Few-Shot Bearing Fault Diagnosis Via Ensembling Transformer-Based Model With Mahalanobis Distance Metric Learning From Multiscale Features

Manh-Hung Vu, Van-Quang Nguyen, Thi-Thao Tran, Van-Truong Pham, and Men-Tzung Lo

***Abstract*— Advanced deep-learning models have shown excellent performance in the task of fault-bearing diagnosis over traditional machine learning and signal-processing techniques. Few-shot learning approach has also been attracting a lot of attention in this task to address the problem of limited training data. Nevertheless, cutting-edge models for fault-bearing diagnosis are often based on convolutional neural networks (CNNs) that emphasize local features of input data. Besides, accurate classification of fault-bearing signals is still nontrivial due to the variations of data, fault types, acquisition conditions, and extremely limited data, leaving space for research on this topic. In this study, we propose a novel end-to-end approach for fault-bearing diagnosis even in the case of limited data with artificial and real faults. In particular, we propose a module for automatic feature extraction from input data namely multiscale large kernel feature extraction. The extracted features are then fed into a two-branch model including a global and a local branch. The global one includes a transformer architecture with cross-attention to handle global context and obtain the correlation between the query and support sets. The local branch is a metric-based model consisting of Mahalanobis distance for separating local features from the support set. The outputs from the two branches are then ensembled for classification purposes. Intensive experiments and ablation studies have been made on the two public datasets including CWRU and PU. Qualitative and quantitative results with different degrees of training samples by the proposed model in comparison with other state-of-the-arts have shown the superior performance of the proposed approach. Our code will be published at https://github.com/HungVu307/Few-shot-via-ensembling-Transformer-with-Mahalanobis-distance**

***Index Terms*— Ensemble classification, fault bearing diagnosis, few-shot learning, Mahalanobis metric learning, multiscale large kernel feature extraction, transformer.**

Manuscript received 13 December 2023; revised 1 February 2024; accepted 20 February 2024. Date of publication 25 March 2024; date of current version 2 April 2024. This work was supported by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant 102.05-2021.34. The Associate Editor coordinating the review process was Dr. Weihua Li. *(Corresponding author: Thi-Thao Tran.)*

Manh-Hung Vu and Van-Quang Nguyen were with the Department of Automation Engineering, School of Electrical and Electronic, Hanoi University of Science and Technology, Hanoi 100000, Vietnam. They are now with the Vingroup BigData Institute, Hanoi 100000, Vietnam.

Thi-Thao Tran and Van-Truong Pham are with the Department of Automation Engineering, School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi 100000, Vietnam (e-mail: thao.tranthi@hust.edu.vn).

Men-Tzung Lo is with the Department of Biomedical Sciences and Engineering, National Central University, Taoyuan 320317, Taiwan.

Digital Object Identifier 10.1109/TIM.2024.3381270

I. Introduction

IN MANY different industrial fields, such as manufacturing, aerospace, and power electronics, electric motors are crucial components of machinery [1], [2], [3], [4]. Since technology has advanced, machines have become more and more necessary in people's lives, enabling them to work more efficiently and with less effort than in the past [5], [6]. As electric motors are frequently the primary parts of machines that produce and sustain electricity, identifying faults in electrical machines is crucial [7], [8]. Approximately 40% of faults in electrical machines are attributed to bearings, according to figures from the IEEE Industrial Society and the Japanese Electrical Machinery Association (JEMA) [9], [10]. Bearings are parts located deep inside an electric machine, an important structural component to operate, consisting of the inner race, outer race, and ball as its three parts. So, with the frequency of continuous use of an electric machine, the bearings are the most sensitive parts [11]. It takes time and requires expert people to disassemble manually using human power for defect detection. Furthermore, this kind of disassembly may have an impact on other electrical machine components, resulting in issues or unforeseen faults [12], [13]. Therefore, artificial intelligence models are developed and applied for diagnosis by measuring the return signal from an electrical machine and identifying that signal, for example, using vibration signals is a common way currently [14], [15].

Before deep-learning models developed strongly, machine-learning models were also widely used to diagnose bearing faults. Zhang et al. [16] proposed a diagnostic method based on principal components analysis (PCA) and support vector machine model (SVM), based on the characteristics extracted from the return signal to make a fault diagnosis. Ye et al. [17] implemented a combination of particle swarm optimization-based support vector machines (PSO-SVMs) to enhance performance, multiscale permutation entropy (MPE), and variational mode decomposition (VMD) techniques. In the process of reconstructing the bearing vibration signal, a novel criterion called the feature energy ratio (FER) is introduced. This criterion is applied after dividing the original bearing vibration signal into multiple intrinsic mode functions (IMFs) using the VMD method. Zhou et al. [18], Yan and Jia [19] also provide similar improvements based on traditional machine-learning models. In addition, other machine-learning methods

such as K-nearest neighbor (KNN) [20], and artificial neural network (ANN) [21] were also used to diagnose faults in signals. However, traditional methods have many limitations: if many hyperparameters need to be adjusted to achieve high performance, to have a good feature set, they need to be extracted manually with the knowledge of experts. This causes a waste of time and expense.

After the strong development of deep-learning models, especially convolutional neural networks (CNNs) [22], the disadvantages of traditional machine-learning models in fault diagnosis have been significantly overcome. Taking advantage of the efficiency of information extraction by convolution, many studies have applied CNN networks and variations and have shown effectiveness, some other studies have combined CNNs and long-short term memory (LSTM) to further take advantage of time-domain continuity [23], [24], [25]. Pan et al. [23] proposed a model that uses 1-D-convolutional layers to extract samples and provide features in the time domain and then takes advantage of LSTM to process information in series before making a diagnosis. Eren et al. [24] and Chen et al. [25] have the same idea of improving convolutional layers by using different kernel sizes to obtain more comprehensive features in the time domain. Besides, improvements in convolution-based models show convenience as well as high efficiency and achieve state-of-the-art on public datasets. Wang et al. [26] proposed adding a squeeze excitation (SE) block to the traditional CNN model to increase the model's generalization ability. Li et al. [27] added an attention mechanism and relies on convolutional layers to capture more information and take advantage of local coherence. Although some of the time and cost drawbacks over traditional models have been overcome by the application of deep learning in bearing defect diagnosis, data remains a major challenge. Large amounts of data and a variety of information are required to create an effective deep-learning model, which is particularly challenging in practice since it is not always possible to gather damaged bearings. If artificial damage is used, it can be more expensive to destroy complete bearings, which may not ensure practicality. Not only that, traditional deep-learning models based on convolution often emphasize local features of input data. The reason is that convolution layers operate according to the sliding window scanning mechanism with a kernel size through the entire sample, the information passing through each layer is local in the sliding window area without contact with other layer sites. Due to the aforementioned drawbacks, few-shot learning, a type of meta-learning, is utilized in the field to address diagnostic issues when there is a shortage of data.

The main distinction between few-shot learning and conventional deep-learning techniques is that few-shot learning demonstrates the model to learn the correlation between samples rather than just the characteristics of each sample, which enables models to learn effectively under low data conditions. In fault diagnosis from signals, a popular few-shot baseline used is the Siamese neural network (SNN). Zhang et al. [28] improved the SNN network by changing CNNs into a deep-learning CNN network with wide first-layer kernels (WDCNN) to better separate features. Once the two input signals have been compared to determine whether they are similar or different, stage 2 proceeds to the diagnosis step. From this baseline, Vu et al. [29] enhanced the Conv-mixer's features to better extract information from the spectrogram image that was created by transforming the original raw signal. In addition, Shen et al.'s [30] research on applying prototype network or Li et al. [31] using Matching Net has also been implemented. It is shown that studies on few-shots in signal recognition, although showing better performance than traditional models, still require many stages to make a diagnosis. Besides, Euclidean distance and cosine distance are the two most popular measures to compare correlation in the above models. In conditions of little data, the above measurements will be difficult to perform well as they are easily affected by noise due to the simplicity of calculation. Following the creation of few-shot models, Li et al. [32] proposed a model named CovaMNet. This model makes use of the covariance metric as a scale, with the covariance matrix serving as a representation of the correlation between all samples in the support set and the query set and there are many studies in many different fields showing the effectiveness of this metric [33], [34], [35], [36]. Another direction is to use cross attention in Transformer—a popular and effective model in the field of natural language processing (NLP)—as a correlation scale also developed by several research groups in few-shot learning [37], [38], [39], [40]. Nevertheless, few-shot models in recognition suffer from some limitations such as follows.

1) Methods that achieve good results are often complex, require more than two stages to diagnose and are expensive to train the model and deploy.
2) Few-shot models mainly rely on a backbone based on CNNs or Transformers pretrained from ImageNet to obtain feature maps, which is not effective for specific datasets such as bearing fault data.
3) Conditions with limited information may not adapt effectively to metrics like Euclidean and Cosine. More advanced metrics, like the covariance metric, are very useful, but they may not be able to detect nonsingular covariance matrices when there are few samples per class available.

To address the weaknesses of traditional deep-learning models, along with addressing the stated disadvantages of current few-shot learning models in the field of fault-bearing diagnosis, an end-to-end diagnostic method is proposed in the current study. First, we use the proposed module, multiscale large kernel feature extraction to extract features from samples in both the query set and the support set. With the model being trained from scratch, for the metrics behind to be learned well, the feature extraction module plays an important role. Next, the model has two branches as follows to capitalize from the metric learning benefits of both covariance metric and cross-attention, as previously mentioned: We use *Transformer* architecture with cross-attention to process global information and obtain correlation between the query and support sets. This solves the local input problem of overreliance on convolution as raised by some traditional CNN-based models. To guarantee that the covariance matrix is a nonsingular matrix, we first proposed and implemented the Mahalanobis module in the field of signal diagnosis using the technique of separating local features from the support set and combining them with

the Mahalanobis metric. To our knowledge, in conditions of limited data, all information is valuable, so the proposed model searches for global and local information from feature maps to help diagnose more accurately and our arguments are also proven through experiments in this study. Our proposed method not only solves the weaknesses of other few-shot learning methods, but also solves the disadvantage of traditional deep-learning models, which is about training data. Our proposed model provides high performance with little training data while ensuring that it is an end-to-end model.

To summarize, our fundamental contributions are as follows.

1) Propose a novel few-shot learning model via an Ensembling Transformer-based model with Mahalanobis distance metric. The proposed model combines both global and local information, solving the problem of deep-learning models of signal diagnosis and recognition in limited data conditions. Besides, the proposed model is an end-to-end model, overcoming the disadvantages of previous multistage models.
2) To ensure that the covariance matrix in Mahalanobis distance is a nonsingular matrix, we propose a method to calculate the covariance matrix of each class in the support set based on local features. Metrics show efficiency compared to traditional metrics.
3) To extract features from the query set and the support set, we also propose the multiscale large kernel feature extraction (MLKFE) module. Because the model is trained from scratch, having a good feature map is very important, especially in limited data conditions. Our module ensures diverse and informative features that help improve the model's diagnostic capabilities.

The study presents the related research in Section II, methodological approaches are presented in Section III, and the experimental findings in Section IV.

II. Related Work

A. Large Kernel Feature Extraction

Guo et al. [41] introduced a method to retrieve features using a larger window that consumes fewer resources, which is large kernel attention (LKA). The structure of LKA is built based on depthwise convolution, dilated depthwise convolution, and pointwise convolution. To obtain spatial information, depthwise convolution is used with a filter in which the weights between channels are equal. Along with that, dilated depthwise convolution helps to obtain long-range information. Pointwise convolution is a 1×1 filter to mix information between channels together. Combining depthwise convolution, dilated depthwise convolution, and pointwise convolution instead of using a convolution with a large window size significantly reduces the number of training parameters. Besides, information can be synthesized from local context and global context through depthwise convolution and dilated depthwise convolution. This module has shown effectiveness on many different tasks such as segmentation or detection [42], [43]. Applied to few-shot, it is found that this is a positive direction when few-shot learning needs a good and informative set of features before comparing them using metrics, so an improvement based on LKA is proposed by us in this study to improve the effectiveness of feature extraction module in diagnosis few-shot model.

B. Transformer

The development of the Transformer model is a significant advancement in NLP. The self-attention mechanism in the Transformer is the main core that helps form the effectiveness of this model. The main structure of the Transformer includes multihead self-attention (MHSA), feed-forward (FFD), layer-norm (LN), linear (LN), and Softmax layer, and they constitute the encoder and the decoder of the model. Not only successful in the field of NLP, but Transformer models also demonstrate effectiveness in a number of fields such as image classification, biomedicine, object detection, and so on [44], [45], [46]. With the efficiency of the Transformer, applying this model to few-shot learning is a promising direction and many researchers have developed this problem. For instance, Han et al. [47] proposed a few-shot Cross-Transformer model that brings good performance in object recognition, and Lu et al. [48] developed a few-shot model applied in the field of semantic segmentation. The Transformer's architecture allows information to be processed back and forth continuously between the encoder and the decoder through cross-attention, so the information is global. Realizing the effectiveness of cross-attention and its ability to process correlation information as a metric, we apply the Transformer model to this study to diagnose signals.

C. Mahalanobis Distance

Mahalanobis distance is a statistical measure used to quantify the similarity or dissimilarity between two points in a multidimensional space [49]. It is especially helpful in conditions when the variables are interrelated and have various scales since it considers the correlations between the variables. The squared Mahalanobis distance between vector $\mathbf{x} \in \mathbb{R}^n$ and the distribution of $Y = [y_1, y_2, \ldots, y_k] \in \mathbb{R}^{k \times n}$ can be defined by the formula

$$d(\mathbf{x}, Q) = \mathbf{x}^T \Sigma_y^{-1} \mathbf{x} \tag{1}$$

where Q defines the distribution of Y, and Σ_y is the covariance matrix calculated from samples $y_1, y_2, \ldots, y_k$. In few-shot learning, the commonly used measures are Euclidean distance and Cosine distance. Although these measures are easy to calculate and set up, they are easily affected by noise, especially in conditions with little data, the outliners can greatly affect the calculation results [50]. In addition, the Euclidean measure calculates the distance in the space between two points, so the scale of the data may affect the results [51]. Mahalanobis distance can overcome that limitation as it measures the distance from a point to a distribution, taking into account many characteristics of the sample in that distribution and Sigma is a covariance matrix that ensures equality between weights.

