Applied Intelligence (2025) 55:484
https://doi.org/10.1007/s10489-025-06361-0

Mixmamba-fewshot: mamba and attention mixer-based method with few-shot learning for bearing fault diagnosis

Nhu-Linh Than[1] · Van Quang Nguyen[1] · Gia-Bao Truong[1] · Van-Truong Pham[1] · Thi-Thao Tran[1]

Accepted: 10 February 2025

Abstract
In recent years, artificial intelligence, particularly machine learning and deep learning has ushered in a new era of technological advancements leading to significant progress across various domains. In the field of computer vision, deep learning has made substantial contributions, impacting everything from daily life to production and industry. When machines, rotating devices, and engines operate, bearing failures are inevitable. Our task is to accurately detect or diagnose these failures. However, a key challenge lies in the lack of sufficient data on bearing faults to train a model capable of delivering highly accurate diagnostic results. To address this issue, in this paper, we propose a new approach named MixMamba-Fewshot, leveraging few-shot learning and using a feature extraction module that integrates an attention mechanism called the Priority Attention Mixer and Mamba - a novel theory that has recently gained considerable attention within the research community. Using Mamba for vision-based feature extraction in classification tasks, particularly in few-shot learning is an innovative approach, and it has shown promising results in improving the accuracy of bearing fault diagnosis. When we tested our model on the datasets provided by Case Western Reserve University (CWRU) and the Paderborn University (PU) Bearing Dataset, we compared it with previously published models. Our proposed approach demonstrated a significant improvement in diagnostic accuracy and clearly outperformed existing approaches. Our code will be available at: https://github.com/linhthan216/MixMamba-Fewshot.

Keywords Bearing fault diagnosis · Few-shot learning · Priority attention mixer · Mamba-based classification · Covariance metric

1 Introduction

Bearings are one of the most important inventions in the history of mechanical engineering, making significant contributions to the development of modern industries. In 1794, Philip Vaughan invented the first rolling bearing, using steel balls placed between the axle and the wheel to reduce friction. Through extensive research and development by scientists [1], bearings have been improved and mass-produced, particularly with the growth of the automotive and aerospace industries. Innovations in materials, manufacturing technologies, and design have made bearings an indispensable component of modern machinery systems [2–4]. Bearings are used to reduce friction between moving parts, which helps lower energy consumption and extend the lifespan of these components. This is crucial in industries where the efficiency and durability of machinery play a key role in maintaining productivity and minimizing maintenance costs. However, after prolonged operation in humid, dirty environments or under high or uneven loads, bearings can become prone to damage. Without regular maintenance procedures, such failures can pose risks to the entire machinery operation and the safety of maintenance engineers. Additionally, bearing failures can lead to significant economic losses for businesses, including repair and replacement costs, as well as equipment downtime. Therefore, monitoring the condition of bearings is essential to ensure stable and reliable system performance. Based on practical experiments, it can be concluded that bearing failures often generate abnormal noise signals, increased vibration, sudden rises in temperature, or a significant decline in machine performance. With the advancement of artificial intelligence, specifically deep learning and machine learning, traditional manual methods for diagnosing bearing failures are increasingly being replaced by advanced technological

✉ Thi-Thao Tran
thao.tranthi@hust.edu.vn

[1] Department of Automation Engineering, School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam

Published online: 27 February 2025

models based on the characteristic signals of faulty bearings [5–7].

Before the development of advanced deep learning models, numerous machine learning techniques were proposed for diagnosing bearing faults based on vibration signals. Kumar et al. [8] employed K-Nearest Neighbors (KNN) to classify various rolling element bearing (REB) faults by utilizing Continuous Wavelet Transform (CWT) on vibration signals. Zhang et al. [6] proposed a diagnostic method for bearing faults based on Principal Component Analysis (PCA) and Support Vector Machine (SVM) to analyze the characteristic features of rolling element vibration signals. Sawaqed et al. [9] implemented Artificial Neural Networks (ANN) with an optimized structure, using genetic algorithms to monitor bearing condition. However, these traditional models often rely heavily on manually engineered features, which require extensive domain expertise and result in time-consuming and costly processes.

Deep learning models have emerged to overcome these limitations, leveraging the powerful capabilities of multilayer neural networks to automatically learn and extract features from data without manual intervention. Zhang et al. [10] employed a method that converts raw signals into two-dimensional images and then applied Convolutional Neural Networks (CNNs) to automatically extract features and diagnose faults, demonstrating effectiveness and meeting the requirements for timely fault diagnosis. Additionally, Chen et al. [11] proposed combining CNNs to extract frequency features from raw data with Long Short-Term Memory (LSTM) networks to identify fault types based on learned characteristics, achieving high accuracy and robust performance even in noisy environments. In 2023, Yang et al. [12] utilized the Transformer as an attention mechanism applied to segmented sub-sequences for feature extraction to determine the fault type of bearings and achieved a high average accuracy.

