Received 19 February 2024, accepted 5 April 2024, date of publication 6 May 2024, date of current version 13 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3397184

RESEARCH ARTICLE

Attention-Aware Meta-Reweighted Optimization for Enhanced Intelligent Fault Diagnosis

GUANG ZHAO[1,2], **SHIQIANG HU**[1], **JIAYUAN FAN**[1], **QIANG GUO**[2], **BO SHEN**[2], **AND LINGKUN LUO**[1]

[1]School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Commercial Aircraft Corporation of China, Ltd., Shanghai 200126, China

Corresponding author: Guang Zhao (zhaoguang@sjtu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773262 and Grant 62006152, and in part by China Aviation Science Foundation 2022Z071057002.

ABSTRACT Due to stringent aircraft safety requirements and the high cost of experiments, there is a scarcity of failure samples, which creates a gap between existing diagnostic models and practical applications. To address this issue, we have developed a small-sample civil aircraft fault diagnosis method. This method combines a meta-learning approach to tackle data imbalance with a channel attention mechanism to enhance feature extraction efficiency. Specifically, our approach integrates the advantages of meta-learning and attention regularization, effectively addressing both the imbalance in training sample distribution and the need for human interaction to enhance feature representation. We then evaluated five data imbalances and introduced a fault diagnosis algorithm based on a one-dimensional convolutional network, which has been successfully applied to solve small sample yield tasks in two datasets. Additionally, we provide baseline accuracy under the same conditions for comprehensive comparison and reference. Through extensive experiments, our method achieves competitive performance and demonstrates its superiority in solving imbalanced distribution experimental configurations.

INDEX TERMS Meta-learning, fault diagnosis, channel attention mechanism, few-shot.

I. INTRODUCTION

Fault diagnosis and health management technology for complex equipment have always been challenging and dynamic areas of research. These technologies leverage collected data to monitor, diagnose, and predict the current state of aircraft, aiding intelligent decision-making. Recently, with advancements in hardware and the availability of massive datasets, data-driven fault diagnosis methods, particularly those based on deep learning [1], [2], have gained significant attention. Their appeal lies in their ability to bypass the need for physical modeling and their remarkable feature extraction capabilities. These methods [3], [39], [40] have seen notable performance enhancements through the analysis of various system parameters and the exploration of feature relationships within the data. However, traditional fault diagnosis algorithms relying on deep learning are heavily reliant on high-quality datasets for effective model training. Their performance tends to degrade or fail altogether when faced with small or unbalanced training datasets. Civil aircraft equipment is mandated to operate under optimal conditions, limiting the availability of failure samples [4]. Moreover, aircraft equipment is typically expensive, making it impractical to gather sufficient fault samples through experiments. Consequently, the proportion of faulty data to healthy data in aircraft systems often remains imbalanced in practice. Traditional algorithms struggle to handle these tasks with limited samples, primarily due to the data requirements of deep learning. To tackle this challenge, Chen et al. [5] and Zhao et al. [6] devised fault diagnosis methods based on data augmentation techniques. These methods generate balanced datasets from imbalanced ones using various sampling strategies. Their straightforward implementation has made them widely adopted in numerous research studies.

Specifically, previous methods relying on data augmentation have effectively transformed imbalanced datasets

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaojun Steven Li.

into balanced ones, thereby significantly enhancing the algorithm's performance. However, they still face the following two challenges:

A. CHALLENGE 1: (BRUTALLY FORCED RE-SAMPLING MECHANISM INDUCED NEGATIVE TRANSFER)

We argue that existing data augmentation methods are unable to fundamentally solve the issue. They often fall short in fully capturing the features of the minority class. Additionally, the integration of borrowed samples with minority class data can disrupt the distribution of the minority data, resulting in negative transfer effects. Motivated by this, it is beneficial to measure the divergence between the unbalanced and balanced datasets, as it can guide the model's generation without drastically altering the training set.

B. CHALLENGE 2: (COMPLEX FEATURE DISTRIBUTION LEADS TO THE DIFFICULTY OF FEATURE EXTRACTION)

In civil aircraft fault diagnosis, the data formats are diverse and intricate, posing challenges to feature extraction. While augmenting the volume of training data is a common strategy, it proves ineffective in scenarios with limited samples. Previous studies have attempted to improve network feature extraction efficiency through techniques like Fourier transforms or specialized feature extraction layers [7]. However, these methods rely on manual design, leading to inconsistent performance, poor generalization, and an inability to handle data from multiple sensor sources.

In analyzing these mentioned challenges, we identify the shortcomings of existing methods as the difficulty in balancing human intervention and model performance. Achieving optimal feature representation and adaptability to imbalanced training often requires significant human priors and meticulous hyperparameter tuning. Consequently, these methods are sensitive to parameters, leading to poor generalizability across different experimental scenarios and unstable functional learning. To address these issues, we propose a hybrid neural network (NN) architecture that integrates meta-learning and attention regularization. Meta-learning enables the model to quickly adapt to new tasks, significantly enhancing its robustness. Concurrently, the channel attention mechanism intelligently learns the importance of different feature extraction channels and automatically applies weighting processing to enhance the network's feature extraction capability. As a result, we focus on these two strategies reinforced NN architecture for our subsequent research.

In response to the aforementioned challenges, we propose a novel fault diagnosis model that combines the following methods: the Channel Attention Enhanced L2R network. Our approach addresses the discrepancy between imbalanced and balanced datasets using meta-knowledge, guiding the model's training process. Additionally, we introduce a channel attention module to enhance the model's sensitivity to critical features. To address challenge 1, we introduce the L2R network, which is grounded in meta-learning principles. This network leverages meta-knowledge to quantify the differences between balanced and imbalanced datasets, allowing for the recalibration of sample weights in the unbalanced training datasets. Importantly, our method autonomously adjusts these weights through meta-learning, eliminating the need for manual parameter configuration. To tackle challenge 2, we introduce a channel attention module that utilizes an auxiliary neural network to evaluate the significance of each feature channel. This module assigns weights to the channels based on their importance, directing more attention to crucial channels. Specifically, we utilize meta-knowledge to dynamically reweight samples from unbalanced datasets and employ additional networks for channel-wise attention allocation.

This paper makes several key contributions, which are summarized as follows:

- This paper introduces an enhanced fault diagnosis algorithm based on convolutional neural networks (CNNs), which automatically extract features from time-series fault data through one-dimensional convolution. This end-to-end fault detection algorithm provides versatility, cost-effectiveness in maintenance, and strong portability. Moreover, it integrates the SE channel attention mechanism to adaptively enhance feature representation, thereby improving the representation capability of these features.
- The optimization strategy, involving meta-learning embedding techniques, enhances model training by improving the model's adaptability to small, unbalanced fault samples. It achieves this by automatically adjusting sample loss weights. Additionally, the strategy enhances gradient updates in L2R through a clipping and lifting approach, resulting in meta-gradient enhancement. This approach increases the effective utilization of fault samples, thereby enhancing the accuracy and stability of fault diagnosis algorithms.
- Extensive experiments conducted on two fault diagnosis datasets demonstrate that our proposed method achieves competitive results in addressing small sample fault diagnosis challenges.

The paper is structured as follows: Section.II delves into the related work, revealing the connections among existing methods. Section.III introduces our proposed method. Section.IV benchmarks the proposed fault diagnosis method and provides an in-depth analysis. Finally, Section.V draws conclusions.

II. RELATED WORKS

A. FAULT DIAGNOSIS FOR IMBALANCED DATA SETS

The issue of dealing with small sample problems [38] is quite common in engineering and has gained significant attention in previous research. Small sample methods in fault diagnosis can be categorized into following three main strategies: data augmentation-based strategy, feature learning-based strategy and classifier design-based strategy [8], [9].In addition, as a new strategy, methods based on meta-learning have gradually received more attention.

In terms of data preprocessing, data augmentation-based strategies aim to address the scarcity of training samples through oversampling or generating synthetic data. Traditional approaches such as Synthetic Minority Oversampling Technique [10] and Adaptive synthetic sampling approach [11] primarily involve linearly interpolating of virtual data into the dataset based on neighborhood relationships, while reducing the quality of augmented samples. More recently, data generation models, exemplified by Generative Adversarial Networks (GAN) [12] or other deep learning neural network [13] have been extensively studied and have shown promising results in augmenting mechanical fault data [14]. However, deep generative models pose challenges in training, requiring substantial computing resources, and are prone to generating low-quality samples.

In the realm of feature extraction, feature learning-based strategies focusing on designing regularized neural networks or feature adaptation without resorting to data augmentation. Researchers design approaches based on these strategies: Yang et al. [15] and Zeng et al. [16] leverage the superior feature extraction capabilities of neural networks to extract useful common features; Li et al. [17] and Yang et al. [18] utilize transfer learning networks to impart knowledge from related datasets. However, extracting effective features from limited and imbalanced data presents a challenge, and obtaining high-quality, relevant datasets for transfer learning networks is also a challenging task.

