Journal of Dynamics, Monitoring and Diagnostics, 2025, 4, 65-75
https://doi.org/10.37965/jdmd.2025.727

RESEARCH ARTICLE

An Interpretable Few-Shot Framework for Fault Diagnosis of Train Transmission Systems with Noisy Labels

Haiquan Qiu,[1] Biao Wang,[2] Yong Qin,[2] Ao Ding,[2] Zhixin He,[3] Jing Liu,[3] and Xin Huang[4]
[1]School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China
[2]State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing, China
[3]Guangzhou Metro Group Co. Ltd., Guangzhou, China
[4]China Railway Jinan Group Co. Ltd., Jinan, China

(Received 06 January 2025; Revised 27 February 2025; Accepted 19 March 2025; Published online 19 March 2025)

Abstract: Intelligent fault diagnosis technology plays an indispensable role in ensuring the safety, stability, and efficiency of railway operations. However, existing studies have the following limitations. 1) They are typical black-box models that lacks interpretability as well as they fuse features by simply stacking them, overlooking the discrepancies in the importance of different features, which reduces the credibility and diagnosis accuracy of the models. 2) They ignore the effects of potentially mistaken labels in the training datasets disrupting the ability of the models to learn the true data distribution, which degrades the generalization performance of intelligent diagnosis models, especially when the training samples are limited. To address the above items, an interpretable few-shot framework for fault diagnosis with noisy labels is proposed for train transmission systems. In the proposed framework, a feature extractor is constructed by stacked frequency band focus modules, which can capture signal features in different frequency bands and further adaptively concentrate on the features corresponding to the potential fault characteristic frequency. Then, according to prototypical network, a novel metric-based classifier is developed that is tolerant to mislabeled support samples in the case of limited samples. Besides, a new loss function is designed to decrease the impact of label mistakes in query datasets. Finally, fault simulation experiments of subway train transmission systems are designed and conducted, and the effectiveness as well as superiority of the proposed method are proved by ablation experiments and comparison with the existing methods.

Keywords: few-shot learning; intelligent fault diagnosis; interpretability; noisy labels; train transmission systems

I. INTRODUCTION

Nowadays, railway transit become a crucial power driving modern urbanization and economic development with its advantages of high transport efficiency, low energy consumption and environmental friendliness. However, with train speed and operational intensity continuing to increase, it is becoming more and more important to ensure the stability and reliability of railway trains. Transmission systems are the vital subsystems of trains, including traction motors, reduction gearboxes, and axle boxes, which determine the efficiency of power transmission and the reliability of train operations. Faults in any one component in the train transmission systems could affect the safe operation of trains and even lead to serious accidents. Thus, it is essential to develop advanced diagnosis techniques for train transmission systems.

In recent decades, intelligent fault diagnosis methods have caught the attention of researchers, due to their powerful adaptive feature extraction capability [1]. For instance, Jin *et al.* [2] applied the grey wolf optimization algorithm to optimize the parameters of deep belief networks, proposing an improved deep belief network method for diagnosing weak faults in axle box bearings. He *et al.* [3] proposed a multi-scale spatiotemporal residual capsule neural network to realize compound fault diagnosis of motors. Zhong *et al.* [4] developed a multilevel discriminative feature learning method to diagnose gearbox compound faults by separating compound faults and single faults using feature-level and decision-level contraction modules. Zhang *et al.* [5] transformed one-dimensional vibration signals into two-dimensional fast spectral kurtosis, and proposed a compound fault diagnosis method for gearboxes combining fast spectral kurtosis and multi-branch CNN. Ding *et al.* [6] design an elastic expandable fault diagnosis method of three-phase motors using continual learning for class-added sample accumulations.

Although the aforementioned studies demonstrate the effectiveness of intelligent diagnosis methods, they lack interpretability, which reduces the credibility and user acceptance of the models, thereby limiting their further application in high-reliability scenarios such as railway transit. Recently, some studies have applied interpretable neural network models to intelligent fault diagnosis. Li *et al.* [7] combined the feature extraction capability of wavelet bases with the learning capability of convolutional kernels, proposing an interpretable WPConvNet wavelet packet kernel-constrained network for bearing fault diagnosis. Qin *et al.* [8] integrated the physical prior knowledge of bearing dynamic models with the parameter identification capability of neural networks, proposing an interpretable inverse physics-informed neural network method that significantly improves diagnosis accuracy and reliability of the diagnosis models. Abid *et al.* [9] fused the advantages of automatic feature extraction in deep learning with the physical interpretability of Sinc functions, proposing an

Corresponding author: Yong Qin (e-mail: yqin@bjtu.edu.cn).

interpretable Deep-SincNet architecture for motor fault diagnosis, thereby making the feature extraction process more transparent. In these studies, the SincNet-based fault diagnosis method utilizes the explicit physical characteristics of the Sinc function, enabling users to intuitively understand the behaviors of the network, thereby enhancing the transparency and credibility of the models. Additionally, they show strong competitiveness since they require learning only a small number of parameters for feature extraction, significantly reducing model complexity. Nonetheless, the SincNet-based [9] fault diagnosis methods typically extract features through by simply stacking Sinc function-based bandpass filters, which overlooks the discrepancies in the importance of different features, significantly compromising the robustness of the extracted features. Hence, it is necessary to design an adaptively feature focus mechanism, aiming to enhance of key features related to the fault characteristic frequency.

Furthermore, existing studies ignore the effects of noisy labels, i.e. potentially mistaken labels in the training datasets. In real scenes, samples are labeled manually according to maintenance logs, making it difficult to eliminate mislabeling caused by subjective experience and misinformation. The presence of noisy labels in the training samples can directly disrupt the ability to learn the true data distribution of the models, causing the classification boundary to shift, which reduces the accuracy and generalization performances of the models. Aiming to address the above problem, it become a new research hotspot in academia and industry to develop deep learning algorithms with noisy labels. For instance, Fir Dunkin *et al.* [10] proposed a diagnosis approach based on multi-granularity information fusion to combat noisy labels, which can efficiently diagnose rotating machinery without requiring prior knowledge of the signal-to-noise ratio. Huang *et al.* [11] designed a novel label noise robust auxiliary classifier generative adversarial network to address mislabeled samples occurred in fault diagnosis of wind turbine gearbox. Wang *et al.* [12] proposed a novel iterative error self-correction for diagnosis of mechanical equipment, which can automatically model the distribution of correct labels as well as gradually identify and automatically correct the mislabeled samples. Cheng *et al.* [13] presented an intelligent fault diagnosis method with noisy labels through a semi-supervised learning architecture, where two deep neural networks are trained simultaneously to filter errors from label noise. He *et al.* [14] introduced a unified label noise-tolerant fault diagnosis framework based on bounded neural network that combined implicit weighted learning and a bounded loss mechanism, enabling valuable features to be extracted from noisy labels. However, these studies rely on massive training samples to mitigate the effects of noise labels. In actual train operation and maintenance, obtaining sufficient high-quality labeled fault samples is challenging. When trained with limited samples, models are more likely to fitting erroneous information, which exacerbates the degradation of the diagnosis performance. Consequently, there is an urgent need to develop a few-show learning mechanism that is tolerant to mislabeled samples.

To address the above items, an interpretable few-shot framework for fault diagnosis with noisy labels is proposed for train transmission systems, which not only provides clear interpretability but also can achieve efficient diagnose from limited samples with noisy labels. As part of the network construction, a frequency band focus (FBF) module is developed to capture signal features in different frequency bands, and adaptively emphasize those features corresponding to the potential fault characteristic frequency by fusing the signal features according to learnable weights. Then, a novel metric-based classifier is introduced so as to reduce the interference of noisy labels in limited support samples by aggregating class prototypes based on the feature median. Besides, a mislabeling insensitive loss function is designed, which combines the strengths of cross-entropy error and mean absolute error, effectively mitigating the disturbance caused by query samples. Last but not least, the effectiveness of the proposed interpretable few-shot diagnosis framework is verified by taking the fault diagnosis study case of traction motors, driving gearboxes, and axle boxes. The main contributions of this article can be summarized as follows:

(1) A frequency band focus module is developed to extract and emphasize signal features across different frequency bands. This module constructs convolution kernels based on the Sinc function, enabling it to capture signal features in different frequency. Then, the features are adaptively fused with weights based on prior knowledge, highlighting the features related to potential faults adjustably.