Nevertheless, a new challenge arises as these standard deep learning models tend to underperform when working with limited data. To address all the issues, one of potential approaches is few-shot learning, which is capable of delivering high performance even with a small amount of data in the task of bearing fault diagnosis. This approach is built upon the foundation of the Siamese Neural Network (SNN) [13], which operates by identifying similarities and differences between the feature vectors of input samples. Unlike traditional neural networks, SNNs require only a few samples to achieve highly accurate predictions. In further improving this model, several research papers have been published, including the Wide-Kernel Deep Convolutional Neural Network (WDCNN) with wide first-layer kernels [6] replaces Convolutional Neural Network [14] and MixerFormer [15] with a Conv-Mixer structure to enhance feature extraction from spectrogram images. Additionally, modern few-shot learning models such as Prototype Network [16] and Matching Net [17] have emerged, offering improved accuracy but still requiring multiple stages to reach the diagnosis phase. From another perspective, Euclidean distance and Cosine distance are two well-known measures for evaluating correlations between samples. However, they are prone to noise interference with little data due to the simplicity of the calculation. Covariance Metric Networks (CovaMNET) [18] was developed to address these limitations by employing the covariance metric to assess the correlation across all images in both the support set and query set. This represents a significant advancement. Building upon the CovaMNET backbone, we propose enhancing the CNN block with a component that can perform feature extraction at a highly advanced level, significantly boosting diagnostic accuracy. By integrating a novel concept from the visual domain, we introduced the Visual State Space Model into the feature extraction block, enabling it to capture exceptionally strong global information, thereby greatly improving model performance. Furthermore, the combination with a Priority Attention Mixer has resulted in our model outperforming previous approaches by a considerable margin.

Although traditional deep learning models have achieved impressive results, they typically require a large amount of data for effective training. In practice, collecting enough data for bearing fault is not always feasible, leading to reduced performance of these models when faced with limited data. To address this issue, we have adopted the few-shot learning (FSL) method, which can deliver high performance even with a small amount of data, specifically in the task of bearing fault diagnosis. One of the popular networks utilizing FSL is the Siamese Neural Network (SNN) [13] which operates by identifying similarities and differences between feature vectors of input samples. Unlike traditional neural networks, SNN requires only a small number of training samples to achieve high prediction accuracy. Based on this foundation, Vu et al. [si] proposed using SNN in combination with Conv-Mixer to extract deeper information from initial signals. To further improve the model, some studies have introduced new models, such as Deep Convolutional Neural Networks with Wide First-layer Kernels (WDCNN) [6] proposed by Zhang et al., which utilize wide kernels in the first layer to replace CNN. Zheng et al. in [19] enhanced traditional relation networks used in few-shot learning by utilizing advanced techniques to better measure the similarity between data points. Additionally, modern FSL models like Prototype Network [16] and Matching Network [17] have been developed, providing higher accuracy but requiring multiple complex stages to achieve an accurate diagnosis. In traditional neural networks, Euclidean distance and Cosine distance are often used to calculate the differences between data. However, when working with limited and highly correlated datasets, these two methods do not accurately reflect the degree of similarity due to their simplicity. To overcome

this limitation, Li et al. developed the Covariance Metric Network (CovaMNET) [18], which utilizes the covariance metric to assess correlations between data points and the mean distribution across both the support and query sets. This approach represents a significant advancement in enhancing performance. Building upon the CovaMNET framework, we propose enhancing the CNN block by incorporating a component capable of advanced feature extraction, significantly improving diagnosis accuracy. Our approach is inspired from the Mamba architecture recently introduced by Gu and Dao [20]. Mamba is recognized as a new attention mechanism that takes advantage of long-range dependency capture through the selective scanning mechanism of State Space Models (SSMs). Building on the success of Mamba in language modeling, researchers have expanded its application to computer vision tasks involving images and videos. The effectiveness of Mamba in this domain was demonstrated through two initial studies, Vision Mamba [21] and VMamba [22], which introduced the ViM block and the VSS block, respectively, laying the foundation for subsequent research. Building on these achievements, we developed an end-to-end model that integrates Mamba into the feature extraction module, combining with Priority Channel Attention (PCA) and Priority Spatial Attention (PSA) mechanisms, to enable the model to better learn the diverse features of input images. Under conditions of limited data, the features of each image become highly valuable, making it necessary to extract as much information as possible for accurate classification. Our proposed model demonstrates very promising performance even with a small amount of training data.

The key contributions of our paper are as follows:

- We propose a new approach called MixMamba-Fewshot, leveraging few-shot learning and integrating the Attention Mixer and Mamba for bearing fault diagnosis.
- We have developed a new feature extraction model called Priority Attention Mixer Mamba. This model combines the strengths of Mamba with attention mechanisms across both spatial and channel dimensions, enhancing its ability to extract robust and diverse features. It is particularly effective in tasks with limited data, significantly improving performance and accuracy.
- We propose a new end-to-end few-shot learning model based on the structure of the Covariance Metric Neural Network. Unlike methods that only learn independent features, this network is designed to learn and exploit the relationships between different features. This enables it to evaluate and capture the correlations and dependencies between features, thereby enhancing performance in scenarios with limited data.
- To demonstrate the effectiveness of the proposed model, we conducted experiments on two datasets: the Case Western Reserve University (CWRU) Bearing Dataset and the Paderborn University (PU) Bearing Dataset. The results show that the model achieves state-of-the-art performance in bearing fault diagnosis, particularly in scenarios with limited data.

