-like token is described by a complex domain formula. Wave-MLP expands each token's complex domain formula with Euler's formula and expresses it in real and imaginary parts, which makes it easy to embed in general vision MLP models. Therefore, a wave-like token with real and imaginary parts can be formulated as:

$$\tilde{z}_j = |z_j| \otimes \cos\theta_j + i|z_j| \otimes \sin\theta_j, j = 1, 2, \cdots, n. \qquad (6)$$

Then, different tokens $\tilde{z}_j$ are aggregated with a token-mixing MLP operation, i.e.,

$$\tilde{b}_j = \text{Token-mixer MLP}(\tilde{Z}, W^t)_j, j = 1, 2, \cdots, n, \qquad (7)$$

where $\tilde{Z} = [\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_n]$ denotes all the wave-like tokens in a layer. The output $\tilde{b}_j$ is a complex-valued representation of the aggregated features.

After that, to obtain the real-valued output b_j, the real and imaginary parts of $\tilde{b}_j$ are summed with their weights. The output b_j can be formulated as:

$$b_j = \sum_k W^t_{jk} z_k \otimes \cos\theta_k + W^i_{jk} z_k \otimes \sin\theta_k, j = 1, 2, \cdots, n, \qquad (8)$$

where W^t and W^i denote the learnable weights.

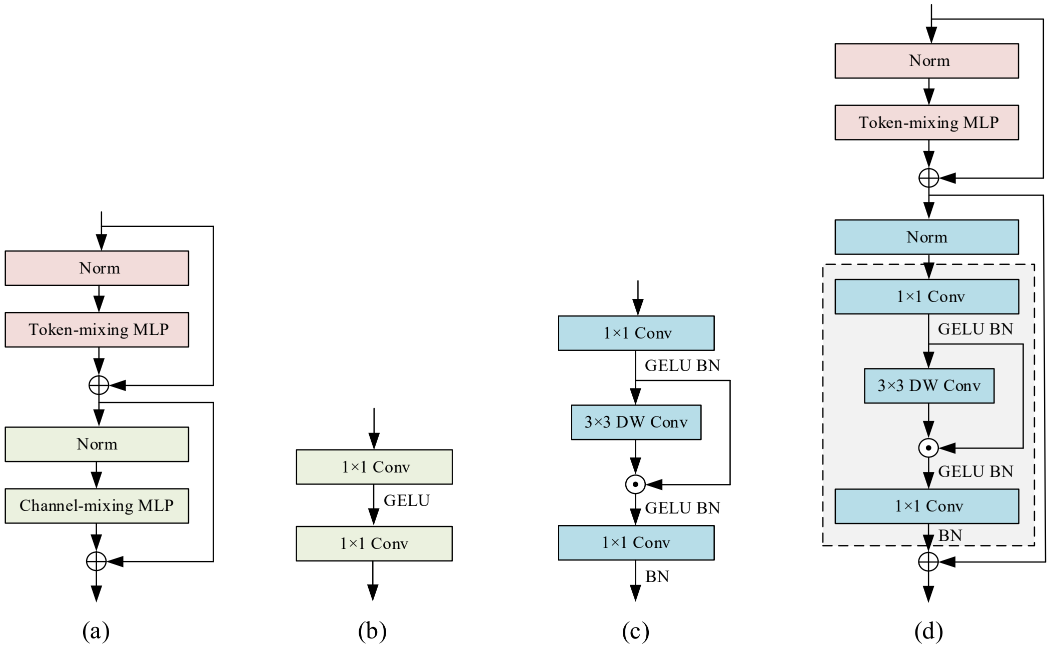


Fig. 2. (a) Wave block in Wave-MLP. (b) Channel-mixing MLP in wave block. (c) Proposed DRFFN. (d) Proposed DWwave.

The final phase-aware token mixing module (PATM) output o_j is obtained by transforming b_j with another channel-mixing MLP operation to enhance the representation capacity of the model, where o_j can be formulated as:

$$o_j = \text{Channel-mixer MLP}(b_j, W^o), \tag{9}$$

where W^o is the channel-mixing MLP weight.

B. Overall Architecture

Recent advances in computer vision [30], [12] have demonstrated that models based on self-attention and pure MLPs can be optimized based on their advantages regarding long-range spatial dependencies. In this work, the MgHNN combines the properties of CNN and vision MLP, aiming to effectively represent local and global features.

Specifically, the MgHNN not only maintains the long-range spatial dependencies among different image patches but also enhances the local feature representation ability of the model at a granular level. The overall architecture of the proposed MgHNN is depicted in Fig. 1. The MgHNN first splits an image into multiple nonoverlapping $H \times W$ patches and projects the flattened patches into $L = HW$ tokens with D dimensions. The basic building block of MgHNN, called MFR block, consists of 1) a split block that splits the input image patches into n patch subsets and 2) a short connection block that can effectively represent local and global features.

C. MFR Block

Unlike most existing vision MLP models that enhance their feature representation abilities in a simple channel-mixing and token-mixing manner, we seek alternative architectures with stronger feature representation abilities at a granular level while utilizing a lower computational load. To achieve this goal, MFR block is developed. We replace the wave block in Wave-MLP containing S patches with a set of n DWwave blocks, each with $S \times i/n$ patches, where $i \in \{1, 2, \cdots, n\}$. As shown in Fig. 1, these n DWwave blocks are further connected in a hierarchical residual-like style to learn local and global feature representations among different image patches that the output features can represent. A DWwave block first extracts features from a group of image patches. Then, this process is repeated several times until all image patches are processed.

Specifically, to obtain fine-grained feature information from different input image patches, we splits the input image patches into n patch subsets, such as the four patch subsets shown in Fig. 1, namely, X_1, X_2, X_3, and X_4. Let X_i denotes the corresponding i-th patch subset, where $i \in \{1, 2, \cdots, n\}$. Each patch subset X_i has $S \times i/n$ patches, each with the same spatial resolution as that of the input image patches, where $i \in \{1, 2, \cdots, n\}$, and each patch subset has a corresponding DWwave block, which is denoted by $H_i()$. We denote the output of $H_i()$ as Y_i. The patch subset X_i is concatenated with the output of $H_{i-1}()$ and then fed into $H_i()$. Thus, Y_i can be written as:

$$Y_i = \begin{cases} H_i(X_i) & i = 1; \\ H_i(X_i \odot Y_{i-1}) & 1 < i \leq n, \end{cases} \tag{10}$$

where $\odot$ denotes the elementwise concatenation operation. Y_1, Y_2, and Y_3 are the learned local information of different scales. Y_4 is the learned global feature representation.

