IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 May 2024, accepted 6 June 2024, date of publication 11 June 2024, date of current version 21 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3412050

RESEARCH ARTICLE

An Improved Lightweight Variant of EfficientNetV2 Coupled With Sensor Fusion and Transfer Learning Techniques for Motor Fault Diagnosis

LIANG JIANG[1], SICHENG ZHU[2], AND NING SUN[1]
[1]School of Automation, Wuxi University, Wuxi, Jiangsu 214105, China
[2]School of Automation, Nanjing University of Information Science and Technology, Nanjing, Jiangsu 210044, China

Corresponding author: Ning Sun (001764@cwxu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 42275156; in part by the "Qing Lan Project" of Jiangsu Universities, the Ministry of Education's Industry-University Cooperation Collaborative Breeding Program under Grant 202102224006; and in part by the Innovation and Entrepreneurship Program for Jiangsu University Students under Grant 202213982017Z.

ABSTRACT Although deep learning methods based on single sensors are widely applied in fault diagnosis, leveraging multi-sensor data to learn useful information remains a challenge. To fully utilize multi-sensor information, this paper proposes a lightweight improvement of the EfficientNetV2 architecture, combined with sensor fusion technology and transfer learning techniques, to develop an efficient and reliable new method specifically for motor fault diagnosis. First, the continuous wavelet transform is utilized to convert the signals from various sensors into time-frequency images, and the Mallat algorithm is employed to decompose each image into sub-band coefficients at different levels. Secondly, a fusion reconstruction method is constructed using coefficient absolute maximum and weighted average fusion rules to integrate the sub-band coefficients of multi-sensor time-frequency images at different levels. Subsequently, EfficientNetV2 is improved to enhance the model's feature extraction capabilities, computational efficiency, and achieve lightweight effects. The EfficientNetV2-M0 network modifies the model's depth and width multiplicity factors, reducing parameters and computational complexity. Furthermore, this network incorporates Diverse Branch Block (DBB) and Multidimensional Collaborative Attention (MCA) to enhance feature extraction under complex backgrounds, and the maximum cross-entropy loss function is improved by using label smoothing and focal loss to dynamically adjust the classification weights for improved accuracy. The network leverages pre-trained models obtained through transfer learning techniques for deployment, combining multi-sensor information fusion and the improved lightweight model for fault diagnosis applications. Finally, a fault diagnosis experiment is conducted using a motor state dataset. The experimental results demonstrate that the proposed method outperforms the control method in terms of diagnostic performance and robustness, with an accuracy of 100%, and it exhibits excellent performance even under conditions of small sample data, with an accuracy of 98.81%.

INDEX TERMS Fault diagnosis, sensor fusion, EfficientNetV2-M0,transfer learning.

I. INTRODUCTORY

During motor operation, faults can negatively impact normal functioning. Physical quantities such as current, temperature, and vibration can be used to characterize motor performance and operating conditions when faults occur. The field of

The associate editor coordinating the review of this manuscript and approving it for publication was Mark Kok Yew Ng.

motor fault diagnosis has seen increased use of multi-sensor fusion methods with the continuous development of artificial intelligence algorithms [1].

Data fusion can be divided into three strategies: data-level, feature-level, and decision-level fusion, depending on the level of abstraction. Data-level fusion has high communication overhead and processing complexity, but it can usually avoid information loss. On the contrary, feature-level and decision-level fusion reduce communication overhead and processing complexity, but may also result in information loss. In their case study, Xia et al. [2] fused multi-channel data from sensors of the same type into a high-order tensor array using sensor fusion techniques. They then extracted design features from this array to detect process faults. Suawa et al. [3] presented a method for integrating different types of multi-sensor data using data-level fusion. They used deep learning models, such as deep convolutional neural networks (DCNN) and long short-term memory (LSTM), for fault classification. Huang and Lee [4] utilized raw data from multiple sensors and extracted features using the convolutional operator commonly used in convolutional neural networks (CNNs). They then performed feature-level fusion to estimate device defects. Previous research has primarily focused on time domain data fusion using raw data, making it difficult to incorporate frequency domain information into the fusion process. This can lead to uncertainty in motor fault diagnosis.

Novel image processing techniques offer alternative methods for fusing sensor data [5]. These techniques can be subdivided into pixel-level fusion, decision-level fusion, and feature-level fusion [6]. Pixel-level fusion is a less computationally expensive technique that relies solely on choosing between different corresponding pixel values of multiple image data based on mathematical formulas or algorithms to avoid information loss. Pixel-level fusion techniques are commonly used in medical image processing [7] and can significantly improve the identification of defects in additive manufacturing by fusing computed tomography images obtained from multimodal sensors, such as optical and acoustic emission [8]. Wang et al [9] proposed an image and sensor fusion technique using a single type of sensor to convert the raw data into a grayscale image, and the sensor channel data are spliced to form a grayscale composite image. Decision-level fusion and feature-level fusion are two equally important concepts in image fusion techniques. In decision-level fusion, each sensor initially processes the data independently and generates its own decision or result. These decisions or results are then fused by some rule or algorithm to produce the final decision or result. Feature-level fusion belongs to the intermediate level, which first extracts representative features from the raw observations provided by each sensor and then fuses these features into a single feature vector. However, these two fusion methods share common disadvantages. Firstly, the computational complexity of the system is high due to the necessity for each sensor to perform independent data processing and decision-making. Secondly, the fusion of the two fusion methods necessitates the development of suitable fusion rules or algorithms to ensure that the decisions or results between different sensors can be effectively fused. This requires a significant amount of experimental and debugging work, which increases the difficulty of system development. Finally, the decision-level fusion process is contingent upon the existence of a high degree of correlation between the sensors. In the event that the correlation between the sensors is insufficient, the fusion process may be adversely affected. In a separate work, Wang et al. [10] proposed a feature-level fusion of time and frequency domain signals using a deep learning method. They introduced an attention mechanism to improve the classification accuracy in motor bearing fault diagnosis.

The technique of transfer learning (TL) [11] is an important approach for dealing with sample data in motor fault diagnosis. It improves model performance by transferring knowledge from the same or related source domains to the target domain [12]. Liu et al. [13] used transfer learning with less data and improved accuracy by 12%. They utilized a CNN model to diagnose faults in the energy system of a building chiller. Mao et al. [14] demonstrated the effectiveness of the migration learning technique in bearing fault diagnosis using a CNN based VGG-16 model pre-trained on a large scale image dataset called ImageNet. Lee et al. [15] acquired short-time Fourier transform (STFT) images of a uniaxial vibration channel. They combined the time-frequency domain of the vibration signals and used the migration learning technique. The resulting image was input into the VGG-19 model for fault diagnosis of gravity acceleration devices. The study demonstrates the advantages of migration learning. The work by Eyup [16] was referenced. The short-time Fourier transform (STFT) is used to convert multi-sensor data into spectrograms, which are then fused at the pixel level. The resulting fused image is input into the Alexnet model for fault diagnosis of electric motors using the migration learning technique. The results demonstrate that the proposed sensor fusion technique can effectively classify motor faults. Additionally, the pre-trained transfer learning (TL) model enhances the training capability of the model.

This paper proposes a lightweight improvement to the EfficientNetV2 architecture and introduces a novel, efficient, and reliable method specifically designed for motor fault diagnosis, which integrates sensor fusion technology and transfer learning. Initially, the continuous wavelet transform is utilized to convert signals from various sensors into time-frequency images, and the Mallat algorithm is employed to decompose each image into sub-band coefficients at different levels. Secondly, fusion rules based on the maximum absolute value method and weighted average method are constructed to integrate and reconstruct the sub-band coefficients from multiple sensor time-frequency images at various levels. Subsequently, improvements are made to the Efficient-NetV2 model to enhance its feature extraction capability,

computational efficiency, and achieve a lightweight effect. The EfficientNetV2-M0 network optimizes the model's depth and width scaling factors, reducing parameters and computational complexity. Furthermore, the network integrates Diverse Branch Block (DBB) and Multidimensional Collaborative Attention (MCA) to enhance feature extraction under complex backgrounds. The maximum cross-entropy loss function is also improved through the application of label smoothing and focal loss, dynamically adjusting classification weights to improve accuracy. The pre-trained model obtained through transfer learning is deployed, combining multi-sensor information fusion with the improved lightweight model for fault diagnosis applications. The principal contributions of this paper are as follows:

(1) The fusion of signals from heterogeneous multimodal sensors enables the comprehensive analysis of anomaly features present in the fusion data obtained from multiple sources.
(2) The proposed model, EfficientNetV2-M0 Transfer Learning, combines the multiplicity factor of the lightweight EfficientNetV2-B0 model with the network structure of EfficientNetV2-M. This model has a low number of model covariates and is designed to learn representative fault features from multisensor fusion data for fault identification.
(3) To address the scarcity of real fault samples, this paper introduces the Diverse Branch Block (DBB) for structural repair in EfficientNetV2-M0. The objective is to improve the feature extraction capability of the backbone network.
(4) An efficient Multidimensional Collaborative Attention (MCA)Â mechanism was introduced in EfficientNetV2-M0 to improve feature learning capability and focus attention on more sensitive features.
(5) The original cross-entropy loss function was improved using Label Smooth and Focal loss to enhance the weight and model generalization ability of hard-to-classify samples.

The paper is organized as follows: Section II describes the proposed method in detail. Section III outlines the experimental setup and data acquisition system, presents the experimental results and dataset analysis, and Section IV provides a summary of the paper and future work.

II. BASIC THEORY

In this section, relevant theoretical and background information is presented. Firstly, the multi-sensor image fusion algorithm based on wavelet transform is introduced, and then the basics of the EfficientNetV2-M network model are described in detail.

A. MULTI-SENSOR IMAGE FUSION ALGORITHM BASED ON WAVELET TRANSFORM

The time and frequency domain characteristics of the sensor data are of great importance for the fault diagnosis of the motor. The time domain characteristics are primarily concerned with the change of the signal on the time axis, including the statistical characteristics of the signal such as the mean, variance, magnitude, waveform, timing relationship, and so forth. Moreover, the effect of faults on the signal is dominated by a few major frequency components, which are significant for monitoring purposes [17]. Consequently, in practical applications, time-domain and frequency-domain characterization are often employed in conjunction to gain a more comprehensive understanding of the signal characteristics. In this paper, the continuous wavelet transform, as depicted in Equation (1), is utilized to process the sensor data, thereby yielding the corresponding two-dimensional time-frequency image.

$$\begin{cases} WT_f(\alpha,\tau) = < f(t), \Psi_{\alpha,\tau}(t) > = \dfrac{1}{\sqrt{\alpha}} \displaystyle\int_R f(t)\Psi^* \left(\frac{t-\tau}{\alpha}\right) dt \\ \Psi_{\alpha,\tau}(t) = \dfrac{1}{\sqrt{\alpha}} \Psi \left(\dfrac{t-\tau}{\alpha}\right), \quad \alpha > 0, \tau \in R \end{cases} \tag{1}$$

where α is the translation factor, determining the position of the time-frequency window in the time domain; $\boldsymbol{\tau}$ is the scale factor, determining the size of the time-frequency window and its position in the frequency domain; as shown in Eq. (1), $\Psi_{\alpha,\tau}(t)$ is the wavelet basis function (also known as the mother wavelet) after the change of the translation and the scale, and the wavelet transform can automatically adjust the factors $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ through the characteristics of the signals, so that the wavelet transform of the vibration signals and current signals of the motor at different intervals of time has adaptive and multi-resolution characteristics. resolution characteristics. Figure 1 shows the time-frequency image transformation of the X-axis vibration sensor under the normal state of the motor.