2 Related work

2.1 Conv-mixer

Attention is a crucial mechanism in neural networks that enables them to focus on important features, thereby improving performance. In the field of computer vision, recent methodologies have primarily centered on extracting information along both spatial and channel dimensions. Trockman and Kolter introduced the Conv-Mixer model in [23], which leverages the combination of Depth-wise and Point-wise convolutions. This approach allows Conv-Mixer to effectively focus on important regions and pixels within each channel of the input features. With its simple yet highly effective architecture, Conv-Mixer significantly reduces computational costs compared to traditional attention models while maintaining high performance. This has garnered considerable attention from the research community, demonstrating Conv-Mixer's potential to compete with more complex attention models.

2.2 Visual state space model

Mamba, also known as the Structured State Space Models with Selective Scan (S6) inspired by the State Space Model (SSM) in control theory, has continuously drawn interest due to its superiority in speed, memory efficiency, and accuracy when applied to sequence data. However, when dealing with image data, where the input is not in sequence form, Mamba faces a significant challenge. Shortly after the release of the first paper "Vision Mamba" [21] applying Mamba to image data, the Visual State Space Model was introduced, promising to become a backbone for many future image processing tasks. The use of the 2D Selective Scan Module (SS2D), an enhancement for S6, allows image inputs to be divided into patches that are scanned in four directions to generate sequences, which are then processed by the S6 block before being reassembled. The implementation of the VSS Block with SS2D has given the model linear complexity while enabling superior global information aggregation, resulting in parameters that are fully dependent on the input.

3 Preliminaries

3.1 State space models (SSMs)

The SSMs are classic tools in control theory, used to describe, analyze, and design multivariate systems. It uses first-order differential equations (ODEs), to represent relationship between input sequence $x(t) \in \mathbb{R}^{D\times 1}$ and output $y(t) \in \mathbb{R}^{D\times 1}$ to be described through a hidden state space $h(t) \in \mathbb{R}^{N\times 1}$ $(N > D)$:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times D}$, and $\mathbf{C} \in \mathbb{R}^{D\times N}$ are the state matrix, input matrix, and output matrix, respectively. However, the disadvantage of this formula is represented in function-to-function form, meaning that depending on the time 't' cannot be learned by a deep learning model. To enable implementation within deep learning frameworks, the continuous-time formula is discretized using a zero-order hold (ZOH) approach with a defined time step $\Delta \in \mathbb{R} > 0$:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}), \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}. \end{aligned} \tag{2}$$

After discretizing $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$, SSMs are computed as follows:

$$\begin{aligned} h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \\ y_t &= \mathbf{C}h_t. \end{aligned} \tag{3}$$

If set $h_{-1} = 0$ correspond to the first input with no hidden state, the output is computed through a global convolution:

$$\begin{aligned} \overline{\mathbf{K}} &= (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}), \\ \mathbf{y} &= \mathbf{x} * \overline{\mathbf{K}}, \end{aligned} \tag{4}$$

where $\overline{\mathbf{K}} \in \mathbb{R}^{L}$ is a convolutional kernel, L is the length of the input sequences, it is made by multiplying the matrices $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$ and $\mathbf{C}$ is independent of the input. Thus, it can be computed in parallel.

3.2 2D selective scan (SS2D)

SS2D is an enhancement of Mamba, first introduced by Liu et al. in [22] to accommodate image data where the input is 2D rather than a 1D sequence like text. The operational principle of SS2D involves using Four-Directional scanning to compute the SSM, also known as Cross-Scan. The input image is divided into patches and scanned in four directions, generating four separate sequences before feeding them into the S6 block (also known as Mamba) for processing. The outputs are then combined to return a 2D feature map to be expanded in the next block:

$$\overline{z} = SS2D(z) = \sum_{i=1}^{4} S6(scan(z, i)), \tag{5}$$

where z is the input feature, $\overline{z}$ is the output feature of the SS2D block, $scan(x, i)$ converts the patches of the input image into sequences based on the direction i. SS2D is the backbone network learning with linear complexity and achieving global receptive fields.

4 Methodology

To mitigate the impact of noise and optimize the extraction of bearing fault characteristics such as timbre, pitch, etc., the input signal is converted into a spectrogram. This process begins by applying a sliding window with a fixed size M, ensuring that M contains at least one complete cycle of the data. The signal is then passed through the Short-Time Fourier Transform (STFT) to convert it from a 1D time-domain representation to a 2D spectral-domain representation. The result is a spectrogram image that fully preserves the frequency and amplitude characteristics [24] Fig. 1.

4.1 The proposed MixMamba-Fewshot model

Figure 2 illustrates the architecture of the proposed MixMamba-Fewshot framework. The framework is designed for few-shot learning in bearing fault diagnosis and consists of two main components: the Priority Attention Mixer Mamba (PAM-Mamba) module and the Covariance Metric Layer. The PAM-Mamba module extracts features from both the support set and the query image, which are processed through multiple stages of Patch Embedding, Path Merging, and PCA-PSA-VSS blocks. Shared weights ensure consistency in feature extraction between the support set and query image. The extracted feature representations are passed into the Covariance Metric Layer, which computes second-order relationships between query and support set features. The layer calculates the local covariance representations for each category and derives a similarity score for classification. This structured approach allows for robust fault diagnosis, even with limited training data.