(2) A novel metric-based classifier is introduced, which establishes robust class prototypes through feature median aggregation from support sets. This design ensures tolerance to mislabeled samples and enables effective diagnosis with limited training data through similarity-based classification.

(3) A mislabeling insensitive loss function is designed to reduce the interference of mislabeled query samples during the optimization process by combining the flexibility of the cross-entropy function with the robustness of the mean absolute error function.

The remainder of this article is organized as follows. Section II introduces the preliminary work including SincNet and few-show learning. Section III presents the proposed framework and key techniques in detail. Section III validates the effectiveness and superiority of the proposed framework through experiments. Finally, Section IV concludes this article.

II. PRELIMINARY

A. SINCNET

SincNet is a deep neural network architecture specifically designed for signal processing applications [15]. Unlike traditional approaches that employ standard convolutional kernels, SincNet utilizes Sinc function-based filters in its initial convolutional layer, which enables direct learning of finite impulse response (FIR) filters in the time domain, effectively capturing critical frequency characteristics. The design philosophy originates from fundamental filter bank theory, where rectangular band-pass filters conventionally decompose signals into discrete frequency components via spectral analysis. Formally, such band-pass operations can be mathematically expressed as the differential combination of two low-pass filter responses:

$$G_{f_1,f_2}(f) = rect\left(\frac{f}{2f_2}\right) - rect\left(\frac{f}{2f_1}\right) \tag{1}$$

where f_1 and $f_2 (f_2 > f_1)$ are the low and high cut-off frequencies, and $rect(\cdot)$ is the frequency response of the rectangular low-pass filter defined as follows:

$$rect(x) = \begin{cases} 0, & if\ |x| > 0.5, \\ 0.5, & if\ |x| = 0.5, \\ 1, & if\ |x| < 0.5, \end{cases} \tag{2}$$

By performing inverse Fourier transform on the filter function G, we can get the impulse response of the filter, represented by the Sinc function:

$$\begin{aligned} g_{f_1,f_2}[n] &= 2f_2 \mathrm{sinc}\left(2\pi f_2 n\right) \\ &\quad - 2f_1 \mathrm{sinc}\left(2\pi f_1 n\right) \end{aligned} \tag{3}$$

In general, the Sinc function is multiplied by a window function to smooth out the abrupt discontinuities of the Sinc function.

$$g^w_{f_1,f_2}[n] = g_{f_1,f_2}[n] \cdot w[n] \tag{4}$$

The window function used is Hamming window. SincNet is increasingly used in various fields since it can efficiently and interpretably process signals.

B. FEW-SHOT LEARNING

Data-driven intelligent methods based on deep learning have been favored by academia and industry since their powerful feature learning and representation capabilities. However, their performance heavily relies on massive labeled datasets, which are difficult to acquire in practical industrial scenarios. The scarcity of training samples is prone to leads to severe overfitting, significantly compromising the generalization capability of intelligent networks. Hence, few-shot learning methods have been developed and examined in recent times. Lu *et al.* [16] proposed a transfer relationship network, effectively addressing the issues of bearing fault diagnosis under inadequate samples. Zhang *et al.* [17] applied Siamese Networks based on metric to bearing fault diagnosis with incomplete samples. Wang *et al.* [18] combined weighted Manhattan distance with prototypical networks, proposing an improved prototypical network method for fault diagnosis of limited data. Among the above methods, prototypical networks have demonstrated superior performance as they do not require extra network structures or auxiliary datasets, and the learning processes are simple. In prototypical networks, class prototype is constructed through a parametric embedding function $f_\phi(\cdot)$. For each class, the prototype vector is computed as the centroid of embedded support samples in the latent space:

$$\boldsymbol{c}_k = \frac{1}{|S_k|} \sum_{(\boldsymbol{x}_i, y_i) \in S_k} f_\phi\left(\boldsymbol{x}_i\right) \tag{5}$$

where, S_k represents the support sets, $\boldsymbol{c}_k \in \mathbb{R}^M$ expresses the class prototypes. The classification mechanism operates by measuring similarity between query features and class prototypes. For a given query sample $\boldsymbol{x}$, the network generates probabilistic class predictions through a Softmax distance metric in the embedding space:

$$p_\phi(y = k|\boldsymbol{x}) = \frac{\exp\left(-d\left(f_\phi(\boldsymbol{x}), \boldsymbol{c}_k\right)\right)}{\sum_{k'} \exp\left(-d\left(f_\phi(\boldsymbol{x}), \boldsymbol{c}_{k'}\right)\right)} \tag{6}$$

where $d(\cdot, \cdot)$ denotes the distance function, and $p_\phi(y = k|\boldsymbol{x})$ represents the predicted probability of $\boldsymbol{x}$ belonging to class k. The optimization objective minimizes the cross-entropy loss through SGD:

$$J(\phi) = -\log p_\phi(y = k|\boldsymbol{x}) \tag{7}$$

where $J(\phi)$ represents the loss function. The training paradigm employs an episodic strategy: Each training episode is constructed by first randomly sampling a class subset from the full training data, then partitioning each selected class's examples into support and query subsets. This approach effectively simulates few-shot learning scenarios during optimization.

III. PROPOSED INTERPRETABLE FEW-SHOT FRAMEWORK

A. INTERPRETABLE FEW-SHOT FRAMEWORK FOR FAULT DIAGNOSIS OF TRAIN TRANSMISSION SYSTEMS WITH NOISY LABELS

Traditional intelligent diagnosis algorithms are typical black-box models, and the lack of interpretability makes it difficult to meet the increasingly strict requirements for model security in practical applications, especially in the railway industry. Besides, in the practical maintenance of railway trains, fault samples are generally limited and may contain mislabeled samples, degrading the generalization performances and diagnosis accuracy of intelligent diagnosis models. Hence, an interpretable few-shot framework for fault diagnosis with noisy labels is proposed in this paper for train transmission systems, as presented in Fig. 1. Firstly, a feature extractor is constructed by stacked frequency band focus modules, which can capture signal features in different frequency bands and further adaptively concentrate on the features corresponding to the potential fault characteristic frequency. After that, a novel metric-based classifier is proposed, which is tolerant to mislabeled support samples by aggregating class prototypes based on the feature median of the support datasets. Last but not least, a new loss function is designed to decrease the impact of label mistakes in query datasets through combining the flexibility of the cross-entropy function with the robustness of the mean absolute error function. The key techniques in the proposed network are detailed below, including the frequency band focus module, metric-based classifier tolerant to mislabeled support samples, and the loss function suppressing mislabeled query samples.