TABLE I
SPECIFICATIONS OF DIFFERENT VARIANTS OF MGHNN. 'DIMENSION' AND 'EXPANSION' REPRESENT THE DIMENSION OF FEATURE AND EXPAND RATIO, RESPECTIVELY.

Layer name	MgHNN-T	MgHNN-S	MgHNN-M	MgHNN-B
Stage 1	$\begin{bmatrix}\text{dimension}=8\\ \text{expansion}=4\end{bmatrix}\times 2$	$\begin{bmatrix}\text{dimension}=16\\ \text{expansion}=4\end{bmatrix}\times 2$	$\begin{bmatrix}\text{dimension}=32\\ \text{expansion}=4\end{bmatrix}\times 2$	$\begin{bmatrix}\text{dimension}=32\\ \text{expansion}=4\end{bmatrix}\times 2$
Conv 1	Conv, 3 × 3, 64, stride 2			
Stage 2	$\begin{bmatrix}\text{dimension}=16\\ \text{expansion}=4\end{bmatrix}\times 1$	$\begin{bmatrix}\text{dimension}=32\\ \text{expansion}=4\end{bmatrix}\times 3$	$\begin{bmatrix}\text{dimension}=64\\ \text{expansion}=4\end{bmatrix}\times 2$	$\begin{bmatrix}\text{dimension}=64\\ \text{expansion}=4\end{bmatrix}\times 3$
Conv 2	Conv, 3 × 3, 64, stride 2			
Stage 3	$\begin{bmatrix}\text{dimension}=32\\ \text{expansion}=4\end{bmatrix}\times 2$	$\begin{bmatrix}\text{dimension}=64\\ \text{expansion}=4\end{bmatrix}\times 3$	$\begin{bmatrix}\text{dimension}=128\\ \text{expansion}=4\end{bmatrix}\times 2$	$\begin{bmatrix}\text{dimension}=128\\ \text{expansion}=4\end{bmatrix}\times 3$
Conv 3	Conv, 3 × 3, 64, stride 2			
Stage 4	$\begin{bmatrix}\text{dimension}=64\\ \text{expansion}=4\end{bmatrix}\times 1$	$\begin{bmatrix}\text{dimension}=128\\ \text{expansion}=4\end{bmatrix}\times 2$	$\begin{bmatrix}\text{dimension}=256\\ \text{expansion}=4\end{bmatrix}\times 2$	$\begin{bmatrix}\text{dimension}=256\\ \text{expansion}=4\end{bmatrix}\times 2$

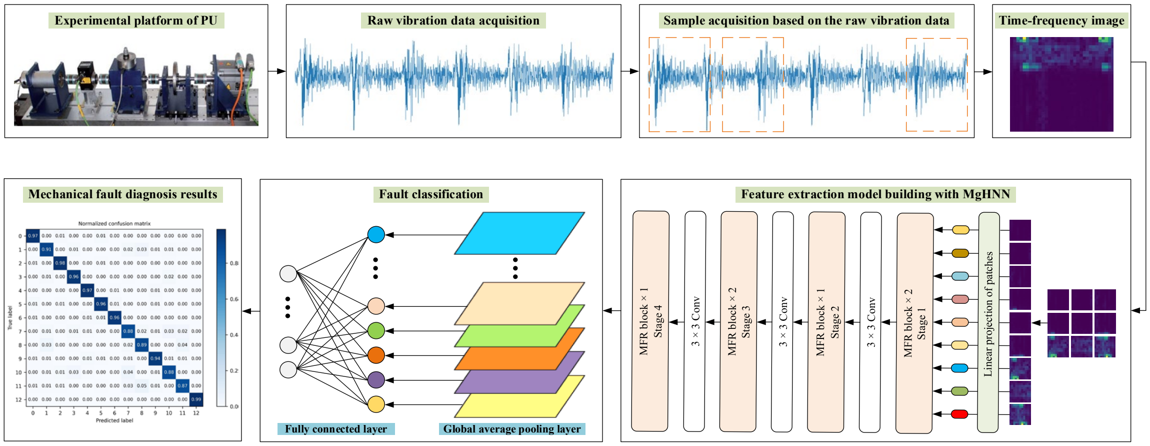

Experimental platform of PU
Raw vibration data acquisition
Sample acquisition based on the raw vibration data
Time-frequency image
Mechanical fault diagnosis results
Normalized confusion matrix
True label
Predicted label
Fault classification
Fully connected layer
Global average pooling layer
Feature extraction model building with MgHNN
MFR block × 1
Stage 4
3 × 3 Conv
MFR block × 2
Stage 3
3 × 3 Conv
MFR block × 1
Stage 2
3 × 3 Conv
MFR block × 2
Stage 1
Linear projection of patches

Fig. 3. General pipeline of fault diagnosis.

In this way, when a group of patches passes through the corresponding DWwave, a concatenation operation is performed to reuse the output features of the previous DWwave. Meanwhile, the spatial resolution of feature maps can be expanded by concatenation operation, which enables the designed architecture to represent fine-grained feature at a granular level and captures local and global information among different image patches.

D. DWwave Block

The DWwave block is developed by integrating depthwise convolution and feature concatenation operations, which can better focus on the local information and effectively overcome vanishing gradient perform. Specifically, the channel-mixing MLP of wave block is replaced by the designed depthwise residual feedforward network (DRFFN), and then the DWwave block is proposed, as shown in Fig. 2(d).

The designed DRFFN is shown in Fig. 2(c). The DRFFN appears similar to inverted residual blocks [36] and [37] consisting of a 1×1 regular convolution acting as expansion, followed by a 3×3 depthwise convolution and a 1×1 regular convolution. Specifically, for any given feature map $I\in\mathbb{R}^{H\times W\times d}$, a 1×1 regular convolution is conducted, which can be formulated as:

$$I_1=\text{BN}(\sigma(\text{Conv}_{1\times 1}(I))), \qquad (11)$$

where σ is the GELU function. BN denotes the batch normalization operation [38]. $\text{Conv}_{1\times 1}$ is a 1×1 regular convolution. Furthermore, a depthwise convolution is utilized to enhance the local feature representation ability of the model with negligible convolution can be calculated as:

$$I_2=\text{BN}(\sigma(\text{DWConv}_{3\times 3}(I_1))), \qquad (12)$$

where $\text{DWConv}_{3\times 3}$ indicates the depthwise convolution operation. DW denotes the depthwise operation [39]. Then, the feature propagation ability of the MFR block is promoted by the concatenation operation rather than the summation operation in our DRFFN. The concatenation operation can be written as:

$$I_3 = I_1 \odot I_2. \tag{13}$$

Finally, a 1×1 regular convolution is used to reduce the dimensionality of the DRFFN output channel. The output of the DRFFN can be formulated as:

$$O = \mathrm{BN}(\mathrm{Conv}_{1\times1}(I_3)). \tag{14}$$

The experimental results in Section V-F show that the proposed DWwave block helps our MgHNN achieve better performance.