To comprehensively obtain feature information of the measured object, it is necessary to fuse the two-dimensional time-frequency images of the heterogeneous multimodal sensor data. This can be achieved by processing the images with continuous wavelet transform. This paper utilizes the Mallat algorithm to decompose the low-frequency and high-frequency components of the wavelet transform. Different criteria are selected for the fusion of the low-frequency and high-frequency features, resulting in the final fused image obtained through the wavelet inverse transform [18], [19], [20].

The Mallat algorithm is a frequently utilized decomposition algorithm in wavelet transform. The primary process of the Mallat image decomposition algorithm and fusion rule is shown in Algorithm 1.

The following fusion rules are applicable to the low-frequency sub-band and high-frequency sub-band coefficients of the image: (1) Coefficient absolute maximum method: This fusion rule is suitable for high-frequency components of the source image that are richer, with higher brightness and contrast. (2) Weighted average method: This method allows for adjustable weight coefficients, a wide range of application, and the elimination of some noise, thus

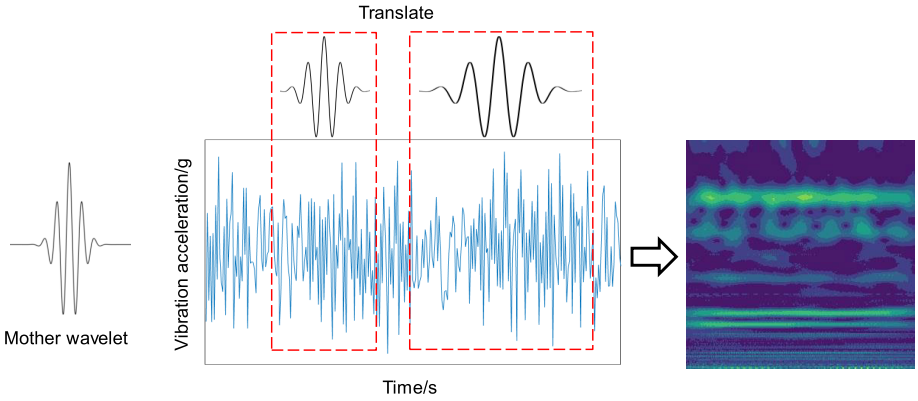

FIGURE 1. Time-frequency image conversion of the *X*-axis vibration sensor in the normal state of the motor.

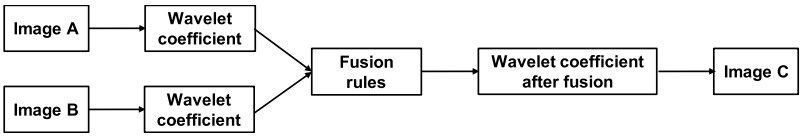

FIGURE 2. Flow of multi-sensor image fusion algorithm based on wavelet transform.

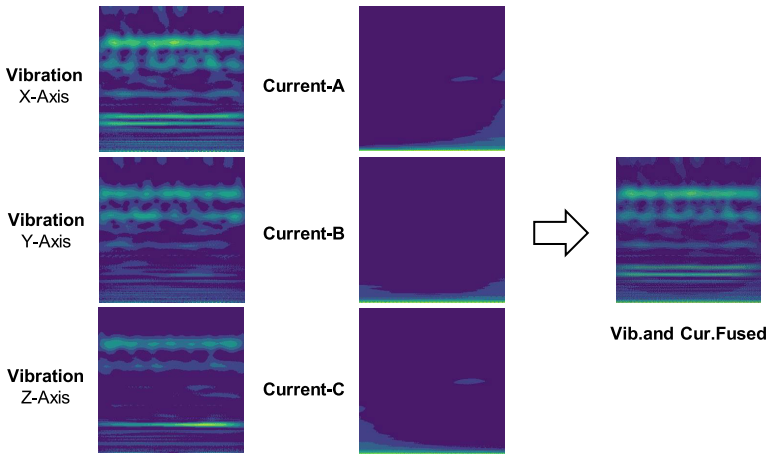

FIGURE 3. Multi-source heterogeneous signal fusion results.

reducing the loss of source image information. The fusion strategy employed in this study is the weighted average method for low-frequency images and the absolute maximum coefficient method for high-frequency images.

The reconstruction process of the 2D Mallat algorithm is shown in Algorithm 2.

The original images to be fused are A and B, and the fused image is C. The steps for image fusion are as follows:

(1) The original images A and B undergo wavelet multi-scale decomposition using the two-dimensional Mallat algorithm to obtain their low-frequency and high-frequency subband coefficients. The low-frequency images are fused using the weighted average method, while the high-frequency images are fused using the method of absolute maximum value of coefficients.

(2) Different fusion rules are applied to the subband coefficients of the image at different levels.

(3) The fused image C is obtained by reconstructing the fused wavelet coefficients using the wavelet inverse transform for the 2D Mallat algorithm.

The flowchart for the multi-sensor image fusion algorithm based on wavelet transform is shown in Figure 2.

For multiple sources of heterogeneous signals (e.g., three-axis vibration and three-phase current signals) acquired by motors using heterogeneous multimodal sensors, the results of fusion using the above process are shown in Figure 3 after processing each signal using continuous wavelet transform.

B. EFFICIENTNETV2-M

EfficientNetv2 [21] is a network model proposed by Quoc V. Le et al. in 2021 that is smaller and faster. It is mainly composed of an inverted fused residual layer (Fuse-MBConv) and an inverted linear bottleneck layer superimposed with a deeply differentiable convolution (MBConv).

Algorithm 1 Image Decomposition and Sub-Image Fusion

1: A two-dimensional image ($f(x, y)$) is subjected to a two-dimensional wavelet filter comprising a low-pass filter ($h(x, y)$) and a high-pass filter ($g(x, y)$) in the horizontal and vertical directions.
2: **repeat**
3: The original image ($f(x, y)$) is convolved with a low-pass filter ($h(x, y)$) in the horizontal direction, and then the convolution result is subjected to a downsampling operation to obtain a low-frequency sub-image ($f_{LL}(x, y)$) in the horizontal direction.
4: The original image ($f(x, y)$) is convolved with a low-pass filter ($h(x, y)$) in the vertical direction, and then the convolution result is subjected to a downsampling operation to obtain a low-frequency sub-image ($f_{HH}(x, y)$) in the vertical direction.
5: The original image ($f(x, y)$) is convolved with a high-pass filter ($g(x, y)$) in the horizontal direction, and then the convolution result is subjected to a downsampling operation to obtain a high-frequency sub-image ($f_{HL}(x, y)$) in the horizontal direction.
6: The original image ($f(x, y)$) is convolved with a high-pass filter ($g(x, y)$) in the vertical direction, and then the convolution result is subjected to a downsampling operation to obtain a high-frequency sub-image ($f_{LH}(x, y)$) in the vertical direction.
7: **until** The requisite number of decomposition layers has been achieved.
8: The low-frequency sub-images are fused using a weighted averaging method, while the high-frequency sub-images are fused using the maximum absolute coefficient method.

Output: Sub-image after fusion

Algorithm 2 Reconstruction of the Mallat Algorithm

1: **repeat**
2: For low-frequency subimages, an upsampling operation should be performed, after which the image should be restored to its original size.
3: The upsampled low-frequency sub-images are convolved with the low-pass filters of the corresponding scales. Concurrently, the corresponding high-frequency sub-images (horizontal, vertical, and diagonal directions) are convolved with the corresponding high-pass filters, respectively. The convolution results of the low-frequency subimages and the high-frequency subimages are then summed at the corresponding positions to obtain the reconstructed image at that scale.
4: **until** The most original image size is reached.
5: Following a series of reconstruction iterations, the inverse wavelet transform result of the fused image can be obtained.

Output: Fused image

The Fused-MBConv structure [22] is shown in Figure 4. The image input undergoes a 3*3 standard convolution, and the output feature maps use a Stochastic Depth type Dropout layer [23]. When the step size is 1 and the input image of the module and the convolution output image are of the same shape, the residuals are used to connect the input and output. When the step size is 2 for the downsampling stage, the convolution output feature map is directly outputted. The Fused-MBConv module can be divided into two different architectures based on the channel expansion multiples. When the channel expansion multiples are not equal to 1, first use a 3×3 standard convolution to increase the channel number, and then use a 1×1 convolution to decrease the channel number. When the channel expansion exponent is equal to 1, use a 3×3 standard convolution directly.

Figure 5 illustrates the MBConv architecture [24], [25]. The image input undergoes a 1×1 convolutional layer to increase the number of channels. Then, deep convolution is applied to the high latitude space. The feature map data is optimized using the SE Attention Mechanism [25]. Finally, a 1×1 convolutional layer with a linear activation function is used to decrease the number of channels. The MBConv module can be divided into two different architectures depending on the step size of the deep convolution layer. If the step size of the deep convolutional layer is 1 and the input feature map has the same shape as the output feature map, a Stochastic Depth Dropout layer is added to the 1×1 convolutional dropout layer to prevent overfitting. Finally, the residuals are connected to the input and output. When the step size of the depth convolution layer is not equal to 1, the Dropout layer and residual connection are not utilized. Instead, the feature map is directly outputted after 1×1 convolution dimensionality reduction.

EfficientNetV2 is available in B0, B1, B2, B3, S, M, and L versions. Among them, EfficientNetV2-B0 is the lightest version, while EfficientNetV2-M has more parameters and better accuracy during training. Therefore, the EfficientNetV2-M network is chosen as the base model. The network structure of EfficientNetV2-M and EfficientNetV2-B0 is shown in Table 1.