III. Methodology

When only a few samples are available, few-shot classification attempts to categorize the unseen samples. Especially in diagnosing bearing faults in electric motors, collecting a large

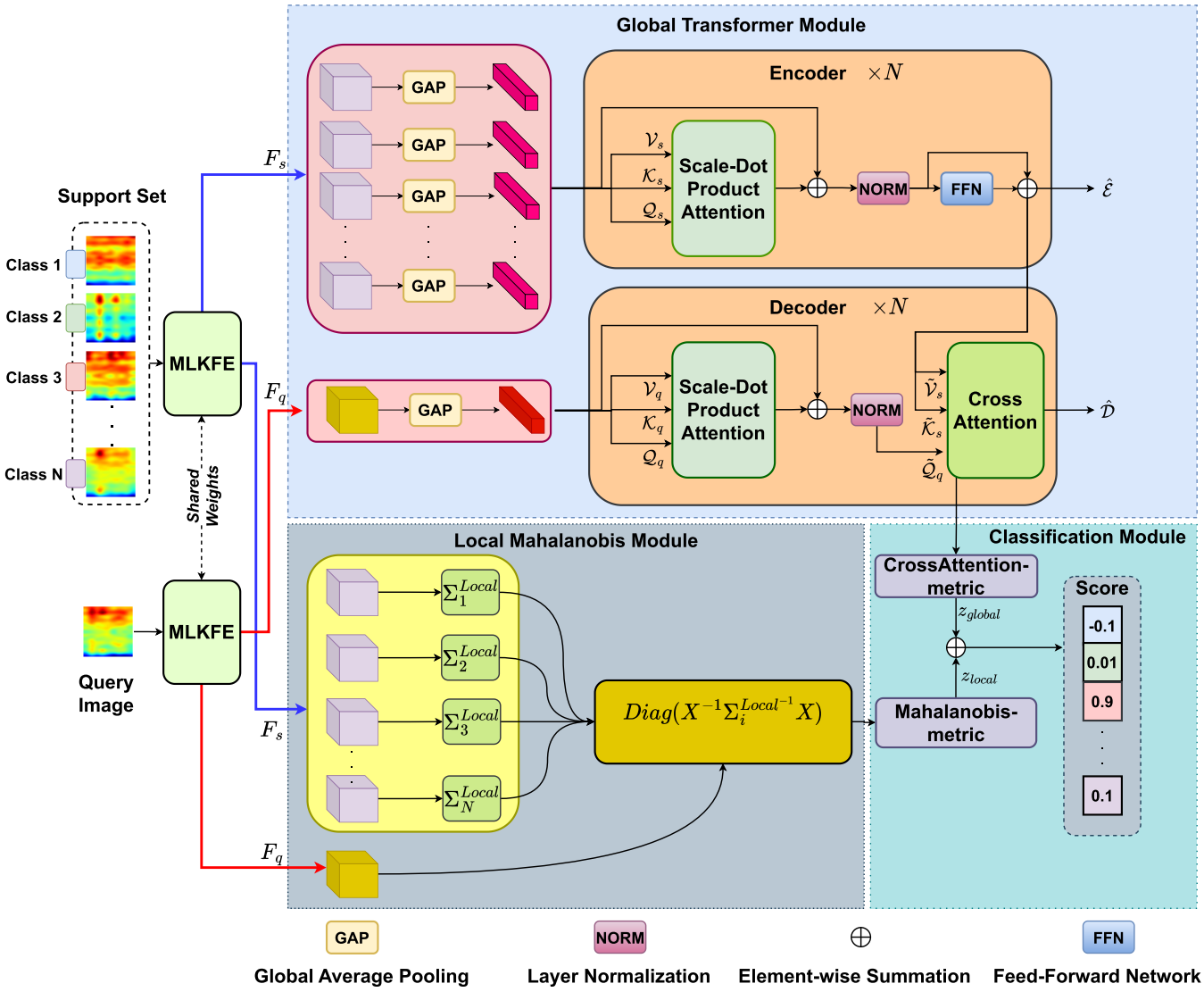


Fig. 1. Overall architecture of our proposed model. The input includes a support set and a query set to be converted to a spectrogram. The proposed module MLKFE is used to extract features before the feature maps are processed through the Global Transformer and Local Mahalanobis modules and finally diagnosed through the classification module.

enough dataset is quite difficult. Following recent approaches [39], [52], [53], [54], we utilize the episode training mechanism, which has been shown successfully in the few-shot learning strategy. In C-way K-shot task, a random sample of C classes and support set $\mathcal{S} = \{(X_i^s, Y_i^s)\}_{i=1}^{n_s}$ ($n_s = C \times K$) of K samples with labels for each class are used to construct each episode. In the meantime, q samples at random are selected from each class to be used for the query, namely query set $\mathcal{Q} = \{(X_j^q, Y_j^q)\}_{j=1}^{n_q}$ with $n_q = q \times K$. The objective is to determine the correlation between samples in scale-like $\mathcal{S}$ and $\mathcal{Q}$. The general architecture of the proposed approach for bearing faults diagnosis is shown in Fig. 1.

In Fig. 1, the dataset's raw signal is first converted to a spectrogram using a short-time Fourier transform (STFT). This would assist in segregating features more effectively across the convolutional layers of the deep-learning network. Instead of using signal preprocessing techniques, the time–frequency graphs are utilized directly because some preprocessing noise filtering methods such as high-pass filters or low-pass filters need to choose the appropriate cut-off frequency. The following may result in a rise in processing demand and data loss. This has been proven through Vu et al.'s [29] experiments and a number of other studies [55], [56], [57]. Then, the support set $\mathcal{S}$ and the query set $\mathcal{Q}$ are trained according to the proposed architecture presented. First, MLKFE was proposed to extract features from both sets $\mathcal{S}$ and $\mathcal{Q}$. Feature extraction is important in the condition that there is only a small amount of training data, so the collected information needs to be diverse and characteristic. Next, to measure the correlation between $\mathcal{S}$ and $\mathcal{Q}$, we, respectively, introduce the module **Global Transformer** inspired by [37] and propose for the first time the **Local Mahalanobis** module. Finally, information is aggregated from both global and local modules to provide bearing fault diagnosis through the **Classification** module. Details of the proposed model parts are presented below.

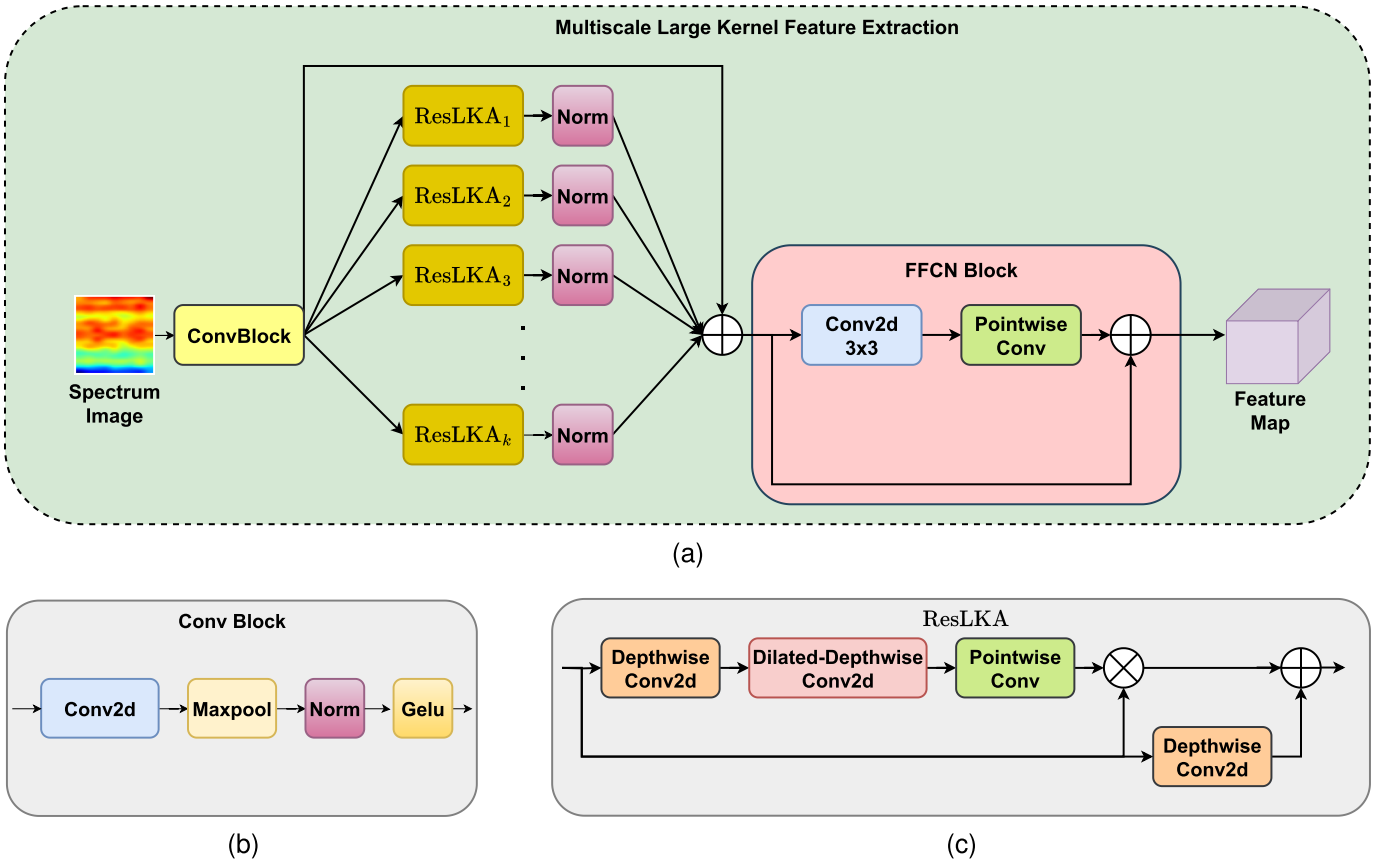


Fig. 2. (a) Proposed architecture of MLKFE. (b) Conv Block. (c) Proposed ResLKA module.

A. Multiscale Large Kernel Feature Extraction

The idea of LKA [41] is to decompose the Large Kernel Convolution $[K \times K]$ to d-dilated depthwise convolution $[(K/d) \times (k/d)]$, depthwise convolution $[(2d-1) \times (2d-1)]$, and pointwise convolution (where K is kernel size and d is dilation rate). Through the above decomposition, LKA can capture long-range relationships with slight computational cost and parameters. While fixed-size LKA has demonstrated its effectiveness in capturing long-range dependencies, it encounters challenges when applied to the problem of the bearing fault. In bearing fault diagnosis, datasets are inherently noisy, diverse, and complex. The fixed configuration of LKA struggles to adapt to varying scales of features, hindering its ability to simultaneously capture fine-grained details and broader contextual information essential for accurate fault detection. Motivated by this limitation, we proposed a multiscale large kernel feature extractor module capable of extracting multiscale features and capturing a salient detail of the spectrogram image. The architecture of the MLKFE module is described in Fig. 2.

The residual large kernel attention (ResLKA) module approximates large kernel convolution with depthwise (DW) convolution, dilated-depthwise (DDW) convolution, and Pointwise convolution (PW) to learn both long-range pixel relationships. Different from the approach proposed in [41], Depthwise convolution with a small kernel size as the skip connection is utilized to enable the model to capture more information about the short-range pixel relationships. Additionally, the skip connection in ResLKA prevents vanishing gradients caused by element-wise products. The formulation of ResLKA is as follows:

$$\text{ResLKA}(X) = \text{DW}(X) + \text{PW}(\text{DDW}(\text{DW}(X))) \otimes X. \quad (2)$$

The multiscale features can be achieved when fusing the features generated by multiple ResLKA modules with different kernel sizes and dilations. Then, the FFD convolutional network (FFCN) [46] is used to create a more nonlinear combination of the features. Note that, we do not feed directly the original spectrogram image to the Multiscale ResLKA module to extract features. Instead, several simple convolution layers (comprised of Conv2d, Maxpooling, LayerNorm, and GELU activation function) are utilized to extract higher-level details of the image. This can enhance the overall performance of the model. The probable reason is that the convolution layers can extract information from various frequency representations of a spectrogram image. This can also be referred to as overlapping patch embedding.

B. Global Transformer Module

Inspired by the very effective *Transformer* model [58] in the field of NLP, along with Wang et al.'s [37] proposal on ImageNet, we, for the first time, apply the Global Transformer for bearing fault diagnosis, which is presented in Fig. 1. Thanks to Wang et al. [37] for an outstanding idea. Similar to the Transformer architecture, the Global Transformer also includes two main components: the encoder and the decoder, with the goal of decoding the characteristic features of $\mathcal{S}$ and providing

a global correlation with $\mathcal{Q}$. The corresponding features after processing through the proposed module **MLKFE** are $\mathcal{F}_s \in \mathbb{R}^{n_s \times h \times w \times c}$ and $\mathcal{F}_q \in \mathbb{R}^{n_q \times h \times w \times c}$ respectively was tokenized via global average pooling to obtain the corresponding token sequence $\hat{\mathcal{F}}_s = \{\hat{F}_s^1, \hat{F}_s^2, \ldots, \hat{F}_s^{n_s}\} \in \mathbb{R}^{n_s \times c}$ for the support set and $\hat{\mathcal{F}}_q = \{\hat{F}_q^1, \hat{F}_q^2, \ldots, \hat{F}_q^{n_q}\} \in \mathbb{R}^{n_q \times c}$ for the query set. The encoder and decoder blocks shown below may process information globally [59] due to the use of global average pooling, which helps to gain global information about the feature.

1) Encoder: Finding characteristic correlations between samples in each class of the support set is the encoder block's objective. First, through linear projection from $\hat{\mathcal{F}}_s$, we obtain $(\mathcal{Q}_s, \mathcal{K}_s, \mathcal{V}_s) \in \mathbb{R}^{n_s \times c}$. Second, a correlation matrix of several support samples is created using a dot-product technique as

$$\text{correlation}_{s \to s}(\mathcal{Q}_s, \mathcal{K}_s) = \text{Softmax}\left(\frac{\mathcal{Q}_s(\mathcal{K}_s)^T}{\sqrt{c}}\right). \tag{3}$$

This attention mechanism helps synthesize information in a focused, selective manner in support samples and the output information of the attention mechanism is shown as follows:

$$\mathcal{E} = \text{LayerNorm}(\hat{\mathcal{F}}_s + \text{correlation}_{s \to s}(\mathcal{Q}_s, \mathcal{K}_s)\mathcal{V}_s). \tag{4}$$

The output of the encoder part can be collected by

$$\hat{\mathcal{E}} = \text{FFN}(\mathcal{E}) + \mathcal{E}. \tag{5}$$

After obtaining the output $\hat{\mathcal{E}}$ of the encoder, this is the input of the next encoder block based on the idea of *Transformer* and is also used to calculate the correlation between support samples and query samples through the decoder part.