In the domain of classifier design, the strategy contends that training a classifier suitable for imbalanced data can obviate requirements for tedious processes such as data augmentation or the design of feature extraction models. However, such classifiers often depend on expert knowledge or auxiliary datasets. Dong et al. [19] and Peng et al. [20] design cost-sensitive loss functions based on expert knowledge for the target scenario, and Li et al. [21] and He et al. [22] fine-tuning the model trained on auxiliary datasets to adapt the target scenario. Consequently, the performance of fault classifiers obtained through this approach is sensitive to human interaction and the quality of the auxiliary dataset.

Recently, inspired by the manner of human learning, meta-learning has been specifically introduced. Meta-learning techniques aim to enhance the network's learning capability for tasks at a higher level, beyond simple classification tasks. Specifically, Model-Agnostic Meta-Learning (MAML) [23] and its derivatives, for example, have shown strong performance in dealing with few-shot instances. As a result, meta-learning-based fault diagnosis methods have found widespread application in recent research. For instance, He et al. [24] introduce MAML-based meta-learning to the fault diagnosis of rolling bearings, enabling end-to-end few-shot sample-based diagnosis of bearing faults under varying working conditions. Wu et al. [25] utilize a Meta Relation Net for intelligent fault diagnosis in rotating machinery. They introduce the Match Net to learn a distance metric function, which is used to match few-shot samples with known categories. Additionally, Dixit et al. [26] design a model called CACGAN, combining Model-Agnostic Meta Learning (MAML) with a GAN framework, leverages MAML to initialize and update network parameters.

B. ATTENTION MECHANISM

Inspired by human perception systems, attention mechanism has demonstrated remarkable effectiveness in various computer vision applications (e.g., [27], [28]) and natural language processing domains (e.g., [29], [30]). In computer vision, Dosovitskiy et al. [27] introduced a novel approach called Vision Transformer (ViT), which applies the Transformer architecture to sequential image patches, leading to improved performance in image classification tasks. Fan et al. [28] made pioneering efforts by using the Transformer model for point cloud video modeling, and applied the Transformer for spatio-temporal modeling in raw point cloud videos. In the realm of natural language processing, Vaswani et al. [31] proposed a groundbreaking work to explore the effectiveness of attention mechanisms in capturing global knowledge within input and output dialogs, significantly enhancing machine translation tasks. More recently, Fan et al. [30] devised a novel recurrent attention network to generate attention-enhanced spatial context for Visual Dialog tasks.

Recently, in the field of intelligent fault diagnosis, attention mechanisms are employed to enhance feature representation. For instance, in the DANDA [32], both channel and spatial attention mechanisms are employed to capture low-level features in fault data. Later on, Zheng et al. [33] utilizes attention mechanisms to filter out extraneous features extracted from the data. In our research, we investigate the efficacy of channel attention in intelligent diagnosis and integrate an attention module into the meta-learning network to enhance the utilization of critical features.

III. PROPOSED METHOD

The fault diagnosis model based on convolutional neural networks possesses robust automatic feature extraction capabilities and offers versatility and portability across various fault data. However, constrained by limited onboard computing resources, the model employs the LeNet lightweight convolutional neural network with fewer parameters as the backbone network for the fault diagnosis model. This design aims to reduce computing costs and meet the real-time requirements of fault diagnosis. Simultaneously, recognizing the enhanced sensitivity of one-dimensional convolution operations to time series samples, the network optimizes the feature extraction module of the backbone network by replacing the two-dimensional convolution layer with a one-dimensional convolution layer. In the classifier section, the final fault diagnosis classification results are determined by comparing the probabilities of different fault categories. This is achieved through two fully connected layers and

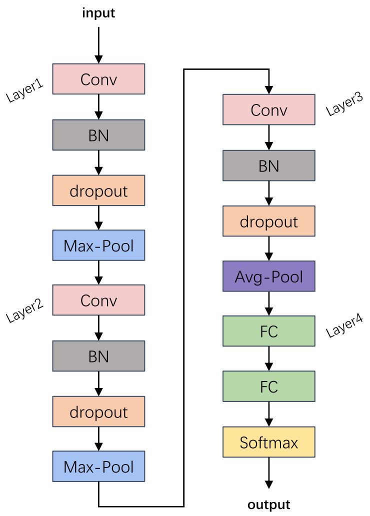


FIGURE 1. Structure of LeNet: The feature extractor of the network consists of three convolutional layers, and the classifier consists of two fully connected layers and a soft-max layer.

one Soft-max layer. Overally, the structure of the backbone network is illustrated in Fig. 1.

However, in solving the task of diagnosing faults in small and unbalanced civil aircraft, the LeNet backbone network still offers significant room for improvement. The limited number of convolution layers fails to provide sufficient feature extraction capabilities, especially given a small size and imbalanced characteristics of the fault samples. This deficiency can potentially lead to overfitting issues in the model. To address these challenges, we employ the channel attention mechanism to adaptively enhance convolution features, thereby lifting the representation ability of fault features. Additionally, we utilize the meta-learning algorithm to learn heavy weights, contributing to improved model training strategies. This chapter focuses on enhancing the network structure and refining training strategies to overcome these limitations.

A. SCENARIO DEFINITION

Our research focuses on the civil aircraft fault diagnosis task within the small sample scenario. First, we introduce the relevant basic symbols. In this hypothetical scenario, due to stringent aircraft safety requirements, the training dataset is defined as follows:

$$D_{Train} = \{x_i, y_i\}_{i=1}^{N_{Train}} \quad x_i \in X_{Train}, y_i \in Y. \tag{1}$$

D_{Train} is the training set. X_{Train} is the set of all training samples, containing N_{Train} samples. x_i represents the i^{th} sample whose label is y_i.

Testing set is defined as follows:

$$D_{Test} = \{x_i, y_i\}_{i=1}^{N_{Test}} \quad x_i \in X_{Test}, y_i \in Y. \tag{2}$$

D_{Test} is the test set. X_{Test} is the set of all test samples, containing N_{Test} samples. x_i represents the i^{th} sample whose label is y_i.

To facilitate quantitative research, we assume that the sample space of the data set only contains two categories: $Y = \{0, 1\}$.

Therefore, the imbalanced data set is defined as follows:

$$U = \sum_{i=1}^{N} \mathbb{I}\left(y_i = 0\right)/N \tag{3}$$

Where $\mathbb{I}(*)$ is the indicator function, and N is the number of samples in dataset, $U \in [0, 1]$.

We aim to find a fault diagnosis model that can exhibit good performance on the balanced test dataset ($U_{Test} = 0.5$) after model training based on an unbalanced training set ($U_{Train} > 0.5$). The fault diagnosis can be defined as:

$$\hat{y} = f(x; \theta) \tag{4}$$

Where $f(.)$ is our neural network model, and θ is the model parameters.

B. CHANNEL ATTENTION MECHANISM

The lightweight convolutional neural network, namely LeNet, with its relatively few convolution layers, possesses limited decoupling and abstraction capabilities for fault features. In this research, structural improvements have been made to enhance its performance.

The Squeeze and Excitation (SE) channel attention mechanism is a neural network module introduced by Hu et al. [34] to augment the capability of extracting image features. Its goal is to allocate distinct weights to various channel features within the channel domain dimension, ensuring the acquisition of crucial feature representation. In this paper, the SE attention mechanism is employed in the time series fault feature extraction module. By enhancing the representation ability of Convolutional Neural Network (CNN), the SE attention mechanism extracts spatial encoding quality throughout the entire feature hierarchy. Its model structure is depicted in Fig. 2.

As illustrated in Fig. 2, for any given transformation F_{tr}, the input x is transformed into the feature space u. Using the convolution operation as an example, the input x undergoes a one-dimensional convolution, mapping it to a feature space of $H \times W \times C$. Subsequently, the feature matrix undergoes global pooling through the squeeze operation $F_{sq}(\cdot)$ to obtain a low-dimensional embedding u of the feature space within the dimension of the global receptive field. The calculation formula is as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{5}$$

The low-dimensional embedding z_c encompasses global information from various feature channels. Then, the excitation operation $F_{ex}(\cdot)$ involves connecting two fully connected layers, with weights generating adaptive channel weights. The calculation is as follows:

$$s = F_{ex}(z, W) = \sigma(g(z, w))\sigma(W_2\delta(W_1 z)) \tag{6}$$

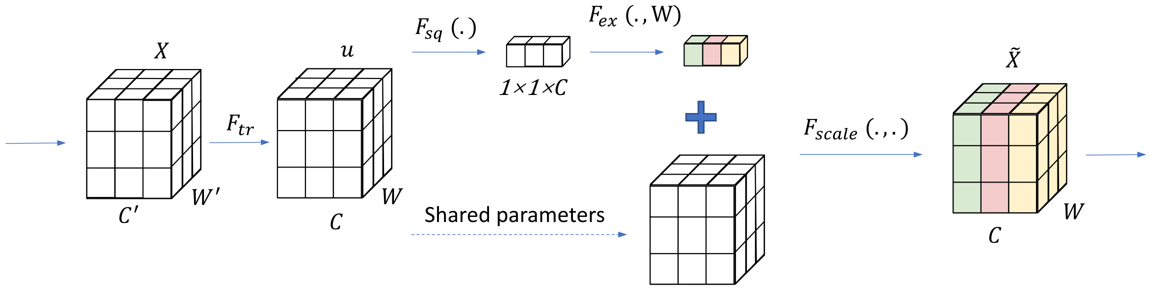


FIGURE 2. Illustration of SE channel attention. The squeeze-and-excitation block explicitly models the channels in the convolutional network (as shown in the upper part of the figure) and feeds back to the network to get the attention-enhanced model (as shown in the right part of the figure).