IV. Fault Diagnosis Method

A new fault diagnosis pipeline is proposed that is able to automatically learn fault signatures and recognize machinery working states directly from the original vibration signals. The proposed machine fault diagnosis pipeline for detecting the working conditions of a mechanical system with high precision is the MLP-based vision model and time-frequency images are used as the input. The general pipeline of fault diagnosis is shown in Fig. 3, including time-frequency imaging, feature extraction model building with MgHNN, and fault classification. The time-frequency domain image generation method used in this paper is discussed in Section IV-A. The detailed information about the feature extraction model is described in Section IV-B. The classifier of proposed fault diagnosis method used is introduced in Section IV-C.

A. Time-frequency Imaging

We first perform a time-frequency analysis on the given raw sensor data to generate a time-frequency image. Time-frequency imaging is a way to convert time series signals into time-frequency domain images. This is a useful technique for analyzing the sensor signals of rotating machinery used for fault diagnosis, as it provides insights that are specific to different machine operating conditions with different time-frequency patterns. The STFT is widely used in feature extraction in rotating machinery fault diagnosis tasks, which can be regarded as a mathematical tool to transform time domain features into time-frequency domain features. In this work, a two-dimensional (2D) time-frequency image with physical significance is generated by the STFT of a one-dimensional (1D) vibration signal, which can fully excavate the time-frequency characteristics of the signal and has the advantages of high robustness and noise resistance. Mathematically, the STFT can be written as:

$$X_{\mathrm{STFT}}[k,n] = \sum_{m=0}^{R-1} x[n+m]w[m]e^{(-j2\pi km)/N}, \tag{15}$$

where x is the signal in the time domain and w is the properly chosen window factor; $0 \le k \le N-1$. After completing the STFT operation, the signal x is projected into a 2D time-frequency space. In this way, 1D time series are converted into time-frequency images.

B. Feature Extraction Model

The feature extraction model is the MgHNN proposed in this paper. Specifications of different MgHNN variants using MFR blocks are summarized in Table I. We begin with a smaller model, namely, MgHNN-T. Then, we introduce three variants of the model (MgHNN-S, MgHNN-M, and MgHNN-B) by simply adjusting the width and depth. For concrete examples of the MgHNN, a detailed description of the MgHNN is presented in Table I. In the MgHNN, the number of patch subsets is set to 4. The patch resolution is set to 7×7 . Moreover, MgHNN-T uses $\{2,1,2,1\}$ MFR blocks, and MgHNN-S, MgHNN-M, and MgHNN-B use $\{2,3,3,2\}$, $\{2,2,2,2\}$, and $\{2,3,3,2\}$ MFR blocks.

C. Classifier

In the classifier of the proposed fault diagnosis method, we take the global average pooling layer and then pass it to a fully connected layer for predicting class labels. Therefore, the classifier is composed of a global average pooling layer and a fully connected layer, which can be formulated as:

$$Q = \mathrm{FC}(\mathrm{GAP}(P)), \tag{16}$$

where FC denotes the fully connected layer and GAP represents the global average pooling layer. P is the input of the classifier.

V. Experimental Verification

To test the performances of the proposed fault diagnosis method and verify its effectiveness, we conduct experiments on two datasets including the Paderborn University (PU) bearing dataset [40] and Southeast University (SEU) gearbox dataset [41]. Comparative experiments are also carried out to compare the diagnostic accuracy and reduce model complexity with existing DL-based methods.

A. Experimental Settings and Evaluation Metrics

All experiments are conducted with PyTorch [42]. The operating system of the computer is Ubuntu 16.04, and the configuration of hardware is Intel Core i7-10870H CPU and NVIDIA GeForce RTX 3060 GPU. During the training, we use Adam [43] optimizer with momentum 0.9, weight decay 0.0001, and a mini-batch size of 8. All models are trained for 50 epochs, whose learning rate is 0.001.

In this paper, the number of parameters (Params) is used to measure the model's space complexity. The floating-point operations per second (FLOPs) is defined as the number of floating-point multiplication-adds, which can be used to measure the model's time complexity [32]. Moreover, accuracy, precision, recall, and F1 score are further used as the evaluation metrics. The four evaluation metrics can be calculated as:

$$\mathrm{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} \tag{17}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{18}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{19}$$

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2023.3301888

TABLE II
THE WORKING CONDITIONS OF THE PU DATASET.

Work condition	Rotating speed (rpm)	Load torque (Nm)	Radial force (N)	Name of setting
A	1500	0.7	1000	N15_M07_F10
B	900	0.7	1000	N09_M07_F10
C	1500	0.1	1000	N15_M01_F10
D	1500	0.7	400	N15_M07_F04

TABLE III
DIAGNOSIS ACCURACY COMPARISON RESULTS OBTAINED BY DIFFERENT MODELS ON THE PU DATASET. ‘PERCENTAGE’ AND ‘ACCURACY’ REPRESENT THE PERCENTAGE OF TRAINING SAMPLES AND AVERAGE TEST ACCURACY, RESPECTIVELY.