2) Decoder: Similar to the encoder part, the input $\hat{\mathcal{F}}_q$ is linearly projected into $(\mathcal{Q}_q, \mathcal{K}_q, \mathcal{V}_q) \in \mathbb{R}^{n_q \times c}$ to collect correlation in each query sample through scale dot attention as

$$\text{correlation}_{q \to q}(\mathcal{Q}_q, \mathcal{K}_q) = \text{Softmax}\left(\frac{\mathcal{Q}_q(\mathcal{K}_q)^T}{\sqrt{c}}\right) \tag{6}$$

$$\mathcal{D} = \text{LayerNorm}(\hat{\mathcal{F}}_q + \text{correlation}_{q \to q}(\mathcal{Q}_q, \mathcal{K}_q)\mathcal{V}_q) \tag{7}$$

where $\mathcal{D} \in \mathbb{R}^{n_q \times c}$ represents the correlation in each query sample. To find global similarity between support samples and query samples, *cross-attention* structure is introduced with the following mechanism: A linear projection of $\mathcal{D}$, which serves as a query in the structure, is used to first obtain $\tilde{\mathcal{Q}}_q$. Following a similar process, linear projection is applied to the encoder's output $\hat{\mathcal{E}}$ to create $\tilde{\mathcal{K}}_s$ and $\tilde{\mathcal{V}}_s$ as the following formula:

$$\text{correlation}_{q \to s}\left(\tilde{\mathcal{Q}}_q, \tilde{\mathcal{K}}_s\right) = \text{Softmax}\left(\tilde{\mathcal{Q}}_q\left(\tilde{\mathcal{K}}_s\right)^T\right) \tag{8}$$

$$\hat{\mathcal{D}} = \text{correlation}_{q \to s}\left(\tilde{\mathcal{Q}}_q, \tilde{\mathcal{K}}_s\right)\tilde{\mathcal{V}}_s. \tag{9}$$

Here, $\hat{\mathcal{D}}$ is the output of the decoder. Similar to the *Transformer* architecture for NLP, output $\hat{\mathcal{D}}$ is also the input of the next decoder block (same as the output of encoder $\hat{\mathcal{E}}$) with N is the depth of the network. Through many encoder and decoder layers, information is extracted back and forth between both the encoder and the decoder in the last decoder layer, leading to a correlation matrix that is considered the global correlation matrix between the query set and the support set. And the special thing here is that after obtaining $\text{correlation}_{q \to s}$ calculated from the Nth block, this is considered a *global metric* between the support set $\mathcal{S}$ and the query set $\mathcal{Q}$, expressed as

$$z_{\text{global}} = \text{correlation}_{q \to s}\left(\tilde{\mathcal{Q}}_q, \tilde{\mathcal{K}}_s\right) \tag{10}$$

where z_{global} carries global information between $\mathcal{Q}$ and each sample in $\mathcal{S}$. This is very different from traditional metrics when information is now continuously synthesized back and forth between $\mathcal{S}$ and $\mathcal{Q}$ through N encoder–decoder blocks.

Different from the traditional Transformer architecture, in the Nth encoder–decoder (i.e., the last layer of the Global Transformer module), the output of the decoder is out of order and only uses the correlation matrix calculated based on the query (represented for the query set) and the key (represented for the support set) via the cross-attention module (8). In this way, the correlation between the query set and the support set is calculated and this is global information because it is processed through N encoder–decoder layers. This helps the proposed overall model have better performance, especially in low data conditions.

C. Local Mahalanobis Module

To make diagnosis even more effective in conditions of limited data, we propose for the first time a novel module named Local Mahalanobis module shown in Fig. 1 that is based on the covariance matrix. Covariance matrix has many applications in many fields such as risk prediction in portfolios [60], [61], [62], [63] or general representation [64], [65], [66]. Applying the covariance matrix to the problem of diagnosing bearing faults to provide the distribution of the support set is a potential direction. However, to train only on a small amount of data, the covariance matrix easily falls into the case of a singular matrix (C with a value of 1 or 5 as usual for each category, and it is much smaller than the dimensions of the feature). Therefore, to apply to the few-shot problem with limited data, an improvement has been proposed as follows.

Similar to the input of the sample Transformer, let $\mathcal{F}_s = \{\mathcal{F}_s^1, \mathcal{F}_s^2, \ldots, \mathcal{F}_s^{n_s}\} \in \mathbb{R}^{n_s \times h \times w \times c}$ be the feature extracted from the support set $\mathcal{S}$ after being processed through **MLKFE**. First, we consider the correlation between samples in each class of the support set with $\mathcal{S}_k = \{\mathcal{F}_{sk}^1, \mathcal{F}_{sk}^2, \ldots, \mathcal{F}_{sk}^C\} \in \mathbb{R}^{C \times h \times w \times c}$ represented for the class with label k in C-shot K-way task ($\mathcal{S}_k \subset \mathcal{F}_s$, $k = 1, 2, \ldots, K$). Next, pixels at the same location in each channel are combined and create a local feature vector as shown in Fig. 3, and after that, we obtain a set of local feature vectors that characterize class k is $\tilde{\mathcal{S}}_k = \{\mathcal{F}_{sk}^1, \mathcal{F}_{sk}^2, \ldots, \mathcal{F}_{sk}^d\} \in \mathbb{R}^{d \times c}$ with $d = C \times h \times w$ denote total number of local feature vectors of a class. Then, the local covariance matrix with the characteristics of the class k is defined as

$$\mathbf{\Sigma}_{\text{Local}}^k = \frac{1}{d-1} \sum_{\mathcal{F}'_{sk} \in \tilde{\mathcal{S}}_k} \left(\mathcal{F}'_{sk} - \mu_k\right) \times \left(\mathcal{F}'_{sk} - \mu_k\right)^T \tag{11}$$

where $\mu_k \in \mathbb{R}^c$ is defined for mean of d local feature vectors in $\tilde{\mathcal{S}}_k$, and $\mathbf{\Sigma}_{\text{Local}}^k \in \mathbb{R}^{c \times c}$ represents local covariance matrix of class k. By considering each local feature vector $\mathcal{F}_{sk}^i$ as an observation in class k, it is shown that the number of observations $|\tilde{\mathcal{S}}_k| = d = C \times h \times w$ is much larger than the

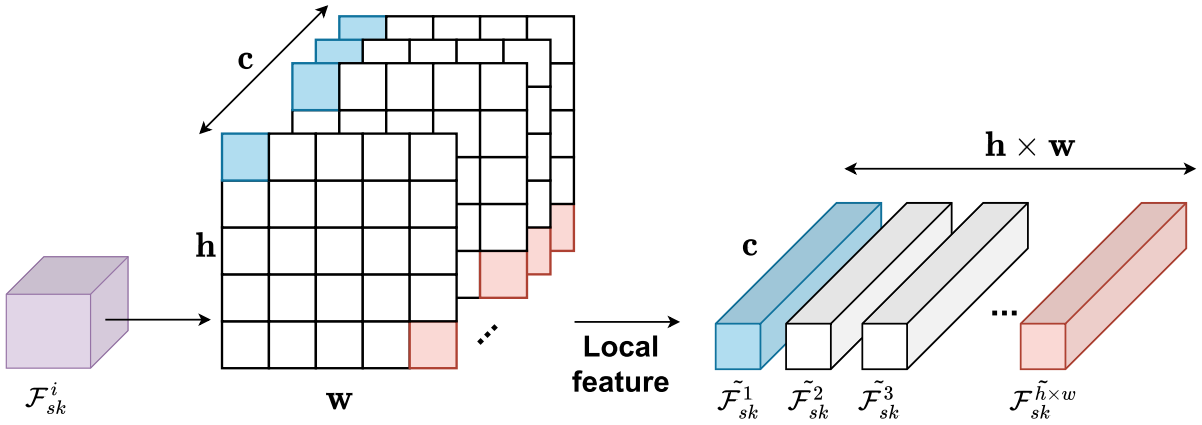


Fig. 3. Illustration of considering each pixel at the same position in the channels as a local vector feature to calculate the local feature covariance matrix.

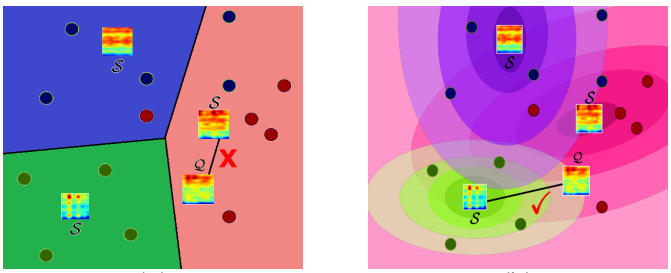


Fig. 4. Example of comparing two measures based on: (a) Euclidean distance and (b) Mahalanobis distribution.

dimensionality of the feature c, so this ensures that $\mathbf{\Sigma}^k_{\text{Local}}$ is the nonsingular matrix. To measure the correlation between $\mathcal{S}_k$ and each sample in $\mathcal{Q}$, the Mahalanobis distance defined by the Mahalanobis metric is applied, represented by the following mathematical formula:

$$f\left(\bar{\mathcal{F}}'_q, \mathbf{\Sigma}^k_{\text{Local}}\right) = \bar{\mathcal{F}}'^{T}_q \times \left(\mathbf{\Sigma}^k_{\text{Local}}\right)^{-1} \times \bar{\mathcal{F}}'_q. \tag{12}$$

Here, $\bar{\mathcal{F}}'_q \in \mathbb{R}^{c\times d}$ represents the local vector obtained from the features set $\mathcal{F}_q$ and with $f(\mathbf{x}, \Sigma)$, we can prove that if $\mathbf{x}$ is in the direction of the smallest eigenvectors of local covariance matrix Σ, the value of $f(\mathbf{x}, \Sigma)$ can reach a maximum based on Theorem 1. This indicates that the distributions of x and this category are consistent and that n is pointing in the direction of the category's major spread. In (12), d local feature vectors also are used to avoid singularities of the matrix $\mathbf{\Sigma}^k_{\text{Local}}$, so the similarity of feature in $\mathcal{Q}$ to $\mathbf{\Sigma}^k_{\text{Local}}$ is formalized by the following formula:

$$z_{\text{local}} = \text{Diag}\left(f\left(\bar{\mathcal{F}}'_q\right)\right) = \text{Diag}\left(\bar{\mathcal{F}}'^{T}_q \times \left(\mathbf{\Sigma}^k_{\text{Local}}\right)^{-1} \times \bar{\mathcal{F}}'_q\right). \tag{13}$$

In contrast to measures like Euclidean (L2) [67], [68], [69], the Mahalanobis measure is particularly appropriate for circumstances where there is just a small quantity of training data, as seen in Fig. 4.

The boundaries between clusters are linear when employing the Euclidean measure, which is equivalent to $\mathbf{\Sigma}^k_{\text{Local}}$ in (12), as shown in Fig. 4. Therefore, it is challenging to categorize samples that are toward the boundaries of the cluster when there is a limited amount of training data. In contrast, for captures characterized by the distribution of $\mathcal{S}_k$, samples in $\mathcal{Q}$ are better classified when compared with various features obtained from $\mathcal{S}_k$.

Theorem 1: Consider $f(\mathbf{x}, \Sigma) = \mathbf{x}^T \times (\Sigma)^{-1} \times \mathbf{x}$ where $\Sigma \in \mathbb{R}^{c\times c}$ is nonsingular covariance matrix and $\mathbf{x} \in \mathbb{R}^c$ is nonzero vector. The value of $f(\mathbf{x}, \Sigma)$ obtains its maximum if $\mathbf{x}$ is in the direction of the n vectors corresponding to the n smallest eigenvalues of Σ. This indicates that the distributions of $\mathbf{x}$ and each category are consistent and that $\mathbf{x}$ is pointed in the direction of the category's major spread.

Proof: Given Σ is a nonsingular matrix, so it is satisfied $\Sigma = V\Lambda V^T$ where $V = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c] \in \mathbb{R}^{c\times c}$ is an orthogonal matrix whose columns are the eigenvectors of Σ, and $V^{-1} = V^T$. $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_c) \in \mathbb{R}^{c\times c}$ is a diagonal matrix containing the eigenvalues of Σ in ascending order ($0 < \lambda_1 < \lambda_2 < \cdots < \lambda_c$). With the space spanned by $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c\}$ orthogonal vectors, we have $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_c]^T$ be the vector of coefficients of the projection of $\mathbf{x}$, so $\mathbf{x} = V\alpha = \sum_{i=0}^{i=c} \mathbf{v}_i \times \alpha_i$ and $\mathbf{x} \times \mathbf{v}_i = ||\mathbf{x}|| \times ||\mathbf{v}_i|| \times \text{Cos}(\psi_i) = \alpha_i$ where ψ_i denote the angle between $\mathbf{x}$ and $\mathbf{v_i}$. The function $f(\mathbf{x}, \Sigma)$ can be rewritten as

$$\begin{aligned} f(\mathbf{x}, \Sigma) &= \mathbf{x}^T \times (\Sigma)^{-1} \times \mathbf{x} \\ &= \mathbf{x}^T \times (V\Lambda V^T)^{-1} \times \mathbf{x} \\ &= \mathbf{x}^T \times (V^T)^{-1} \times (\Lambda)^{-1} \times (V)^{-1} \times \mathbf{x} \\ &= \mathbf{x}^T \times V \times (\Lambda)^{-1} \times V^T \times \mathbf{x} \\ &= (V \times \alpha)^T \times V \times (\Lambda)^{-1} \times V^T \times (V \times \alpha) \\ &= \alpha^T (\Lambda)^{-1} \alpha. \end{aligned} \tag{14}$$

With λ_i's condition, (14) can be rewritten as follows:

$$\begin{aligned} f(\mathbf{x}, \Sigma) &= \alpha^T \times \text{Diag}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_c}\right) \times \alpha \\ &= \sum_{i=1}^{c} \alpha_i^2 \times \frac{1}{\lambda_i}. \end{aligned} \tag{15}$$

Considering first n smallest eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, if $\mathbf{x}$ is in the same direction as the corresponding n eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ then with $\mathbf{v}_j$, where $n < j \leq c$ then $\mathbf{x}^T \times \mathbf{v}_j = 0$, equivalent to $\alpha_j = 0$. Therefore, the value of $f(\mathbf{x}, \Sigma)$ is

calculated as

$$f(\mathbf{x}, \Sigma) = \sum_{k=1}^{n} \alpha_k^2 \times \frac{1}{\lambda_k} = ||\mathbf{x}||^2 \sum_{k=1}^{n} ||\mathbf{v}_k||^2 \times \mathrm{Cos}(\psi_k)^2 \times \frac{1}{\lambda_k}. \quad (16)$$

So, with nonzero vector $\mathbf{x}$, the value of $f(\mathbf{x}, \Sigma)$ obtains its maximum is $||\mathbf{x}||^2 \sum_{k=1}^{n} ||\mathbf{v}_k||^2 \times \mathrm{Cos}(\psi_k)^2 \times (1/\lambda_k)$ if $\mathbf{x}$ is in the direction of the n vectors corresponding to the n smallest eigenvalues of Σ, the theorem is proven.