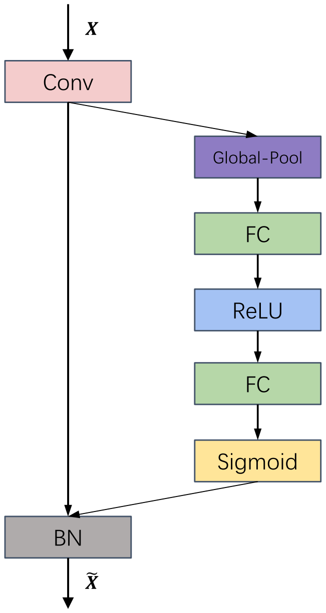


FIGURE 3. Illustration of convolutional channel attention. Convolutional layers (left) are enhanced with attention mechanisms by inserting channel attention modules (right).

W_1 and W_2 are the weight parameters of the two fully connected layers, and R is the squeeze hyperparameter, whose size represents the degree of channel weight squeezing. After the above operations, a generative model of adaptive channel weights is constructed. Finally, the feature generated by the SE channel attention module is obtained through channel weighting:

$$\widetilde{x_c} = F_{scale}(u_c, s_c) = s_c u_c \tag{7}$$

Combined with a one-dimensional convolution layer, the SE channel attention algorithm structure is illustrated in Fig. 3.

In a series of previous research, the attention mechanism has been proven effective in lifting image feature representation. In the context of the fault time series data, modifying the weight distribution among different feature channels through the SE attention mechanism can also enhance the impact of crucial features, thereby improving feature representation capability. In this study, the SE attention mechanism is employed in fault diagnosis technology, significantly enhancing the accuracy and effectiveness of fault diagnosis. The enhanced one-dimensional convolutional neural network structure is depicted in Fig. 4:

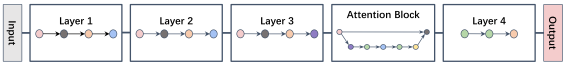


FIGURE 4. Channel enhanced LeNet structure diagram: The feature extraction module of the network can be enhanced by embedding the attention module in the third convolutional layer.

C. LEARNING TO RE-WEIGHT

The enhanced LeNet network, which coupled with the channel attention mechanism, has enhanced its feature extraction capability, it still facing the challenge of model overfitting due to small and unbalanced data distribution. In scenarios where training samples are limited and imbalanced, manifesting as noisy and biased data samples. During forward propagation, the loss weight calculated by the model influences the model's parameter update direction through backpropagation. The greater the loss weight, the more pronounced the impact of the sample on updating the model parameters. To address potential overfitting caused by these samples, we opt to assign smaller loss weights to high-noise samples and higher loss weights to samples from smaller classes, mitigating the adverse effects of noise and bias on model training. However, manually designing loss weights is unreliable to solving the experimental scenarios adaptively. To tackle this weight assignment issue, we employ a meta-learning algorithm based on learning heavy weights to enhance the training process of the fault diagnosis model.

Assume there is a biased and noisy training set sample pair (x_i, y_i), $1 \leq i \leq N$, and we also have a small unbiased and clean validation set (x_j, y_j), $1 \leq j \leq N$, where $M \gg N$ is the total number of validation set samples. The validation set samples are usually derived from the training set. $\phi(x, \theta)$ denotes the neural network model with θ as the model parameters. Our objective is to minimize the loss function $L(\hat{y}, y)$, where $\hat{y}$ is the output of the neural network. In previous model training, assuming that the loss function of the training set is: $\frac{1}{N}\sum_{i=1}^{N} L(\hat{,}y) = \frac{1}{N}\sum_{i=1}^{N} f_i\theta$ where the weights of each input sample are equal, our goal is to reduce the loss of the training set through gradient descent. By calculating the similarity between the training set and the

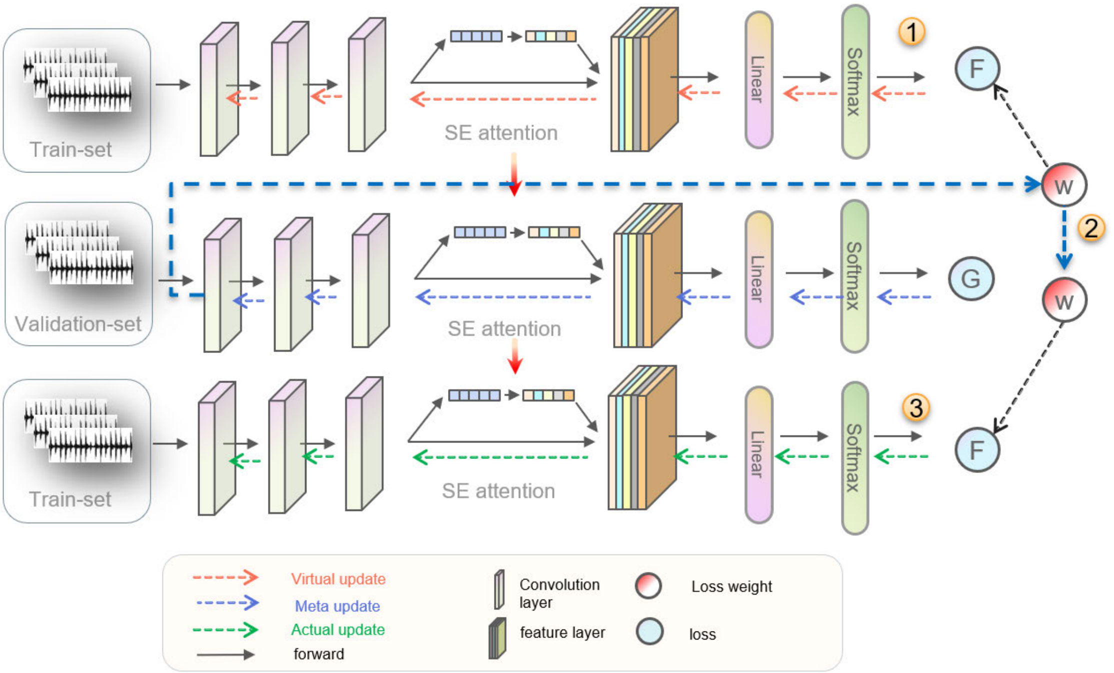


FIGURE 5. Framework of proposed LRS: The model mainly consists of three steps: 1). Use the training set (imbalanced) to update the network and obtain a set of updated model parameters. 2). Obtain verification loss based on the verification set (small and balanced) and use the loss to update the weight of the samples. 3). Use the updated weight to retrain the model parameters based on the training set.

validation set gradient, we optimize the weight of the sample to better minimize the weighted loss. The model parameter update formula is as follows:

$$\theta^*(w) = \arg\min_{\theta} \sum_{i=1}^{N} \epsilon_i f_i(\theta) \tag{8}$$

where ϵ_i represents the loss weight of the sample, and w is the network parameter. Our objective is to optimize the loss weight of the original sample by minimizing the loss of the validation set. The update formula for ϵ is as follows:

$$\epsilon^* = \arg\min_{\epsilon \geq 0} \frac{1}{M} \sum_{i=1}^{M} f_i^v(\theta^*(w)) \tag{9}$$

Due to the negative impact of sample loss, it often leads to significant fluctuations in the loss during the model training process, reducing the stability of the model. Therefore, in Formula 9, all ϵ^* weights of negative samples are excluded by gradient clipping. Updating weights online requires two nested optimization loops. Initially, we employ gradient descent to optimize the loss of the training set, using Stochastic Gradient Descent (SGD) as the optimization algorithm. At each training step, a sample (x_i, y_i) is randomly chosen from the training set, where $1 \leq i \leq n$ and n is the mini-batch size. In the case of the SGD optimizer, a virtual update is performed, and the formula for the virtual update is as follows:

$$\theta_{t+1} = \theta_t - \alpha \nabla(\frac{1}{n} \sum_{i=1}^{n} f_i(\theta_t)) \tag{10}$$

where α is the learning rate. Then, a clean and balanced sample from the validation set is fed into the updated network, and its sample loss is calculated. We aim to ensure that the loss of the original network on the validation set is also sufficiently small, thereby guaranteeing that the model is better suited for training with clean and unbiased samples. By using the loss, the original sample weight is updated using one-step gradient descent, empowering the network with the ability to classify clean and balanced samples. The formula for the meta-update loss weight is as follows:

$$\begin{aligned} & f_{i,\epsilon}(\theta) = \epsilon_i f_i(\epsilon) \\ & \hat{\theta}_{t+1}(\epsilon) = \theta_t - \alpha \nabla \sum_{i=1}^{n} f_{i,\epsilon}(\theta)\big|_{\theta=\theta_t} \end{aligned} \tag{11}$$

However, multiple weight updates using gradient descent is still a time-consuming process. Therefore, we choose to employ a single-step gradient descent.