The EfficientNetV2-M network consists of 9 stages. Stage 0 is an ordinary convolutional layer with a 3×3 kernel size and a step size of 2, which includes the BN and Swish activation functions. Stages 1 to 3 are repetitions of the Fused-MBConv structure, while Stages 4 to 7 are repetitions of the MBConv structure. Stage 8 comprises a 1×1 ordinary convolutional layer, an average pooling layer, and a fully connected layer. In the Fused-MBConv structure, Fused-MBConv1 and Fused-MBConv4 refer to the first 1×1 convolutional layer expanded with the number of eigenchannels of the input matrix by a factor of 1 and 4, respectively. The size of the convolution kernel used by Depthwise Conv in Fuse-MBConv is denoted by

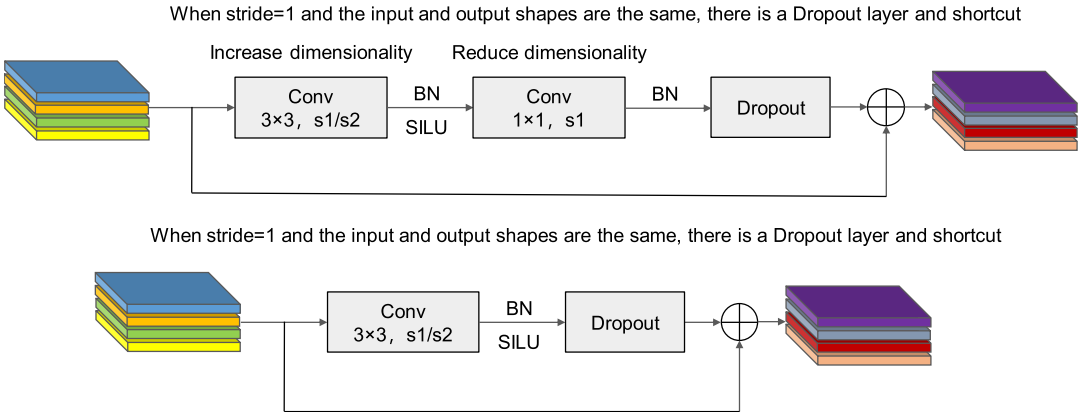

FIGURE 4. Fused-MBConv structure.

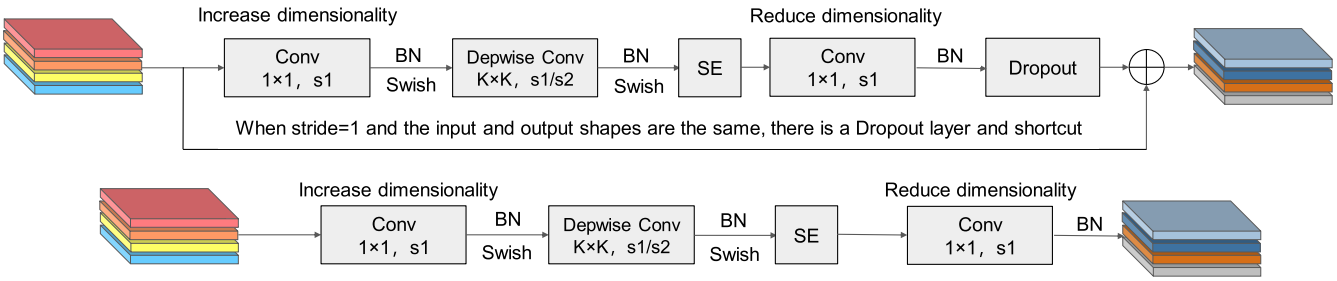

FIGURE 5. MBConv structure.

TABLE 1. EfficientNetV2-M and EfficientNetV2-B0 network structure table.

Stage	Operator	EfficientNetV2-M			EfficientNetV2-B0		
		Stride	Channels	Layers	Stride	Channels	Layers
0	Conv3x3	2	24	1	2	16	1
1	Fused-MBConv1, k3x3	1	24	3	1	16	1
2	Fused-MBConv 4, k3x3	2	48	5	2	32	2
3	Fused-MBConv4, k3x3	2	80	5	2	48	2
4	MBConv4, k3x3, SE0.25	2	160	7	2	96	3
5	MBConv6, k3x3, SE0.25	1	176	14	1	112	3
6	MBConv6, k3x3, SE0.25	2	304	18	2	192	4
7	MBConv6, k3x3, SE0.25	1	512	5	-	-	-
8	Conv1x1 & Pooling & FC	-	2560	1	-	960	1

k3 × 3 or. MBConv4 and MBConv6 refer to the initial 1 × 1 convolutional layer in the MBConv structure, which is expanded by a factor of 4 and 6, respectively, based on the number of feature channels in the input matrix. 'Channels' refers to the number of feature channels, which represents the number of channels in the output matrix after passing through this stage. 'SE0' is unclear and requires further context. The Squeeze-and-Excitation (SE) module has 25 channels, with the number of nodes in the first fully-connected layer being 1/4 of the number of channels input to the MBConv module. The modules are repeated a certain number of times, as indicated by the Layers.

C. IMPROVED EFFICIENTNETV2-M0

1) OPTIMIZATION OF THE MULTIPLICITY FACTOR

EfficientNetV2 utilizes the Neural Architecture Search (NAS) technique to optimize the configuration of three network parameters: image input resolution (r), network depth, and channel width. This optimization improves network

TABLE 2. EfficientNetV2-M0 network structure table.

Stage	Operator	Stride	Channels	Layers
0	Conv3x3	2	16	1
1	Fused-MBConv 1, k3x3	1	16	1
2	Fused-MBConv4, k3x3	2	32	2
3	Fused-MBConv4, k3x3	2	48	2
4	MBConv4, k3x3, SE0.25	2	96	3
5	MBConv6, k3x3, SE0.25	1	112	3
6	MBConv6, k3x3, SE0.25	2	192	4
7	MBConv6, k3x3, SE0.25	1	320	1
8	Conv1x1 & Pooling & FC	-	1600	1

performance by modifying these parameters. To adjust the width of the network, the number of convolutional kernels must be adjusted, which will also change the output feature matrix of the channels. Similarly, adjusting the depth of the network means adjusting the number of times each stage repeatedly stacks the network structure.

EfficientNetV2-B0 has a multiplicity factor of 1.0 on both the channel dimension and depth, resulting in a significant reduction in the number of parameters and computational complexity. Its network structure consists of only 8 layers, which is one less stage than EfficientNetV2-M's MBConv network structure. As a result, it has lower accuracy than EfficientNetV2-M in training experiments. In the training experiment, EfficientNetV2-B0 was found to have lower accuracy than EfficientNetV2-M. This is due to the fact that the multiplicity factor on the channel dimension in EffcientNetV2-M is 1.4, and the multiplicity factors on the depth are all 1.8, resulting in a larger and more complex model. To fully utilize the advantages of the EfficientNetV2-M network, the multiplicity factor in B0 was compared to that of the EfficientNetV2-M network structure. The combination of the multiplicity factor in B0 with the network structure of EfficientNetV2-M results in the improved EfficientNetV2-M0 network. This network enhances the model's accuracy while reducing the number of parameters and computational complexity. The improved network structure is illustrated in Table 2.

It is noteworthy that the pooling layer utilized by the enhanced EfficientV2-M0 network is the average pooling layer. This is due to the fact that the average pooling structure of the original EfficientNetV2 network exhibited the most optimal performance among the various pooling layers [26] (e.g., the maximum pooling layer, the AAD pooling layer, and the average pooling layer). Consequently, the average pooling structure of the original network was adopted.

2) INTRODUCTION OF THE DBB MODULE FOR STRUCTURAL REPARAMETERIZATION

To enhance the feature extraction capability of EfficientNetV2-M0, this paper introduces the Diverse Branch Block (DBB) [27] to the Fused-MBConv structure, achieving structure reparameterization. The DBB module is an innovative approach to over-parameterized convolution, following in the footsteps of ACNet [28] and RepVGG [29]. It combines Inception's multi-branching and multi-scale concepts with over-parameterization ideas to create the DBB module proposed in this paper. The Diverse Branch Block (DBB) utilizes a complex multi-branching microstructure while maintaining the overall network structure during training. This allows for efficient inference or deployment. The DBB's complex structure can be converted into a single convolution during inference, resulting in minimal loss of accuracy and reduced inference time. DBB can be directly embedded into any existing architecture as an equivalent embedding module. This reflects the diversity and flexibility of over-referencing modules and can significantly improve the feature extraction capability of various backbone feature extraction networks.

The DBB module comprises six transformations: branch-add combining, depth-splicing combining, multi-scale operations, mean pooling, and convolutional sequences. During the inference/deployment phase, it also includes multi-branch module merging, such as conv layer and BN layer merging, branch merging, convolutional sequence merging, depth splicing merging, mean pooling transformation, and multi-scale convolutional transformation. Figure 6 shows the structure of the six transformations of the DBB module. Among them, K represents the convolutional kernel size, and 1×1 represents a convolutional kernel of size 1×1.

The introduction of the DBB module solves the problem of slower inference caused by increased network width. The DBB module uses the above six transformations, and its structure for model transformation at deployment/inference time is shown in Figure 7.

In this paper, the DBB module is added to the Fused-MBConv structure to improve accuracy and ensure faster inference speed. The improved Fused-MBConv structure is shown in Figure 7.

3) INTRODUCING THE MCA ATTENTION MECHANISM

Extracting effective fault features without noise interference is a major concern in academia. The attention mechanism has been identified as a means to increase the acceptance of potential features through the attention graph. This is an effective way to suppress irrelevant information and enhance representative features. The EfficientNetV2-M0 network employs the SE module as its attention mechanism, which comprises a global average pooling and two fully connected layers, as illustrated in Figure 8. The first fully connected layer has a number of nodes equal to 1/4 of the input to the MBConv feature matrix channels and uses the Swish activation function. The second fully connected layer has several nodes equal to the output feature matrix channels of the Depthwise Conv layer and uses the Sigmoid activation function. The purpose of the two fully connected layers in

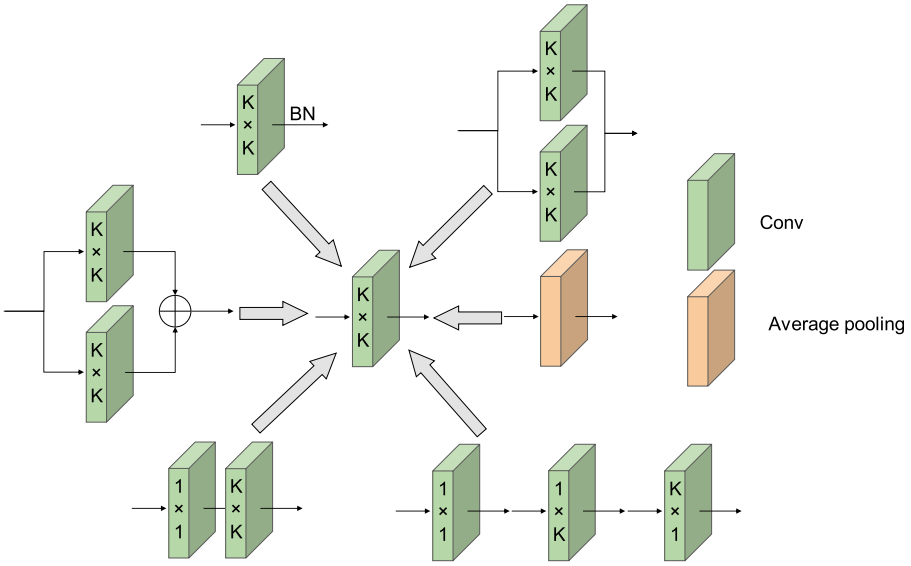

FIGURE 6. DBB module conversion structure.