D. Classification Module

After capturing the correlation both globally through the **Global Transformer** module and locally through the **Local Mahalanobis** module, the classification block is introduced to synthesize and classify, creating an end-to-end architecture for the proposed model. In (10), it is argued that z_{global} is a matrix that maintained global information between $\tilde{\mathcal{Q}}_q$ taken from the feature of $\mathcal{Q}$ and $\tilde{\mathcal{K}}_s$ taken from the feature of $\mathcal{S}$, respectively. Attention weight enables the model to concentrate on specific data, especially when there is a restricted amount of input, and $\tilde{\mathcal{Q}}_q$ transports global information both ways through repeated information synthesis (N is the depth of the encoder and the decoder in the **Global Transformer**). The global metric is achievable as follows:

$$M_{\text{global}} = z_{\text{global}} \times \mathbf{w}^T \quad (17)$$

where $\mathbf{w}$ is the learnable weight matrix implemented as a linear layer. The local correlation matrix z_{local} in (13) captures the correlation locally as presented in Section III-C. All local similarities between the query image and C categories are concatenated in sequence, so a 1-D convolution layer with a stride of $d = h \times w$ is utilized to compute the local metric

$$M_{\text{local}} = \mathrm{conv1d}(z_{\text{local}}, \text{kernel\_size} = d, \text{stride} = d). \quad (18)$$

The overall metric between $X_j^q \in \mathcal{Q}$ and $\mathcal{S}$ can be determined as

$$M_{\text{total}} = \mu \times M_{\text{global}} + (1 - \mu) \times M_{\text{local}} \quad (19)$$

where $\mu \in (0, 1)$ is a hyperparameter. The nearest neighbor strategy [52] is used for classification, with $X_j^q \in \mathcal{Q}$ having the highest similarity to $X_i^s \in \mathcal{S}$, the corresponding label is classified as Y_i^s. The contrastive loss is employed to train with M_{global} as proposed in [70], shown as

$$L_c = -\log \frac{\sum_{Y_j^s = Y_i^q} e^{M_{\text{global}}(X_j^q, X_i^s)}}{\sum_{Y_j^s = Y_i^q} e^{M_{\text{global}}(X_j^q, X_i^s)} + \sum_{Y_j^s \neq Y_i^q} e^{M_{\text{global}}(X_j^q, X_i^s)}}. \quad (20)$$

The objective of (20) is to push samples with negative labels (i.e., $Y_j^s \neq Y_i^q$) away while bringing samples with the positive label (i.e., $Y_j^s = Y_i^q$) closer together. For M_{local} obtained by optimizing the cross entropy loss function (L_{ce}) [71], the total loss function is presented as follows:

$$L_{\text{total}} = \mu \times L_c + (1 - \mu) \times L_{\text{ce}}(\hat{Y}^q, Y^q) \quad (21)$$

where $\hat{Y}^q$, Y^q denoted predicted label and true label of input query samples.

TABLE I
DETAILED STRUCTURAL PARAMETERS AND COMPUTATIONAL COMPLEXITY (FLOPS) OF THE PROPOSED MODEL

Module	Parameters	FLOPs
Multiscale Large Kernel Feature Extraction	37,440	33.12×10^6
Global Transformer	46,016	82.18×10^3
Local Mahalanobis	0	34.08×10^6
Classification	4,354	9.23×10^3
Total	87,810	67.29×10^6

To specify the size for each module, Table I gives specific parameters and computational complexity measured in the number of floating-point operations (FLOPs) of key blocks in the proposed model.

In Table I, the model parameters show that the total parameters of the entire model are 87 810. This is equivalent to the model taking up only 0.33 MB of memory in float32. Thus, deploying to real-world devices with constrained capacity is quite possible. The computational complexity of each module is also given through the flops. Compared with some other CNN-based models such as VGG16 or Resnet50 under the same setting conditions, our model has better flops. While VGG16 is 1.37×10^9 and Resnet50 is 293.5×10^6, respectively, our model has flops of just 67.29×10^6. The suggested model has an advantage in terms of computation cost because it uses transformer architecture and depthwise convolutions rather than conventional convolutions in the feature extraction module. In contrast, models like VGG16 or Resnet50 include many traditional convolution layers, which increases computational complexity.

The training process based on our proposed method is presented in the pseudo-code described in Algorithm 1.

IV. EXPERIMENT

A. Dataset

Implement and compare the effectiveness of the proposed method, the datasets used for training and evaluation are the Case Western Reverse University dataset (CWRU) [72] and the Paderborn University dataset (PU) [73]. These are two famous, reputable datasets recognized by the research community in the field of bearing fault diagnosis through many studies [26], [28], [29], [74], [75]. Besides, these two datasets have good quality through the characteristics of a number of errors, size, and diversity.

1) CWRU Dataset: The CWRU dataset [72] is a well-known dataset in the field of bearing fault diagnosis, with artificial faults. The main structure of the bearing includes the inner race, outer race, and ball. These are three main faults that cause bearing faults, each main fault includes three small faults related to the size of the crack on that part. Specifically, the fault labels and locations are shown in Table II.

In total, ten labels including nine fault labels and one nonfault label are obtained. Each label is loaded from three different load types (1, 2, and 3 hp corresponding to motor

Algorithm 1 Algorithm for Training With Our Proposed Model

Require: **Support Set** $\mathcal{S} = \{(X_i^s, Y_i^s)\}_{i=1}^{n_s}$
Query Set $\mathcal{Q} = \{(X_j^q, Y_j^q)\}_{j=1}^{n_q}$
Base Model f_θ with random initial weights
Step size hyper parameters α, β
Number of Iterations T
Ensure: **Class Probabilities** $p(Y^q|X^q, \mathcal{S})$ for each example X^q in $\mathcal{Q}$

Initialize f_θ with random weights
for $t = 1$ to T **do**
  for each class c in $\mathcal{S}$ **do**
    /* Random choose samples for sub-task set */
    Sample a few examples from $\mathcal{S}$ for class c: $\mathcal{S}_c$
    /* Feature extraction */
    Compute $\hat{\mathcal{F}}_s$ and $\hat{\mathcal{F}}_q$ using **MLKFE** module.
    /* Global Transformer */
    Compute M_{global} using Eq (17)
    Compute contrastive loss $L_{c,\mathcal{S}_c}$ using Eq (20)
    /* Local Mahalanobis */
    Compute M_{local} using Eq (18)
    Compute cross-entropy loss $L_{ce,\mathcal{S}_c}(\hat{Y}^q, Y^q)$
    /* Gradient descent */
    Compute $L_{total,\mathcal{S}_c}$ using Eq (21)
    $\theta_i' = \theta - \alpha \nabla_\theta L_{total,\mathcal{S}_c}(f_\theta)$
  end for
  /*Update weights */
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{S}_c \in \mathcal{S}} L_{total,\mathcal{S}_c}(f_{\theta'})$
  /*Evaluation */
  for each example X^q in $\mathcal{Q}$ **do**
    Compute M_{total} using (19)
    Predict the class probability distribution: $p(Y^q|X^q, \mathcal{S}) = \text{Softmax}(M_{total})$
  end for
  $\rightarrow$ *Save the best weight if needed*
end for

TABLE II
OVERVIEW OF THE CWRU DATASET USED IN EXPERIMENT

Location Of Fault	**Size of Fault** (in Mils)	**Label**
Healthy	-	0
Ball	0.007	1
Ball	0.014	2
Ball	0.021	3
Inner Race	0.007	4
Inner Race	0.014	5
Inner Race	0.021	6
Outer Race	0.007	7
Outer Race	0.014	8
Outer Race	0.021	9

speeds of 1772, 1750, and 1730 rpm). Each sample is taken from two vibration signals as illustrated in [28]. We create training samples from half of the vibration signal and test samples from the other half. The 2048-point sliding window with an 80-point shift step is utilized to create the train samples. The test samples are conducted with a sliding window that is the same size and does not overlap. To demonstrate the efficacy of the suggested model in several other circumstances, we randomly selected 750 samples for testing and modified the number of training samples in many different scenarios (from 30 to 19 800 samples). This dataset focuses on building diverse error sizes for each part of the bearing: inner race, outer race, and ball. The complexity of the classes is similar.

TABLE III
OVERVIEW OF THE PU DATASET USED IN THE EXPERIMENT

Location Of Fault	**Cause of Fault**	**Severity**	**Type**	**Code**	**Label**
Healthy	-	-	-	K001	0
Outer Race	Electrical discharge machining	1	Artificial	KA01	1
Outer Race	Electric engraver	2	Artificial	KA03	2
Outer Race	Pitting	1	Real	KA04	3
Outer Race	Drilling	1	Artificial	KA07	4
Outer Race	Plastic deform	1	Real	KA15	5
Outer Race	Pitting	2	Real	KA16	6
Inner Race	Electrical discharge machining	2	Artificial	KI01	7
Inner Race	Electric engraver	1	Artificial	KI03	8
Inner Race	Pitting	1	Real	KI04	9
Inner Race	Electric engraver	2	Artificial	KI07	10
Inner Race	Pitting	3	Real	KI16	11
Inner Race	Pitting	2	Real	KI18	12

2) PU Dataset: In contrast to the CWRU dataset, the PU dataset [73] contains both artificial and real faults that are brought about by electrical machine operation. In our experiments, to prove that the proposed model is applicable in practice, we perform diagnostics on the PU dataset with more complex faults than the CWRU dataset.

Table III shows 13 labels selected out of a total of 32 labels provided by the author. Of these, half of the fault labels are artificial and the other half are real faults. The severity levels of faults are also displayed in column Severity, where 1 denotes a defect length of less than 2 mm, l2 denotes a defect length of between 2 and 4.5 mm, and 3 denotes a defect length of 4.5–13.5 mm. It can be seen that the complexity of the PU dataset is much larger than the CWRU set, as the size of the cracks is also larger and the faults have different severity. The sampling method is similar to that presented with the CWRU set, but both training samples and testing samples are nonoverlap. There are 1950 samples in the test set.

B. Implementation Details

For model setting: Three *resLKA* modules with the convolution dimension of 64 are included in the MLKFE settings for the suggested model. The kernel size and dilated rate parameters for each module are $(7, 2)$, $(11, 3)$, and $(15, 4)$, respectively, for depthwise convolution and DDW. The number of encoder–decoder in the **Global Transformer** is chosen to be $N = 3$. Hyperparameters μ is the ratio between the global branch and the local branch chosen based on the ablation study, with $\mu = 0.3$ in the experiment with the CWRU dataset and $\mu = 0.4$ in the experiment with the PU dataset.

To train the proposed model presented in Section III, the Adam optimization method is used to optimize the loss function in (19). The learning rate is initially set to 0.001 and divided into half every ten epochs until it reaches 0.00001, then it is stable. In both experiments, the proposed model was

TABLE IV
COMPARISON OF THE TEN-WAY ONE-SHOT ACCURACY (IN %) OF THE DIFFERENT METHODS ON THE CWRU DATASET UNDER DIFFERENT DATA CONDITIONS

Method	Metric	Training samples					
		30	60	90	300	600	19800
ProtoNet [30]	Euclide	61.13	65.14	66.32	78.77	96.61	97.62
Cosine Classifier [38]	Cosine	78.26	85.87	88.13	92.67	99.13	99.73
MatchingNet [77]	Cosine	69.43	86.27	93.23	95.93	99.46	99.86
RelationNet [78]	Euclide	65.14	82.23	83.14	88.05	96.65	98.12
Siamese-WDCNN [28]	Euclide	57.13	82.88	91.84	96.48	97.35	99.53
Siamese-ConvMixer [29]	Euclide	61.12	83.17	95.48	96.14	99.63	99.84
Cross Attention Network [39]	CA	72.17	81.13	92.28	93.14	96.83	99.29
CovaMNet [32]	Cov	79.73	92.12	96.51	98.91	99.25	99.62
SA-CovaMNet [34]	Cov	80.93	92.56	95.14	98.62	99.67	99.69
MF-Net [76]	Cov	83.21	94.13	97.51	99.51	99.58	99.63
QS-Former [37]	CA + EMD	79.88	92.14	96.13	98.21	99.06	99.63
Proposed	CA + Mah	**84.87**	**96.82**	**98.17**	**99.61**	**99.78**	**99.89**

TABLE V
COMPARISON OF THE TEN-WAY FIVE-SHOT ACCURACY (IN %) OF THE DIFFERENT METHODS ON THE CWRU DATASET UNDER DIFFERENT DATA CONDITIONS

Method	Metric	Training samples				
		60	90	300	600	19800
ProtoNet [30]	Euclide	67.14	67.42	79.26	99.13	98.82
Cosine Classifier [38]	Cosine	88.67	90.13	94.71	99.26	99.73
MatchingNet [77]	Cosine	90.53	89.77	97.13	99.76	99.86
RelationNet [78]	Euclide	83.41	86.74	89.75	97.65	99.12
Siamese-WDCNN [28]	Euclide	86.12	91.84	96.48	98.71	99.13
Siamese-ConvMixer [29]	Euclide	87.14	93.64	95.70	97.63	99.07
Cross Attention Network [39]	CA	82.44	91.64	94.14	94.83	99.87
CovMNet [32]	Cov	93.41	96.59	99.01	99.14	99.67
SA-CovaMNet [34]	Cov	92.01	94.23	98.72	99.77	99.81
MF-Net [76]	Cov	95.61	98.78	99.56	99.71	99.73
QS-Former [37]	CA+EMD	94.11	96.25	98.14	99.21	99.53
Proposed	CA+Mah	**98.23**	**98.98**	**99.14**	**99.53**	**99.76**

trained for 100 epochs with NVIDIA TeslaT4 16 GB GPU and the training time was approximately 30 min.