$$\begin{aligned} & u_{i,t} = -\eta \frac{\partial}{\partial \epsilon_{i,t}} \frac{1}{m} \sum_{j=1}^{m} f_{ji}^v(\theta_{t+1}(\epsilon))\Big|_{\epsilon_{i,t}=0} \\ & \widetilde{w}_{i,t} = \max(u_{i,t}, 0) \end{aligned} \tag{12}$$

Lately, in applying gradient clipping, the weights remain positive, mitigating the adverse effects of negative sample loss on training stability. Subsequently, we normalize all sample weights to ensure their sum is one, preventing the occurrence of the explosion and disappearance phenomena in the final weighted loss:

$$w_{i,t} = \frac{\widetilde{w}_{i,t}}{(\sum_j \widetilde{w}_{j,t}) + \delta(\sum_j \widetilde{w}_{j,t})} \tag{13}$$

The symbol δ in Formula 13 represents a discrete function, taking the value 1 when $\sum_j \widetilde{w}_{j,t}$ is 0, and 0 otherwise. This is employed to prevent the denominator from becoming zero. At this stage, the updated weights inherently incorporate directional information aimed at minimizing the validation set loss. Consequently, the network adjusts in the direction of better adaptation to small and balanced samples. The formula for gradient descent with the updated weights in the last step is expressed as follows:

$$\theta_{t+1} = \theta_t - \alpha\nabla(\frac{1}{N}\sum_{i=1}^{N} w_{i,t}f_i(\theta_t)) \tag{14}$$

Given that the updated weights incorporate gradient information aimed at minimizing the validation set loss, we observe the reduction in both training and validation set losses. While ensuring full utilization of training set sample information, a small fraction of clean and balanced weights is proposed for the training process, thereby enhancing network generalization and improving stability in handling class imbalance and noise samples. The detailed parameter update process is outlined in Algorithm 1.

Algorithm 1 Meta-Reweighted Optimization Enforced Civil Fault Diagnosis

```
Require: $\theta_0$, $D_{Train}$, Iteration number $K$, Step size $\alpha$
while $k < K$ do
    Get $\mathcal{T}_t = \{x_t, y_t\}_{t=1}^{N_t}$ from $D_{Train}$
    Get $\mathcal{T}_v = \{x_v, y_v\}_{v=1}^{N_v}$ from $D_{Train}$
    Forward propagation on $\mathcal{T}_t \longrightarrow \hat{y}_t = f(\mathcal{T}_t; \theta_k)$
    Initial the weight $\epsilon \longrightarrow 0$
    Calculate $L_t = \sum_{i=1}^{N_t} \epsilon \times l(y_{t,i}, \hat{y}_{t,i})$
    Get the grad $\nabla\theta_k$
    Calculate $\hat{\theta}_k = \theta_k - \alpha\nabla\theta_k$
    Forward propagation on $\mathcal{T}_v \longrightarrow \hat{y}_v = f(\mathcal{T}_v; \hat{\theta}_k)$
    Calculate $L_v = \frac{1}{N_v}\sum_{i=1}^{N_v} l(y_{v,i}, \hat{y}_{v,i})$
    Get the grad $\nabla\epsilon$
    Update the weight $\tilde{w}_k$ according to Formula.13
    Recalculate $\hat{L}_t = \sum_{i=1}^{N_t} \tilde{w}_k \times l(y_{v,i}, \hat{y}_{v,i})$
    Update the model parameters $\theta_{k+1}$ based on $\hat{L}_t$
end while
```

D. META GRADIENT BOOSTING

Finally, the fault diagnosis model, incorporating meta-learning and convolutional neural network, is depicted in Fig. 5. The training steps primarily involve virtual update, meta-update, and actual update.

However, there are certain challenges in the practical application of the algorithm. The weight clipping process eliminates negative sample weights, particularly in the later stages of training, where samples exhibit high imbalance. This leads to the disappearance of weights, even though the sum of sample weights is 1, a significant proportion of samples are assigned zero weight after reweighting. This substantially reduces the utilization efficiency of training set samples, equivalent to a drastic reduction in the effective size of the training samples. Even in the case of an imbalanced training set, the sample information is valuable, and the disappearance of sample weights introduces issues such as network overfitting. To address these issues, we aim to enhance the algorithm's stability by refining the weight clipping method. Specifically, the original gradient clipping formula is expressed as follows:

$$\begin{aligned} u_{i,t} &= -\eta\frac{\partial}{\partial\epsilon_{i,t}}\frac{1}{m}\sum_{j=1}^{m} f_{ji}^{v}(\theta_{t+1}(\epsilon))\Big|_{\epsilon_{i,t}=0} \\ \widetilde{w}_{i,t} &= \max(u_{i,t}, 0) \end{aligned} \tag{15}$$

By incorporating a positive bias into the gradient, we enhance the effectiveness of the updated weights. This modification preserves the original distribution of different sample weights while mitigating the issue of weight disappearance. Consequently, it enables the efficient utilization of sufficient number of training samples. This enhancement contributes to improved stability and accuracy in the fault diagnosis algorithm. The refined gradient clipping formula is presented as follows:

$$\hat{u}_{i,t} = u_{i,t} - c * \min(u_{i,t}) \tag{16}$$

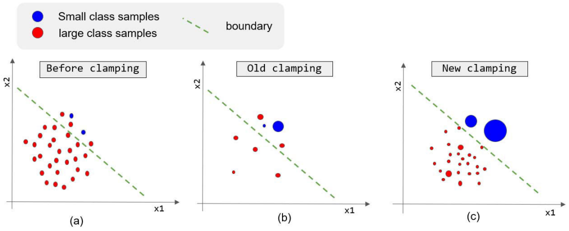


FIGURE 6. Algorithm contrast between old gradient clamping and new gradient clamping.

Fig. 6 illustrates the principle of the improved gradient clipping. As depicted in Fig. 6, the blue circles represent the losses of small-class samples, the black circles represent the losses of large-class samples, and the green dotted line represents the classification boundary. The size of the sample loss area corresponds to the weight assigned to the sample loss, and the number of circles indicates the number of sample losses involved in the model update. While the original gradient clipping strategy in Fig. 6.(b) distributes the loss weights reasonably, it discards a significant portion of the original large-class samples. This leads to a low utilization rate of the original samples, reducing the number of available training samples and making the model susceptible to overfitting issues. In contrast, as shown in Fig. 6.(c), the improved gradient clipping strategy maximizes the utilization of the original fault samples. The sample losses still receive a reasonable weight distribution, ensuring the effective utilization of the original fault sample information. Consequently, the accuracy and stability of the fault diagnosis algorithm are significantly enhanced.

IV. EXPERIMENTS AND DISCUSSION

A. DATASETS

In this research, we conducted experiments on two datasets, including the bearing fault dataset and the wing beam fault dataset. Additionally, contrast experiments were implemented to compare the performance of the proposed method with the standard CNN.

1) CWRU BEARING DATASET

As depicted in Fig. 7, the Case Western Reserve University (CWRU) datasets [18] are obtained from the bearing fault simulation machine. The testing machine primarily includes the dynamometer, torque transducer, and encoder, drive and bearing system, electric motor, fan, and bearing system. The motor speed is controlled through the controller. Single-point faults were induced in the test bearings using electro-discharge machining with fault diameters of 7 mils, 14 mils, 21 mils, 28 mils, and 40 mils (1 mil = 0.001 inches).

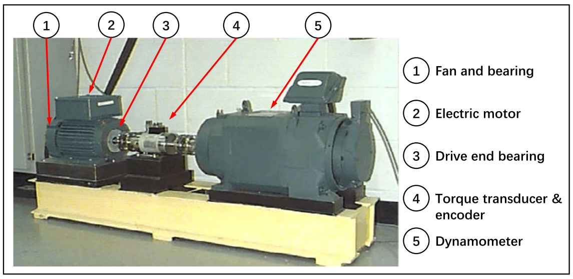


FIGURE 7. Experimental facility of CWRU dataset.

Vibration data were collected using accelerometers, which were affixed to the housing with magnetic bases. Accelerometers were positioned at the 12 o'clock at both the drive end and fan end of the motor housing. In some experiments, the accelerometer was attached to the motor supporting base plate as well. Vibration signals were captured using a 16-channel DAT recorder. Digital data were collected at 12,000 samples per second, and data were also collected at 48,000 samples per second for drive end bearing faults. Speed and horsepower data were obtained using the torque transducer/encoder. In Table 1, all states, including health state, inner raceway fault, ball fault, and the outer raceway fault, were categorized into seven classes (two health states and five fault states) based on different fault modes.

TABLE 1. Detailed description of CWRU datasets.