Model	Percentage (%)	Accuracy (%)
LSTM [25]	70	87.07
LFGRU [11]	80	92.80
DEFT-CNN [11]	80	96.32
ICN [27]	70	98.54
GCN [28]	80	95.60
SGCN [28]	80	98.12
GraphSage [28]	80	98.80
MgHNN-T	70	98.45
MgHNN-S	70	98.00
MgHNN-M	70	98.46
MgHNN-B	70	98.77

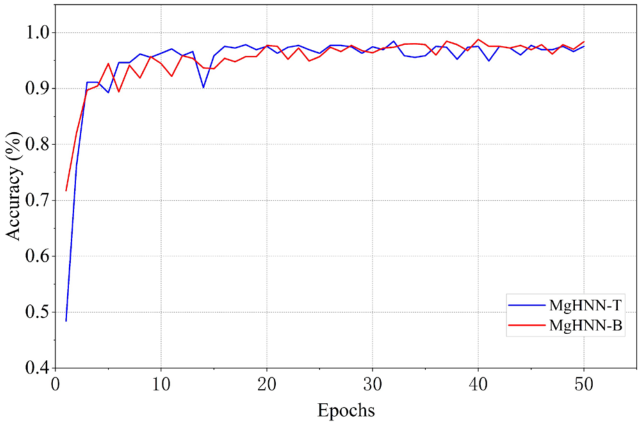


Fig. 4. Test accuracy curve of MgHNN-T and MgHNN-B on the PU dataset.

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{20}$$

where TP denotes true-positive which is the number of positive samples that are correctly predicted. TN denotes true-negative which is the number of negative samples that are correctly predicted. FP denotes false-positive which is the number of negative samples that are incorrectly predicted. FN denotes false-negative which is the number of positive samples that are incorrectly predicted.

B. Normalization

Normalization is the basic step in data preparation, which can promote the time-frequency domain analysis and feature extraction of subsequent data, and accelerate the convergence of DL models.

The normalization method used in this paper can be implemented in the following way:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{21}$$

where x is the input sample. $x_{\min}$ and $x_{\max}$ are the minimum and maximum values in x, respectively. x^* denotes the normalized data. All normalized data are distributed in the interval [0, 1].

C. Experimental Verification on Bearing Dataset

Experimental data for the PU bearing dataset is provided by the Paderborn University. The PU dataset includes the synchronous measurement of current signals and vibration signals, thus, enabling the fusion of multi-source signals to improve the accuracy of bearing fault diagnosis. The current signals and vibration signals are acquired on 6 healthy bearings and 26 damaged bearings. Bearings of the PU dataset are divided into: (1) 6 undamaged bearings; (2) 12 artificially damaged bearings; (3) 14 bearings with real damages caused by accelerated life tests. Each dataset is collected under four working conditions as shown in Table II. In this paper, in order to verify the effectiveness of the proposed method, we use the data collected from real damaged bearings under the working condition A.

On the PU bearing dataset, the proposed MgHNN is compared with one traditional model, namely LSTM [25]. In addition, the state-of-the-art models, i.e., DEFT-CNN [11], LFGRU [11], ICN [27], GCN [28], SGCN [28], and GraphSage [28], are used to be compared with the proposed MgHNN. The experimental results are shown in Table III. These experimental results are all obtained under the working condition A as shown in Table II. The test accuracy curves produced by MgHNN-T and MgHNN-B on the PU dataset are shown in Fig. 4. LSTM, a variant of recurrent neural network (RNN), uses time-frequency domain data as model input, but only uses simple hidden recurrent units to extract features, which is difficult to capture the high-level fault information. Therefore, LSTM only achieves an average test accuracy of 87.07%. Due to the characteristics of joint manual feature

TABLE IV
DETAILED DESCRIPTION OF THE SEU DATASET.

Label	Fault mode	RS-LC	Label	Fault mode	RS-LC
0	Rolling element	20 Hz–0 V	10	Chipped tooth	20 Hz–0 V
1	Inner + Outer ring	20 Hz–0 V	11	Health gear	20 Hz–0 V
2	Health bearing	20 Hz–0 V	12	Missing tooth	20 Hz–0 V
3	Inner ring	20 Hz–0 V	13	Root fault	20 Hz–0 V
4	Outer ring	20 Hz–0 V	14	Surface fault	20 Hz–0 V
5	Rolling element	30 Hz–2 V	15	Chipped tooth	30 Hz–2 V
6	Inner + Outer ring	30 Hz–2 V	16	Health gear	30 Hz–2 V
7	Health bearing	30 Hz–2 V	17	Missing tooth	30 Hz–2 V
8	Inner ring	30 Hz–2 V	18	Root fault	30 Hz–2 V
9	Outer ring	30 Hz–2 V	19	Surface fault	30 Hz–2 V

TABLE V
DIAGNOSIS ACCURACY COMPARISON RESULTS OBTAINED BY DIFFERENT MODELS ON THE SEU DATASET.

Model	Bearing		Gear	
	20 Hz/ 0V	30 Hz/ 2V	20 Hz/ 0V	30 Hz/ 2V
1D MLP [21]	86.50	90.50	84.30	90.60
RNN [21]	92.90	92.00	92.30	89.30
LFGRU [21]	93.20	94.00	94.80	95.80
HVG-GIN [29]	97.45	99.00	-	-
WHVG-GIN+ [29]	99.65	98.43	-	-
MgHNN-T	97.86	96.91	97.92	97.54
MgHNN-S	98.17	97.58	98.25	97.62
MgHNN-M	98.68	98.29	98.71	98.53
MgHNN-B	98.85	99.37	98.76	98.98

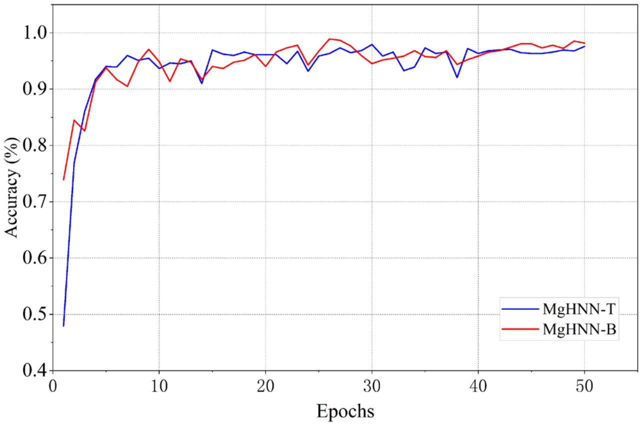


Fig. 5. Test accuracy curve of MgHNN-T and MgHNN-B on the SEU dataset.