C. Results on the CWRU Dataset

This section presents the results of the proposed model compared with other few-shot learning-based models, which are very prominent and widely used in the field of bearing fault diagnosis. All reported models are reimplemented from the open-source code published by the authors. First, the experiment is performed in the one-shot ten-way case with varying conditions of training data, ranging from a very small amount of only 30 samples (three samples per class) to a large amount of training data (total 19 800 samples), the results are given in Table IV.

In the metric column, CA denotes the cross-attention metric, Cov denotes the covariance metric, EMD denotes the Earth'mover distance metric, and Mah denotes the Mahalanobis metric. First, we observe that when there is only a very limited quantity of training data, few-shot models based on Euclidean measures perform poorly, such as Siamese-WDCNN [28] only achieves 57.13% accuracy and the highest is only 65.14% with RelationNet [77]. With models using cosine metrics such as the Cosine Classifier [38] and MatchingNet [76], the reported results were better when the highest level was 78.26%. As has been proven, models using uniform distributions achieve the best results, while models based on the CovaMNet [32] baseline all give positive results (around 80%). And especially, with our proposed model when combining both global metrics and local metrics as presented in the theory section, the result was outstanding, with 84.87% when only 30 samples of training were available. The suggested model not only performs well when trained on small quantities of data, but it also performs exceptionally well under optimum training conditions (more training samples). This is also reported in Table IV when in other cases the proposed model still shows out performance with the remaining state-of-the-art models. Confusion matrix and visualization on 2-D by TSNE are also shown in Figs. 5 and 6 to make the comparisons more intuitive.

Similarly, experiments with five-shot ten-way were also performed and reported in Table V. When the number of samples in classes is increased, models using distributional features still show superiority in limited data conditions. With different data conditions from 60 samples to 19 800 samples, upon reaching an accuracy of 98.23% under the scenario of 60 training samples, the proposed model outperformed the others (top 2 is Mixer-Former Net reach 95.61%), while models using Euclidean or cosine measures only reach approximately less than 90%. This has proven the effectiveness of the proposed model when combining local features of each sample in the class (local metric) with cross-attention similarity (global metric), making the model highly generalizable, especially in conditions where there is limited data, finding important information plays an even greater role. Confusion matrix and visualization by TSNE in 2-D space are also presented in Figs. 7 and 8 to visualize the effectiveness of the proposed model compared to the top-score models in Table V.

D. Results on the PU Dataset

To demonstrate the effectiveness of the proposed model, we conducted experiments on the PU dataset. Different from the CWRU dataset, the PU dataset includes many complex faults when there are both artificial faults and faults encountered in real life by the operated electrical machine itself, and the faults also have different severity as shown in Table III. Table VI shows the results of our work and the comparisons it made with various models.

The findings indicate that with just 195 training samples, models like ProtoNet [30], Cosine classifier [38], MatchingNet [76], RelationNet [77], Siamese-WDCNN [28], and Siamese-ConvMixer [29] that use basic measures such as Euclidean or cosine are not very effective due to the complexity of big data (only approximately 70% accuracy or less). The cross-attention network [39] achieves 77.72% accuracy, however, with the model with the restriction of ignoring locality, it cannot achieve too high performance, while CovMNet [32], SA-CovaMNet [34], and MF-Net [75] uses covariance metric to achieve approximately 80% accuracy but does not have global properties, so the desired results have not been achieved. By considering both local and global characteristics, our proposed model has demonstrated its effectiveness when applying Mahalanobis distance with local features and cross attention to capturing global features. It has achieved high efficiency of 80.08% and 86.15% corresponding to one shot and five shots. To more clearly represent and compare each class

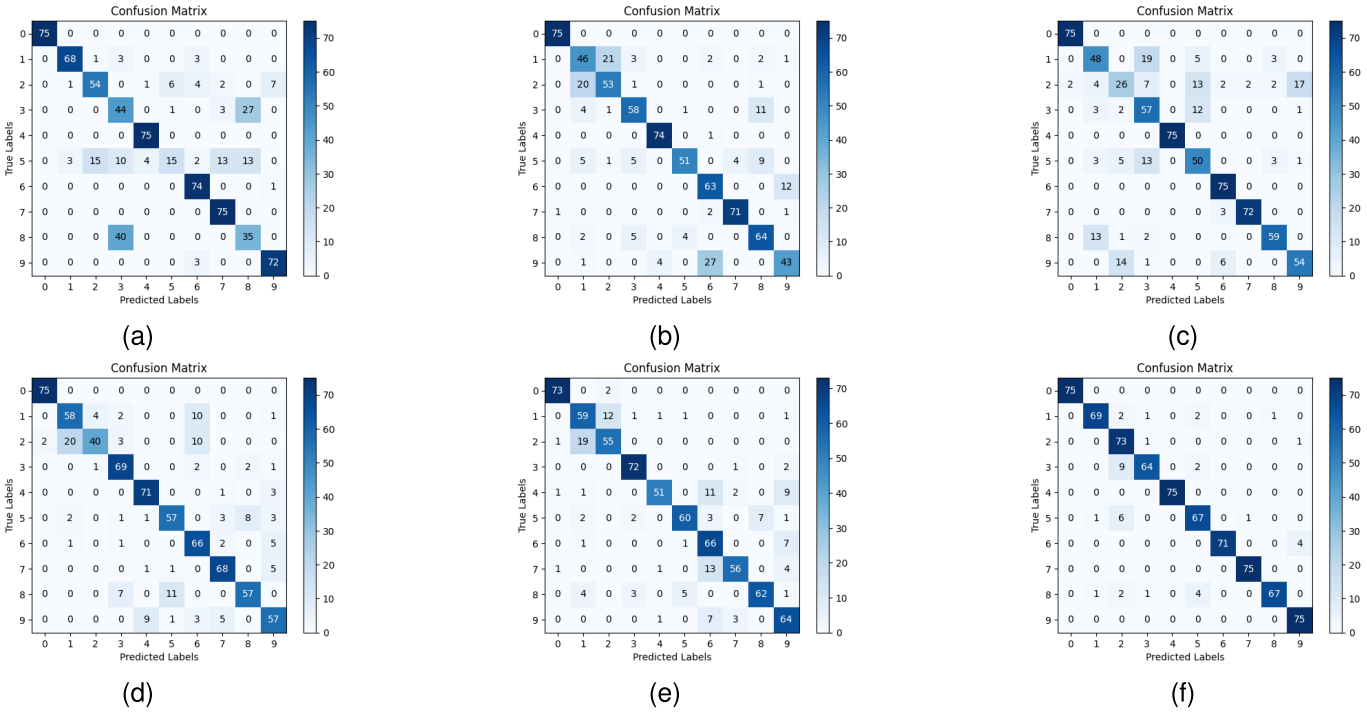


Fig. 5. Confusion matrix of the six methods with the highest accuracy in Table IV under the condition of 30 training samples of CWRU data. (a) Cosine classifier. (b) CovaMNet. (c) QS Former. (d) SA CovaMNet. (e) MF Net. (f) Our proposed.

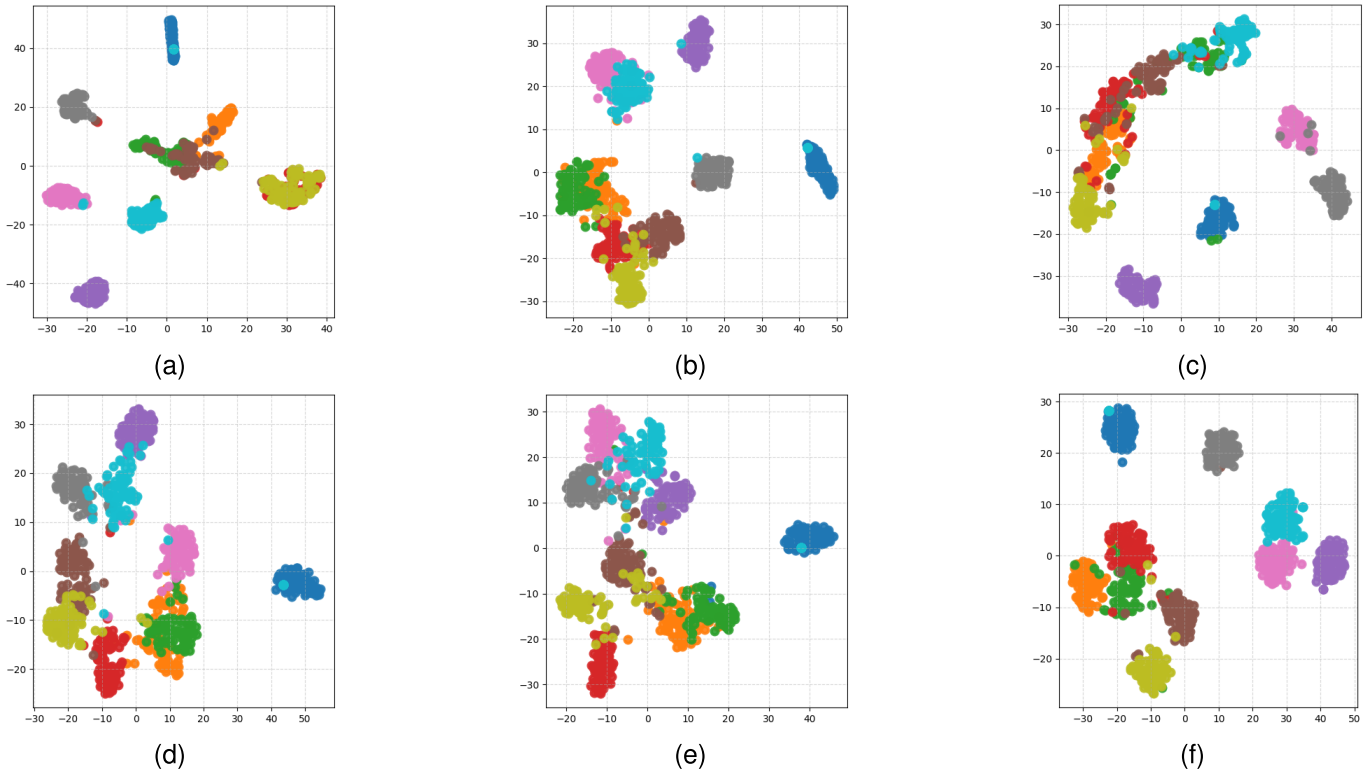


Fig. 6. Visualization (TSNE) of the six methods with the highest accuracy in Table IV under the condition of 30 training samples of CWRU data. (a) classifier. (b) CovaMNet. (c) QS Former. (d) SA CovaMNet. (e) MF Net. (f) Our proposed.

specifically, since there are both real and artificial samples, we present a confusion matrix and visualize it through TSNE presented in Fig. 9.

The confusion matrix of CovMNet [32], QS-Former [37], and MF-Net [75] demonstrates that while these models perform admirably on artificial faults, their performance is poor for real faults, particularly those that are extremely complicated like class 6, 10, 11, 12 (Severity = 2 or 3 in Table III). When there are not too many incorrect samples—especially when the incorrectly diagnosed samples do not fall into other

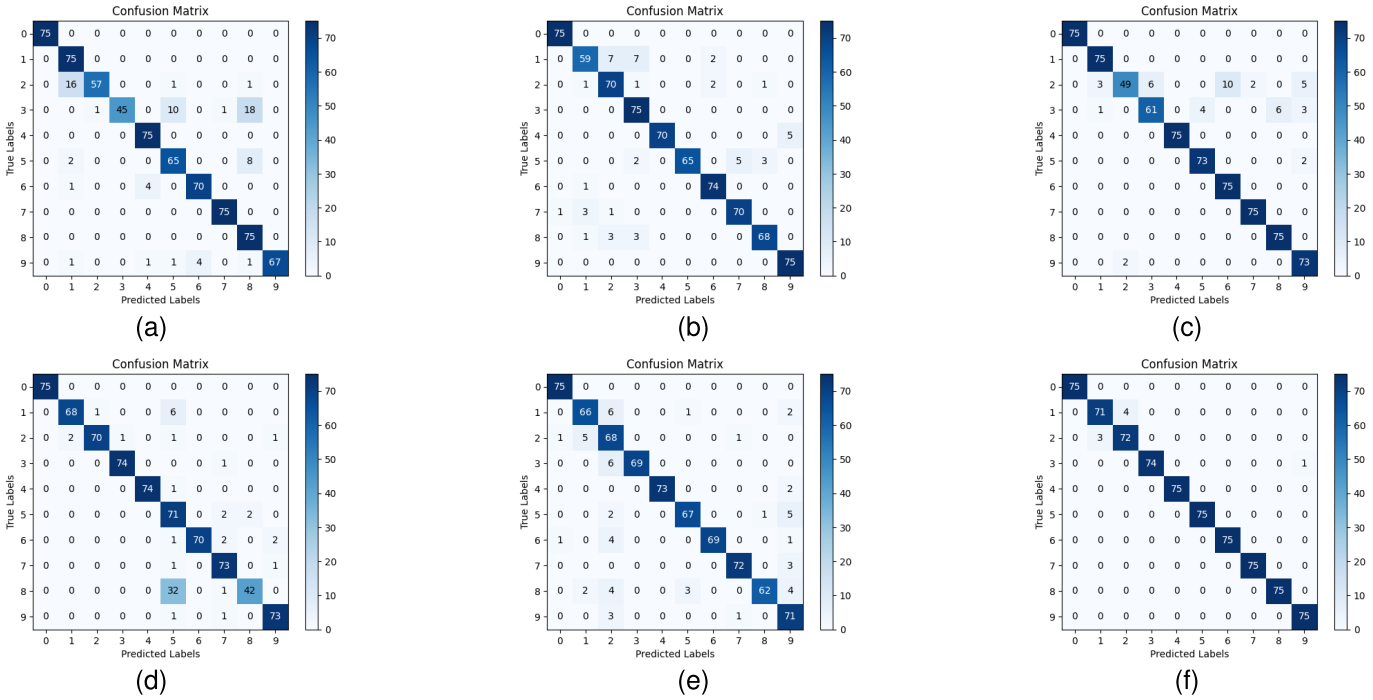


Fig. 7. Confusion matrix of the six methods with the highest accuracy in Table V under the condition of 60 training samples of CWRU data. (a) Matching Net. (b) CovaMNet. (c) QS Former. (d) SA CovaMNet. (e) MF Net. (f) Our proposed.