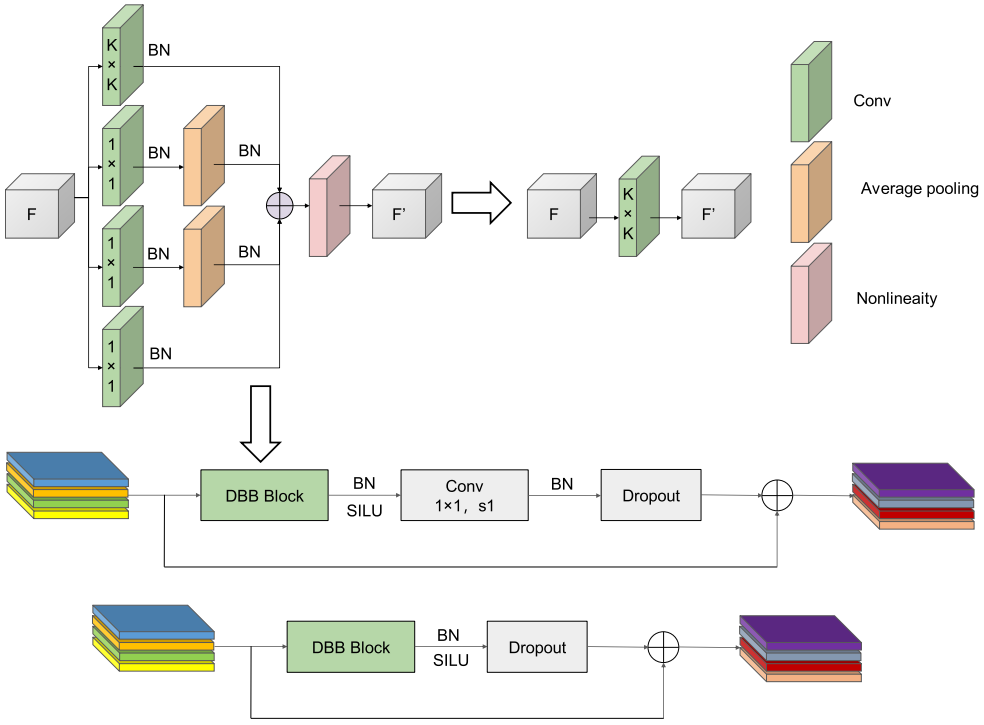

FIGURE 7. DBB model conversion and improved Fused-MBConv structure.

this module is to capture nonlinear cross-channel interactions. These interactions produce channels that do not correspond to weights due to dimensionality reduction, resulting in a loss of feature details. Capturing the dependencies between all channels is inefficient and unnecessary.

To enhance the attention mechanism of the original EfficientNetV2, this paper [30] employs the Multidimensional Collaborative Attention (MCA) mechanism. MCA is a lightweight and efficient attention mechanism that improves the representation of learned features and identifies objects of interest. It uses a three-branch architecture to model complementary attention in the channel, height, and width dimensions simultaneously. Figure 9 shows the MCA module, which comprises three branches. The left and middle branches capture feature interdependencies on the spatial dimensions W and H, respectively. The right branch

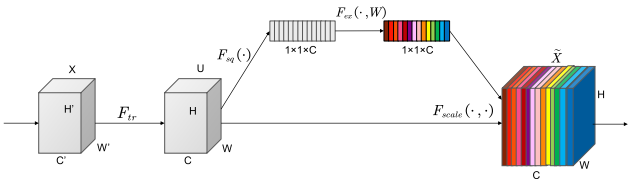

FIGURE 8. SE attention mechanism.

is used to capture inter-channel interactions. Finally, the outputs of the three branches are averaged and aggregated in an integration phase. The attention weights generated in different dimensions are recalibrated to derive the final refined feature map. The core components of MCA are constructed using the squeezing transform and the excitation transform, which improve the SE channel. In the squeezing transform, an adaptive combining mechanism is developed to merge the global mean and standard deviation pool features, enhancing the representation of feature descriptors. The excitation transform captures local feature interactions in a lightweight manner, rather than using inefficient dimensionality reduction strategies in SE, to overcome the trade-off between performance and computational overhead.

This paper also replaces the SE module in MBConv with the MCA module. the MCA module enables the feature map not to lose information due to the dimension reduction operation, and also the network is able to extract the image features more adequately. The structure of the improved MBConv module is shown in Figure 10.

4) LOSS FUNCTION IMPROVEMENT

During the training process, the time-frequency image presents a large target and stable shape features, which are easy to learn. However, the time-frequency signal generated during faults presents diverse features, smaller targets, and more difficult-to-train features. Additionally, the imbalance in the number of samples of different classes can also make it difficult for the model to learn other features, which can affect the direction of the gradient update of the loss function. This paper proposes the use of Label Smooth and Focal loss to enhance the original cross-entropy loss function. The Focal loss function reduces the weight of easy-to-classify examples and focuses on hard examples, improving the model's performance on difficult examples. Label Smooth improves the model's generalization ability and prevents overfitting during training by assigning a label coefficient with a higher probability for the target class and lower probabilities for other classes.

The true probability distribution is given as equation (2). The probability distribution after label smoothing is shown in equation (3).

$$P_i = \begin{cases} 1, & if\,(i = y) \\ 0, & if\,(i \neq y) \end{cases} \tag{2}$$

$$P_i = \begin{cases} (1-\varepsilon), & if\,(i = y) \\ \dfrac{\varepsilon}{K-1}, & \text{if } (i \neq y) \end{cases} \tag{3}$$

where K denotes the number of classifications and ε is a small hyperparameter, the updated distribution is equivalent to adding noise to the true distribution, which obeys a simple uniform distribution for ease of computation.

Focal Loss dynamically adjusts the weights of different classes of samples in the loss function, reduces the weight of easy-to-categorize samples, increases the weight of difficult-to-categorize samples, and alleviates the problem of sample imbalance. The loss function of the original network is the summation of the cross-entropy of each training sample, i.e., different classes have the same weight in the loss function. Equation (4) is as follows:

$$CE(p, y) = \begin{cases} -\log(p), & \text{if } y = 1 \\ -\log(1-p), & \text{otherwise} \end{cases} \tag{4}$$

p denotes the probability that the predicted sample belongs to 1 (in the range $0 - 1$), y denotes the label, and y takes the values $(+1, -1)$. Multiclassification and so on For convenience, p_t is used instead of p.

$$p_t = \begin{cases} p, & \text{if } y = 1 \\ 1-p, & \text{otherwise} \end{cases} \tag{5}$$

For simplicity of representation, the probability that the sample belongs to TRUE class is denoted by p_t. Therefore, equation (6) can be written as:

$$CE(p, y) = CE\,(p_t) = -\log\,(p_t) \tag{6}$$

The shared weight of positive and negative samples to the total loss is controlled by setting the value of α.

$$CE\,(p_t) = -\alpha_t \log\,(p_t) \tag{7}$$

To realize the control of the weights of easy and difficult-to-classify samples, the focal loss formula is as follows:

$$FL\,(p_t) = -\,(1-p_t)^{\gamma} \log\,(p_t) \tag{8}$$

where $(1-p_t)^{\gamma}$ is called the modulation factor. When p_t tends to 0, the modulation factor tends to 1 and contributes a lot to the total loss. When p_t tends to 1, the modulation coefficient tends to 0 and contributes little to the total loss.

D. TRANSFER LEARNING

Transfer Learning is a Machine Learning (ML) method that improves model performance by transferring knowledge from the same or related source domain to the target domain. Unlike training a model from scratch, a pre-trained model can be used to improve a specific target task [12].

For the formal definition: Let D define the domain of an ML classification problem that consists of two parts consisting of $D = \{X, P(X)\}$, where X denotes the feature space and P(X) denotes the marginal probability distribution. The feature vector $x = \{x_1, \cdots, x_n\} \in X$ is a specific element of the feature space and y is the corresponding class labeling belonging to the labeling space y. For domain D, the task can be defined as $T = Y, f()$, where $f(.)$ is the prediction function learned from $\{x_i, y_i\}$ pair. The source domain dataset

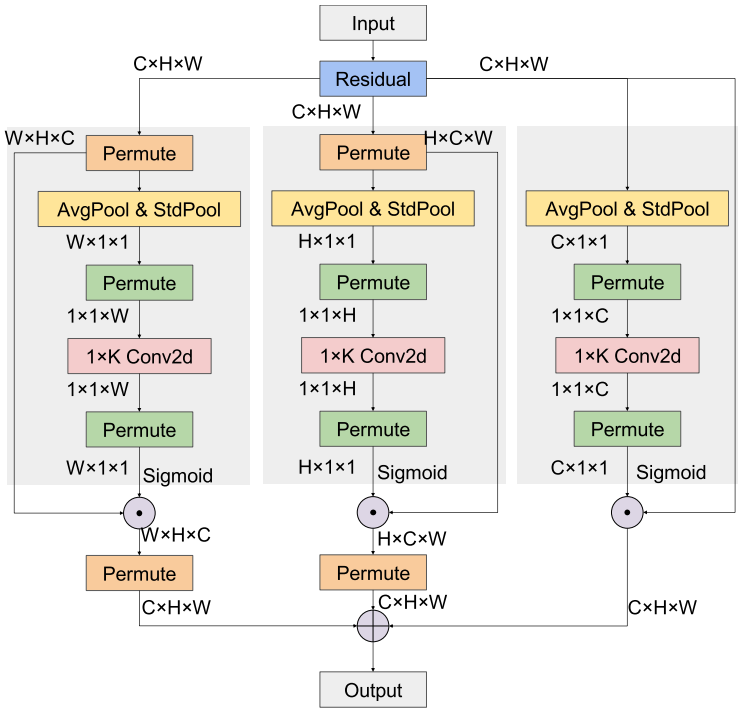

FIGURE 9. MCA attention mechanism.