B. FREQUENCY BAND FOCUS MODULE

In general, traditional convolution filters lack interpretability and fail to make full use of prior knowledge, making it difficult to extract fault-related features, which affects the generalization performances of the models. Aiming to solve this problem, a frequency band focus module is proposed, which captures signal features from different frequency bands by adopting convolution kernels based on the Sinc function, as shown in Fig. 2. According to the working conditions of training samples, an indicator is assigned to each convolution kernel, which represents its significance to a specific frequency band, so as to adaptively highlight the features corresponding to the fault characteristic frequency. Firstly, a convolution kernel based on the Sinc function is used to perform convolution operations on the

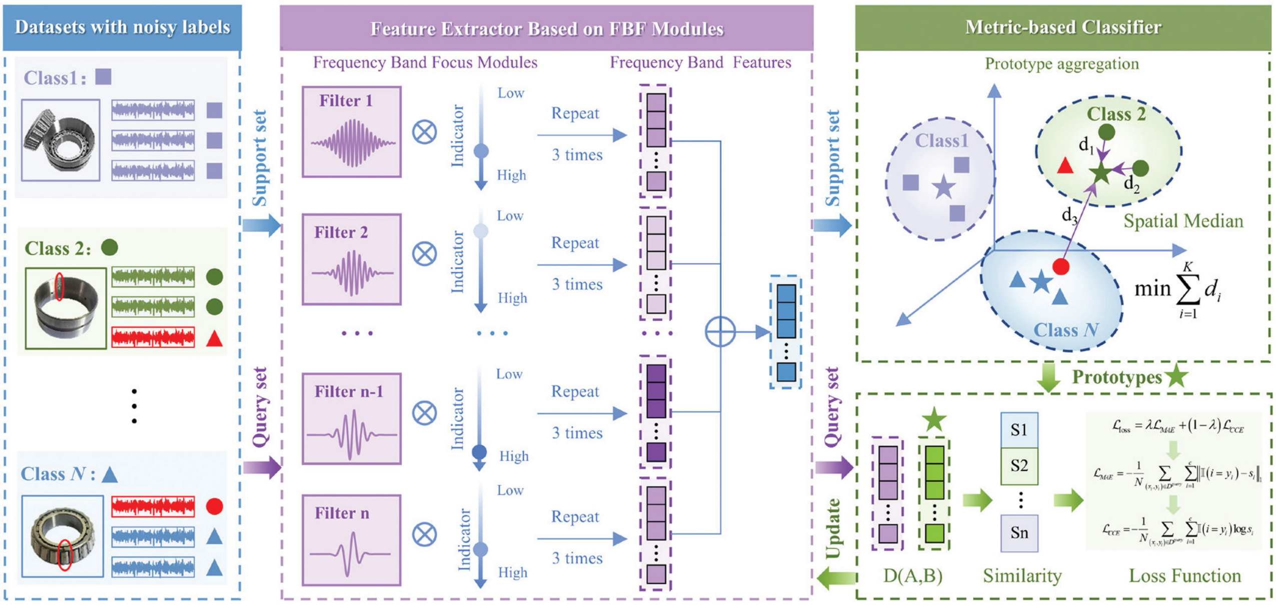
Datasets with noisy labels
Class1 :
Class 2:
Class N :
Support set
Query set
Feature Extractor Based on FBF Modules
Frequency Band Focus Modules
Frequency Band Features
Filter 1
Filter 2
Filter n-1
Filter n
Indicator
Low
High
Repeat
3 times
Metric-based Classifier
Prototype aggregation
Class1
Class 2
Class N
d_1
d_2
d_3
Spatial Median
$\min \sum_{i=1}^{K} d_i$
Prototypes
Support set
Query set
Update
S1
S2
Sn
D(A,B)
Similarity
Loss Function
$\mathcal{L}_{\text{loss}} = \lambda \mathcal{L}_{MAE} + (1-\lambda)\mathcal{L}_{CCE}$
$\mathcal{L}_{MAE} = -\frac{1}{N}\sum_{(x_i,y_i)\in D^{query}} \sum_{i=1}^{c} \lVert \mathbb{I}(i=y_i) - s_i \rVert_1$
$\mathcal{L}_{CCE} = -\frac{1}{N}\sum_{(x_i,y_i)\in D^{query}} \sum_{i=1}^{c} \mathbb{I}(i=y_i)\log s_i$

Fig. 1. Proposed interpretable few-shot framework for fault diagnosis of train transmission systems with noisy labels.

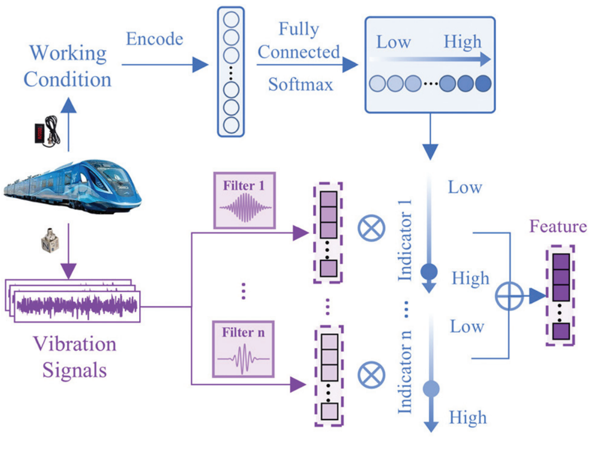

Fig. 2. The structure and working mechanism of a frequency band focus module.

input data. The time-domain function of an ideal low-pass filter is approximately the Sinc function. Convolution kernels based on the Sinc function can learn well-defined filters with clear frequency bandwidths, improving the physical interpretability and generalization of the models. The k-th filter in the l-th module can be expressed as:

$$\begin{aligned} \boldsymbol{W}_k^{(l)}\left(t,\boldsymbol{f}_l,\boldsymbol{f}_h\right) \\ = \left[g\left(\boldsymbol{f}_l\right) - g\left(\boldsymbol{f}_h\right)\right] \cdot w_{\text{window}}\ (t) \end{aligned} \tag{8}$$

$$g(\boldsymbol{f}) = 2\boldsymbol{f}\,\text{sinc}(2\pi\boldsymbol{f}t) \tag{9}$$

where, t denotes time, $\boldsymbol{f}_l$ and $\boldsymbol{f}_h$ is the upper and lower cutoff frequencies of the bandpass filter, which depends on the signal sampling rateFs with a preset number of convolution N. Thus the bandpass is $[(k-1)\frac{Fs}{2N}, k\frac{Fs}{2N}]$. $w_{\text{window}}\ (t)$ is the hamming window function, which is used to smooth the Sinc function and reduce edge effects. The Sinc function sinc(t) is the time domain function of the low pass filter, which can be described as:

$$\text{sinc}(2\pi\boldsymbol{f}t) = \frac{\sin(2\pi\boldsymbol{f}t)}{2\pi\boldsymbol{f}t} \tag{10}$$

The signal features from different frequency bands are captured by the convolution process, which can be described as:

$$\begin{aligned} \boldsymbol{y}_k^{(l)}(t) &= \left(\boldsymbol{x}^{(l)} * \boldsymbol{W}_k^{(l)}\right)(t) \\ &= \int_{-\infty}^{\infty} \boldsymbol{x}^{(0)}(\tau)\boldsymbol{W}_k^{(1)}(t-\tau,\boldsymbol{f}_l,\boldsymbol{f}_h)d\tau \end{aligned} \tag{11}$$

where, $*$ denotes the convolution operator. It is mentioned that the Sinc function is a symmetric function, whether or not the filter is flipped during convolution does not affect the result. $\boldsymbol{x}^{(l)}$ indicates the input of layer l, $\boldsymbol{W}_k^{(l)}$denotes the k-th filter in the l-th module, $\boldsymbol{y}_k^{(l)}(t)$ represents the output features of the k-th filter. After the features of different frequency bands are obtained, an indicator is assigned to each frequency band feature to dynamically enhance the critical features. Traditional intelligent fault diagnosis networks typically summarize features by averaging them after extraction. However, due to the varying amounts of information contained in features extracted by different filters, this simple averaging method struggles to effectively capture fault-related critical frequency information under different working conditions. Therefore, in the FBF module, the band features are multiplied by the indicators to fully utilize the working condition information of train transmission systems, so as to adjustably enhance the features in frequency ranges related to potential faults. Firstly, the working condition tensor $\boldsymbol{p} \in \mathbb{R}^{I\times 1}$ is encoded, which records rotation speeds, loads, etc. I denotes the number of working condition indicators. Then, the $\boldsymbol{p}$ is fully connected and activated to obtain the indicator value tensor $\boldsymbol{\alpha} = (\alpha_1,\alpha_2 \ldots \alpha_N) \in \mathbb{R}^{N\times 1}$, as presented in:

$$\boldsymbol{\alpha}^{(l)} = \sigma\left(\boldsymbol{U}_\alpha^1\left(\boldsymbol{U}_\alpha^2\boldsymbol{p} + \boldsymbol{b}_\alpha^2\right) + \boldsymbol{b}_\alpha^1\right) \tag{12}$$