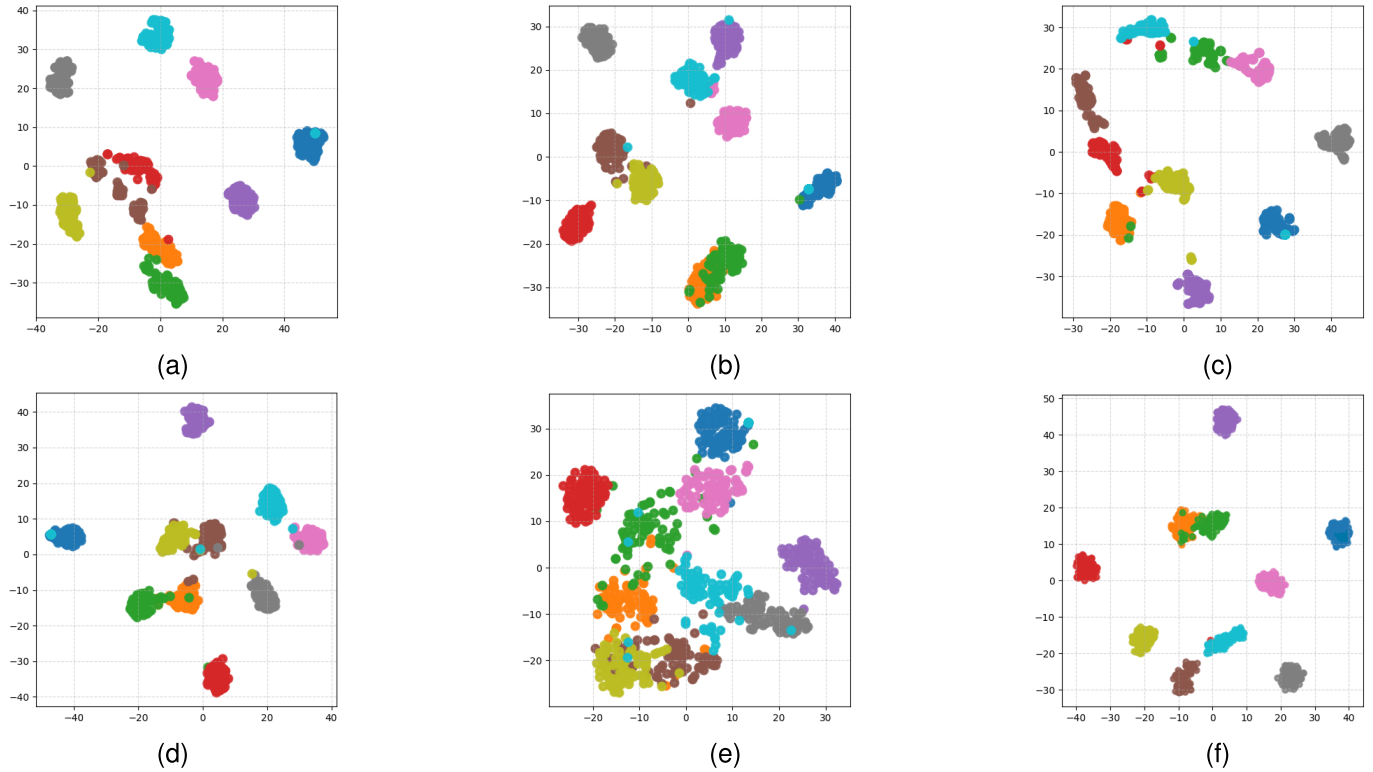


Fig. 8. Visualization (TSNE) of the six methods with the highest accuracy in Table V under the condition of 60 training samples of CWRU data. (a) Cosine classifier. (b) CovaMNet. (c) QS Former. (d) SA CovaMNet. (e) MF Net. (f) Our proposed.

fault categories, which primarily distinguish between Inner Race and Outer Race—our suggested approach solves this issue. In addition, Table VI also shows the superiority of the proposed model with other cases of data when with more or less training data, the proposed model provides high diagnostic accuracy.

E. Ablation Study

To clearly demonstrate the improvements of the proposed model, we conduct ablation experiments. Baseline was initially built from CovaMNet [32] with a CNN network (Conv64) to extract features from the query set $\mathcal{Q}$ and the support set $\mathcal{S}$. Then, the covariance metric is applied to compare

TABLE VI

COMPARISON OF THE 13-WAY K-SHOT ($k = 1, 5$) ACCURACY (IN %) OF THE DIFFERENT METHODS ON THE PU DATASET UNDER DIFFERENT DATA CONDITIONS

Method	Metric	195		260		650		1300		25844	
		1shot	5shot	1shot	5shot	1shot	5shot	1shot	5shot	1shot	5shot
ProtoNet [30]	Euclide	53.14	55.51	58.16	58.23	69.07	69.88	89.23	92.14	96.68	97.23
Cosine Classifier [38]	Cosine	61.24	69.53	73.14	78.02	85.12	85.25	93.07	93.81	97.24	98.01
MatchingNet [77]	Cosine	56.72	64.14	67.12	67.06	78.07	78.13	77.25	77.28	87.06	90.92
RelationNet [78]	Euclide	63.14	71.28	72.28	76.05	79.11	79.68	87.74	89.06	91.16	97.22
Siamese-WDCNN [28]	Euclide	66.72	74.59	79.81	79.13	82.06	84.62	90.01	93.42	98.07	98.76
Siamese-ConvMixer [29]	Euclide	70.28	76.81	82.14	84.01	87.13	87.46	95.06	95.73	99.43	99.17
Cross Attention Network [39]	CA	75.17	77.72	79.06	79.63	83.14	84.32	90.07	91.26	97.23	97.46
CovMNet [32]	Cov	76.38	80.23	84.25	88.46	89.92	93.67	97.13	97.21	99.06	99.35
SA-CovaMNet [34]	Cov	72.75	79.92	83.12	89.14	90.06	90.17	92.83	93.14	98.07	99.15
MF-Net [76]	Cov	77.13	83.25	87.06	89.17	91.23	92.14	92.06	95.87	99.23	99.37
QS-Former [37]	CA+EMD	80.17	82.28	88.13	88.61	94.02	95.17	96.28	96.66	98.13	99.52
Proposed	CA+Mah	**80.08**	**86.15**	**87.13**	**89.46**	**94.68**	**96.72**	**98.14**	**99.23**	**99.62**	**99.76**

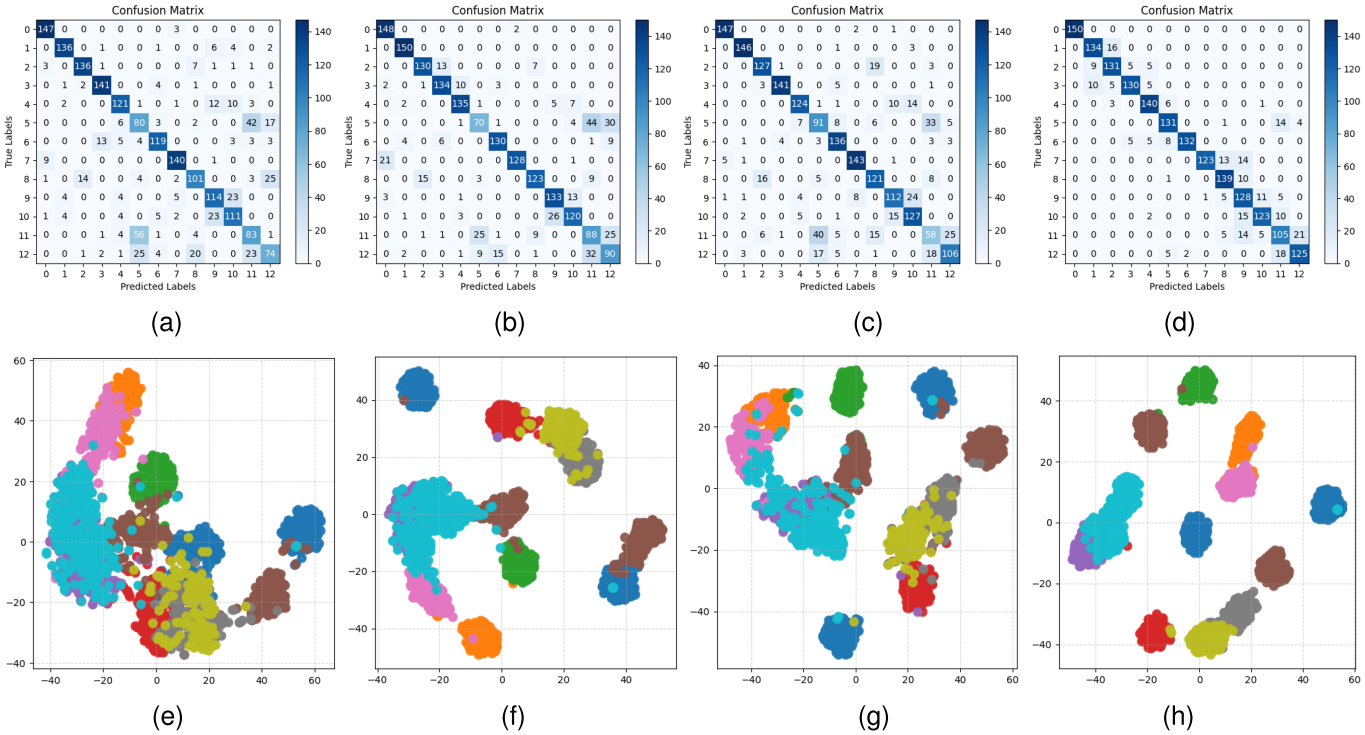


Fig. 9. Confusion matrix and visualization (TSNE) of the four methods with the highest accuracy in Table VI under the condition of 195 training samples of PU data. (a) QS Former. (b) CovaMNet. (c) MF Net. (d) Our proposed. (e) QS Former. (f) CovaMNet. (g) MF Net. (h) Our proposed.

correlation and classification. In our model with three main improvements as follows: MLKFE was proposed to replace CNN (Conv64) to create features with more information, the **Global Transformer** module was developed to obtain global features, based on cross attention as a metric in the transformer architecture and the **Local Mahalanobis** module are proposed by us to retrieve local features, increasing computational information, especially in conditions of limited training data.

Table VII presents eight test cases to demonstrate the effectiveness of the proposed modules equivalent to $\Delta 1$–$\Delta 8$. In $\Delta 1$, only the baseline based on CovaMNet is used. The lowest results are given when only reaching 79.73% and 80.23% accuracy above in conditions of limited data as set out above the CWRU dataset and the PU dataset. After changing the feature extraction block from CNN to the **MLKFE** proposal module in $\Delta 2$, the accuracy increased by approximately 3% in low data conditions. Besides, if we only use global features and local features as in case $\Delta 7$, the results are only slightly better than the baseline. This has proven the effectiveness of **MLKFE** when the information is more condensed. In the cases $\Delta 5$, $\Delta 6$, and $\Delta 7$, testing with the **Global Transformer** module and the **Local Mahalanobis** module is given. The results show that the proposed metrics are better than the baseline. Thus, we can observe that the covariance metric is not very useful when little data is provided since it is susceptible to nonsingular covariance matrices, which makes

TABLE VII
ABLATION STUDY FOR DIFFERENT COMPONENTS OF THE PROPOSED MODEL ON THE CWRU DATASET AND THE PU DATASET UNDER DIFFERENT DATA CONDITIONS

	Baseline	MLKFE	Global Transformer	Local Mahalanobis	*CWRU Dataset*			*PU Dataset*		
					30	**600**	**19800**	**195**	**650**	**25844**
$\Delta 1$	✓				79.73	99.01	99.67	80.23	93.97	99.35
$\Delta 2$	✓	✓			81.17	99.26	99.71	83.17	94.32	99.46
$\Delta 3$	✓		✓		80.97	99.11	99.54	81.02	94.26	99.25
$\Delta 4$	✓			✓	82.23	99.37	99.62	81.78	93.78	99.14
$\Delta 5$	✓	✓	✓		82.16	99.23	99.71	84.06	95.12	99.56
$\Delta 6$	✓	✓		✓	83.01	99.63	99.45	84.79	96.23	99.71
$\Delta 7$	✓		✓	✓	80.79	98.89	99.21	83.29	95.79	99.62
$\Delta 8$	✓	✓	✓	✓	**84.87**	**99.78**	**99.89**	**86.15**	**96.72**	**99.76**

it impossible to identify the nonsingular region characteristic space of each class in $\mathcal{S}$. Additionally, case $\Delta 6$ shows that our proposed local feature covariance in the Mahalanobis module yields a very high efficiency of 83.01%. However, in general, in case 8 with all proposals, it has been shown that the combination of proposed blocks is reasonable when achieving the best results (84.87%). With spectrogram images, information has both time and frequency meanings, and especially in limited data conditions, each pixel can be considered extremely important information. Therefore, combining global and local information from good features provides high accuracy in bearing fault diagnosis. To clarify the influence of global information and local information, we conducted an ablation study with the μ variable in (19) and (21) shown in Fig. 10.

Since μ is the balance coefficient, (19) and (21) state that the higher the influence of M_{local}, the smaller μ is, and inversely. The best result obtained in Fig. 10(a) with only 30 training samples on the CWRU dataset is 84.87% with $\mu = 0.3$. This indicates how successfully the suggested local feature works in increasing the number of samples in each class of $\mathcal{S}$ while figuring out the covariance matrix. Global features are also useful in improving accuracy when $\mu = 0$, which corresponds to utilizing just local features, greatly impairs the model's ability to diagnose faults. Similarly, with the bar plot in Fig. 10(b) showing the ablation study of μ on the PU dataset in the condition of 195 training samples, we see that the number of training samples is more than CWRU in the case of 30 training samples training, but the complexity of the PU dataset is much greater as described. Therefore, with $\mu = 0.4$, the proposed diagnostic model provides the highest accuracy of 86.15%. That shows a balance between global information and local information and they are well combined and complement each other in finding and synthesizing the best information for the model in limited data conditions.