Frequency	Fault Mode	Description
10*12kHz	Health State	the normal bearing
	Ball 1	0.007 inch ball fault
	Ball 2	0.014 inch ball fault
	Ball 3	0.021 inch ball fault
	Inner Raceway 1	0.007 inch inner raceway fault
	Inner Raceway 2	0.014 inch inner raceway fault
	Inner Raceway 3	0.021 inch inner raceway fault
	Outer Raceway 1	0.007 inch outer raceway fault
	Outer Raceway 2	0.014 inch outer raceway fault
	Outer Raceway 3	0.021 inch outer raceway fault
7*48kHz	Health State	the normal bearing
	Ball 1	0.007 inch ball fault
	Ball 2	0.014 inch ball fault
	Ball 3	0.021 inch ball fault
	Inner Raceway 1	0.007 inch inner raceway fault
	Inner Raceway 2	0.014 inch inner raceway fault
	Inner Raceway 3	0.021 inch inner raceway fault

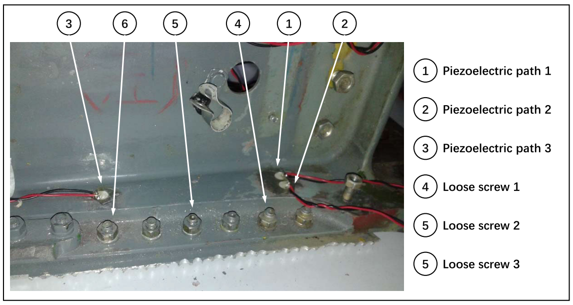


FIGURE 8. Beam and sensor location.

TABLE 2. Detailed description of SEU datasets.

Working Condition	Fault Mode
20HZ-0V	Health State
	Ball fault
	Outer Raceway Fault
	Inner Raceway Fault
	Combination Fault

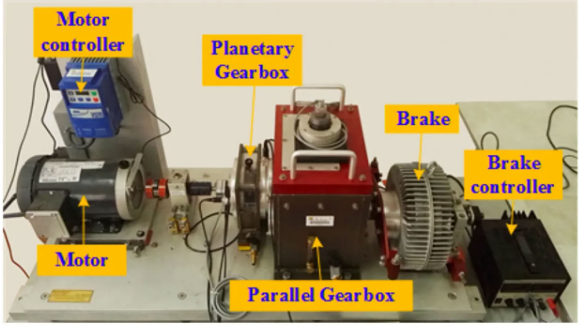


FIGURE 9. Experimental facility of SEU dataset.

2) AWB FAULT DATASET

As depicted in Fig. 8, the Aircraft Wing Beam (AWB) fault dataset [18] was obtained from the beam fault simulation machine. The test beam comprises three piezoelectric sensors. In our experiments, piezoelectric patch 1 served as the excitation source, and piezoelectric patch 3 as the receiver, with a space of 120mm between the two sensors. The experimental excitation signal frequency is 200kHz, the number of wave points is 40000, the sampling rate is fixed at 10MHz, the number of sampling points is 10000, the average number is 100, and the wave amplitude is ±70V. Screw loosening and simulated damage were introduced to the test. In this paper, we focus solely on screw loosening

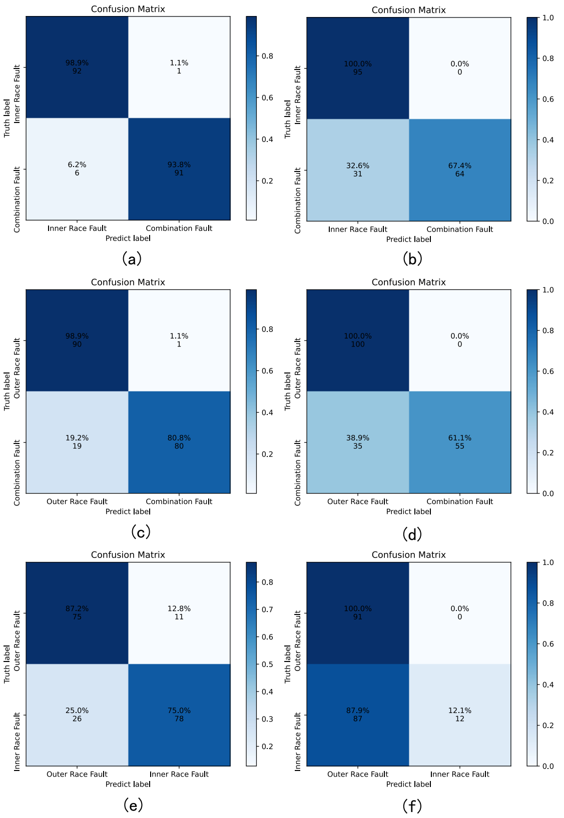


FIGURE 10. Confusion matrix on SEU dataset with a 0.95 data imbalance: (a) Confusion matrix of our method on task C-IR/Comb. (b) Confusion matrix of CNN on task C-IR/Comb. (c) Confusion matrix of our method on task C-OR/Comb. (d) Confusion matrix of CNN on task C-OR/Comb. (e) Confusion matrix of our method on task C-OR/IR. (f) Confusion matrix of CNN on task C-OR/IR.

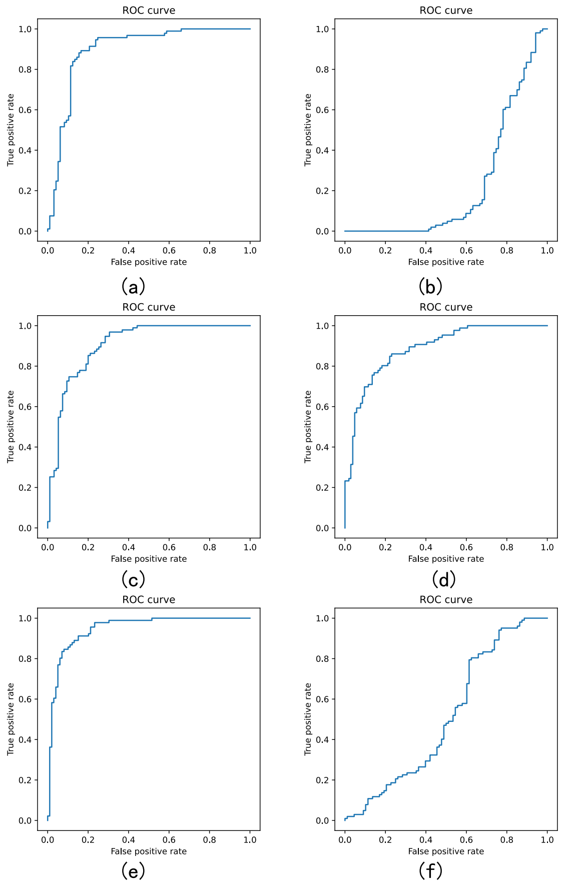


FIGURE 11. ROC curve of our method and CNN on SEU dataset with 0.95 data imbalance: (a) Confusion matrix of our method on task C-BALL/OR. (b) Confusion matrix of CNN on task C-BALL/OR. (c) Confusion matrix of our method on task C-BALL/IR. (d) Confusion matrix of CNN on task C-BALL/Comb. Each color.(e)Confusion matrix of our method on task C-BALL/Comb. (e) Confusion matrix of CNN on task C-BALL/Comb.

faults, and these faults are classified into three categories based on the location of the loosened screw (screws 1/2/3 are loose, respectively).

3) SEU FAULT DATASET

As illustrated in Fig. 9, the Southeast University (SEU) datasets [18] were acquired from the drive-train dynamic simulator. A comprehensive dataset comprising 8 channels of data was systematically collected. Two specific channels were meticulously chosen, encompassing datasets pertaining to bearing and gear data. Each sub-dataset encapsulated operational data associated with a single healthy state and four fault states under two distinct working conditions, namely 20Hz-0v and 30Hz-2v. The specifics of the dataset under the 20Hz-0v condition are clarified in Table 2.

B. EVALUATION METRICS

In our experiments, it is challenging to effectively reflect the performance of the classifier under conditions of limited sample imbalance. Therefore, to accurately evaluate the classifier's performance with imbalanced training samples, following previous studies [35], we propose following quantitative indicators to measure the test effect of the models. They are the average fault identification accuracy (Acc), precision for class one (P1), precision for class two (P2), recall for class one (R1), recall for class two (R2), confusion matrix, and the ROC curve.

TABLE 3. Fault diagnosis task based on CWRU dataset.

Task	State	Abbreviation
Detection	Normal / Ball Fault	D-12-N/B
	Normal / Inner Raceway Fault	D-12-N/IR
	Normal / Outer Raceway Fault	D-12-N/OR
	Normal / Ball Fault	D-48-N/B
	Normal / Inner Raceway Fault	D-48-N/IR
Classification	Ball fault / Inner Raceway Fault	C-12-B/IR
	Inner / Outer Raceway Fault	C-12-IR/OR
	Ball / Outer Raceway Fault	C-12-B/OR
	Ball / Inner Raceway Fault	C-48-B/IR

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

$$P1 = \frac{TP}{TP + FP} \tag{18}$$

$$R1 = \frac{TP}{TP + FN} \tag{19}$$

TABLE 4. Fault diagnosis task based on AWB dataset.

Task	State	Task number
Detection	Normal / No.1 screw loosing	D-N/S1
	Normal / No.2 screw loosing	D-N/S2
	Normal / No.3 screw loosing	D-N/S3
Classification	No.1 screw / No.2 screw loosing	C-1/S2
	No.2 screw / No.3 screw loosing	C-2/S3
	No.3 screw / No.1 screw loosing	C-1/S3

$$P2 = \frac{TN}{TN + FN} \tag{20}$$

$$R2 = \frac{TN}{FP + TN} \tag{21}$$

where TP is the true positive (positive samples are successfully predicted as positive), TN denotes the true negative (negative samples are correctly predicted to be negative), FP represents the false positive (negative samples are incorrectly predicted to be positive), and FN is the false negative (positive samples are incorrectly predicted to be negative).