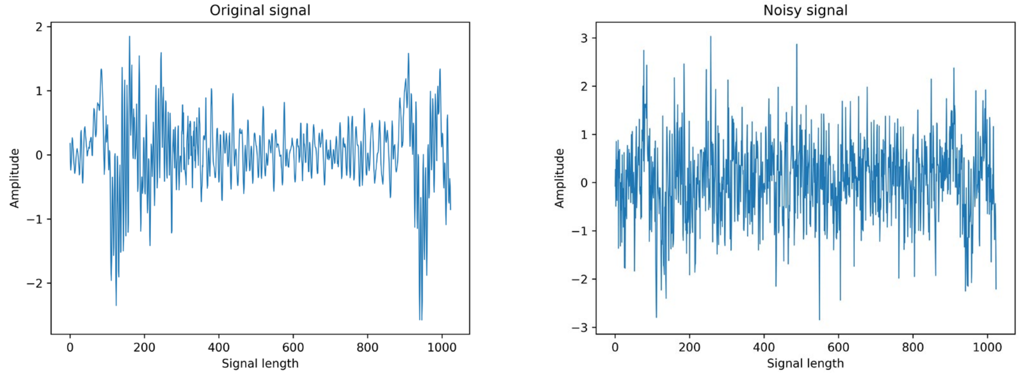
design and automatic feature learning, LFGRU obtains a better diagnostic performance compared with LSTM. MgHNN-T achieves a 98.45% average test accuracy, which outperforms LSTM and LFGRU by 11.38% and 5.65%, respectively. While DEFT-CNN achieves a 96.32% diagnostic accuracy by incorporating feature information and temporal information, MgHNN-B outperforms DEFT-CNN by 2.45%. The diagnostic accuracy of MgHNN-T, MgHNN-S, and MgHNN-M with smaller model complexity is slightly lower than that of ICN, but MgHNN-B outperforms ICN by achieving a 98.77% average test accuracy. Three GNN-based node-level fault diagnosis models, GCN, SGCN, and GraphSage, use frequency domain data as model input and employ the PathGraph to construct the sensor network. Meanwhile, the above GNN models can learn latent feature representations between time series signals from the perspective of graph topology, which also enables these models to achieve good diagnostic performance. For example, GraphSage achieves an average test accuracy of 98.80%, which outperforms ICN by 0.26%. Different from GNN-based models, our MgHNN divides the time-frequency domain image into nonoverlapping image patches and uses the designed MFR block to capture fine-grained features, enabling the four MgHNN variants to achieve competitive diagnostic performance. The results obtained on the PU bearing dataset show that MgHNN can provide more effective information for learning data representations, and the excellent performance of the proposed model is demonstrated.

D. Experimental Verification on Gearbox Dataset

To further verify the performance of our proposed MgHNN on different fault diagnosis tasks, the proposed model is compared with other DL-based models on the SEU gearbox dataset. The configurations of rotating speed-load

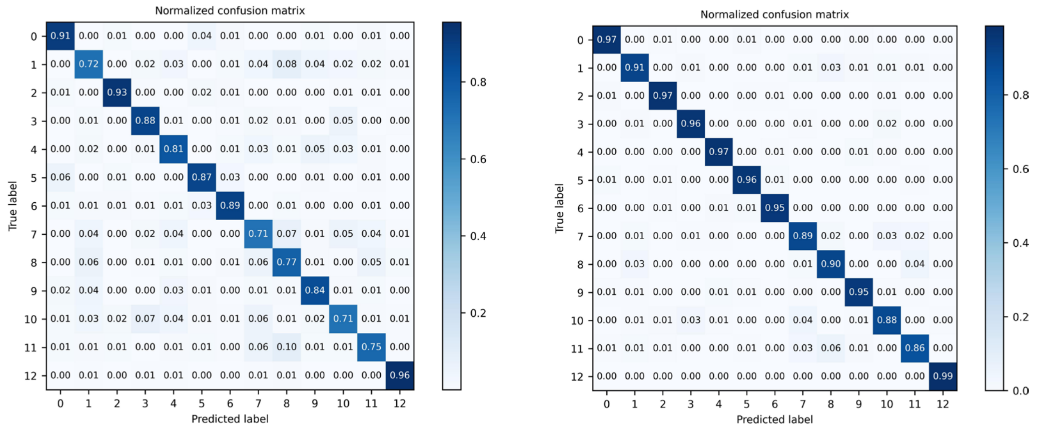
TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT MODELS WITH MODEL PARAMETERS, FLOPS, ACCURACY, PRECISION, RECALL AND F1 SCORE ON THE PU DATASET.

Model	Params (M)	FLOPs (G)	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
LeNet-5	0.12	0.35	71.43	71.67	71.59	71.63
AlexNet	57.03	1.34	90.32	90.58	90.45	90.51
ResNet-18	11.18	3.98	97.54	97.76	97.63	97.69
ResNet-50	23.53	9.20	97.70	97.88	97.92	97.90
MgHNN-T	0.30	0.15	98.45	98.59	98.51	98.55
MgHNN-S	1.92	0.74	98.00	98.26	98.17	98.21
MgHNN-M	6.52	2.14	98.46	98.65	98.53	98.59
MgHNN-B	7.27	2.62	98.77	99.04	98.86	98.94



(a) (b)

Fig. 6. (a) A piece of original vibration signal from the PU dataset. (b) The Gaussian white noise. (c) The noisy signal with SNR of −2.



(a) (b)

Fig. 7. Confusion matrix of the PU dataset based on MgHNN-B. (a) SNR=−2. (b) SNR=10.

configuration (RS-LC) of the experimental platform are 20 Hz-0 V and 30 Hz-2 V, respectively. The SEU dataset includes one normal state and four bearing faults under two working conditions. The description of the SEU dataset is showed in Table IV.

On the SEU gearbox dataset, the proposed MgHNN is compared with two traditional models, namely 1D MLP [21] and RNN [21]. In addition, the state-of-the-art models, i.e., LFGRU [21], HVG-GIN [29], and WHVG-GIN+ [29], are used to be compared with the proposed MgHNN. The experimental results are shown in Table V. These experimental results are all obtained under the two working conditions of 20 Hz-0 V and 30 Hz-2 V as shown in Table IV. The test accuracy curves produced by MgHNN-T and MgHNN-B on the SEU dataset are shown in Fig. 5. From these experimental results, it can be seen that LFGRU obtains a better performance compared with 1D MLP and RNN. HVG-GIN uses the HVG to transform time series into graph data, and adopts the sum aggregator to aggregate node features. Different from HVG-GIN, WHVG-GIN+ uses the improved HVG to obtain the graph-structured data. The average test accuracy of HVG-GIN are superior to LFGRU

TABLE VII
THE EFFECTIVENESS OF THE DWWAVE BLOCK.