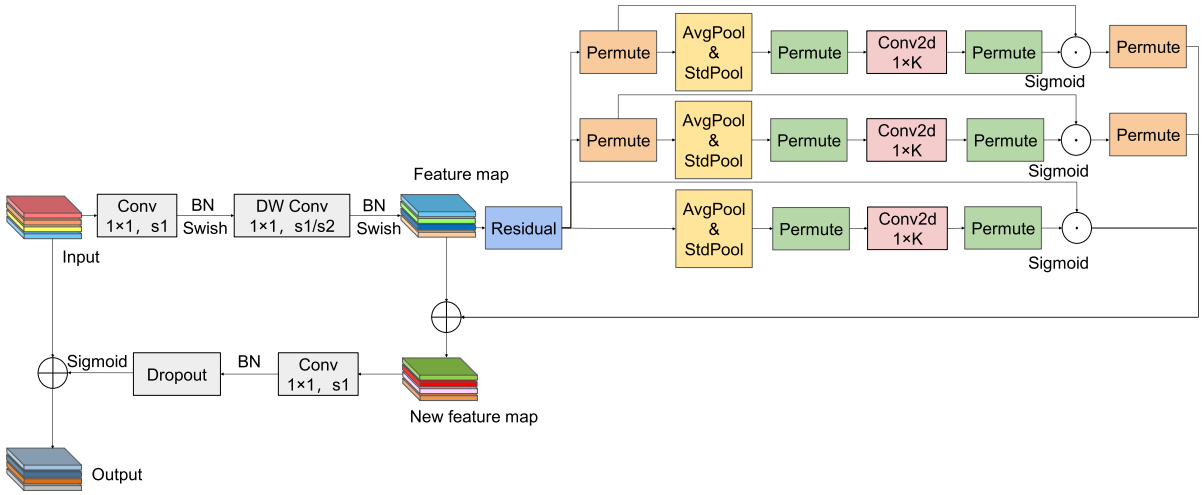

FIGURE 10. Improved MBConv structure.

can be defined as $D_S = \{(x_{s1}, y_{s1}), \cdots, (x_{sn}, y_{sn})\}$, where D_S and $y_{s1} \in Y_s$ of $x_{s1} \in X_s$ are the corresponding class labels. Similarly, the target domain can be defined as $D_T = \{(x_{T1}, y_{T1}), \cdots, (x_{Tn}, y_{Tn})\}$. The tasks of the source and target domains are T_S and T_T, respectively. If the prediction functions of the source and target tasks are $f_s(.)$ and $f_T(.)$, respectively, then Transfer Learning can be formally defined as utilizing the knowledge acquired by D_S and T_S at $D_S \neq D_T$ or $T_S \neq T_T$ to improve $f_T(.)$.

This paper uses the popular ImageNet [31] datasets to obtain D_S and T_S, and utilizes a lightweight and improved EfficientNetV2-M0 model for time-frequency image classification training. The model was pre-trained on images from the ImageNet dataset and frozen prior to initiating training with fused sensor images. During training, only the final fully-connected layer with the final fully-connected layer of the classification output was modified to align with the number of fault categories in our experiments. The optimized source model is capable of recognizing features from over one million images and is designed to contribute to the task of fault diagnosis in the target domain.

III. EXPERIMENTATION AND ANALYSIS

To validate and analyze the effectiveness of the proposed application of a lightweight and improved EfficientNetV2-M0 based on transfer learning and sensor fusion for motor fault diagnosis, a comprehensive simulation testbed for mechanical faults is built, and a series of experiments are carried out for different faults. In the experiments, this paper uses Pycharm for the training and testing operations of the migration learning model and Pytorch 1.10.1 framework for the migration learning model training. All programs were run on a computer with the following configuration: AMD Ryzen 7 5800H, NVIDIA RTX 3070, 16GB RAM.

All the weights of the deep learning models are transferred from Transfer Learning from previously optimized and pre-trained models using the well-known ImageNet dataset. Model training integrates fused time-frequency images on top of a previously trained network. All models are trained with the same hyperparameters using the same data preprocessing. An exponential decay strategy is used to adjust the size of the learning rate, the Batch Size is 8, the maximum number of Epoch is limited to 100, the initial learning rate is set to be 0.01, and the learning rate is scaled down to 0.0001 after every 100 rounds of iteration, and the model with stable convergence of the loss function is selected as the final classification model. The pre-training models explored were the improved EfficientNetV2-M0 and other comparative models using a Transfer Learning approach, where all models were pre-trained on images from the ImageNet dataset and the weights of the remaining layers except for the last layer where the network structure has the classification output were frozen before starting the formal training process using the fused sensor images.

A. EXPERIMENTAL SETUP AND DATA ACQUISITION

The mechanical failure simulation test bench is comprised of a 1.5 kW three-phase asynchronous motor of the Y132S1-2 model, a rotor supported by bearings at both ends, a planetary gearbox, and a series magnetic brake. The bearing model is 6203. Figure 11 shows the layout of the test bench. In this paper, we use motor current analysis and vibration spectrum analysis techniques to test and verify the fundamental characteristics of electrically and mechanically faulty motors, obtaining valuable experimental data under the same operating conditions. The vibration acceleration sensor with three axes collects vibration signals from both faulty and normal rolling bearings. Similarly, the three-phase current detection sensor collects current fluctuation signals from both faulty and normal motors. The sampling frequency of this experiment is 12.8 kHz, the motor load is 1.5 kW, the motor speed is 2600 rpm, and the sampling time is 20 seconds. The experimental dataset comprises vibration and current data, which have been divided into six categories: normal state (NS), inter-turn short-circuit (ITSC), broken rotor bar (BR), eccentricity fault (EF), bearing inner-ring fault (IRF), and bearing outer-ring fault (ORF). In the experiment, the one-dimensional time series signals in each category were partitioned into multiple time series signals with a sample number of 300. These time series signals with a sample number of 300 were then transformed into time-frequency maps using the continuous wavelet transform, which yielded 1750 time-frequency images for each category. The total number of samples for all fault modes was 10500 time-frequency images. Furthermore, the dataset was divided into three distinct subsets: 70% for training, 20% for model selection and cross-validation, and 10% for final testing. This approach ensures a comprehensive evaluation of the model's performance. Additionally, each training and testing dataset was randomly divided.

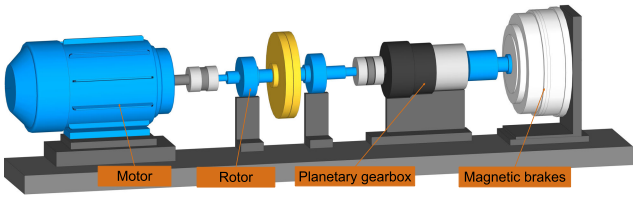

FIGURE 11. Comprehensive simulation test bed for mechanical failures.

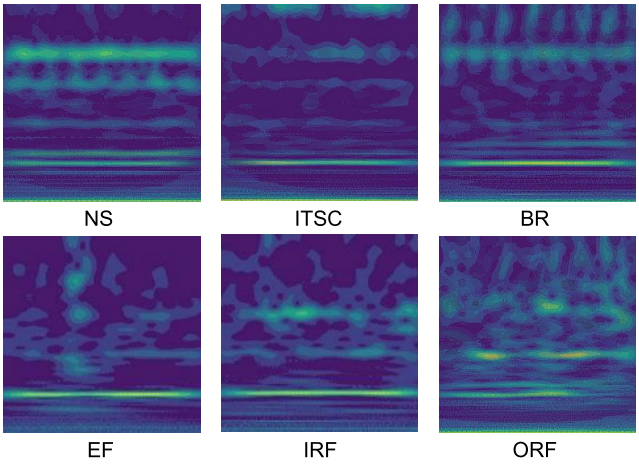

FIGURE 12. Fused images in different failure modes.

B. DATA PRE-PROCESSING

In this paper, Continuous Wavelet Transform (CWT) is used to process the raw data, selecting non-overlapping sliding windows to obtain the time-frequency maps of various types of signals with an image size of 300 × 300. After completing the generation of all the time-frequency maps, the time-frequency maps of various types of signals are fused by wavelet transform-based image fusion technique, and the size is adjusted according to the size of the image inputs of the particular Transfer Learning model. The fused images for different fault modes are shown in Figure 12, and these images will be used as inputs for the proposed fault diagnosis model.

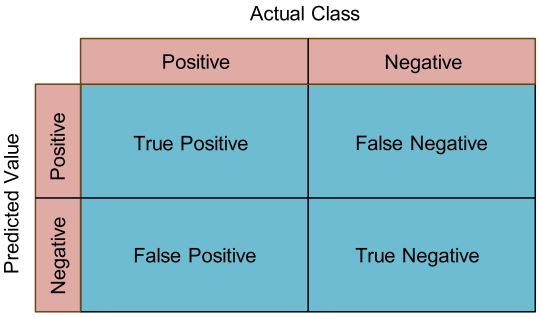

FIGURE 13. Confusion matrix.

C. INDICATORS FOR MODEL EVALUATION

When combining machine learning techniques, accuracy assessment is an integral part of examining the performance of the model, which is measured using various matrices derived from a 2×2 confusion matrix, as shown in Figure 13.

In this experiment, four parameters, Accuracy (ACC), Precision, Recall/Sensitivity curve (*recall/sensitivity*), and Params are chosen to evaluate the model [32]. Where *accuracy* is the proportion of all correct predictions (positive and negative categories) to the total, as shown in equation (9); Precision is the proportion of correct predictions to all positive predictions, as shown in equation (10); and *recall/sensitivity* is the proportion of correct predictions to all actual positive predictions [33], as shown in equation (11). Params refers to the number of parameters in the model's size, which is used here as a measure of model complexity.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$precision = \frac{TP}{TP + FP} \tag{10}$$

$$recall/sensitivity = \frac{TP}{TP + FN} \tag{11}$$

where TP is the number of diagnostic results and detection samples are true, that is, the detection of faulty motor detection results for the number of faults and the results of the correct judgment; FP is the number of false diagnoses, that is, the number of fault diagnostic results does not match the actual motor samples; TN is the number of the actual discrimination results and motor samples are false, that is, the number of the motor has not been detected as a fault and the results of the judgment are true; FN is the number of the actual data FN is the number of positive samples but negative samples detected, i.e., the number of motor faults not detected.

This paper also uses validation accuracy and loss to further evaluate the training model. In addition, continuous monitoring is used to detect any apparent deviations in training and validation performance when it comes to validation accuracy and loss.

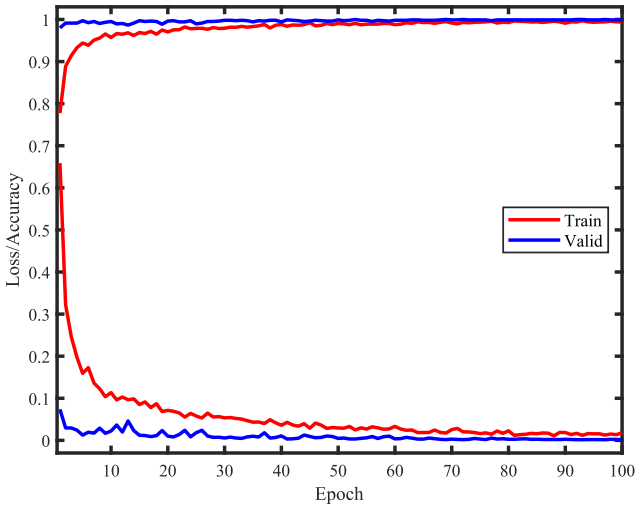

FIGURE 14. Improved EfficientNetV2-M0 training curve.