The proposed Priority Attention Mixer Mamba Module is inspired by the architecture of MedMamba [25]. Firstly, the input image with size $H \times W \times C$ (Height $\times$ Width $\times$ Channel) is passed through the Patch Embedding layer.

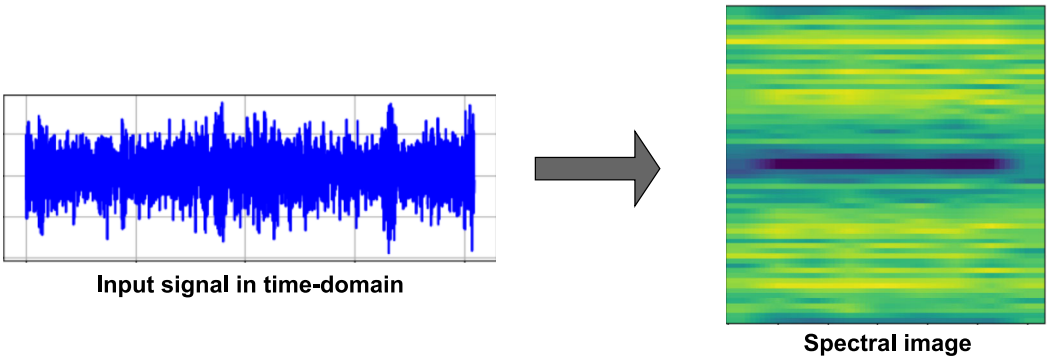


Fig. 1 Signals in the time-domain are transformed into spectral form

This layer utilizes 3×3 convolutions with a stride of 1 to generate overlapping patches, transforming the image into a set of embedding vectors suitable for the training processing. The feature maps obtained after the embedding layer retain their original length and width, while the number of channels increases to 8 ($H \times W \times 8$) aligning with the subsequent training process. The use of overlapping patches helps capture continuous and more detailed feature information in the image. Next, the embedded image is fed into PCA-PSA-VSS block for feature extraction and complex representations learning. After each PCA-PSA-VSS block, Path Merging layers are applied to gradually reduce the height and width of the input feature maps by half, while doubling the number of channels. This step aims to reduce computational complexity in later stages while enhancing the model's ability to capture more distinct features. The model employs [2, 3, 2] PCA-PSA-VSS blocks across three stages, with the number of channels progressively increasing in the order [8, 16, 64] throughout these stages. The final feature block obtained has a size of $\frac{H}{4} \times \frac{W}{4} \times 64$. This final feature block is then passed into the classification module, which uses the Covariance metric method to accurately classify the image label.

4.2 PCA-PSA-VSS block

The key factor contributing to the success of our model is the proposed PCA-PSA-VSS block as illustrated in Fig. 3. The PCA-PSA-VSS consists of two parallel branches: the PCA-PSA branch developed based on the Attention Mixer, and the VSS branch inspired by the SS-Conv-SSM block in the MedMamba [25]. This architecture is designed to capture the complex spatio-temporal features in bearing fault signals.

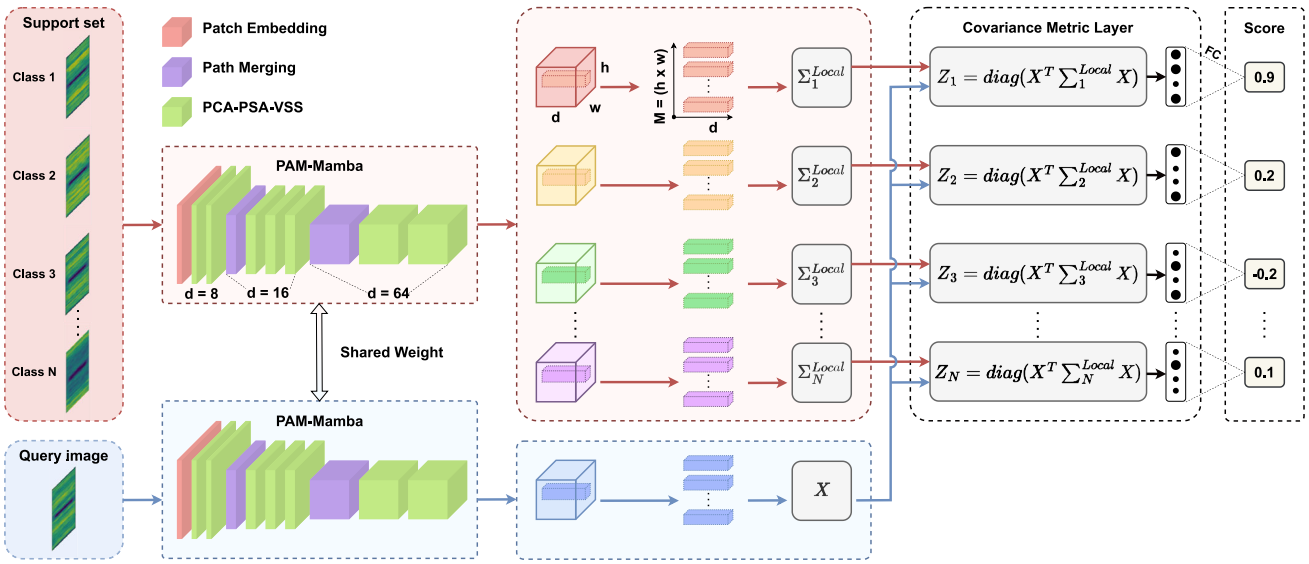


Fig. 2 The architecture of the proposed MixMamba-Fewshot. The PAM-Mamba module extracts features from the support set and query image. The extracted feature representations are passed into the Covariance Metric Layer, which calculates local covariance representations and derives classification scores