Model	Params (M)	FLOPs (G)	Accuracy (%)
MgHNN-T (‘With wave blocks’)	0.10	0.21	97.70
MgHNN-S (‘With wave blocks’)	1.38	0.52	97.70
MgHNN-M (‘With wave blocks’)	4.79	1.55	97.54
MgHNN-B (‘With wave blocks’)	5.32	1.89	97.39
MgHNN-T (‘With DWwave blocks’)	0.30	0.15	98.45
MgHNN-S (‘With DWwave blocks’)	1.92	0.74	98.00
MgHNN-M (‘With DWwave blocks’)	6.52	2.14	98.46
MgHNN-B (‘With DWwave blocks’)	7.27	2.62	98.77

and comparable to those of WHVG-GIN+. Through the combination of CNN and vision MLP, our models with four different parameters achieve competitive diagnostic performance under different working conditions. For instance, on the SEU bearing dataset with a working condition of 20 Hz-0 V, MgHNN-B achieves an average test accuracy of 98.85%, which is much higher than that of 1D MLP, RNN, LFGRU, and HVG-GIN. Moreover, MgHNN-B obtains a 99.37% average test accuracy under a working condition of 30 Hz-2 V, which outperforms HVG-GIN and WHVG-GIN+ by 0.37% and 0.94%, respectively. On the SEU gear dataset with a working condition of 30 Hz-2 V, MgHNN-B achieves a 98.98% average test accuracy, which outperforms LFGRU by 3.18%. The superiority of the MgHNN means that the proposed method with a fine-grained feature extraction capability has great potential. Moreover, the above experimental results show that our method can effectively deal with gearbox fault diagnosis under different operating conditions.

E. Complexity/accuracy Trade-offs

To further verify the effectiveness of the MgHNN, we construct four fault diagnosis methods based on the LeNet-5, AlexNet, ResNet-18, and ResNet-50 feature extraction models and compare these methods with the MgHNN in terms of their numbers of parameters, numbers of FLOPs, average test accuracy, precision, recall, and F1 score on the PU dataset. The specifications of the LeNet-5, AlexNet, ResNet-18, and ResNet-50 feature extraction models are summarized in [44], [45], and [46].

The experimental results are shown in Table VI. As can be seen in these results, the smaller MgHNN variants often outperform CNN-based models even when the models employ significantly more parameters. For example, MgHNN-T, MgHNN-S, MgHNN-M, and MgHNN-B outperforms AlexNet by achieving 98.45%, 98.00%, 98.46%, and 98.77% average test accuracy with lower model complexity. ResNet-50 provides significantly better results than LeNet-5 and AlexNet, and its average test accuracy, precision, recall, and F1 score are 97.70%, 97.88%, 97.92%, and 97.90%, respectively. However, the metrics of ResNet-50 is still around 0.90% lower than that of MgHNN-T. Note that, MgHNN-T outperforms LeNet-5 by 27.02% with almost 1/2 the FLOPs, which is extremely efficient. Another finding is that ResNet-18 has almost half the number of parameters as ResNet-50, but ResNet-18 and ResNet-50 achieve the same accuracy. Compared with ResNet-18, MgHNN-M improves average test accuracy by 0.92% while reducing the number of parameters by 41.68%. Notably, MgHNN-B outperforms ResNet-18 by 1.07% with almost 1/3 the number of parameters. The above experimental results show that MgHNN has better complexity/accuracy trade-offs.

F. The Effectiveness of the DWwave Block

The designed DWwave block plays an active role in aggregating local representations and long-range dependencies; its effectiveness is investigated on the PU dataset. If the proposed DWwave block is used, the model can transfer and aggregate features more efficiently, achieving much better performance. The experimental results are shown in Table VII. As can be seen in these results, MgHNN-T (‘With DWwave blocks’) outperforms MgHNN-B (‘With wave blocks’) by achieving a 98.45% average test accuracy with lower model complexity. MgHNN-S (‘With DWwave blocks’) achieves an accuracy of 98.00%, which is much higher than that of MgHNN-S (‘With wave blocks’). In addition, MgHNN-M (‘With DWwave blocks’) model outperforms MgHNN-M (‘With wave blocks’) by 0.40% while having similar FLOPs. MgHNN-S (‘With DWwave blocks’) outperforms MgHNN-B (‘With wave blocks’) by 0.61% with almost 1/3 the number of parameters. Since the designed DWwave block can effectively focus on the local information, our MgHNN with the lowest complexity achieves excellent diagnostic accuracy. For instance, MgHNN-T (‘With DWwave blocks’) outperforms MgHNN-S (‘With wave blocks’) by achieving a 98.45% average test accuracy with almost 1/4 the number of parameters. These experimental results are consistent with those obtained on the PU dataset; that is, the DWwave block helps MgHNN achieve better performance.

G. Noise Interference

We argue that if the designed fault diagnosis method can extract fault features from noise signals with high accuracy, then it should have a strong ability to resist noise interference. Therefore, the diagnostic accuracies achieved by the LeNet-5, AlexNet, ResNet-18, ResNet-50, MgHNN-T, and MgHNN-B under different noise cases are compared. Specifically, the Gaussian white noise with different intensities is added to the original vibration signals of the PU dataset to simulate noisy vibration signals. The signal-to-noise (SNR is defined as $\mathrm{SNR\,(dB)} = 10\log 10\,(\mathrm{Psignal/Pnoise})$, where Psignal and Pnoise represent the power of the original signal and additional Gaussian noise. The smaller the SNR means the stronger noise. Especially, the noise has the same power as the signal when SNR = 0 dB. In this work, to comprehensively evaluate the anti-noise interference ability of the proposed method, four noise levels (i.e., SNR = −2, 2, 6, and 10 dB) are

TABLE VIII
ACCURACY COMPARISON OF SIGNALS WITH DIFFERENT SNR VALUES.