C. DATA IMBALANCE EXPERIMENTS

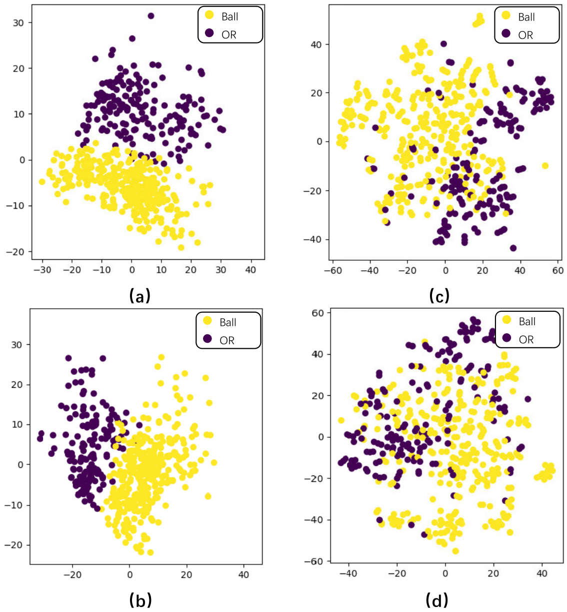
The fault diagnosis task is formulated based on the aforementioned datasets. Three types of tasks are designed using the three datasets, namely fault detection and fault classification. In Table 3, Table 4 and Table 5, task numbers are assigned to distinguish different tasks. The letters in the task number represent the task category (D and C), and the system state to be diagnosed as (N, B, IR, OR, S and Comb), respectively. Additionally, the numbers (12 and 48) in the task number based on the CWRU dataset denote the sampling frequency and the number (20) in the task number based on the SEU dataset represents the working condition. Adhering to these principles, we meticulously devised twenty five sets of fault diagnosis experiments based on the three datasets. The datasets were strategically subsampled to generate a class imbalance binary classification task, where one of the two classes predominates the training data distribution. The imbalance degree of the data is systematically varied as 0.6, 0.7, 0.8, and 0.9, respectively. To prevent our method from gaining an unfair advantage by training on more data, we carefully separated the balanced validation set from the training set.

The final accuracy results on the three datasets are presented in Table 6, Table 7 and Table 8, respectively. It is evident that our method consistently achieves satisfactory results across almost all tasks. In the CWRU dataset and SEU dataset, the algorithm maintains a high accuracy, and even with increasing data imbalance, the algorithm's accuracy is remain robust. Similarly, in the AWB dataset, the algorithm demonstrates commendable performance. In comparison, CNN also achieves good performance when the data imbalance is low, but as the imbalance increases, the performance of the algorithm drops quickly.

Beyond comparative experiments on accuracy, we conducted additional experiments based on tasks in the SEU data set. We generated confusion matrices and ROC curves, while also calculating recall and precision rates. As shown in Fig. 10, in analyzing the confusion matrix, we observed that our method effectively copes with the fault diagnosis task under imbalanced training samples. In contrast, the CNN is heavily biased toward the majority class. Therefore, although it can recognize the majority class, it struggles when a minority fault occurs (as shown in Fig. 10(e)).

TABLE 5. Fault diagnosis task based on SEU dataset.

Task	State	Task number
Detection	Normal / Ball Fault	D-20-Ball
	Normal / Outer Raceway Fault	D-20-OR
	Normal / Inner Raceway Fault	D-20-IR
	Normal / Combination Fault	D-20-Comb
Classification	Ball fault / Outer Raceway Fault	C-Ball/OR
	Ball fault / Inner Raceway Fault	C-Ball/IR
	Ball fault / Combination Fault	C-Ball/Comb
	Outer Raceway fault / Inner Raceway Fault	C-OR/IR
	Outer Raceway Fault / Combination Fault	C-OR/Comb
	Inner Raceway Fault / Combination Fault	C-IR/Comb



(a) (c) (b) (d)

FIGURE 12. The t-SNE visualization of feature representation on C-12-B/OR task with 0.6 and 0.9 data imbalance: (a) Our method trained with 0.6 data imbalance. (b) Our method trained with 0.9 data imbalance. (c) CNN trained with 0.6 data imbalance. (d) CNN trained with 0.9 data imbalance. Each color in the graphs stands for a category of fault state.

As shown in Fig. 11, we drew the ROC curve based on SEU dataset. In analyzing the ROC curve, we observe that the curve of the proposed method completely improves performance of the CNN, indicating supriority of our proposed approach among these tasks.

In comparing with CNN, Fig. 12 additionally provides the visualization results. It is evident that both methods perform well when the training data is nearly balanced. However, when the training set becomes severely imbalanced, our method maintains its performance (99.06% $\longrightarrow$ 99.01%), while CNN experiences degradation (99.04% $\longrightarrow$ 90.72%). This discrepancy indicates that our method utilizes meta-knowledge to assess the distinctions between

TABLE 6. Accuracy of different tasks on CWRU datasets(percentage).

Imbalance	D12-N/B		D-12-N/IR		D-12-N/OR		D-48-N/B	
	Proposed	CNN	Proposed	CNN	Proposed	CNN	Proposed	CNN
0.6	99.99±0.01	99.79±0.05	99.99±0.01	98.05±0.14	99.99±0.01	98.01±0.16	99.99±0.01	95.35±0.36
0.7	99.99±0.01	98.64±0.25	99.97±0.02	98.35±0.29	99.99±0.01	97.15±0.21	99.99±0.01	98.23±0.17
0.8	99.98±0.01	99.88±0.14	99.98±0.02	99.98±0.01	99.98±0.02	99.98±0.01	99.98±0.01	97.33±0.25
0.9	99.81±0.11	99.12±0.26	99.98±0.11	98.69±0.32	99.96±0.02	98.88±0.03	99.96±0.03	98.22±0.01

Imbalance	C-12-B/IR		C-12-IR/OR		C-12-B/OR		C-48-B/IR	
	Proposed	CNN	Proposed	CNN	Proposed	CNN	Proposed	CNN
0.6	99.81±0.01	99.88±0.04	99.25±0.43	98.65±0.13	99.06±0.21	99.04±0.15	91.57±0.23	92.35±0.17
0.7	99.04±0.32	99.64±0.21	98.31±0.13	98.25±0.15	99.71±0.03	98.45±0.24	91.72±0.11	90.15±0.26
0.8	99.41±0.27	99.18±0.14	98.84±0.13	96.28±0.21	98.91±0.12	94.98±0.21	90.91±0.21	85.52±0.15
0.9	99.41±0.13	94.12±0.21	97.44±0.17	91.63±0.27	99.01±0.31	90.72±0.43	89.05±0.33	84.93±0.41

TABLE 7. Accuracy of different tasks on AWB datasets(percentage).

Imbalance	D-N/S1		D-N/S2		D-N/S3		C-S1/S2		C-S2/S3		C-S1/S3	
	Proposed	CNN	Proposed	CNN	Proposed	CNN	Proposed	CNN	Proposed	CNN	Proposed	CNN
0.6	95.28±0.21	96.22±0.34	98.29±0.12	98.21±0.19	99.55±0.21	99.99±0.01	95.53±0.31	96.02±0.14	89.65±0.17	90.33±0.25	91.21±0.18	92.23±0.22
0.7	94.32±0.17	93.24±0.27	96.38±0.24	91.12±0.34	97.14±0.11	93.15±0.23	97.19±0.15	90.23±0.16	90.35±0.22	88.12±0.31	98.14±0.07	90.36±0.13
0.8	93.52±0.31	89.94±0.15	96.68±0.23	88.45±0.21	97.79±0.39	94.26±0.31	95.38±0.21	90.37±0.18	71.61±0.32	60.23±0.43	95.93±0.29	91.01±0.14
0.9	89.95±0.23	60.17±0.31	98.39±0.14	53.12±0.42	97.99±0.21	68.14±0.34	91.66±0.27	78.31±0.39	72.66±0.25	53.12±0.38	95.23±0.12	78.61±0.18

TABLE 8. Accuracy of different tasks on SEU datasets(percentage).