The input $X \in \mathbb{R}^{B\times H\times W\times C}$ representing a batch of feature maps, where B is the batch size, H and W are the spatial dimensions, and C is the number of channels is split into two parts along the channel dimension. In our model, C represents the total number of channels in the input feature map (e.g., 64 channels in our specific implementation). These channels are evenly split into two groups for processing in parallel branches: the PCA-PSA branch and the VSS branch. Each branch processes $C/2$ channels independently, extracting complementary features. The choice of two branches (rather than more) strikes a balance between computational efficiency and feature diversity. Splitting the channels helps reduce the computational complexity within each branch while maintaining sufficient representational capacity to capture complex features. After processing, the feature maps from both branches are concatenated and passed through a shuffle layer. The propose of the shuffle layer is to enhance the model's generalization ability by reducing its dependence on a fixed feature ordering. This prevents the model from being constrained by specific feature groups, allowing it to learn more generalized and flexible relationship between the features in the data.

In the PCA-PSA branch, we build upon the Attention ConvMixer architecture proposed by Le et al. in [26]. The main idea is to replace the Depthwise Convolution (DW-Conv) and Pointwise Convolution (PW-Conv) layers in ConvMixer with PCA and PSA layers, aiming to enhance the model's ability to capture data variations across both spatial and channel dimensions. First, in the PCA block, the input feature map is passed through a DW-Conv layer. In addition to reducing computations compared to regular convolution, DW-Conv allows the model to learn channel features specific to each channel. Next, the mean value along the channel dimension is computed from both the original feature map C and the feature map after DW-Conv layer C'. The difference between C' and C $(C' - C)$ is then processed through softmax and sigmoid operations to compute and normalize the channel attention, ensuring that important features are amplified. In the PSA layer, the feature map is passed through a PW-Conv layer to extract individual pixel features. Then, the mean value along the spatial dimension is computed from both the input feature map S and the feature map after PW-Conv layer S'. Similar calculations as in the PCA block are applied to enhance spatial attention and normalize it using the difference $(S' - S)$. Through these two blocks, the model is able to focus more effectively on channels and spatial regions with significant features, thereby strengthening its ability to extract robust characteristics from the input data.

In the VSS branch, the input is first normalized using a LayerNorm layer to stabilize the training process of the subsequent layers. Next, the input is divided into two parallel branches. The first branch (Upper Branch) includes a linear layer followed by a SiLU activation layer. Meanwhile, in the second branch (Lower Branch), after passing through a linear layer, the feature map proceeds through a DW-Conv layer to extract features. The selective scanning mechanism of S6 ensures that the model focuses on regions containing important information and ignore less relevant areas. Once completed, the sequences are merged back into a 2D feature map with the same size as the input. The output from the SS2D layer is then normalized again via a Layernorm layer. After that, the results from the Upper Branch and Lower Branch are element-wise multiplied. Finally, a linear layer is used to combine the features, completing the processing in the VSS branch.

4.3 Covariance metric layer

The Covariance Metric Layer in the proposed MixMamba-Fewshot framework is essential for evaluating the relationships between features extracted from the support set and query image. This layer builds upon the Covariance Metric Network (CovaMNET) methodology, which was introduced by Li et al. [18] as a robust approach for classification under limited data conditions. Unlike traditional methods relying on first-order statistics, such as Euclidean or cosine similarity, CovaMNET leverages covariance matrices to exploit second-order statistics. This enables a deeper understanding of the data distribution structure, reducing noise and improving stability and reliability, especially in challenging environments.

In real-world classification tasks, data scarcity is a common challenge, making it difficult for models to achieve high performance using traditional methods. To address this, Few-Shot Learning has emerged, enabling models to learn effectively with only a few samples. Inspired by CovaMNET, the Covariance Metric Layer in our framework computes second-order relationships between query and support set features, allowing for robust fault diagnosis in bearing applications.