where $\boldsymbol{U}_\alpha^1$, $\boldsymbol{U}_\alpha^2$, $\boldsymbol{b}_\alpha^1$, $\boldsymbol{b}_\alpha^2$ respectively represent the weights and biases of the fully connected layers, and $\sigma(\cdot)$ denotes the activation function. Finally, the indicator value tensor $\boldsymbol{\alpha}^{(l)}$ is assigned to each convolution kernel, as shown in:

$$z^{(l)}(t) = \sigma\left(\left(\sum_{i=1}^{K} \boldsymbol{\alpha}_i \boldsymbol{W}_i^{(l)}\right) * \boldsymbol{x}^{(l)} + \boldsymbol{b}^{(l)}\right) \tag{13}$$

where, $*$ denotes the convolution operator, $z_k^{(l)}(t)$ represents the signal features from different frequency bands, and $\boldsymbol{W}^{(l)} = \{\boldsymbol{W}_1^{(l)}, \boldsymbol{W}_2^{(l)} \cdots \boldsymbol{W}_K^{(l)}\}$ is the k-th filter in the l-th module. The above steps can be summarized in as follows:

$$z^{(l)} = \psi\left(\boldsymbol{x}^{(l)} \middle| \boldsymbol{\theta}^{(l)}\right) \tag{14}$$

where, $\psi(\cdot|\cdot)$ indicates the frequency bands feature extraction operation, $\boldsymbol{\theta}^{(l)}$ is the trainable parameters, including the fully connected layers. In the FBF module, convolution kernels based on the Sinc function act as band-pass filters to extract information from specific frequency bands. Representative speeds and loads can be summarized according to the actual working conditions of the train transmission systems. After obtaining training samples from various typical working conditions, an indicator is assigned to each convolution kernel to represent its attention to a specific frequency band, thereby enhancing the features related to fault-relevant frequency bands. The filter design in the FBF module directly corresponds to frequency domain features and assigns an indicator to each filter, making the extracted features more robust while clearly demonstrating the dynamic focusing process on fault-related frequency bands of the models, thereby enhancing the credibility and interpretability in practical applications of the models.

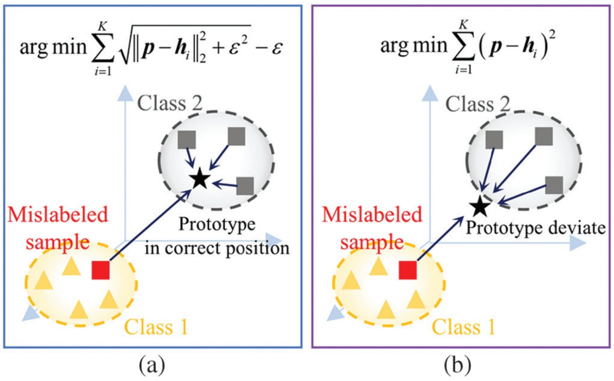
$\arg\min \sum_{i=1}^{K} \sqrt{\|p-h_i\|_2^2+\varepsilon^2}-\varepsilon$
Class 2
Mislabeled sample
Prototype in correct position
Class 1
(a)
$\arg\min \sum_{i=1}^{K} (p-h_i)^2$
Class 2
Mislabeled sample
Prototype deviate
Class 1
(b)

Fig. 3. Schematic diagram of prototype aggregation. (a) Spatial median. (b) Spatial mean.

C. METRIC-BASED CLASSIFIER TOLERANT TO MISLABELED SUPPORT SAMPLES

A metric-based classifier tolerant to mislabeled support samples is further proposed to cope with the issue of mislabeled samples in the support sets. In traditional prototypical networks, the class center is typically represented by calculating the spatial mean of the support set sample features, achieving clear classification decisions. However, the spatial mean is calculated by assigning equal weights to all samples, making it susceptible to being affected by outlier values. When the support set contains mislabeled samples, the spatial features of mislabeled samples may significantly deviate from the correct range, causing the mean-based class prototype to shift from its true position and thus reducing the diagnosis performance of the network. To solve this issue, the proposed classifier calculates the more robust vector median instead of the mean, effectively mitigating the impact of outlier features during class prototype aggregation, as shown in Fig. 3. Actually, the spatial median merely relies on the middle value after sorting, which can greatly reduce the influence of outlier values. For example, suppose there are 5 support samples of a certain class whose feature values are 1,2,3,4,1000. Among them, 1000 is the feature of a mislabeled sample. In this case, the mean is affected by the outlier value of 1000, resulting in the mean-based class prototype close to 500, shifting from its actual position. The representational capability of the median feature vector is demonstrated via visualization validation. Correspondingly, the median is the value in the middle of the sort, which in this example is 3, effectively mitigating the impact of outlier and representing the true center of the class. For scalar data, the median is the middle value of an ordered dataset. Whereas, in vector spaces, the median is the vector that minimizes the total absolute error between itself and all other points. Therefore, the process of aggregating class prototypes can be understood as finding a vector that minimizes the total absolute error between the vector and the features of the same class. Specifically, this can be achieved by constructing a median iterative objective function, which is defined as follows:

$$\begin{aligned} \boldsymbol{p} &= \arg\min \mathrm{R}(\boldsymbol{p}) \\ &= \arg\min \textstyle\sum_{i=1}^{K} \sqrt{\|\boldsymbol{p}-\boldsymbol{h}_i\|_2^2+\epsilon^2}-\epsilon \end{aligned} \tag{15}$$

where, $\mathrm{R}(\boldsymbol{p})$ represents the total distance between the class prototype and the sample feature vectors, K represents the number of vectors in the set, $\|\cdot\|_2$ denotes the $L2$, ϵ is a small constant to prevent division by zero when the estimated median coincides with a vector in the support set, and $\boldsymbol{h}_i$ represents the sample features of the support set, automatically generated by the feature extractor and calculated using the following formula:

$$\boldsymbol{h}_i = F(\boldsymbol{X}_i|\boldsymbol{\theta}) \tag{16}$$

where, $F(\cdot,\cdot)$ represents the feature extraction network, $\boldsymbol{X}_i$ denotes the i-th training sample, and $\boldsymbol{\theta}$ refers to the trainable parameters in the network. Since the median iterative objective function does not have a closed-form solution, Newton's method is used for iterative solving. The iterative formula is as follows:

$$\boldsymbol{p}(t+1) = \boldsymbol{p}(t) - \mathrm{H}^{-1}(\boldsymbol{p}(t)) \cdot \nabla\mathrm{R}(\boldsymbol{p}(t)) \tag{17}$$

where, $\nabla\mathrm{R}(\boldsymbol{p})$ represents the gradient of function $\mathrm{R}(\boldsymbol{p})$, and $\mathrm{H}(\boldsymbol{p})$ is the Hessian matrix of $\boldsymbol{p}$. Both the gradient $\nabla\mathrm{R}(\boldsymbol{p})$ and the Hessian matrix $\mathrm{H}(\boldsymbol{p})$ are derived using the matrix calculus method of numerator layout:

$$\nabla\mathrm{R}(\boldsymbol{p}) = \sum_{i=1}^{K} \frac{\boldsymbol{p}-\boldsymbol{h}_i}{\sqrt{\|\boldsymbol{p}-\boldsymbol{h}_i\|_2^2+\epsilon^2}} \tag{18}$$

$$\mathrm{H}(\boldsymbol{p}) = \left(\textstyle\sum_{i=1}^{K} \frac{1}{\sqrt{\|\boldsymbol{p}-\boldsymbol{h}_i\|_2^2+\epsilon^2}}\right)\boldsymbol{I}_{D\times D} - \boldsymbol{U}\boldsymbol{U}^T \tag{19}$$

where, D represents the dimensionality of the spatial vector, $\boldsymbol{I}_{D\times D}$ is the unit vector, and $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_K]$ is a $D \times K$ matrix composed of sequentially stacked vectors