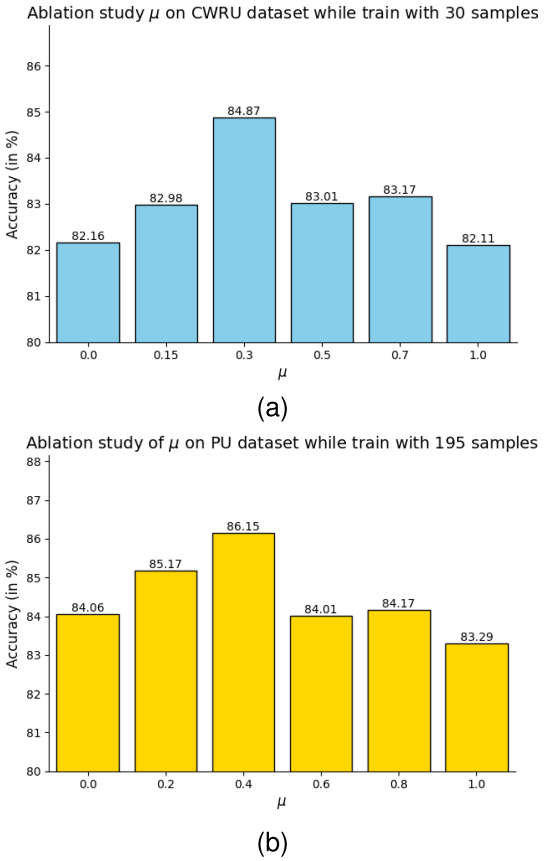


Fig. 10. Influence of μ on: (a) CWRU dataset while trained with 30 samples and (b) PU dataset while trained with 195 samples.

V. DISCUSSION

A. Global and Local Information

In conditions where there is only little training data, finding information is important to help the bearing fault diagnosis model produce good results. Instead of learning directly from samples like traditional deep-learning models, the few-shot model learns the correlation between samples, so information from both global and local is important, respects and complements each other. With our proposed method, the input signal is directly converted to spectral form and processed through the MLKFE block to create information-rich features that are selected before filtering global and local information through two blocks Global Transformer and Local Mahalanobis.

Global Transformer design is based on the architectural *Transformer*. Two main components encoder and decoder play the role of information processing of the support set $\mathcal{S}$ and the query set $\mathcal{Q}$. The input is the feature that has been handled through **MLKFE** as be tokenizer through global pooling. To obtain the correlation globally between $\mathcal{Q}$ and $\mathcal{S}$, one can take advantage of the matrix correlation through cross-attention in the decoder. This correlation matrix's information globally is handled through the connected block encoder–decoder and with mechanisms of attention characteristic based on the architecture.

In addition to information processing in the global information bureau tonic also be handled at the **Local Mahalanobis**. Different from the traditional methods mentioned above when applying to conditions where there is only a little training data and to ensure the covariance matrix is a nonsingular matrix, we propose to calculate the covariance matrix based on local features. First, we consider the correlation between samples in each class of the support set $S_k = \{\mathcal{F}_{sk}^1, \mathcal{F}_{sk}^2, \ldots, \mathcal{F}_{sk}^C\} \in \mathbb{R}^{C\times h\times w\times c}$ represented for class with label k in *C-shot K-way* task. Next, pixels at the same location in each channel are combined and create a local feature vector (Fig. 3). A set of local feature vectors that describe class k is obtained: $\tilde{S}_k = \{\mathcal{F}_{sk}^1, \mathcal{F}_{sk}^2, \ldots, \mathcal{F}_{sk}^d\} \in \mathbb{R}^{d\times c}$. Here, $d = C \times h \times w$ denotes the total number of local feature vectors of a class. With our proposed method, two problems are solved. First, the covariance matrix representing class k is guaranteed to be a nonsingular matrix because the number of observed samples in the support feature set $\tilde{S}_k$ is d (i.e., $256 \leq d \leq 1280$) much larger than the dimensionality of the feature c (i.e., $c = 64$—the number of filters in common neural networks), so the Mahalanobis distance is eligible to apply. Second, in conditions of limited data, all information is very important to be able to identify and distinguish signals. Therefore, the covariance matrix Σ calculated according to 11 carries correlation information between each pixel of the feature in class k, helping to enrich information and increase diagnostic accuracy for the proposed model. Thus, Mahalanobis is calculated according to 12 and ensures convergence through Theorem I, which has been proven.

In the suggested study, there is a stronger correlation between local and global information. When Transformer and Mahalanobis distance are combined, the suggested strategy offers a way to help increase the model's accuracy and help it generalize to a large extent when it comes to the limits of the training data.

B. Limitations and Considerations

This research aims to build an end-to-end few-shot learning model without relying on pretraining to diagnose bearing faults. The proposed approach has solved one of the major problems of modern deep-learning models: with only limited training data, the proposed model still provides diagnostic results with high accuracy. Besides, the proposed model has no higher computational complexity than other traditional methods and also does not take up much memory.

However, for end-to-end few-shot learning models, when diagnosing complex conditions such as adapting to the varying speeds of rotating machinery, they may not give high results. Because the data distribution of the training set and the testing set is significantly different, it is difficult for the model to learn general knowledge [3]. It is another challenge in fault-bearing diagnosis, especially in few-shot learning without transfer learning and fine-tuning.

The following are some approaches that can be utilized to address this issue. First, generative models can be a solution. This strategy is used to overcome data limitations, which can be used to increase the amount of data or create new samples of varying complexity. Lee et al. [78] use generative adversarial networks (GANs) to generate more training data from the original data, and this data contains a lot of noise intending to allow the model to learn from noisy samples. Zhang et al. [79] suggested the A2CNN model inspired by GAN, which is made up of a label classification algorithm, a domain discriminator, and an extractor for source and target features. Yan et al. [80] combined the diffusion model and the support vector machine to create realistic data under different conditions.

Besides applying generative methods, transfer learning is also a possible solution. With the model being pretrained on a sufficiently large and diverse dataset, taking advantage of fine-tuning on specific tasks makes it easier for the model to learn new knowledge and may improve diagnostic performance under different conditions. Wu et al. [81] proposed a deep-learning model based on fine-tuned CNN and LSTM to adapt to faults in a number of different conditions. Wang et al. [82] take advantage of pretraining from ImageNet—a large and famous dataset in the field of computer vision classification—to learn low-level features effectively. In addition, there are also many other studies applying transfer learning to adapt to different working conditions [83], [84], [85].

In addition, meta-learning applied to domain adaptation is also an interesting approach. In this method, the model will be updated with weights based on both the parameters of the training set and when forwarding with the validation set, provided that the training set and validation set have significantly different distributions. Li et al. [31] proposed a meta-learning-based method to find the most optimal set of parameters during training based on tasks with different complexity. Lin et al. [86] established a distance metric between different operations to determine the commonalities among domain tasks. Che et al. [87] used gradient consistency as an enhanced gradient-based meta-learning technique, training the base learner and optimizing the meta-learner through large task datasets to establish a new learner with strong generalization capabilities for defect diagnosis under varying working conditions. Also with the thought of meta-learning on domain adaptation, a number of other methods are also given [88], [89], [90].

Nevertheless, most methods to solve problems about domain adaptation can increase the computational complexity and consume memory capacity. Generative model methods might complicated, have limited capacity to produce samples that closely resemble reality, or produce forms that are extremely flawed and that learn from variations in the target model. With transfer learning, timing issues could be considered when the need for training on a data general, at the same time

the number of parameters may need to increase as want to generalize on the big data. Meta-learning, however, can solve problems with multitask different, but the volume of calculations is also relatively high due to the need to find a set of weights to suit every task, and pipeline training quite complex. Therefore, to be able to domain adaptation should consider tradeoffs on factors such as volume calculation, memory capacity, and duration of training.

VI. Conclusion

In the current study, we have introduced a novel approach for the few-shot classification of faults in bearing data. The input signal data including the support set and query images are converted to a spectrogram for enriching information before feeding to the feature extraction and classification framework. We propose a new MLKFE with the ResLKA built upon the depthwise-pointwise architecture of various kernel sizes and dilations. The output features are then fed in parallel to the Global Transformer module and the Local Mahalanobis module before going through the Classification module. In this way, our approach can learn global information from the Transformer module and local features from the Mahalabonis metric module. Experiments on the two data including the CWRU and PU datasets show excellent performance of the proposed approach especially in dealing with limited data. In the future, several directions can be considered to improve the proposed model. Transfer learning or meta-learning can be employed so that the model can adapt to different working conditions. In addition, error signals from other sources can also be applied such as electrical signals, and thermal signals. Also, other fault types besides bearing fault signals can be tried such as those from turbines or power systems.

References

[1] Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3757–3767, Jun. 2015.

[2] Z. Zhao et al., "Applications of unsupervised deep transfer learning to intelligent fault diagnosis: A survey and comparative study," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–28, 2021.

[3] G. Niu, X. Dong, and Y. Chen, "Motor fault diagnostics (don't short) based on current signatures: A review," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–19, 2023.

[4] X. Chen, R. Yang, Y. Xue, M. Huang, R. Ferrero, and Z. Wang, "Deep transfer learning for bearing fault diagnosis: A systematic review since 2016," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–21, 2023.

[5] Y. Lei, J. Lin, M. J. Zuo, and Z. He, "Condition monitoring and fault diagnosis of planetary gearboxes: A review," *Measurement*, vol. 48, pp. 292–305, Feb. 2014.

[6] X. Yu, Y. Wang, Z. Liang, H. Shao, K. Yu, and W. Yu, "An adaptive domain adaptation method for rolling bearings' fault diagnosis fusing deep convolution and self-attention networks," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.

[7] G. Bramerdorfer, J. A. Tapia, J. J. Pyrhönen, and A. Cavagnino, "Modern electrical machine design optimization: Techniques, trends, and best practices," *IEEE Trans. Ind. Electron.*, vol. 65, no. 10, pp. 7672–7684, Oct. 2018.

[8] X. Yang, Y. Zheng, Y. Zhang, D. S. Wong, and W. Yang, "Bearing remaining useful life prediction based on regression shapalet and graph neural network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[9] IEEE Motor Reliability Working Group, "Report of large motor reliability survey of industrial and commercial installations, Part I," *IEEE Trans. Ind. Appl.*, vol. IA-21, no. 4, pp. 853–864, Jul. 1985.

[10] *On Recommended Interval of Updating Induction Motors*, JEMA (in Japanese), Tokyo, Japan, 2000.

[11] W. Zhou, T. G. Habetler, and R. G. Harley, "Bearing condition monitoring methods for electric machines: A general review," in *Proc. IEEE Int. Symp. Diag. Electr. Mach., Power Electron. Drives*, Sep. 2007, pp. 3–6.

[12] V. Niskanen, A. Muetze, and J. Ahola, "Study on bearing impedance properties at several hundred kilohertz for different electric machine operating parameters," *IEEE Trans. Ind. Appl.*, vol. 50, no. 5, pp. 3438–3447, Sep. 2014.

[13] N. Uzhegov, A. Smirnov, C. H. Park, J. H. Ahn, J. Heikkinen, and J. Pyrhönen, "Design aspects of high-speed electrical machines with active magnetic bearings for compressor applications," *IEEE Trans. Ind. Electron.*, vol. 64, no. 11, pp. 8427–8436, Nov. 2017.

[14] F. Immovilli, A. Bellini, R. Rubini, and C. Tassoni, "Diagnosis of bearing faults in induction machines by vibration or current signals: A critical comparison," *IEEE Trans. Ind. Appl.*, vol. 46, no. 4, pp. 1350–1359, Jul. 2010.

[15] I. Sadeghi, H. Ehya, J. Faiz, and H. Ostovar, "Online fault diagnosis of large electrical machines using vibration signal—A review," in *Proc. Int. Conf. Optim. Electr. Electron. Equip. (OPTIM) Int. Aegean Conf. Electr. Mach. Power Electron. (ACEMP)*, May 2017, pp. 470–475.

[16] X. Zhang, Y. Liang, J. Zhou, and Y. Zang, "A novel bearing fault diagnosis model integrated permutation entropy, ensemble empirical mode decomposition and optimized SVM," *Measurement*, vol. 69, pp. 164–179, Jun. 2015.

[17] M. Ye, X. Yan, and M. Jia, "Rolling bearing fault diagnosis based on VMD-MPE and PSO-SVM," *Entropy*, vol. 23, no. 6, p. 762, Jun. 2021.

[18] J. Zhou, M. Xiao, Y. Niu, and G. Ji, "Rolling bearing fault diagnosis based on WGWOA-VMD-SVM," *Sensors*, vol. 22, no. 16, p. 6281, Aug. 2022.

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

[20] J. Tian, C. Morillo, M. H. Azarian, and M. Pecht, "Motor bearing fault detection using spectral kurtosis-based feature extraction coupled with K-nearest neighbor distance analysis," *IEEE Trans. Ind. Electron.*, vol. 63, no. 3, pp. 1793–1803, Mar. 2016.

[21] Y. Yu, D. Yu, and J. Cheng, "A roller bearing fault diagnosis method based on EMD energy entropy and ANN," *J. Sound Vib.*, vol. 294, nos. 1–2, pp. 269–277, Jun. 2006.

[22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[23] H. Pan, X. He, S. Tang, and F. Meng, "An improved bearing fault diagnosis method using one-dimensional CNN and LSTM," *J. Mech. Eng.*, vol. 64, Jul. 2018.

[24] L. Eren, T. Ince, and S. Kiranyaz, "A generic intelligent bearing fault diagnosis system using compact adaptive 1D CNN classifier," *J. Signal Process. Syst.*, vol. 91, no. 2, pp. 179–189, Feb. 2019.

[25] X. Chen, B. Zhang, and D. Gao, "Bearing fault diagnosis base on multi-scale CNN and LSTM model," *J. Intell. Manuf.*, vol. 32, no. 4, pp. 971–987, Apr. 2021.

[26] H. Wang, J. Xu, R. Yan, and R. X. Gao, "A new intelligent bearing fault diagnosis method using SDP representation and SE-CNN," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 52, pp. 2377–2389, May 2019.

[27] X. Li, S. Wan, S. Liu, Y. Zhang, J. Hong, and D. Wang, "Bearing fault diagnosis method based on attention mechanism and multilayer fusion network," *ISA Trans.*, vol. 128, pp. 550–564, Sep. 2022.