D. TRAINING RESULTS

This section presents the performance of the proposed model in terms of training and validation accuracy and loss, as depicted in Figure 14. The model was trained for 100 iterations. As shown in the figure, the training accuracy of the model starts at 77.75% in the first iteration and increases significantly. During the 3rd iteration, the model achieved an accuracy of over 90%, with a training accuracy of 99.68% after 100 iterations. The training loss rate decreased from 0.6586 in the first iteration to 0.1 in the 12th iteration. After approximately 40 iterations, the accuracy and loss values of the training dataset stabilized, indicating that the model had begun to converge.

Overall, the majority of models perform well during training. However, the model's performance in the validation phase is poor due to being trained solely on supervised data. It is crucial to validate the model's performance to gain a better understanding of its capabilities. Figure 12 displays the accuracy and loss of the proposed model on the validation set.

The paper describes a model with a validation accuracy exceeding 98% in the first round of iterations and a loss rate of 0.08. The model achieves 100% accuracy during 100 iterations. Additionally, the accuracy and loss values of the validation dataset are largely determined in the first iteration, indicating strong convergence ability of the improved EfficientNetV2-M0 model. Note that during the initial stages of training, the training accuracy may be lower than the validation accuracy due to the use of Dropout, which limits the model's training ability. As the training progresses, Dropout encourages the network to learn more resilient features. Eventually, the training and validation accuracies stabilize at the same value, indicating that the network has good generalization capabilities. The regularization technique employed ensures the network's robust generalization.

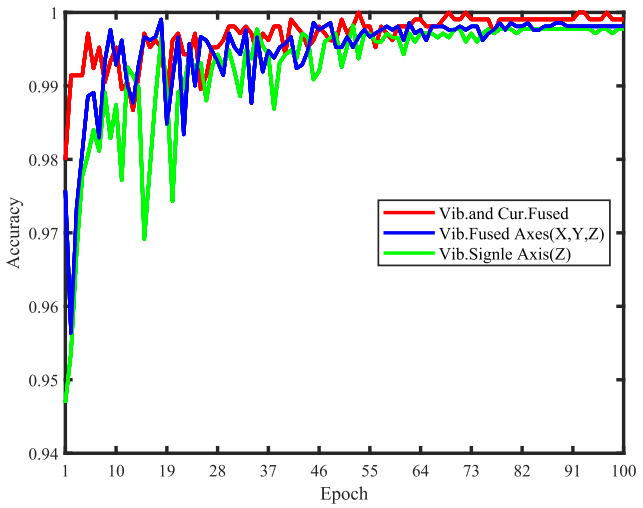

FIGURE 15. Different sensor fusion vs. single sensor training curve.

To demonstrate the application of image fusion techniques in sensor fusion methods for motor fault diagnosis, this paper performs controlled experiments using three types of sensor channels and shows in Figure 15 the percentage of accuracy obtained in 100 iterations using the EfficientNetV2-M0 model architecture. The green line indicates the classification accuracy after fusing all sensor information. This includes the three axes (X,Y, and Z) of the vibration channel and the three-phase current sensors of the motor. The purple line indicates that only three axes (X, Y, and Z) of the vibration channel are fused using the proposed method. The blue line indicates the single vibration axis Z used by the model.

Figure 15 shows that the accuracy of the single vibration axis signal in the first iteration is much lower than that of the other two signals. The accuracy curve of the single vibration axis signal fluctuates unstably, showing a trend of significant ups and downs until the 22nd iteration, after which the accuracy percentage tends to stabilize. The accuracy of the three-axis signal fused with the vibration channel is higher than that of the single-axis signal. However, the convergence of these two signals is slower than that of the three-axis vibration channel fused with the three-phase current channel. On the other hand, the accuracy of the signal fused with three-axis vibration channels and three-phase current channels exceeds 98% in the first round of iterations and begins to converge after the fourth round of iterations. The overall accuracy is higher than that of the other two signals. Specifically, the accuracy of the signal fused with three-axis vibration channels and three-phase current channels reaches up to 100%. In contrast, the fusion signal from the triaxial vibration channel and the three-phase current channel not only achieves high recognition accuracy but also remains stable, reaching 100% quickly. This study proves that the proposed sensor fusion method has good convergence, stable prediction, and high diagnostic efficiency.

TABLE 3. Comparison experiments of improved EfficientNetV2-M0 and EfficientNetV2-B0.

Method	Mean accuracy	Best accuracy	Params num
Improved EfficientNetV2-M0	**99.68%**	**100%**	**4,927,847**
EfficientNetV2-B0	95.20%	98.95%	4,465,984

TABLE 4. Comparison experiments of improved EfficientNetV2-M0 and EfficientNetV2-M.

Method	Mean accuracy	Best accuracy	Params num
Improved EfficientNetV2-M0	**99.68%**	**100%**	**4,927,857**
EfficientNetV2-M	99.57%	100%	54,139,356

To demonstrate the advancement of the improved EfficientNetV2-M0 network, this paper compares the improved network with the original EfficientNetV2-B0 and EfficientNetV2-M, respectively, as shown below.

As can be seen from Table 3 and Figure 16, the verification accuracy of the improved EfficientNetV2-M0 network in this paper is significantly higher than that of the EfficientNetV2-B0 network with a similar number of model parameters, and the highest accuracy of the improved network can reach 100%; moreover, the accuracy and loss rate of the improved network tends to be stabilized in the first round of iterations, with a very good convergence effect, while the fluctuation of the accuracy and loss rate curves of the EfficientNet-B0 network is not unstable and prone to sudden changes.

As can be seen from Figure 17 and Table 4, the accuracy and loss rate of the improved EfficientNetV2-M0 network in this paper is approximate to that of the EfficientNetV2-M network, but the number of model parameters of the improved network is dramatically reduced to only 1/10 of the number of model parameters of the EfficientNet-M network, which makes the improved network, while maintaining high accuracy, to This enables the improved network to maintain high accuracy while having a smaller size, reaching the standard of lightweight neural networks. The improved model makes it possible to run neural network models on mobile terminals and embedded devices. To demonstrate the superiority and feasibility of the improved model in this paper, as well as its robustness to motor fault recognition, this paper also conducted experiments comparing the model with some common classification network models and used the trained model to diagnose the test set of data to evaluate the diagnostic performance. Furthermore, to demonstrate the lightness of the improved EfficientNetV2-M0 model proposed in this paper, which is favorable for the deployment of mobile terminals and embedded devices, this paper employs three classic lightweight models for experimental comparison. The models under consideration are MobileNetV3, ShuffleNet,

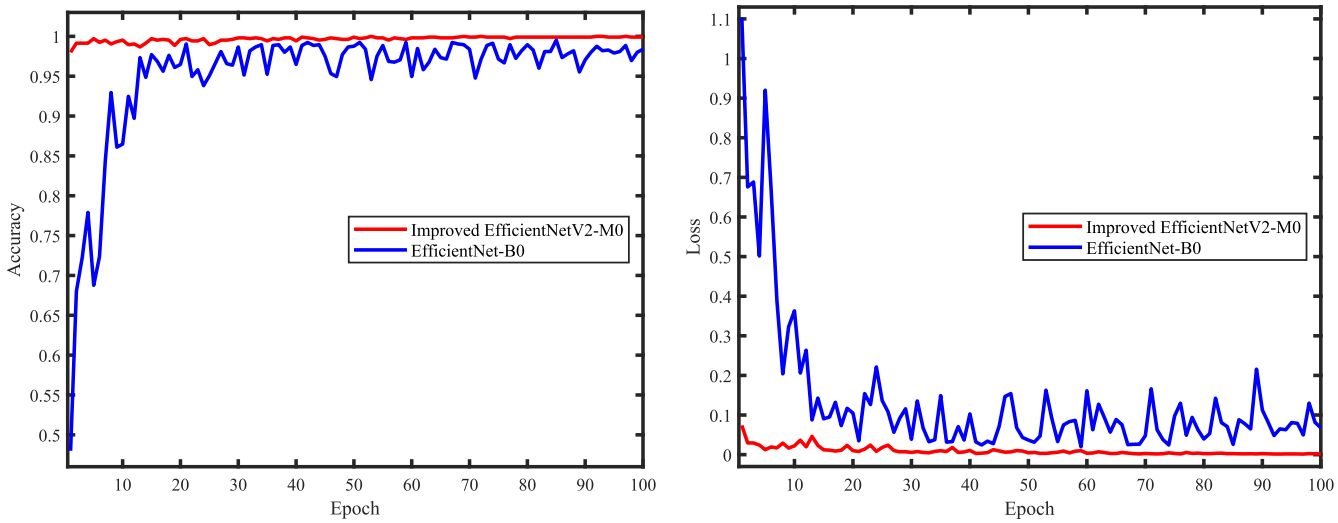

FIGURE 16. Comparison training curves of improved EfficientNetV2-M0 and EfficientNetV2-B0.

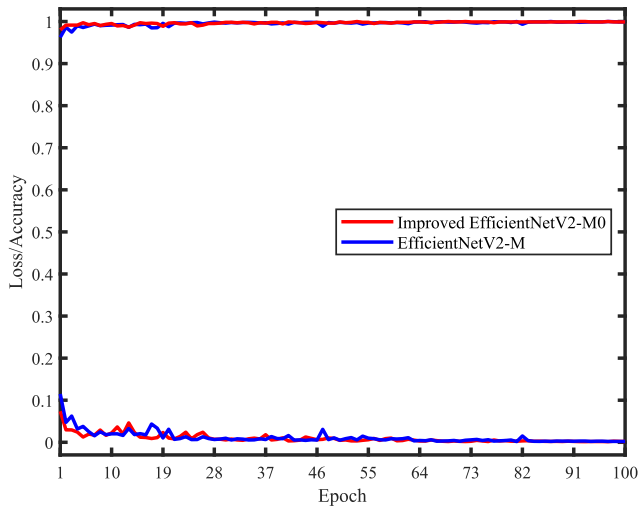

FIGURE 17. Comparison training curves of improved EfficientNetV2-M0 and EfficientNetV2-M.

TABLE 5. Comparison of different model runtime.