$\boldsymbol{u}_i = \frac{\boldsymbol{p}-\boldsymbol{h}_i}{(\|\boldsymbol{p}-\boldsymbol{h}_i\|_2^2+\epsilon^2)^{\frac{3}{4}}}$. As an approximation, the second off-diagonal term in the Hessian matrix can be ignored. Therefore, the iterative process can be expressed as:

$$\boldsymbol{P}(t+1) = \boldsymbol{P}(t) - \frac{\sum_{i=1}^{K}\left(\frac{\boldsymbol{p}(t)-\boldsymbol{h}_i}{\sqrt{\|\boldsymbol{p}(t)-\boldsymbol{h}_i\|_2^2+\epsilon^2}}\right)}{\sum_{i=1}^{K}\left(\frac{1}{\sqrt{\|\boldsymbol{p}(t)-\boldsymbol{h}_i\|_2^2+\epsilon^2}}\right)} \tag{20}$$

Through the above iterative steps, the median prototype $\boldsymbol{P} = [p_1, p_2, \cdots, p_n]$ is obtained. The predicted class of the query sample is the class corresponding to the nearest class prototype. Subsequently, the distance between the feature $\boldsymbol{Q}_j$ of the query sample and each class prototype $\boldsymbol{P}_i$ is calculated, using Euclidean distance as the metric:

$$d_{ji} = \|\boldsymbol{Q}_j - \boldsymbol{P}_i\|_2 \tag{21}$$

where, $\boldsymbol{Q}_j$ represents the feature of the j-th query sample, $\boldsymbol{P}_i$ denotes the prototype of the i-th state class, $\|\cdot\|_2$ refers to the $L2$ norm, and d_{ji} indicates the spatial distance between the feature of the j-th query sample and the i-th prototype. In addition, by inputting the spatial distance d_{ji} into the Softmax function, the similarity s_{ji} can be calculated. The mathematical expression for this process is as follows:

$$s_{ji} = \frac{d_{ji}}{\sum_{i=1}^{C} d_{ji}} \tag{22}$$

where C denotes the health status class. The class corresponding to the maximum similarity value is the predicted class of the query sample.

In summary, the metric-based classifier tolerant to mislabeled support samples replaces the noise-sensitive mean with a robust median during class prototype aggregation, effectively mitigating the impact of outlier features in the support sets on class prototypes. This classifier not only enhances the adaptability of the classification process and improves the noisy label tolerance of the models but also makes the aggregation process more intuitive through median calculation, clearly demonstrating the classification mechanism of the model, which significantly enhancing its interpretability.

D. LOSS FUNCTION SUPPRESSING MISLABELED QUERY SAMPLES

During model training, samples are divided into support sets and a query sets. Although the issue of mislabeled samples in the support sets has been suppressed, mislabeled samples may also exist in the query sets. These mislabeled samples can introduce erroneous information into the optimization objective, significantly impairing the classification performance of the models. To deal with this issue, a loss function suppressing mislabeled query samples is proposed, which combines the advantages of cross-entropy error and mean absolute error, effectively mitigating the interference caused by mislabeled samples during training. In traditional prototypical networks, the objective function used to update model parameters is based on the cross-entropy function, which is defined as:

$$\mathrm{L}_{CCE} = -\frac{1}{N}\sum_{(\boldsymbol{x},y)\in D^q}\sum_{i=1}^{C}\mathrm{I}(i=y_i)\log s_{ji} \tag{23}$$

where N represents the number of samples in the query set, D^q denotes the query set, $\mathrm{I}(\cdot)$ is the indicator function, and s_{ji} represents the similarity between the j-th query sample and the i-th health status class. However, the cross-entropy function represents high sensitivity to mislabeled samples, making the models prone to interference when dealing with mislabeled, affecting the diagnosis performance of the models. In contrast, the mean absolute error function demonstrates greater robustness, effectively mitigating the impact of noisy or mislabeled samples. The mean absolute error function is defined as:

$$\mathrm{L}_{MAE} = -\frac{1}{N}\sum_{(\boldsymbol{x},y)\in D^q}\sum_{i=1}^{c}\|\mathrm{I}(i=y_i) - s_{ji}\|_1 \tag{24}$$

where, $\|\cdot\|_1$ represents the $L1$ norm. Nevertheless, the mean absolute error function suffers from gradient sparsity, degrading the performance of the models during model training, especially when there are no mislabeled samples in the query sets. To address this limitation, a loss function suppressing mislabeled query samples is introduced, which incorporates a parameter λ, combining the flexibility of the cross-entropy function with the robustness of the mean absolute error function, achieving a better balance in handling mislabeled samples. Specifically, the cross-entropy function amplifies the loss of mislabeled samples due to its logarithmic operation on classification probabilities. In comparison, the mean absolute error function only computes the linear difference between the predicted and true values, resulting in a more moderate influence from mislabeled samples. As a result, it significantly enhances the robustness and generalization ability of the models. The loss function suppressing mislabeled query samples is defined as:

$$\mathrm{L}_{\mathrm{loss}} = \lambda\mathrm{L}_{MAE} + (1-\lambda)\mathrm{L}_{CCE} \tag{25}$$

where λ is an adjustable parameter, typically set to 0.5, which balances the sensitivity and robustness of the model. L_{MAE} denotes the mean absolute error function, while L_{CCE} represents the cross-entropy function. By appropriately tuning the parameter λ based on the proportion of mislabeled samples, the loss function can dynamically adjust its sensitivity to incorrect labels, ensuring that the gradients derived from such labels are constrained within a reasonable range. λ is typically set to 0.5 according to the cross-validation. This mechanism effectively reduces the impact of mislabeled samples on model optimization, enhancing both the robustness and generalization ability of the model when applied to noisy query samples.

IV. EXPERIMENTAL RESULT AND ANALYSIS

A. DATA ACQUISITION AND NETWORK CONFIGURATION

To verify the proposed interpretable few-shot framework for fault diagnosis of train transmission systems with noisy labels, fault simulation experiments are designed and carried out. As presented in Fig. 4, the experimental platform is designed and adjusted on the basis of the real bogies of subway trains. The signal collected has a sampling frequency of 12 kHz, with 15 signal channels, including triaxial acceleration signal from the drive end and non-drive end of motor, the large gear end and small gear end of the gearbox, and the left and right sides of the axle box. Table I

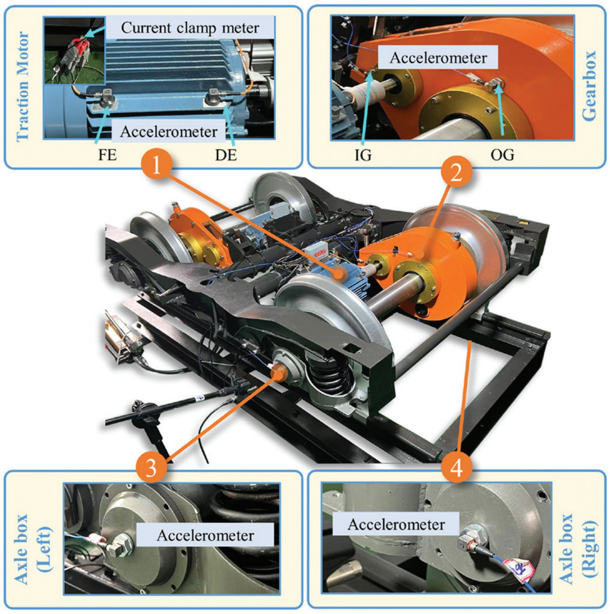

Fig. 4. Experimental platform of subway train transmission systems.