Imbalance	D-20-N/B		D-20-N/OR		D-20-N/IR		D-20-N/Comb		C-B/OR	
	Proposed	CNN	Proposed	CNN	Proposed	CNN	Proposed	CNN	Proposed	CNN
0.6	98.89±0.24	99.01±0.10	98.68±0.11	98.72±0.19	92.83±0.12	95.65±0.20	94.95±0.15	94.32±0.21	97.78±0.15	97.35±0.24
0.7	99.29±0.24	87.34±0.29	98.99±0.13	89.66±0.27	97.27±0.26	73.35±0.18	88.18±0.14	72.33±0.13	97.27±0.18	86.34±0.17
0.8	98.69±0.14	65.33±0.26	98.99±0.24	63.15±0.29	92.22±0.16	69.25±0.16	93.64±0.12	63.93±0.18	97.17±0.17	57.35±0.18
0.9	98.89±0.10	56.76±0.12	99.29±0.21	57.59±0.17	86.97±0.18	54.26±0.24	91.82±0.27	57.11±0.10	95.25±0.21	53.32±0.28

Imbalance	C-B/IR		C-B/Comb		C-OR/IR		C-OR/Comb		C-IR/Comb	
	Proposed	CNN	Proposed	CNN	Proposed	CNN	Proposed	CNN	Proposed	CNN
0.6	97.11±0.26	98.12±0.12	99.09±0.13	99.05±0.12	97.78±0.14	98.01±0.18	95.96±0.29	95.35±0.29	97.17±0.15	96.32±0.23
0.7	96.57±0.18	92.64±0.18	99.29±0.15	98.35±0.11	98.18±0.20	93.15±0.19	92.63±0.25	90.23±0.14	94.99±0.28	92.33±0.05
0.8	96.67±0.12	73.29±0.21	98.18±0.23	69.26±0.27	98.08±0.14	79.34±0.09	92.93±0.14	75.33±0.03	89.61±0.22	80.38±0.21
0.9	95.96±0.14	61.02±0.27	94.14±0.10	55.26±0.25	97.47±0.16	53.14±0.03	93.64±0.19	62.02±0.16	81.21±0.28	63.15±0.31

TABLE 9. Comparison of results between proposed method and CNN on SEU data set with 0.95 data imbalance.

Task	P1		P2		R1		R2	
	Ours	CNN	Ours	CNN	Ours	CNN	Ours	CNN
C-Ball/OR	0.88	0.63	0.98	1.00	0.98	1.00	0.88	0.38
C-Ball/IR	0.94	0.83	1.00	1.00	1.00	1.00	0.93	0.83
C-Ball/Comb	0.91	0.78	1.00	1.00	1.00	1.00	0.89	0.71
C-OR/IR	0.96	0.81	0.97	1.00	0.97	1.00	0.96	0.88
C-OR/Comb	0.88	0.75	0.97	1.00	0.98	1.00	0.8	0.61
C-IR/Comb	0.67	0.57	0.83	1.00	0.84	1.00	0.65	0.13

imbalanced and balanced datasets, and then employs this information to adapt the model. In contrast, CNN tends to demonstrate overfitting in imbalanced conditions.

The recall and precision performances of imbalanced ratios ($U_{Test} = 0.95$) of proposed method and CNN are summarized in Table 9. Through comparison, we observed that CNN tends to severely overfit to categories with more samples in the training data when it is extremely unbalanced. This leads to a significant decline in diagnostic performance. In contrast, our method continues to maintain good diagnostic performance.

D. ABLATION STUDIES

To assess the contributions of various modules in the model, we conducted ablation studies, with a specific emphasis on experimenting and analyzing the data processing module and the channel attention module.

1) DATA PROCESSING MODULE

Fig. 13 illustrates the diagnostic results of the proposed method on different tasks for the two datasets, utilizing time and frequency-domain samples as model inputs. Notably, models based on frequency-domain inputs generally exhibit superior performance on the CWRU dataset, while time-domain input-based models perform better on the AWB dataset. In relatively balanced tasks with a small category gap, the performances of the two input types are nearly identical. Additionally, the performance of the two inputs is almost similar on the AWB dataset. However, on the CWRU dataset, the frequency-domain input yields higher performance, especially in challenging small-sample fault diagnosis tasks. This is because frequency-domain data characterize signals on a high-level scale, making it easier for deep models to automatically learn valuable features. Nevertheless, experiments on the AWB dataset reveal that such an approach is not universally effective. Therefore, in civil aircraft fault diagnosis, an appropriate preprocessing method should be selected based on the type and characteristics of the input data.

2) CHANNEL ATTENTION MODULE

To underscore the effectiveness of our proposed approach, we selected challenging tasks for analysis. Fig. 14 presents

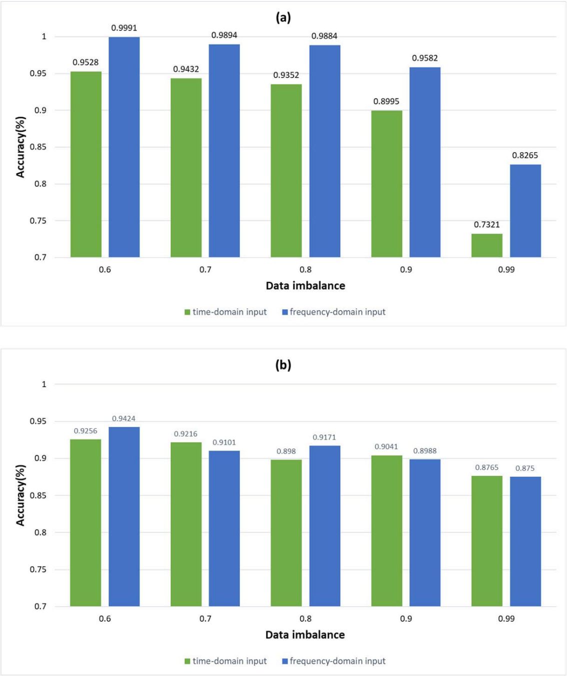


FIGURE 13. Influences of data processing module on (a) CWRU data set; (b) AWB data set.

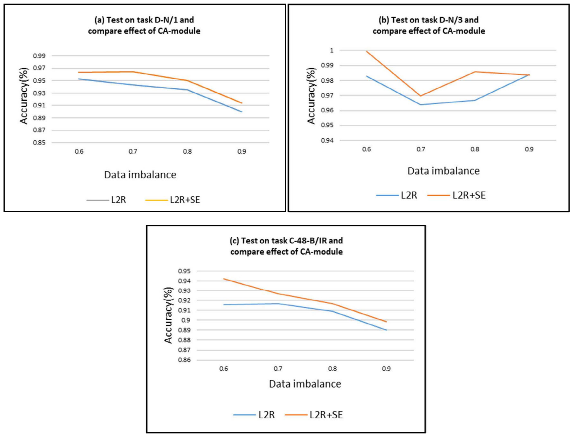


FIGURE 14. Influence of channel attention module on (a) task D-N/1; (b) task D-N/3; (c) task C-48-B/IR.

the results of ablation experiments on three demanding tasks, comparing the performance of the proposed method with and without the Channel Attention (CA) module. It is evident that the channel attention module enhances the model's performance, although the degree of performance improvement diminishes as the task difficulty increases. This suggests that the channel attention block, as a plug-and-play module, serves an effective solution to enhance the performance of fault diagnosis. When confronted with straightforward tasks where the model can extract abundant and excellent features, the module can allocate attention to more crucial features, thereby improving diagnostic accuracy. However, as the task complexity rises, and the model struggles to extract high-quality features. The reason is that the module can only allocate attention to the already extracted features and cannot contribute to feature extraction, leading to a decrease in its impact accordingly.

V. CONCLUSION

In this study, we tackled the challenge of small-sample fault diagnosis in civil aircraft, a common issue in engineering applications. We developed an L2R model based on meta-learning and the channel attention algorithm for effective fault diagnosis under limited sample conditions. Our method adopts the essence of meta-learning to approach the L2R optimization framework by dynamically reweighting training samples based on the degree of data imbalance. Unlike previous approaches, our method automatically adjusts sample weights during model training, reducing reliance on expert experience and minimizing human intervention. Additionally, the embedded channel attention module enhances feature representation, thereby improving the model's fault diagnosis performance. We conducted case studies on CWRU and AWB datasets, considering various levels of data imbalance. We used the traditional Convolutional Neural Network (CNN) as a baseline for comparison. Specifically, the key findings are as follows: (1) The L2R model shows advantages in addressing small-sample challenges, particularly evident with increasing data imbalance. (2) While the performance of the L2R model tends to decline with higher task difficulty, it achieves comparable performance to CNN in experiments with balanced data distribution. Thus, our method is specifically designed for scenarios characterized by significant data imbalance.

Despite the demonstrated effectiveness of our proposed approach in addressing experimental settings with imbalanced data distribution, its rationale lies in the hybridized meta-learning mechanism and attention regularization strategy. However, the introduced functional learning inevitably increases the model's complexity and reduces its generalization ability. Similarly, popular meta-learning models, such as MAML, also encounter challenges like local minimizers and saddle points during optimization [36]. Therefore, our future research direction aims to explore strategies to reduce local minimizers [37] to further enhance the robustness of our proposed method in solving small-sample fault diagnosis tasks and proactively improve its performance across a wide range of imbalance degrees.

REFERENCES

[1] H. Zhu, J. Cheng, C. Zhang, J. Wu, and X. Shao, "Stacked pruning sparse denoising autoencoder based intelligent fault diagnosis of rolling bearings," *Appl. Soft Comput.*, vol. 88, Mar. 2020, Art. no. 106060.

[2] W. Zhang, G. Biswas, Q. Zhao, H. Zhao, and W. Feng, "Knowledge distilling based model compression and feature learning in fault diagnosis," *Appl. Soft Comput.*, vol. 88, Mar. 2020, Art. no. 105958.

[3] F. Xu, W. T. P. Tse, and Y. L. Tse, "Roller bearing fault diagnosis using stacked denoising autoencoder in deep learning and GathGeva clustering algorithm without principal component analysis and data label," *Soft Comput.*, vol. 73, pp. 898–913, Dec. 2018.