[28] 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.

[29] M.-H. Vu, V.-Q. Nguyen, T.-T. Tran, and V.-T. Pham, "A new convmixer-based approach for diagnosis of fault bearing using signal spectrum," in *Proc. Conf. Inf. Technol. Appl.* Cham, Switzerland: Springer, 2023, pp. 3–14.

[30] H. Shen, D. Zhao, L. Wang, and Q. Liu, "Bearing fault diagnosis based on prototypical network," in *Proc. Int. Conf. Mechatronics Eng. Artif. Intell. (MEAI)*, Mar. 2023, pp. 79–84.

[31] C. Li, S. Li, A. Zhang, Q. He, Z. Liao, and J. Hu, "Meta-learning for few-shot bearing fault diagnosis under complex working conditions," *Neurocomputing*, vol. 439, pp. 197–211, Jun. 2021.

[32] W. Li, J. Xu, J. Huo, L. Wang, Y. Gao, and J. Luo, "Distribution consistency based covariance metric networks for few-shot learning," in *Proc. Conf. AAAI Artif. Intell.*, vol. 33, no. 1, Jul. 2019, pp. 8642–8649.

[33] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7260–7268.

[34] J. Zhai, L. Han, Y. Xiao, M. Yan, Y. Wang, and X. Wang, "Few-shot fine-grained fish species classification via sandwich attention CovaMNet," *Frontiers Mar. Sci.*, vol. 10, Mar. 2023, Art. no. 1149186.

[35] H. Huang, J. Zhang, J. Zhang, J. Xu, and Q. Wu, "Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification," *IEEE Trans. Multimedia*, vol. 23, pp. 1666–1680, 2021.

[36] H. Huang, Z. Wu, W. Li, J. Huo, and Y. Gao, "Local descriptor-based multi-prototype network for few-shot learning," *Pattern Recognit.*, vol. 116, Aug. 2021, Art. no. 107935.

[37] X. Wang, X. Wang, B. Jiang, and B. Luo, "Few-shot learning meets transformer: Unified query-support transformers for few-shot classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7789–7802, Dec. 2023.

[38] Y. Zhao, H. Ding, H. Huang, and N.-M. Cheung, "A closer look at few-shot image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9130–9140.

[39] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[40] D. Kang, H. Kwon, J. Min, and M. Cho, "Relational embedding for few-shot classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8822–8833.

[41] M. H. Guo, C. Z. Lu, Z. N. Liu, M. M. Cheng, and S. M. Hu, "Visual attention network," *Comp. Vis. Media*, vol. 9, pp. 733–752, Jul. 2023.

[42] H. Li, Y. Nan, J. Del Ser, and G. Yang, "Large-kernel attention for 3D medical image segmentation," *Cognit. Comput.*, vol. 2023, pp. 1–15, Feb. 2023.

[43] Y. Fan, J. Liu, R. Yao, and X. Yuan, "COVID-19 detection from X-ray images using multi-kernel-size spatial-channel attention network," *Pattern Recognit.*, vol. 119, Nov. 2021, Art. no. 108055.

[44] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[45] H. Lai, T. Tran, and V. Pham, "Axial attention MLP-mixer: A new architecture for image segmentation," in *Proc. IEEE 9th Int. Conf. Commun. Electron. (ICCE)*, Jul. 2022, pp. 381–386.

[46] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, Sep. 2022.

[47] G. Han, J. Ma, S. Huang, L. Chen, and S.-F. Chang, "Few-shot object detection with fully cross-transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5321–5330.

[48] Z. Lu, S. He, X. Zhu, L. Zhang, Y.-Z. Song, and T. Xiang, "Simpler is better: Few-shot semantic segmentation with classifier weight transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8741–8750.

[49] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The Mahalanobis distance," *Chemometrics Intell. Lab. Syst.*, vol. 50, no. 1, pp. 1–18, 2000.

[50] H. Ghorbani, "Mahalanobis distance and its application for detecting multivariate outliers," *Facta Universitatis, Math. Informat.*, vol. 34, no. 3, pp. 583–595, Jun. 2019.

[51] S. Xiang, F. Nie, and C. Zhang, "Learning a Mahalanobis distance metric for data clustering and classification," *Pattern Recognit.*, vol. 41, no. 12, pp. 3600–3612, 2008.

[52] C. Zhang, Y. Cai, G. Lin, and C. Shen, "DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 12203–12213.

[53] S. Gidaris and N. Komodakis, "Generating classification weights with GNN denoising autoencoders for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 21–30.

[54] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10657–10665.

[55] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. 16th Eur. Signal Process. Conf.*, Aug. 2008, pp. 1–4.

[56] W. Wang, G. Zhang, L. Yang, V. S. Balaji, V. Elamaran, and N. Arunkumar, "Revisiting signal processing with spectrogram analysis on EEG, ECG and speech signals," *Future Gener. Comput. Syst.*, vol. 98, pp. 227–232, Sep. 2019.

[57] S. Ö. Arik, H. Jun, and G. Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 94–98, Jan. 2019.

[58] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[59] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[60] S. Deshmukh and A. Dubey, "Improved covariance matrix estimation with an application in portfolio optimization," *IEEE Signal Process. Lett.*, vol. 27, pp. 985–989, 2020.

[61] Y. Zhang, J. Tao, Z. Yin, and G. Wang, "Improved large covariance matrix estimation based on efficient convex combination and its application in portfolio optimization," *Mathematics*, vol. 10, no. 22, p. 4282, Nov. 2022.

[62] Z. Zhu, A. Thavaneswaran, A. Paseka, J. Frank, and R. Thulasiram, "Portfolio optimization using a novel data-driven EWMA covariance model with big data," in *Proc. IEEE 44th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Jul. 2020, pp. 1308–1313.

[63] B. Jiang, C. Liu, and C. Y. Tang, "Dynamic covariance matrix estimation and portfolio analysis with high-frequency data," *J. Financial Econometrics*, vol. 2023, Feb. 2023, Art. no. nbad003.

[64] H. Wu, W. Wang, Z. Xia, Y. Chen, Y. Liu, and J. Chen, "A discriminative multiple-manifold network for image set classification," *Appl. Intell.*, vol. 53, no. 21, pp. 25119–25134, Nov. 2023.

[65] X. Chen, G. Zhu, and J. Wei, "MMML: Multimanifold metric learning for few-shot remote-sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023, Art. no. 5618714.

[66] X. Chen, G. Zhu, M. Liu, and Z. Chen, "Few-shot remote sensing image scene classification based on multiscale covariance metric network (MCMNet)," *Neural Netw.*, vol. 163, pp. 132–145, Jun. 2023.

[67] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[68] D. Das and C. S. G. Lee, "A two-stage approach to few-shot learning for image recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 3336–3350, 2020.

[69] B. Oreshkin, P. Rodríguez López, and A. Lacoste, "TADAM: Task dependent adaptive metric for improved few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[70] Z. Yang, J. Wang, and Y. Zhu, "Few-shot classification with contrastive learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 293–309.

[71] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[72] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," *Mech. Syst. Signal Process.*, vols. 64–65, pp. 100–131, Dec. 2015.

[73] (2014). *Paderborn University Bearing Data Center*. Accessed: Jan. 2021. [Online]. Available: https://mb.uni-paderborn.de/kat/forschung/datacenter/bearing-datacenter

[74] C. Li, S. Li, H. Wang, F. Gu, and A. D. Ball, "Attention-based deep meta-transfer learning for few-shot fine-grained fault diagnosis," *Knowl.-Based Syst.*, vol. 264, Mar. 2023, Art. no. 110345.

[75] M.-H. Vu and V.-T. Pham, "MixerFormer-covariance metric neural network: A new few-shot learning model for bearing fault diagnosis," in *Proc. 12th Int. Conf. Control, Autom. Inf. Sci. (ICCAIS)*, Nov. 2023, pp. 639–644.

[76] J. Chang and Y. Chen, "Pyramid stereo matching network," in *Proc. CVPR*, Jun. 2018, pp. 5410–5418.

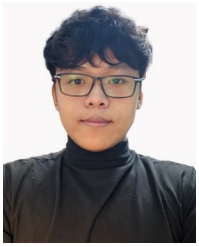
[77] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[78] Y. O. Lee, J. Jo, and J. Hwang, "Application of deep neural network and generative adversarial network to industrial maintenance: A case study of induction motor fault detection," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 3248–3253.

[79] B. Zhang, W. Li, J. Hao, X.-L. Li, and M. Zhang, "Adversarial adaptive 1-D convolutional neural networks for bearing fault diagnosis under varying working condition," 2018, *arXiv:1805.00778*.

[80] L. Yan, Z. Pu, Z. Yang, and C. Li, "Bearing fault diagnosis based on diffusion model and one-class support vector machine," in *Proc. Prognostics Health Manage. Conf. (PHM)*, May 2023, pp. 307–311.

[81] Z. Wu, H. Jiang, K. Zhao, and X. Li, "An adaptive deep transfer learning method for bearing fault diagnosis," *Measurement*, vol. 151, Feb. 2020, Art. no. 107227.
[82] X. Wang, C. Shen, M. Xia, D. Wang, J. Zhu, and Z. Zhu, "Multi-scale deep intra-class transfer learning for bearing fault diagnosis," *Rel. Eng. Syst. Saf.*, vol. 202, Oct. 2020, Art. no. 107050.
[83] P. Ma, H. Zhang, W. Fan, C. Wang, G. Wen, and X. Zhang, "A novel bearing fault diagnosis method based on 2D image representation and transfer learning-convolutional neural network," *Meas. Sci. Technol.*, vol. 30, no. 5, May 2019, Art. no. 055402.
[84] J. Shao, Z. Huang, and J. Zhu, "Transfer learning method based on adversarial domain adaption for bearing fault diagnosis," *IEEE Access*, vol. 8, pp. 119421–119430, 2020.
[85] H. Zhong, Y. Lv, R. Yuan, and D. Yang, "Bearing fault diagnosis using transfer learning and self-attention ensemble lightweight convolutional neural network," *Neurocomputing*, vol. 501, pp. 765–777, Aug. 2022.
[86] J. Lin et al., "Cross-domain fault diagnosis of bearing using improved semi-supervised meta-learning towards interference of out-of-distribution samples," *Knowl.-Based Syst.*, vol. 252, Sep. 2022, Art. no. 109493.
[87] C. Che, H. Wang, M. Xiong, and X. Ni, "Few-shot fault diagnosis of rolling bearing under variable working conditions based on ensemble meta-learning," *Digit. Signal Process.*, vol. 131, Nov. 2022, Art. no. 103777.
[88] D. Zhang, K. Zheng, Y. Bai, D. Yao, D. Yang, and S. Wang, "Few-shot bearing fault diagnosis based on meta-learning with discriminant space optimization," *Meas. Sci. Technol.*, vol. 33, no. 11, Nov. 2022, Art. no. 115024.
[89] J. Zhao et al., "Adaptive meta transfer learning with efficient self-attention for few-shot bearing fault diagnosis," *Neural Process. Lett.*, vol. 55, no. 2, pp. 949–968, Apr. 2023.
[90] L. Ma, B. Jiang, L. Xiao, and N. Lu, "Digital twin-assisted enhanced meta-transfer learning for rolling bearing fault diagnosis," *Mech. Syst. Signal Process.*, vol. 200, Oct. 2023, Art. no. 110490.



Manh-Hung Vu received the B.S. degree in automation and control from Hanoi University of Science and Technology, Hanoi, Vietnam, in 2023.

He is working with VinBigData, Hanoi, as an AI Engineer. His research interests include computer vision, deep learning, few-shot learning, and optical character recognition.

Van-Quang Nguyen received the B.S. degree in automation and control from Hanoi University of Science and Technology, Hanoi, Vietnam, in 2023.

He is working with VinBigData, Hanoi, as an AI Engineer. His research interests include computer vision, deep learning, few-shot learning, and speech recognition.

Thi-Thao Tran received the B.S. and M.S. degrees in electrical engineering from Hanoi University of Science and Technology, Hanoi, Vietnam, in 2003 and 2005, respectively, and the Ph.D. degree in electrical engineering from National Central University, Taoyuan, Taiwan, in 2016.

She is currently an Associate Professor with the Department of Automation Engineering, School of Electrical and Electronics, Hanoi University of Science and Technology. Her research interests include computer vision, image processing, deep learning, and signal processing.

Van-Truong Pham received the B.S. and M.S. degrees in electrical engineering from Hanoi University of Science and Technology, Hanoi, Vietnam, and the Ph.D. degree in electrical engineering from National Central University, Taoyuan, Taiwan, in 2013.

From 2013 to 2016, he was a Post-Doctoral Position with the National Central University. He is currently an Associate Professor with the Department of Automation Engineering, School of Electrical and Electronics, Hanoi University of Science and Technology. His research interests include computer vision, image processing, deep learning, signal processing, and applied control theory.

Men-Tzung Lo received the Ph.D. degree in communication engineering from National Taiwan University, Taoyuan, Taiwan, in 2004, with a focus on biomedical signal and image processing, as well as biomedical imaging and drug delivery systems.

He serves as a Professor with the Department of Biomedical Sciences and Engineering, National Central University, Taoyuan. Following this, he undertook post-doctoral training with Taipei Veterans General Hospital, Taipei, Taiwan, and Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. During this period, he applied innovative nonlinear signal analysis techniques to various biomedical signals from diverse disease groups, aiming to uncover their inherent properties and quantify deviations from normalcy as indicators of disease severity or prognosis. He has collaborated extensively with clinicians, resulting in numerous publications covering a broad spectrum of clinical topics, in addition to contributions to physics and engineering literature. Beyond research, he has played a pivotal role in developing miniaturized medical hardware tailored for diverse clinical applications. Leading an autonomous portable healthcare system group in the Qualcomm Tricorder X Prize international competition, he contributed significantly to the Taiwan team winning the second-place prize. Several of his technological innovations have been successfully commercialized, including the world's smallest 24-h Holter monitor designed for wrist-worn electrocardiography, which has received approval from regulatory bodies such as the FDA, European authorities, and Taiwan's health regulators. His research expertise spans specific areas of digital healthcare within biomedical engineering, including portable medical devices, the design and application of artificial intelligence algorithms, noninvasive cardiovascular assessment, and nonlinear complex signal and image analysis.