Table I. Overview of working conditions

Serial Number	Speed/Load	Serial Number	Speed/Load
WC1	20Hz/0kN	WC6	40Hz/–10kN
WC2	20Hz/10kN	WC7	60Hz/0kN
WC3	20Hz/–10kN	WC8	60Hz/10kN
WC4	40Hz/0kN	WC9	60Hz/–10kN
WC5	40Hz/10kN		

gives a total of 9 working conditions considered in the experiments. The motor speed is set at 20Hz, 40Hz, and 60Hz to simulate the different train speed. The horizontal load is applied using an electro-hydraulic loading device to simulate the train going straight or round corners, with three types of settings: 0kN,+10kN, and –10kN. The positive horizontal load is directed towards the motor side, while the negative horizontal load is directed towards the gearbox side. As shown in Fig. 5, various various health states of key components in the train transmission systems are taken into account in the experiments, covering 4 health states of the three-phase current of the motor, 8 health states of the gearbox, and 4 health states of the axle box. For each health state of the collected signals, there are 10 samples per working condition, with each channel containing 3200 sampling points. Noisy labels are generated through a label corruption process. Concretely, different proportions of training samples are selected randomly. Then, the original labels of the selected samples are randomly modified toother categories. Finally, these selected samples are mixed back into the original training datasets. The outcomes of selecting network hyperparameters via cross-validation are presented in Table II. Learnable parameters are iteratively updated using the Adam optimizer, with an initial learning rate of 0.001. The total number of training epochs is set to 500.

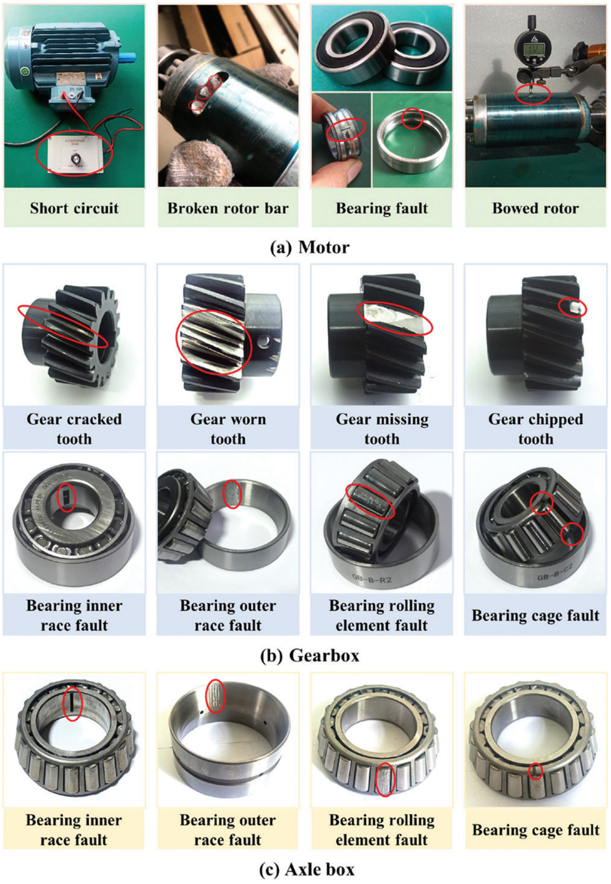

Fig. 5. Photographs of fault simulations.

Table II. Network configuration

Layer	Parameters
Input	–
Frequency Band Focus Module	Size:7 Number:256
Batch Normalization	–
Pooling	Size = 4
Frequency Band Focus Module	Size:7 Number:128
Batch Normalization	–
Pooling	Size = 4
Frequency Band Focus Module	Size:7 Number:64
Batch Normalization	–
Pooling	Size = 4
Metric-based Classifier	–

B. BENEFITS OF THE FREQUENCY BAND FOCUS MODULE AND THE INTERPRETABILITY ANALYSIS

To validate the benefits of the frequency band focus module in fault diagnosis of train transmission systems, ablation experiments are conducted in this section. Three methods are designed and compared with the proposed method. Method I employs traditional convolutional kernels without

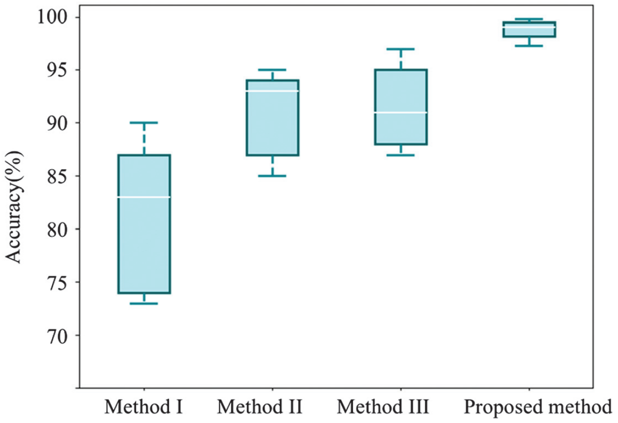

Fig. 6. Fault diagnosis accuracy of four methods.

frequency band focus mechanism. Method II utilizes Sinc-based convolutional kernels but lacks frequency band focus capability. Method III retains frequency band focus mechanism while abandoning Sinc-based kernel construction. The training dataset consisted of 120 samples per fault class. After network training, they are tested using samples that outside of the training datasets. This process of training and testing was repeated ten times in the experiments. Fig. 6 illustrates the comparative diagnosis accuracy across the four methods. It can be seen that the proposed method performs better than the comparative method. This is because the Sinc function-based convolutional kernels explicitly extract frequency band features with clear physical meaning. Building upon this foundation, the frequency band focus mechanism adaptively focuses the frequency bands associated with the fault characteristic frequencies. By contrast, the traditional convolutional kernels exhibit limited interpretability in feature extract and may also dilute critical fault-related frequency components. In conclusion, the benefits of the frequency band focus module is demonstrated through the above ablation experiments.

To verify the interpretability of the proposed diagnosis framework, a visual analysis of the feature selection process of the frequency band focus module is conducted. According to Section III, the frequency band focus module captures signal features in different frequency bands and assigns different weights to each frequency band feature, which adaptively highlights the vital frequency bands that play a role in the decision-making process during the training process. From the signal processing point of view, the constructed convolution kernel based on the Sinc function essentially functions as an FIR filter, where the convolution operation performs filtering while the pooling operation performs spectral folding. The three-layer convolution architecture progressively refines frequency bands through hierarchical processing. Based on prior knowledge, the frequency bands corresponding to the fault characteristic frequencies contain important fault information. A superior fault diagnosis model should prioritize these frequency bands to enhance the diagnosis performance of the models. Therefore, the envelope spectrum of the test samples, the fault characteristic frequencies of the samples, and the frequency bands with high weights during the feature selection process are illustrated in Fig. 7. Specifically, three outer ring fault test samples of axle box bearings were selected, corresponding to motor speeds of 20Hz, 40Hz, and 60Hz, and the fault characteristic parameters are shown in Table III. After inputting these samples into the trained network, the weights for each frequency band in the first layer were calculated, which reflect the importance of each frequency band in the decision-making process. The top 5 frequency bands with the highest weights are shown in Fig. 7, and the shade of the color indicates the value of the

Table III. Fault characteristic parameters of axle box bearings

Motor Rotation Frequency	Output Shaft Rotation Frequency	Outer Ring Fault Characteristic Frequency
20Hz	2.99Hz	24.14Hz
40Hz	5.98Hz	48.28Hz
60Hz	8.97Hz	72.42Hz