In **N**-way **K**-shot tasks, during each training episode, the support set is formed by randomly selecting **K** labeled samples from each class. With **N** classes, a total of **N** x **K** samples are selected from the training set. The query set is formed by randomly selecting **Q** unlabeled samples from the **N** classes. The objective is to learn an optimal mapping function to predict the labels of the samples in the query set, determining which of the **N** categories they belong to. The local covariance representation of the *j-th* category $\Sigma_j^{\text{local}} \in \mathbb{R}^{d\times d}$, where d represents the dimensionality of each descriptor, is calculated as follows:

$$\Sigma_j^{\text{local}} = \frac{1}{MK-1}\sum_{i=1}^{K}(\mathbf{X}_i - \mu)(\mathbf{X}_i - \mu)^\top, \tag{6}$$

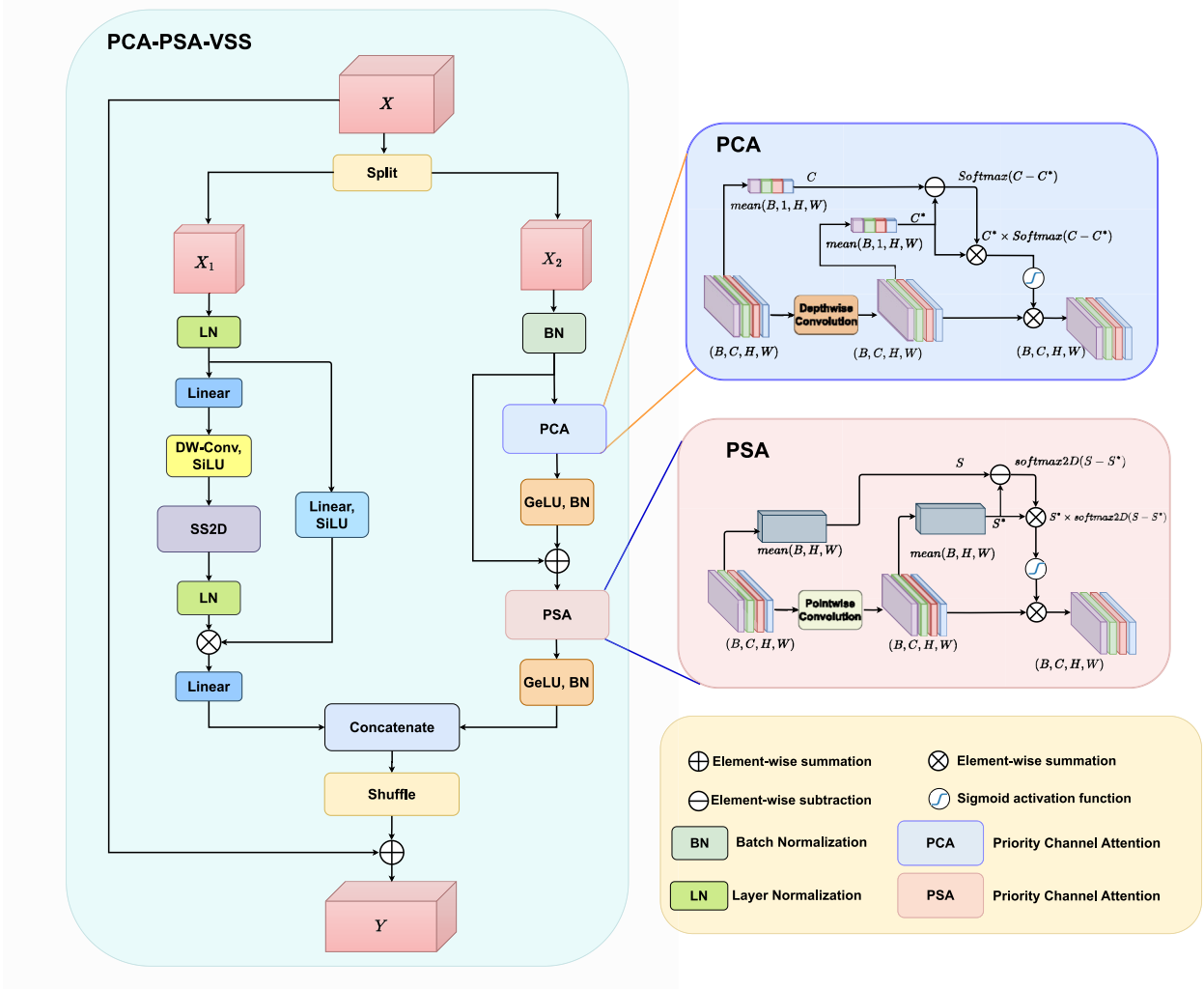


Fig. 3 The detailed structure of the proposed PCA-PSA-VSS block

where $\mathbf{X}_i \in \mathbb{R}^{d \times M}$ is the input feature vector, M is the number of local deep descriptors, K denotes the total number of samples in the *j-th* class, and μ is the mean vector matrix.

The covariance metric measures the correlation between the query image and the support image set. Its measure function is defined as:

$$f(\mathbf{x}, \Sigma) = \mathbf{x}^\top \Sigma \mathbf{x}, \tag{7}$$

where $\mathbf{x} \in \mathbb{R}^d$ is a sample with zero-mean over the output of the query image. When the value of (7) reaches its maximum, the feature vector of the query image aligns closely with the principal components corresponding to the largest eigenvalues of the covariance matrix Σ. It is important to note that the term "maximum" in (7) refers to identifying the category j for which $f(\mathbf{x}, \Sigma_j^{\text{local}})$ is the highest among all categories in the support set. This ensures that the query image is assigned to the category most similar to its features.