Method	Accuracy	Params num	Flops
Improved EfficientNetV2-M0	**100%**	**4.9M**	**0.817GFlops**
EfficientNetV2-B0	98.92%	4.5M	0.724GFlops
EfficientNetV2-M	100%	54.1M	10.342GFlops
MobileNetV3 [34]	98.17%	8.2M	0.760GFlops
ShuffleNet [35]	97.33%	5.4M	0.524GFlops
GhostNet	97.33%	5.4M	0.608GFlops
VGG-19 [36]	97.04%	143.6M	19.643GFlops
ResNet-50 [37]	95.23%	25.6M	3.798GFlops
AlexNet [38]	96.76%	60.9M	7.190GFlops

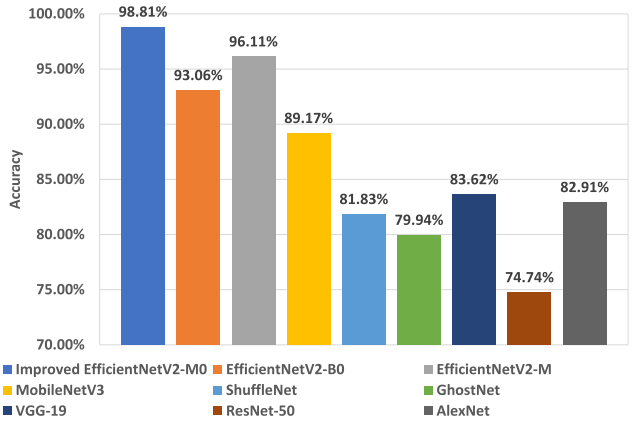

FIGURE 18. Model comparison tests with small samples.

and GhostNet. Among these, MobileNetV3 is smaller in size and higher in accuracy than the original MobileNet network. This is achieved through the use of NAS and NetAdapt algorithms, which enable the optimal model structure to be identified automatically, resulting in a smaller size, less computation, and higher accuracy in the task. To reduce the chance of experimentation, each method is tested 10 times. The experimental results are shown in Table 5.

From the data in Table 5, it can be seen that the classification detection results of this paper's method on the motor fault sensor fusion image test set are significantly better than other methods. In terms of classification accuracy, the accuracy of this paper's method is improved by about 3.3% compared with AlexNet, and about 4.8% compared with ResNet-50, but the accuracy improvement is lower compared with the EfficientNetV2-M network. The results show that the method is effective in classifying and recognizing

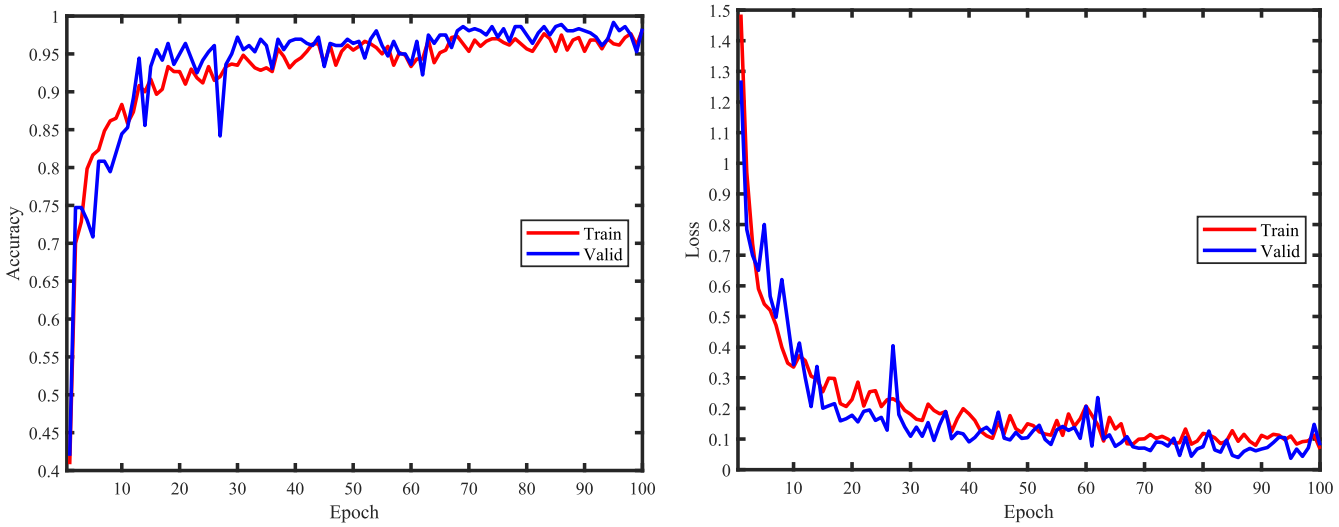

FIGURE 19. Improved EfficientNetV2-M0 training curves in small sample dataset.

TABLE 6. Four different models for ablation experiments.

Export	EfficientNetv2-M0	DBB	MCA	Accuracy	Params num
1	✓			99.14%	5,344,288
2	✓	✓		99.81%	5,560,576
3	✓		✓	99.71%	4,711,569
4	✓	✓	✓	100%	4,927,857

TABLE 7. Precision, recall, and specificity for four different models under six fault categories: (a)EfficientNetv2-M0-MCA-DBB; (b) EfficientNetv2-M0; (c) EfficientNetv2-M0-DBB; (d)EfficientNetv2-M0-MCA.

	Precision	Recall/Sensitivity	Specificity
NS	1.0	1.0	1.0
EF	1.0	1,0	1,0
BR	1.0	1.0	1.0
IRF	1.0	1.0	1.0
ORF	1.0	1.0	1.0
ITSC	1.0	1.0	1.0

(a)

	Precision	Recall/Sensitivity	Specificity
NS	1.0	0.989	1.0
EF	0.962	1,0	0.992
BR	1.0	1.0	0.998
IRF	0.988	0.96	1.0
ORF	1.0	1.0	1.0
ITSC	1.0	1.0	1.0

(b)

	Precision	Recall/Sensitivity	Specificity
NS	1.0	1.0	1.0
EF	0.994	0.994	0.999
BR	1,0	1.0	1.0
IRF	0.994	0.994	0.999
ORF	1.0	1.0	1.0
ITSC	1.0	1.0	1.0

(c)

	Precision	Recall/Sensitivity	Specificity
NS	1.0	1.0	1.0
EF	0.989	0.994	0.998
BR	1.0	1.0	1.0
IRF	0.994	0.994	0.999
ORF	1.0	0.994	1.0
ITSC	1.0	1.0	1.0

(d)

motor fault signals. Meanwhile, with the improvement of accuracy, the Params num of the model of the method is reduced from 54.1M in EfficientNetv2-M to 4.9M, which greatly reduces the complexity of the model. Although MobileNetV3, ShuffleNet, and EfficientNetV2-B0 have very low Params num, their accuracies are still lower than the final results of the method proposed in this paper on the test set, and the experiments show that the improved

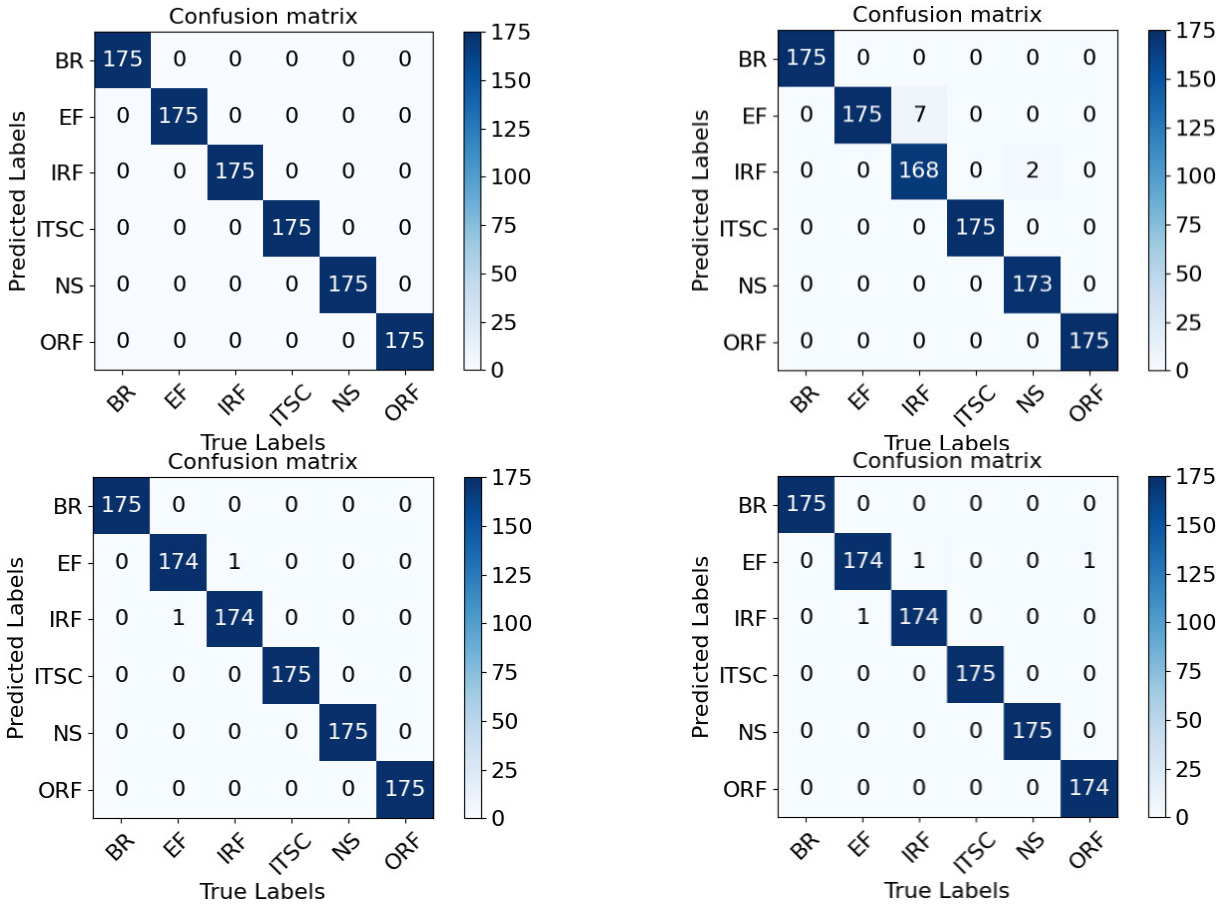

FIGURE 20. Confusion matrices for four different models: (a) EfficientNetv2-M0-MCA-DBB; (b) EfficientNetv2-M0; (c) EfficientNetv2-M0-DBB; (d) EfficientNetv2-M0-MCA.

EfficientNetV2-M0 proposed in this paper is favorable for mobile model deployment.

To ascertain the advantages of feature extraction afforded by the proposed enhanced EfficientNetV2-M0 network and the diagnostic efficacy of its migration learning technique in the context of limited sample data, the original dataset was subjected to a significant reduction in size in this experiment, creating a sparse sample size condition, which was then analysed for fault diagnosis. A total of five fault classes and one normal condition are included: normal state (NS), inter-turn short-circuit (ITSC), rotor broken bar (BR), eccentricity fault (EF), bearing inner-ring fault (IRF), and bearing outer-ring fault (ORF). The number of training data samples for each fault class is 100, while the number of test data samples for each fault class is 60, for a total of 600 and 360 data samples for training and testing, respectively. Furthermore, each data sample comprises 300 data points. The experimental results are presented in Figure 18 and 19:

As illustrated in the figure, the proposed novel approach to fault diagnosis, based on migration learning with an enhanced EfficientNetV2-M0 network and sensor fusion, demonstrates robust performance in the context of limited sample sizes. The method exhibits superior diagnostic accuracy compared to other models, with an accuracy of 98.81% on the test set. This indicates that the feature extraction module of the enhanced EfficientNetV2-M0 network exhibits notable advantages, and the migration learning technique of the method demonstrates enhanced diagnostic performance on limited sample data.