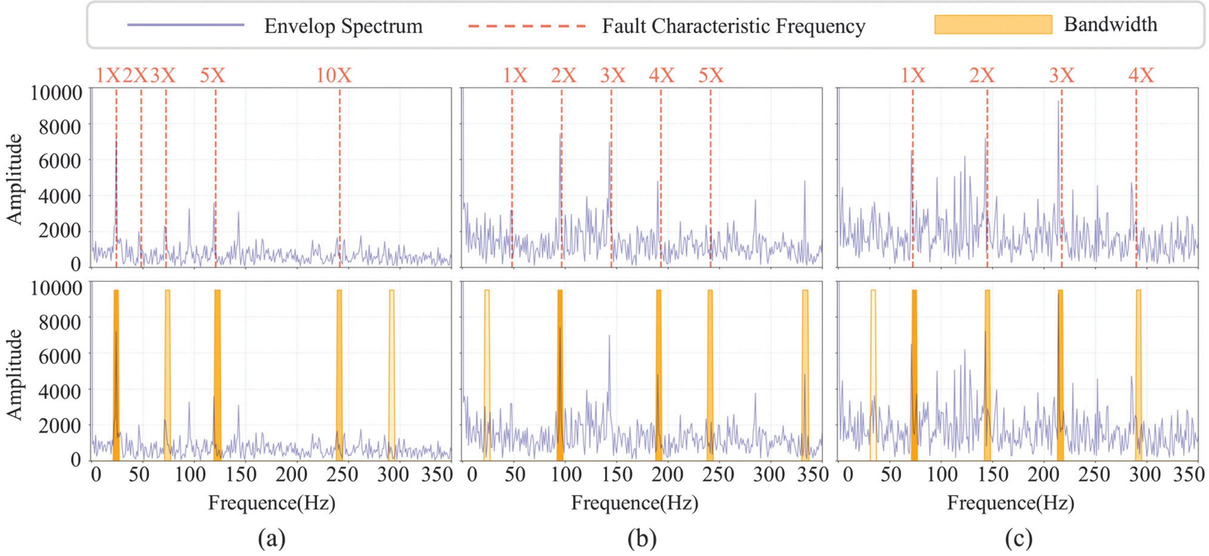

Fig. 7. Fault characteristic frequencies and the frequency bands with high weights. (a) Motor rotation speed of 20Hz. (b) Motor rotation speed of 40Hz. (c) Motor rotation speed of 60Hz.

weights. It can be seen that at a motor rotation speed of 20Hz, the top 5 frequency bands with the highest weights cover the 1 times, 3 times, 5 times, and 10 times frequencies of the outer ring fault characteristic frequency. At a motor rotation speed of 40Hz, the top 5 frequency bands with the highest weights cover the 2 times, 4 times, and 5 times frequencies of the outer ring fault characteristic frequency. At a motor rotation speed of 60Hz, the top 5 frequency bands with the highest weights cover the fundamental frequency, 2 times, 3 times, and 4 times frequencies of the outer ring fault characteristic frequency. It indicates that the frequency band focus module effectively highlights the features corresponding to the potential fault characteristic frequency during the training process, thus enabling the model to identify potential fault modes of the samples. More importantly, this strategy accords with the working principles of signal processing-based fault diagnosis methods, demonstrating excellent interpretability.

C. BENEFITS OF THE METRIC-BASED CLASSIFIER TOLERANT TO MISLABELED SUPPORT SAMPLES AND THE LOSS FUNCTION SUPPRESSING MISLABELED QUERY SAMPLES

Ablation experiments are performed to prove the benefits of the metric-based classifier tolerant to mislabeled support samples and the loss function suppressing mislabeled query samples. Concretely, three similar diagnosis methods are constructed according to the proposed method, and then the diagnosis performances of these methods are compared in train transmission systems. In the experiments, four networks are constructed, which extracted features by the proposed frequency band focus module while adapting different classification and training strategies. Specifically, the proposed method adopts the metric-based classifier tolerant to mislabeled support samples and the loss function suppressing mislabeled query samples. Method A employs Softmax classifier and the loss function suppressing mislabeled query samples. Method B utilizes the traditional metric-based classifier and the loss function suppressing mislabeled query samples. Method C adopts the metric-based classifier tolerant to mislabeled support samples and cross-entropy loss function. In training datasets, each fault categories of key components in the train transmission systems only contains 10 samples, with 2 mislabeled samples, which are constructed by mixing samples from all 9 working conditions, and randomly divided into support and query sets. Table IV illustrates the diagnosis performance of these method of motors, gearboxes, and axle box. It can be observed that Method A has the worst performance because it does not take into account the limitation of insufficient samples, which leads to severe overfitting of the models. Methods B, C, and the proposed method all employ corresponding strategies to address limited samples. However, Method B ignores the mislabeled samples in support sets, causing the prototype to shift from its true position. Method C overlooks the mislabeled samples in query sets, contributing to the erroneous information in the optimization process of the models. In contrast, in the proposed method, the metric-based classifier tolerant to mislabeled support samples does not rely on complex model structures and is able to handle the case of insufficient training samples while alleviating the interference of mislabeled support samples. The loss function suppressing mislabeled query samples can restrain misleading due to mislabeled query samples. Therefore, the proposed method has the best diagnosis performance on each component. Further, confusion matrices of the proposed method are plotted to intuitively show the excellent diagnostic accuracy of each class, as shown in Fig. 8 To sum up, the experimental results show that the proposed classifier and training strategy can realize effective learning even from limited samples with noisy labels.

Table IV. Results of ablation experiments for metric-based classifier and loss function suppressing mislabeled query samples

	Diagnosis Object		
Method	**Traction Motor**	**Gearbox**	**Axle Box**
Method A	66.1% ±1.9%	64.5% ±2.4%	72.7% ±1.6%
Method B	82.1% ±1.2%	74.8% ±1.4%	81.9% ±1.2%
Method C	80.7% ±1.1%	75.2% ±1.2%	81.2% ±1.2%
Proposed Method	**85.6%** **±0.9%**	**80.5%** **±1.1%**	**89.4%** **±0.8%**

D. COMPARISON WITH STATE-OF-THE-ART METHODS

To demonstrate the superiority of the proposed interpretable few-shot diagnosis framework for train transmission systems in view of noisy labels, four existing methods are employed to perform the diagnosis tasks of train transmission systems with different proportions of noisy labels. Method 1 fuses an interpretable SincNet [9] diagnosis network with the symmetric cross entropy loss [19] based on the prototypical network. Method 2 replaces the learning with noisy labels strategy of Method 1 with co-teaching mechanism [20]. Method 3 combines the proposed frequency band focus module with the symmetric cross entropy loss based on the prototypical network. Method 4 replaces the learning with noisy labels strategy of Method 3 with co-teaching mechanism. In training datasets, each fault category of key components in the train transmission systems only contains 10 samples with different proportions of noisy labels, so as to simulate the scenarios where training samples are scarcely and mixed with noisy labels. Fig. 9 demonstrates the comparative results of diagnosis accuracy for train transmission system with different proportions of mislabeled samples. It can be seen that the diagnosis accuracy of Method 1 is lower than Method 3 and Method 2 is lower than Method 4. This is because Method 1 and Method 2 adopts diagnosis network based on SincNet, which ignore the discrepancies in the importance of different features. Method 3 and Method 4 employ the proposed frequency band focus module to highlight the features corresponding to the potential fault characteristic frequency. However, Method 3 fails to fit training sets that have a high proportion of noisy labels. Method 4 requires comprehensive training samples to support the model in learning data distribution. Therefore, these existing

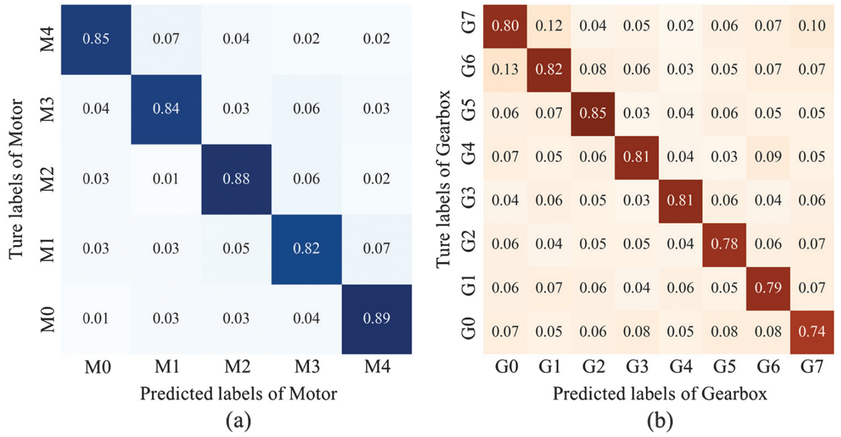

Fig. 8. Confusion matrix of the proposed method. (a) Motor diagnosis. (b) Gearbox diagnosis. (c) Axle box diagnosis.