To evaluate the local covariance metric similarity between the query image $\mathbf{X}$ and one category Σ_j^{local}, the following formula is used:

$$\mathbf{z} = \text{diag}\, f(\mathbf{X}, \Sigma_j^{\text{local}}) = \text{diag}(\mathbf{X}^\top \Sigma_j^{\text{local}} \mathbf{X}), \tag{8}$$

where $\mathbf{z} \in \mathbb{R}^K$ contains K local similarities between the query image and one category. The global similarity score $\mathbf{Z}$ is derived by aggregating $\mathbf{z}$ using a fully connected layer and a 1D-convolution layer. These layers ensure that all local similarities contribute effectively to the final classification decision. Finally, a softmax layer normalizes the global similarity scores across all categories, assigning the query image to the most likely fault type in the support set.

5 Experiment

5.1 Datasets

To ensure objectivity and accurately assess the effectiveness of the proposed model, we selected two of the leading benchmark datasets in the field of bearing fault diagnosis: the Case Western Reserve University (CWRU) [27] and Paderborn University (PU) [28] Bearing datasets. Both datasets provide diverse information on faulty bearings at various positions under different load conditions and rotational speeds. With their high quality and comprehensive nature, they meet the research requirements and have been widely recognized and utilized in numerous studies [6, 15, 29–31].

Algorithm 1 Training algorithm for our MixMamba-Fewshot approach

Require: Support Set $\mathcal{S} = \{(X_i^s, Y_i^s)\}_{i=1}^{n_s}$, **Query Set** $\mathcal{Q} = \{(X_j^q, Y_j^q)\}_{j=1}^{n_q}$,
The Base Model f_θ is initialized with random weights
Step size hyper parameters γ
Number of epochs $\boldsymbol{E}$
Number of iterations $\boldsymbol{I}$
Number of categories $\boldsymbol{C}$
Ensure: Optimized weight parameter θ
Categories Probabilities $p(Y^q|X^q, \mathcal{S})$ for each query sample X^q in $\mathcal{Q}$
Initialize the model's weights θ with random values.
for each epoch $\boldsymbol{e}$ from 1 to $\boldsymbol{E}$ **do**
  for each iteration $\boldsymbol{i}$ from 1 to $\boldsymbol{I}$ **do**
    Create support set $\mathcal{S}$ and query set $\mathcal{Q}$ from the training dataset
    for each category $\boldsymbol{c}$ from 1 to $\boldsymbol{C}$ **do**
      /* Feature extraction */
    Using PAM-Mamba module to extract feature maps $\hat{\mathcal{F}}_s$ and $\hat{\mathcal{F}}_q$
      /*Covariance Metric Layer*/
    Calculate Σ_j^{local} using (6)
    end for
    Compute correlation vector $f(\hat{\mathcal{F}}_q, \Sigma_c^{\text{local}})$ using (7)
    Evaluate the local covariance metric similarity between the query image and ***c-th*** category using (8)
    /* Gradient Descent & Update weights*/
    Compute contrastive loss $L_{cl,S_c}(\hat{y}^q, y^q)$
    Update $\theta_i' = \theta - \gamma \nabla_{\theta_i} L_{\text{CL},S_c}(f_\theta)$
    $\theta \leftarrow \theta - \gamma \nabla_\theta \sum_{(X,Y)\in\mathcal{Q}} \mathcal{L}_{CL}(P(X), Y)$
  end for
  → *Save the best weights*
end for

5.1.1 CWRU bearing dataset

The CWRU Bearing Dataset [32] is a benchmark dataset developed in a controlled laboratory environment to collect and study common bearing defects, including single-point defects at the drive-end (DE) and fan-end (FE). The experiments were conducted on a 2-HorsePower Reliance Electric motor, focusing on defects in the outer race, inner race, and rolling element, with defect sizes of 0.007 inch, 0.014 inch, and 0.021 inch. Vibration data was collected under various load conditions (0, 1, 2, and 3 HP) and different rotational speeds (1730, 1750, 1772 and 1797 rpm). The signals were recorded at two sampling frequencies (12 kHz and 48 kHz) and processed in MATLAB, with all data stored in MATLAB (.mat) format [32]. To extract samples for training, we applied a sliding window of 2048 points with an 80-point step size. The dataset includes 10 category labels, of which 9 are defect labels and 1 is normal. Detailed information about the defect locations and label naming is presented in Table 1, and the label distribution of the CWRU dataset is shown in Fig. 4a, highlighting the proportion of each class and their corresponding sample counts. To evaluate the effectiveness of the proposed model, we tested with various training sample sizes: 60, 90, 300, 600, and 19,800 samples. For testing, we used 750 images (with 75 images per category), randomly selected from the test set. Given the complexity and diversity of the CWRU Bearing Dataset, applying signal preprocessing techniques is crucial for better feature extraction and improved classification performance.