Model	SNR (dB)			
	−2	2	6	10
LeNet-5	45.31	56.53	62.06	65.75
AlexNet	60.52	62.37	78.34	76.50
ResNet-18	73.27	86.02	92.78	96.31
ResNet-50	76.96	84.49	92.78	94.78
MgHNN-T	78.80	89.40	93.24	96.31
MgHNN-B	80.18	91.24	93.55	96.96

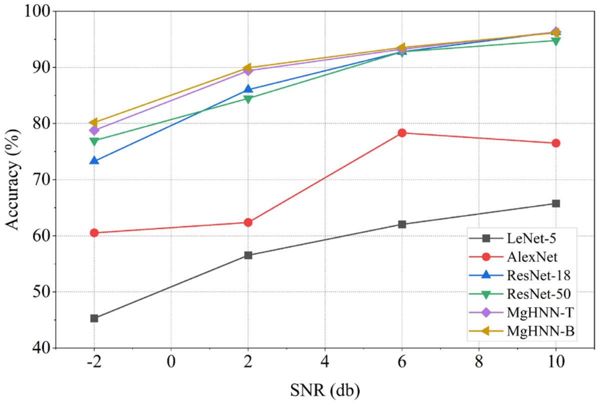


Fig. 8. Test accuracy curve of signals on the PU dataset with different SNR values.

considered, which range from strong to weak noise.

As shown in Fig. 6(b), after adding noise with the SNR of −2 to the original signals of the PU dataset, the periodic impact component of the fault signals is weakened, which inevitably leads to model performance degradation. To explore the separability of the proposed MgHNN for the normal and faulty data, a confusion matrix is used to demonstrate the detailed diagnostic results obtained on the PU dataset; this matrix is shown in Fig. 7. The experimental results of noise interference are shown in Table VIII. The test accuracy curve produced for the signals in the PU dataset with different SNR values is shown in Fig. 8. As seen in these results, the fault diagnosis accuracies of ResNet-18 and ResNet-50 exceed 94.00% when the SNR value is 6 dB, and the accuracies gradually decrease as the noise intensity increases. For example, the fault diagnosis accuracy of ResNet-50 only reaches 76.96% when the SNR value is −2 dB. In addition, in the case where the SNR = 2, LeNet-5 only achieves an average test accuracy of 56.53%. In contrast, the fault diagnosis accuracy of the MgHNN under strong noise is higher than that of the above models. For example, in the case where the SNR = 10, MgHNN-T and MgHNN-B achieve 96.31% and 96.96% average test accuracies, respectively, which are better than those of the other models. Moreover, when the SNR value is 2 dB, MgHNN-B can achieve an average test accuracy of 91.24%. The above experimental results show that the proposed model has stronger model robustness under strong noise interference.

VI. CONCLUSION

In this paper, a simple yet efficient multigrained hybrid neural network, i.e., MgHNN, is proposed for fault feature extraction, whose basic building block (MFR) consists of: 1) a split block that splits input image patches into n patch subsets; 2) a short connection block that can effectively represent local and global features. The proposed model is applied to diagnose the experimental bearing and gearbox datasets. Fault classification results on the PU and SEU datasets show that the MgHNN achieves a good accuracy/complexity trade-off with various architectures and can effectively deal with gearbox fault diagnosis under different operating conditions. Ablation studies obtained on the PU dataset demonstrate that the DWwave block plays an active role in transferring and aggregating features more efficiently. Moreover, experimental results in the case of Gaussian white noise interference indicate that the MgHNN is more robust to large range noise than other four methods based on DL. Especially, when SNR value is 2dB, the MgHNN-B can achieve 91.24% average test accuracy. In the future work, we are going to explore the performance of our framework in fault diagnosis. For example, MFR block proposed in this paper can be embedded into the vision transformer model to extract and classify fault features from time-frequency domain data.

REFERENCES

[1] X. Zhao, M. Jia and Z. Liu, “Semisupervised Deep Sparse Auto-Encoder With Local and Nonlocal Information for Intelligent Fault Diagnosis of Rotating Machinery,” *IEEE Instrum. Meas. Mag*., vol. 70, pp. 1-13, Aug. 2021.

[2] Z. Li, T. Zheng, Y. Wang, Z. Cao, Z. Guo and H. Fu, “A Novel Method for Imbalanced Fault Diagnosis of Rotating Machinery Based on Generative Adversarial Networks,” *IEEE Instrum. Meas. Mag*., vol. 70, pp. 1-17, Jul. 2021.

[3] Z. Huo, M. Martínez-García, Y. Zhang and L. Shu, “A Multisensor Information Fusion Method for High-Reliability Fault Diagnosis of Rotating Machinery,” *IEEE Instrum. Meas. Mag*., vol. 71, pp. 1-12, Dec. 2022.

[4] Z. Zhao, S. Wu, B. Qiao, S. Wang and X. Chen, “Enhanced Sparse Period-Group Lasso for Bearing Fault Diagnosis,” *IEEE Trans. Ind. Electron*., vol. 66, no. 3, pp. 2143-2153, Mar. 2019.

[5] S. Wang, X. Chen, C. Tong and Z. Zhao, “Matching Synchrosqueezing Wavelet Transform and Application to Aeroengine Vibration Monitoring,” *IEEE Instrum. Meas. Mag*., vol. 66, no. 2, pp. 360-372, Feb. 2017.

[6] H. Habbouche, Y. Amirat, T. Benkedjouh and M. Benbouzid, “Bearing Fault Event-Triggered Diagnosis Using a Variational Mode Decomposition-Based Machine Learning Approach,” *IEEE Trans. Energy Convers*., vol. 37, no. 1, pp. 466-474, Mar. 2022.

[7] M. Ma, C. Sun and X. Chen, “Discriminative Deep Belief Networks with Ant Colony Optimization for Health Status Assessment of Machine,” *IEEE Instrum. Meas. Mag*., vol. 66, no. 12, pp. 3115-3125, Dec. 2017.

[8] S. Xie, Y. Li, H. Tan, R. Liu and F. Zhang, “Multi-scale and multi-layer perceptron hybrid method for bearings fault diagnosis,” *Int. J. Mech. Sci*., vol. 235 p.107708, Dec. 2022.

[9] Y. -L. He, K. Li, N. Zhang, Y. Xu and Q. -X. Zhu, “Fault Diagnosis Using Improved Discrimination Locality Preserving Projections Integrated With Sparse Autoencoder,” *IEEE Instrum. Meas. Mag*., vol. 70, pp. 1-8, Nov. 2021.

[10] S. Ye, J. Jiang, J. Li, Y. Liu, Z. Zhou and C. Liu, “Fault Diagnosis and Tolerance Control of Five-Level Nested NPP Converter Using Wavelet Packet and LSTM,” *IEEE Trans. Power Electron*., vol. 35, no. 2, pp. 1907-1921, Feb. 2020.

[11] C. Zhu, Z. Chen, R. Zhao, J. Wang and R. Yan, "Decoupled Feature-Temporal CNN: Explaining Deep Learning-Based Machine Health Monitoring," *IEEE Instrum. Meas. Mag*., vol. 70, pp. 1-13, May. 2021.