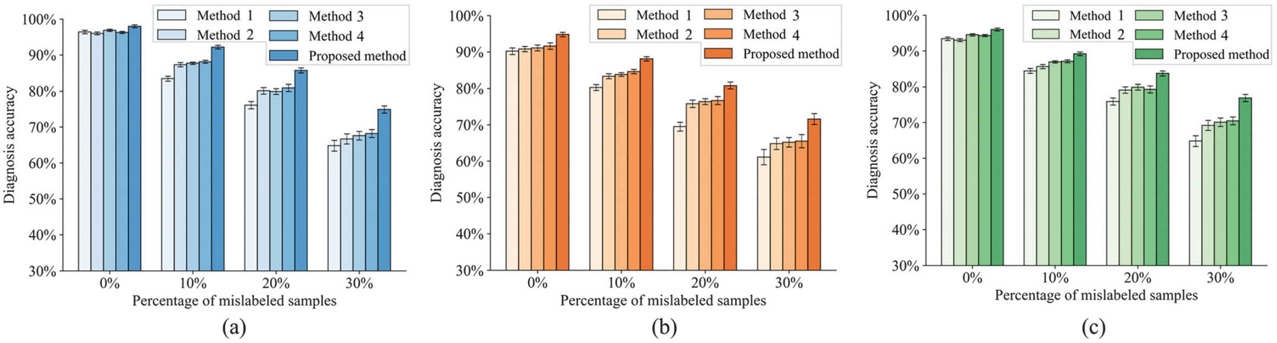

Fig. 9. Comparison results of the proposed method and the other existing methods. (a) Motor diagnosis. (b) Gearbox diagnosis. (c) Axle box diagnosis.

methods cannot effectively reduce the interference noisy labels with limited samples. In contrast, the method proposed systematically combines the frequency band focus module and learning with noisy labels strategy from the aspect of the support set and query set. In this framework, the signal features corresponding to the potential fault characteristic frequency are captured and highlighted, and the noisy labels are suppressed by specific mechanisms, thereby obtaining outstanding diagnosis performance. In summary, the superiority of the proposed method can be demonstrated by comparison with these state-of-the-art methods.

V. CONCLUSION

This article proposes a few-shot framework for fault diagnosis of train transmission systems with noisy labels, which are interpretable by fusing signal processing techniques into the neural network architecture. In the proposed framework, a feature extractor is constructed by stacked frequency band focus modules, which can capture signal features in different frequency bands and further adaptively concentrate on the features corresponding to the potential fault characteristic frequency. Then, according to prototypical networks, a novel metric-based classifier is developed that is tolerant to mislabeled support samples. Besides, a new loss function is designed to decrease the impact of label mistakes in query datasets. Finally, fault simulation experiments of subway train transmission systems are designed and conducted, and the effectiveness as well as superiority of the proposed method are proved by ablation experiments and comparison with the existing methods. In future research, the proposed framework will be continued to improve in two aspects, including physics-informed neural network and automatic correction of wrong labels, so as to make the work more engineering value.

ACKNOWLEDGEMENTS

This research was supported in part by the National Key R&D Program of China under Grant 2022YFB4300601 and in part by the State Key Laboratory of Advanced Rail Autonomous Operation under Grant RAO2023ZZ003.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

[1] T. Zhou, D. Yao, J. Yang, C. Meng, A. Li, and X. Li, "DRSwin-ST: an intelligent fault diagnosis framework based on dynamic threshold noise reduction and sparse transformer with shifted windows," *Reliab. Eng. Syst. Saf.*, vol. 250, p. 110327, 2024.

[2] Z. Jin, D. He, and Z. Wei, “Intelligent fault diagnosis of train axle box bearing based on parameter optimization VMD and improved DBN,” *Eng. Appl. Artif. Intell.*, vol. 110, p. 104713, 2022.

[3] Y. He and W. Shen, “MSRCN: a cross-machine diagnosis method for the CNC spindle motors with compound faults,” *Expert Syst. Appl.*, vol. 233, p. 120957, 2023.

[4] Z. Baihong, Z. Minghang, Z. Shisheng, L. Lin, and W. Lin, “Mechanical compound fault diagnosis via suppressing intra-class dispersions: a deep progressive shrinkage perspective,” *Measurement*, vol. 199, p. 111433, 2022.

[5] J. Zhang, B. Xu, Z. Wang, and J. Zhang, “An FSK-MBCNN based method for compound fault diagnosis in wind turbine gearboxes,” *Measurement*, vol. 172, p. 108933, 2021.

[6] A. Ding, Y. Qin, B. Wang, X. Cheng, and L. Jia, “An elastic expandable fault diagnosis method of three-phase motors using continual learning for class-added sample accumulations,” *IEEE Trans. Ind. Electron.*, vol. 71, no. 7, pp. 7896–7905, 2024.

[7] S. Li, T. Li, C. Sun, X. Chen, and R. Yan, “WPConvNet: an interpretable wavelet packet kernel-constrained convolutional network for noise-robust fault diagnosis,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 14974–14988, 2024.

[8] Y. Qin, H. Liu, Y. Wang, and Y. Mao, “Inverse physics–informed neural networks for digital twin–based bearing fault diagnosis under imbalanced samples,” *Knowl.-Based Syst.*, vol. 292, p. 111641, 2024.

[9] F. B. Abid, M. Sallem, and A. Braham, “Robust interpretable deep learning for intelligent fault diagnosis of induction motors,” *IEEE Trans. Instrum. Meas.*, vol. 69, no. 6, pp. 3506–3515, 2020.

[10] F. Dunkin, X. Li, H. Li, G. Wu, C. Hu, and S. S. Ge, “MgCNL: a sample separation approach via multi-granularity balls for fault diagnosis with the interference of noisy labels,” *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 7748–7761, 2024.

[11] N. Huang, Q. Chen, G. Cai, D. Xu, L. Zhang, and W. Zhao, “Fault diagnosis of bearing in wind turbine gearbox under actual operating conditions driven by limited data with noise labels,” *IEEE Trans. Instrum. Meas.*, vol. 70, p. 3502510, 2021.

[12] H. Wang and Y.-F. Li, “Iterative error self-correction for robust fault diagnosis of mechanical equipment with noisy label,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[13] C. Cheng, X. Liu, B. Zhou, and Y. Yuan, “Intelligent fault diagnosis with noisy labels via semisupervised learning on industrial time series,” *IEEE Trans. Ind. Inf.*, vol. 19, no. 6, pp. 1–10, 2023.

[14] S. He, W. K. Ao, and Y.-Q. Ni, “A unified label noise-tolerant framework of deep learning-based fault diagnosis via a bounded neural network,” *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–15, 2024.

[15] M. Ravanelli and Y Bengio, “Interpretable Convolutional Filters with SincNet,” *arXiv preprint arXiv:1811.09725v2*, 2019.

[16] N. Lu, H. Hu, T. Yin, Y. Lei, and S. Wang, “Transfer relation network for fault diagnosis of rotating machinery with small data,” *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 11927–11941, 2022.

[17] A. Zhang, S. Li, Y. Cui, W. Yang, R. Dong, and J. Hu, “Limited data rolling bearing fault diagnosis with few-shot learning,” *IEEE Access*, vol. 7, pp. 110895–110904, 2019.

[18] C. Wang, J. Yang, and B. Zhang, “A fault diagnosis method using improved prototypical network and weighting similarity-Manhattan distance with insufficient noisy data,” *Measurement*, vol. 226, p. 114171, 2024.

[19] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, “Symmetric cross entropy for robust learning with noisy labels,” *arXiv preprint arXiv:1908.06112v1*, 2019.

[20] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. W. Tsang, and M. Sugiyama, “Co-teaching: robust training of deep neural networks with extremely noisy labels,” *arXiv preprint arXiv:1804.06872v3*, 2018.