[4] L. Chen, Q. Li, C. Shen, J. Zhu, D. Wang, and M. Xia, "Adversarial domain-invariant generalization: A generic domain-regressive framework for bearing fault diagnosis under unseen conditions," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 1790–1800, Mar. 2022.

[5] R. Chen, J. Zhu, X. Hu, H. Wu, X. Xu, and X. Han, "Fault diagnosis method of rolling bearing based on multiple classifier ensemble of the weighted and balanced distribution adaptation under limited sample imbalance," *ISA Trans.*, vol. 114, pp. 434–443, Aug. 2021.

[6] C. Zhao, G. Liu, and W. Shen, "A balanced and weighted alignment network for partial transfer fault diagnosis," *ISA Trans.*, vol. 130, pp. 449–462, Nov. 2022.

[7] T. Li, Z. Zhao, C. Sun, L. Cheng, X. Chen, R. Yan, and R. X. Gao, "WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 52, no. 4, pp. 2302–2312, Apr. 2022.

[8] J. Shu, Z. Xu, and D. Meng, "Small sample learning in big data era," 2018, *arXiv:1808.04572*.

[9] T. Zhang, J. Chen, F. Li, K. Zhang, H. Lv, S. He, and E. Xu, "Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions," *ISA Trans.*, vol. 119, pp. 152–171, Jan. 2022.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[11] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1322–1328.

[12] I. Goodfellow, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–11.

[13] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 113–123.

[14] S. Shao, P. Wang, and R. Yan, "Generative adversarial networks for data augmentation in machine fault diagnosis," *Comput. Ind.*, vol. 106, pp. 85–93, Apr. 2019.

[15] J. Yang, G. Xie, and Y. Yang, "An improved ensemble fusion autoencoder model for fault diagnosis from imbalanced and incomplete data," *Control Eng. Pract.*, vol. 98, May 2020, Art. no. 104358.

[16] Y. Zeng, X. Wu, and J. Chen, "Bearing fault diagnosis with denoising autoencoders in few labeled sample case," in *Proc. 5th IEEE Int. Conf. Big Data Anal. (ICBDA)*, May 2020, pp. 349–353.

[17] Q. Li, B. Tang, L. Deng, Y. Wu, and Y. Wang, "Deep balanced domain adaptation neural networks for fault diagnosis of planetary gearboxes with limited labeled data," *Measurement*, vol. 156, May 2020, Art. no. 107570.

[18] B. Yang, Y. Lei, F. Jia, and S. Xing, "A transfer learning method for intelligent fault diagnosis from laboratory machines to real-case machines," in *Proc. Int. Conf. Sensing,Diagnostics, Prognostics, Control (SDPC)*, Aug. 2018, pp. 35–40.

[19] X. Dong, H. Gao, L. Guo, K. Li, and A. Duan, "Deep cost adaptive convolutional network: A classification method for imbalanced mechanical data," *IEEE Access*, vol. 8, pp. 71486–71496, 2020.

[20] P. Peng, W. Zhang, Y. Zhang, Y. Xu, H. Wang, and H. Zhang, "Cost sensitive active learning using bidirectional gated recurrent neural networks for imbalanced fault diagnosis," *Neurocomputing*, vol. 407, pp. 232–245, Sep. 2020.

[21] X. Li, H. Jiang, K. Zhao, and R. Wang, "A deep transfer nonnegativity-constraint sparse autoencoder for rolling bearing fault diagnosis with few labeled data," *IEEE Access*, vol. 7, pp. 91216–91224, 2019.

[22] Z. He, H. Shao, X. Zhang, J. Cheng, and Y. Yang, "Improved deep transfer auto-encoder for fault diagnosis of gearbox under variable working conditions with small training samples," *IEEE Access*, vol. 7, pp. 115368–115377, 2019.

[23] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 1126–1135.

[24] Y. He, C. Zang, P. Zeng, M. Wang, Q. Dong, and Y. Liu, "Rolling bearing fault diagnosis based on meta-learning with few-shot samples," in *Proc. 3rd Int. Conf. Ind. Artif. Intell. (IAI)*, Nov. 2021, pp. 1–6.

[25] J. Wu, Z. Zhao, C. Sun, R. Yan, and X. Chen, "Few-shot transfer learning for intelligent fault diagnosis of machine," *Measurement*, vol. 166, Dec. 2020, Art. no. 108202.

[26] S. Dixit, N. K. Verma, and A. K. Ghosh, "Intelligent fault diagnosis of rotary machines: Conditional auxiliary classifier GAN coupled with meta learning using limited data," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[28] H. Fan, Y. Yang, and M. Kankanhalli, "Point 4D transformer networks for spatio-temporal modeling in point cloud videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14204–14213.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[30] H. Fan, L. Zhu, Y. Yang, and F. Wu, "Recurrent attention network with reinforced generator for visual dialog," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 3, pp. 1–16, Aug. 2020.

[31] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[32] S. Zhang, S. Wu, X.-Y. Jing, and F. Wu, "Domain adaptation network with parameter shared domain-specific attention for fault diagnosis," in *Proc. 2nd Int. Conf. Artif. Intell. Comput. Eng. (ICAICE)*, Nov. 2021, pp. 590–593.

[33] X. Zheng, J. Wu, and Z. Ye, "An end-to-end CNN-BiLSTM attention model for gearbox fault diagnosis," in *Proc. IEEE Int. Conf. Prog. Informat. Comput. (PIC)*, Dec. 2020, pp. 386–390.

[34] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[35] J. Wu, Z. Zhao, C. Sun, R. Yan, and X. Chen, "Learning from class-imbalanced data with a model-agnostic framework for machine intelligent diagnosis," *Rel. Eng. Syst. Saf.*, vol. 216, Dec. 2021, Art. no. 107934.

[36] M. Abbas, Q. Xiao, L. Chen, P.-Y. Chen, and T. Chen, "Sharp-MAML: Sharpness-aware model-agnostic meta learning," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 10–32.

[37] D. Bahri, H. Mobahi, and Y. Tay, "Sharpness-aware minimization improves language model generalization," 2021, *arXiv:2110.08529*.

[38] Y. Xu, S. Li, X. Yan, J. He, Q. Ni, Y. Sun, and Y. Wang, "Multiattention-based feature aggregation convolutional networks with dual focal loss for fault diagnosis of rotating machinery under data imbalance conditions," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–11, 2024.

[39] H. Nizam, S. Zafar, Z. Lv, F. Wang, and X. Hu, "Real-time deep anomaly detection framework for multivariate time-series data in industrial IoT," *IEEE Sensors J.*, vol. 22, no. 23, pp. 22836–22849, Dec. 2022.

[40] C. Cheng, X. Liu, B. Zhou, and Y. Yuan, "Intelligent fault diagnosis with noisy labels via semi-supervised learning on industrial time series," *IEEE Trans. Ind. Informat.*, vol. 19, no. 6, pp. 7724–7732, Jan. 2023.

GUANG ZHAO received the master's degree in computer science (distributed systems and applications) from Pierre and Marie Curie University Paris VI (Sorbonne University since 2018). He is currently pursuing the Ph.D. degree with Shanghai Jiao Tong University, Shanghai, China. He was an Avionics System Integration Engineer of COMAC's C919 Program. He is also in charge of international cooperation in research and innovation. His research interests include machine learning, safety, and reliability of civil aircraft.

SHIQIANG HU received the Ph.D. degree from Beijing Institute of Technology. He is currently the Dean of the School of Aeronautics and Astronautics, Shanghai Jiao Tong University. He is also a Full Professor with Shanghai Jiao Tong University. He has been at the helm of numerous research projects, including those funded by the National Science Foundation and the 863 National High Technology Plan. With over 300 publications to his name, he has also effectively mentored more than 20 Ph.D. students. His research interests include machine learning, image understanding, and nonlinear filters.

JIAYUAN FAN received the B.S. degree from Northwestern Polytechnical University, in 2021. He is currently pursuing the master's degree with Shanghai Jiao Tong University. His current research interests include machine learning and machinery fault diagnostic.

QIANG GUO received the Ph.D. degree from Shanghai Jiaotong University, in 2005. He is currently an Engineer with Commercial Aircraft of China (COMAC). His current research interests include reliability, safety, and maintenance of civil aircraft.

BO SHEN received the B.S. degree from Nanjing University of Aeronautics and Astronautics. He is currently the Vice President of Commercial Aircraft of China (COMAC) and the President of the COMAC Shanghai Aircraft Design and Research Institute (SADRI). His research interests include design and development of civil aircraft.

LINGKUN LUO was a Research Assistant and a Postdoctoral Researcher with the Department of Mathematics and Computer Science, École Centrale de Lyon, and a member with the LIRIS Laboratory. He is currently a Research Fellow with Shanghai Jiao Tong University. He has authored over 30 research articles, including publications in *International Journal of Computer Vision*, ACM-CS, IEEE Transactions on Image Processing, IEEE Transactions on Cybernetics, and IEEE Transactions on Information Forensics and Security. His research interests include machine learning, pattern recognition, and computer vision.