To demonstrate the effectiveness of the improved modules, this paper does ablation experiments on the improved EfficientNetV2-M0 model, four different models are designed for each of the improved modules, as shown in Table 6, and the confusion matrices of the models are generated separately, as shown in Figure 20, where the darker the color the higher the value in the confusion matrix.

From the ablation experiments in Table 6, it can be seen that each module improved in this paper for EfficientNetV2-M0 is effective, the introduction of the DBB module for structural reparametrization significantly improves the accuracy of the model prediction, and the introduction of the MCA attention mechanism significantly reduces the number of model parameters while improving the accuracy. From the confusion matrix in Figure 20 and Table 7, it is intuitively obvious that the original model is not precise in classifying eccentric faults (EF) and bearing inner-ring faults (IRF), whereas the improved EfficientNetV2-M0 achieves precise

classification of all types of motor fault signals and the model achieves the maximum values of accuracy, recall, and specificity in each category.

The experimental results indicate that the proposed method, which combines sensor fusion and migration learning techniques with the EfficientNetV2-M0 architecture, is an effective approach for classifying motor fault signals. The test set achieved a 100% recognition accuracy, and the method's complexity is lower than that of most other models.

In order to verify the advantages of the proposed sensor information fusion technology, this section utilizes single sensor and multi-sensor fusion data sets for fault diagnosis to conduct comparison tests. The experimental results demonstrate that the diagnostic effect based on multi-sensor fusion signals is significantly superior to that of single sensor fusion signals. The diagnostic effect of multi-source heterogeneous sensor fusion is more accurate than that of single sensor fusion signals, particularly in the context of multi-sensor signal co-diagnosis. This proves the feasibility and effectiveness of collaborative diagnosis of multi-sensor signals.

In order to validate the improved model proposed in this paper and the superiority of utilizing the migration learning technique, this section compares the improved EfficientNetV2-M0 neural network with different CNN deep learning architectures. The experimental results demonstrate that the proposed method can effectively mine fault-sensitive features, more fully utilize multi-sensor information, and improve diagnostic effectiveness and stability. The reduced complexity of the enhanced EfficientNetV2-M0 model employed in this study facilitates its deployment on mobile terminals and embedded devices. Additionally, a series of small-sample experiments are conducted to demonstrate that the method exhibits robust generalization capabilities.

IV. CONCLUSION AND FUTURE WORK

This paper presents lightweight improvements to the EfficientNetV2 architecture and introduces a novel, efficient, and reliable method designed for motor fault diagnosis that integrates sensor fusion techniques and migration learning. First, signals from different sensors are converted into time-frequency images using the continuous wavelet transform, and then each image is decomposed into different levels of subband coefficients using the Mallat algorithm. Subsequently, fusion rules based on the maximum absolute value method and weighted average method are constructed to integrate and reconstruct the subband coefficients at all levels in time-frequency images from multiple sensors. Next, the EfficientNetV2 model is enhanced to improve its feature extraction capability and computational efficiency, thereby achieving a lightweight effect. The EfficientNetV2-M0 network optimizes the depth and width scaling factors of the model to improve the accuracy and reduce the parameters and computational complexity. Furthermore, the network incorporates Diverse Branch Block (DBB) and Multidimensional Collaborative Attention (MCA) to enhance detection efficiency and augment feature extraction in the context of sparse sample sizes. The maximum cross-entropy loss function is also enhanced through the application of label smoothing and focus loss, which serve to prevent model overfitting and dynamically adjust the classification weights, thereby improving accuracy. The network is deployed using a pre-trained model obtained from a transfer learning technique that combines multi-sensor information fusion and an improved lightweight model for application in fault diagnosis.

This paper presents four comparison experiments designed to validate the advantages of the proposed method. The data fusion experiments demonstrate that the diagnostic effect based on the fused signals from multiple sources of heterogeneous sensors reaches 100%, which provides evidence of the feasibility and effectiveness of the collaborative diagnosis of multi-sensor signals. In the model comparison experiment, the accuracy of this model reached 100% and the model complexity reached the lightweight standard, which demonstrated the superiority of this paper's method in terms of lightweight design and diagnostic accuracy. In the small sample experiment, it was demonstrated that the feature extraction module of the improved EfficientNetV2-M0 network exhibited notable advantages, and the migration learning technique of this method demonstrated excellent diagnostic performance on small sample data. Finally, the ablation experiments demonstrated the necessity of the improved modules and the confusion matrix was employed to identify the fault categories that were challenging to distinguish in the dataset.

Future research will extend the proposed method to other motor faults, such as stator, misalignment, or winding faults. Additionally, the consideration of additional sensors, such as torque or electromagnetic sensors, will be explored. Finally, one of the future goals of this research is to transfer the trained model to embedded edge AI real-time applications.

REFERENCES

[1] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, "Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges," *Inf. Fusion*, vol. 35, pp. 68–80, May 2017.

[2] Z. Xia, F. Ye, M. Dai, and Z. Zhang, "Real-time fault detection and process control based on multi-channel sensor data fusion," *Int. J. Adv. Manuf. Technol.*, pp. 795–806, Mar. 2021.

[3] P. Suawa, T. Meisel, M. Jongmanns, M. Huebner, and M. Reichenbach, "Modeling and fault detection of brushless direct current motor by deep learning sensor data fusion," *Sensors*, vol. 22, no. 9, p. 3516, May 2022.

[4] P.-M. Huang and C.-H. Lee, "Estimation of tool wear and surface roughness development using deep learning and sensors fusion," *Sensors*, vol. 21, no. 16, p. 5338, Aug. 2021.

[5] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.

[6] H. Kaur, D. Koundal, and V. Kadyan, "Image fusion techniques: A survey," *Arch. Comput. Methods Eng.*, vol. 28, pp. 4425–4447, Dec. 2021.

[7] N. Tawfik, H. A. Elnemr, M. Fakhr, M. I. Dessouky, and F. E. A. El-Samie, "Survey study of multimodality medical image fusion methods," *Multimedia Tools Appl.*, vol. 80, no. 4, pp. 6369–6396, Feb. 2021.

[8] J. Petrich, Z. Snow, D. Corbin, and E. W. Reutzel, "Multi-modal sensor fusion with machine learning for data-driven process monitoring for additive manufacturing," *Additive Manuf.*, vol. 48, Dec. 2021, Art. no. 102364.
[9] H. Wang, S. Li, L. Song, and L. Cui, "A novel convolutional neural network based fault recognition method via image fusion of multi-vibration-signals," *Comput. Ind.*, vol. 105, pp. 182–190, Feb. 2019.
[10] D. Wang, Y. Li, L. Jia, Y. Song, and Y. Liu, "Novel three-stage feature fusion method of multimodal data for bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.
[11] Z. Zheng, J. Fu, C. Lu, and Y. Zhu, "Research on rolling bearing fault diagnosis of small dataset based on a new optimal transfer learning network," *Measurement*, vol. 177, Jun. 2021, Art. no. 109285.
[12] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big data*, vol. 3, no. 1, pp. 1–40, May 2016.
[13] J. Liu, Q. Zhang, X. Li, G. Li, Z. Liu, Y. Xie, K. Li, and B. Liu, "Transfer learning-based strategies for fault diagnosis in building energy systems," *Energy Buildings*, vol. 250, Nov. 2021, Art. no. 111256.
[14] W. Mao, L. Ding, S. Tian, and X. Liang, "Online detection for bearing incipient fault based on deep transfer learning," *Measurement*, vol. 152, Feb. 2020, Art. no. 107278.
[15] S. Lee, H. Yu, H. Yang, I. Song, J. Choi, J. Yang, G. Lim, K.-S. Kim, B. Choi, and J. Kwon, "A study on deep learning application of vibration data and visualization of defects for predictive maintenance of gravity acceleration equipment," *Appl. Sci.*, vol. 11, no. 4, p. 1564, Feb. 2021.
[16] E. Cinar, "A sensor fusion method using transfer learning models for equipment condition monitoring," *Sensors*, vol. 22, no. 18, p. 6791, Sep. 2022.
[17] K. Khalil, O. Eldash, A. Kumar, and M. Bayoumi, "Machine learning-based approach for hardware faults prediction," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 11, pp. 3880–3892, Nov. 2020.
[18] J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, and R. Arbiol, "Multiresolution-based image fusion with additive wavelet decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1204–1211, May 1999.
[19] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.
[20] G. Liu and W. Yang, "Image fusion method based on wavelet decomposition and performance evaluation," *Acta Automatica Sinica*, vol. 28, no. 6, pp. 927–934, 2002.
[21] M. Tan and Q. V. Le, "EfficientNetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
[22] S. Gupta and M. Tan, "EfficientNet-EdgeTPU: Creating accelerator-optimized neural networks with AutoML," *Google AI Blog*, vol. 2, no. 1, 2019.
[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[24] 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.*, Jun. 2018, pp. 4510–4520.
[25] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.
[26] K. Khalil, O. Eldash, A. Kumar, and M. Bayoumi, "Designing novel AAD pooling in hardware for a convolutional neural network accelerator," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 30, no. 3, pp. 303–314, Mar. 2022.
[27] X. Ding, X. Zhang, J. Han, and G. Ding, "Diverse branch block: Building a convolution as an inception-like unit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10881–10890.
[28] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1911–1920.
[29] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style ConvNets great again," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13728–13737.
[30] Y. Yu, Y. Zhang, Z. Cheng, Z. Song, and C. Tang, "MCA: Multidimensional collaborative attention in deep convolutional neural networks for image recognition," *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 107079.
[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[32] A. R. Abd-Elhay, W. A. Murtada, and M. I. Youssef, "A reliable deep learning approach for time-varying faults identification: Spacecraft reaction wheel case study," *IEEE Access*, vol. 10, pp. 75495–75512, 2022.
[33] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Inf. Sci.*, vols. 340–341, pp. 250–261, May 2016.
[34] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.
[35] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–7.

LIANG JIANG received the B.S. degree from Jilin University, Changchun, China, in 2008, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2015. He currently is an Associated Professor and a Senior Engineer with the School of Automation, Wuxi University. His research interests include intelligent equipment fault diagnosis and industrial control systems development.

SICHENG ZHU received the B.S. degree from the School of Mechanical and Electrical Engineering, Suqian University, Suqian, China, in 2022. He is currently pursuing the master's degree in electronic information with Nanjing University of Information Science and Technology, Nanjing, China. His research interests include fault diagnosis, deep learning, and image processing techniques.

NING SUN received the B.S. and M.S. degrees from Nanjing University of Information Science and Technology, Nanjing, China, in 2004 and 2007, respectively. He is currently a Professor with the School of Automation, Wuxi University. His research interests include data-driven fault detection and diagnosis, fault prediction, and digital twin technology.