[12] I. O. Tolstikhin *et al*., "Mlp-mixer: An all-mlp architecture for vision," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2021, vol.34, pp. 24261-24272.

[13] H. Touvron *et al*., "ResMLP: Feedforward Networks for Image Classification With Data-Efficient Training," *IEEE Trans. Pattern Anal. Mach. Intell*., Sep. 2022, pp. 1-9, doi: 10.1109/TPAMI.2022.3206148.

[14] D. Lian, Z. Yu, X. Sun and S. Gao, "As-mlp: An axial shifted mlp architecture for vision," 2021, *arXiv:2107.08391*. [Online]. Available: http://arxiv.org/abs/2107.08391

[15] Y. Tang, K. Han, J. Guo, C. Xu, Y. Li, C. Xu and Y. Wang, "An Image Patch is a Wave: Phase-Aware Vision MLP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., Jun. 2022, pp. 10925-10934.

[16] Y. M. Yeap and A. Ukil, "Fault detection in HVDC system using Short Time Fourier Transform," in *IEEE Power and Energy Society General Meeting*, pp. 1-5, Jul. 2016.

[17] S. -H. Gao, M. -M. Cheng, K. Zhao, X. -Y. Zhang, M. -H. Yang and P. Torr, "Res2Net: A New Multi-Scale Backbone Architecture," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 43, no. 2, pp. 652-662, 1 Feb. 2021.

[18] Z. Chen and W. Li, "Multisensor Feature Fusion for Bearing Fault Diagnosis Using Sparse Autoencoder and Deep Belief Network," *IEEE Instrum. Meas. Mag*., vol. 66, no. 7, pp. 1693-1702, Jul. 2017.

[19] X. Pei, X. Zheng and J. Wu, "Rotating Machinery Fault Diagnosis Through a Transformer Convolution Network Subjected to Transfer Learning," *IEEE Instrum. Meas. Mag*., vol. 70, pp. 1-11, Oct. 2021.

[20] A Vaswani *et al*., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2017, pp. 6000–6010.

[21] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen and J. Wang, "Machine Health Monitoring Using Local Feature-Based Gated Recurrent Unit Networks," *IEEE Trans. Ind. Electron*., vol. 65, no. 2, pp. 1539-1548, Feb. 2018.

[22] T. Li *et al*., "WaveletKernelNet: An Interpretable Deep Neural Network for Industrial Intelligent Diagnosis," *IEEE Trans. Syst. Man Cybern*., vol. 52, no. 4, pp. 2302-2312, Apr. 2022.

[23] X. Zhao *et al*, "Intelligent Fault Diagnosis of Gearbox Under Variable Working Conditions With Adaptive Intraclass and Interclass Convolutional Neural Network," *IEEE Trans Neural Netw Learn Syst*., pp. 1-15, Jan. 2022, doi: 10.1109/TNNLS.2021.3135877.

[24] J. Sun, J. Wen, C. Yuan, Z. Liu and Q. Xiao, "Bearing Fault Diagnosis Based on Multiple Transformation Domain Fusion and Improved Residual Dense Networks," *IEEE Sens. J*., vol. 22, no. 2, pp. 1541-1551, Jan. 2022.

[25] Z. Zhao, T. Li, J. Wu, C. Sun, S. Wang, R. Yan and X. Chen, "Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study," *ISA Trans*., vol. 107, pp. 224-255, Dec. 2020.

[26] L. Chen, N. Qin, X. Dai, and D. Huang, "Fault diagnosis of high-speed train bogie based on capsule network," *IEEE Trans. Instrum. Meas*., vol. 69, no. 9, pp. Sep. 2020.

[27] Z. Zhu, G. Peng, Y. Chen, and H. Gao, "A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis," *Neurocomputing*, vol. 323, pp. 62–75, 2019.

[28] C. Li, L. Mo and R. Yan, "Fault Diagnosis of Rolling Bearing Based on WHVG and GCN," *IEEE Instrum. Meas. Mag*., vol. 70, pp. 1-11, Jun. 2021.

[29] T. Li, Z. Zhou, S. Li, C. Sun, R. Yan, and X. Chen, "The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study," *Mech. Syst. Signal Process*., vol. 168, Apr. 2022, Art. no. 108653.

[30] A Dosovitskiy *et al*., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*. [Online]. Available: http://arxiv.org/abs/2010.11929

[31] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve and H. Jégou, "Going deeper with image transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., pp. 32-42, Oct. 2021.

[32] Y. Rao, W. Zhao, Z. Zhu, J. Lu and J. Zhou, "Global filter networks for image classification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, pp. 980-993, Dec. 2021.

[33] H. Liu, Z. Dai, D. So and Q. V. Le, "Pay attention to mlps," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, pp. 9204-9215, Dec. 2021.

[34] T. Yu, X. Li, Y. Cai, M. Sun and P. Li, "Rethinking token-mixing mlp for mlp-based vision backbone," 2021, *arXiv:2106.14882*. [Online]. Available: http://arxiv.org/abs/2106.14882

[35] S. Chen, E. Xie, C. Ge, D. Liang and P. Luo, "CycleMLP: A MLP-like architecture for dense prediction," 2021, *arXiv:2107.10224*. [Online]. Available: http://arxiv.org/abs/2107.10224

[36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. -C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., pp. 4510-4520, Jun. 2018.

[37] J. Guo *et al*., "CMT: Convolutional Neural Networks Meet Vision Transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., pp. 12165-12175, Jun. 2022.

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, vol. 15, pp. 448-456, Jul. 2015.

[39] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. Conf. Comput. Vis. Pattern Recognit*., Honolulu, HI, USA, Jul. 2017, pp. 1800–1807.

[40] 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 *PHM Society European Conference*. vol. 3, no. 1, pp. 05–8. Jul. 2016.

[41] S. Shao, S. McAleer, R. Yan and P. Baldi, "Highly Accurate Machine Fault Diagnosis Using Deep Transfer Learning," *IEEE Trans. Ind. Inform*., vol. 15, no. 4, pp. 2446-2455, Apr. 2019.

[42] A. Paszke *et al*., "Automatic differentiation in pytorch," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2017, pp. 1-4.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[44] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.

[45] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, Jun. 2017.

[46] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., pp. 770-778, Jun. 2016.