Table 1 Overview of the CWRU Dataset Used in Experiment

Fault Position	Fault Size (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

5.1.2 PU Bearing Dataset

The PU Bearing Dataset [28] provided by the KAT Data Center at Paderborn University contains a range of faults from simple to complex, including both artificial and real faults in industrial environments. This dataset includes 32 type 6203 bearings, of which 6 are in a normal state, 12 have artificial faults, and 14 have faults resulting from accelerated life testing [32]. The faults are divided into three main categories: outer race faults, inner race faults, and rolling element faults, with severity levels classified into three ranges: below 2 mm, from 2-4.5mm, and from 4.5-13.5mm [30]. Data were collected under four different operating conditions, combining rotational speed, torque, and axial force. To assess the applicability of the proposed model to more complex fault types

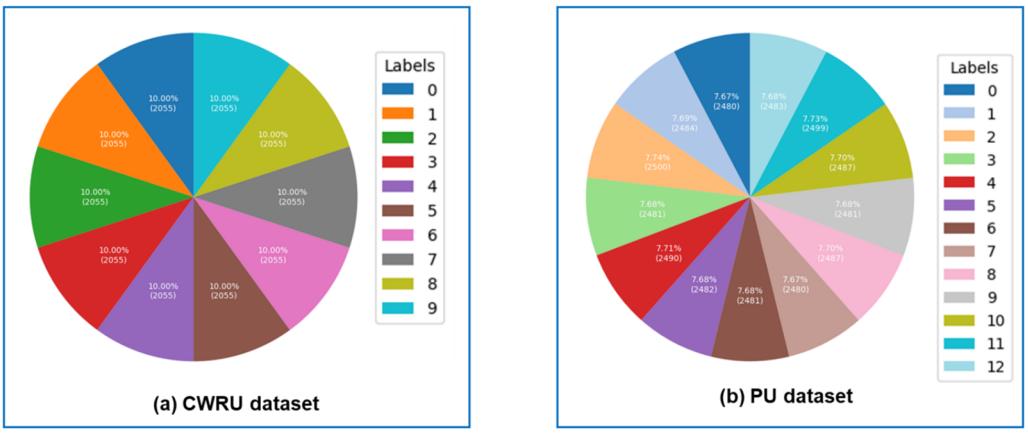


Fig. 4 Data visualization and label distributions of (a) the CWRU dataset, and (b) the PU dataset

and the impact of environmental factors, we selected 13 representative categories out of the 32 available, including 1 normal bearing, 6 with artificial faults, and 6 with real faults. Detailed information about the faults is presented in Table 2, and data visualization of the PU data is given in Fig. 4b. The sampling method applied to the PU Bearing Dataset is similar to that used for the CWRU Bearing Dataset. We conducted training with varying sample sizes: 195, 260, 650, 1300, and 25,844 samples. For testing, we used 1950 images, with each class containing 150 randomly selected images from the test set. Due to the greater complexity of the PU Bearing Dataset, our preprocessing involved additional steps compared to the CWRU Bearing Dataset, to ensure the data were better suited for training and more accurate fault diagnosis.

5.2 Implementation details

In this study, we aimed to accurately classify a query image into its correct category based on a support set with limited training samples per category. To achieve this, we used the Cross-Entropy Loss function to quantify the difference between the model's predicted probability distribution and the true label of the query image. For optimization, we applied the Adam algorithm with an initial learning rate of 0.001, which was reduced by a factor of 0.1 every 10 epochs. The model was trained on an NVIDIA Tesla T4 16GB GPU, demonstrating strong performance after approximately 20-30 epochs. On average, each epoch took around 8 min when using the CWRU Bearing Dataset and about 16 min with the PU Bearing Dataset. The training process of our proposed approach is detailed in Algorithm 1. To evaluate the classification performance of the proposed methods, we compared them with other state-of-the-art few-shot learning models, primarily designed for bearing fault diagnosis. The evaluation metrics include accuracy and confusion matrices. Additionally, for statistical significance testing, we used the paired t-test to compute the p-value between the results of the proposed method and the compared models.

Table 2 Overview of the PU Dataset Used in the Experiment

Fault Position	Fault Cause	Severity	Fault 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	Real	KA07	4
Outer Race	Plastic deform	1	Real	KA15	5
Outer Race	Pitting	2	Real	KA16	6
Inner Race	Electrical discharge machining	2	Artificial	K101	7
Inner Race	Electric engraver	1	Artificial	K103	8
Inner Race	Pitting	1	Real	K104	9
Inner Race	Electric engraver	2	Artificial	K107	10
Inner Race	Pitting	3	Real	K116	11
Inner Race	Pitting	2	Real	K118	12

Table 3 Comparison of accuracy (in %) of the proposed model with other models under different data conditions in the CWRU Bearing Dataset

Few-shot models	60 samples		90 samples		300 samples		19800 samples	
	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
Cosine Classifier [33]	85.87	88.67	88.13	90.13	92.67	94.71	99.73	99.73
Cross Attention Network [35]	81.13	82.44	92.28	91.64	93.14	94.14	99.29	99.87
RelationNet [36]	82.23	83.41	83.14	86.74	88.05	89.75	98.12	99.12
MatchingNet [34]	86.27	90.53	93.23	89.77	95.23	97.13	99.86	99.86
Siamese-WDCNN [6]	82.88	86.12	91.84	91.84	95.22	96.48	99.13	99.53
Siamese-Convmixer [37]	83.17	87.14	95.48	93.64	96.14	95.70	99.84	99.07
CovaMNET [18]	92.13	93.44	96.51	96.59	98.91	99.01	99.62	99.67
SA-CovaMNet [38]	92.56	94.23	95.14	98.72	98.62	98.72	99.69	99.81
MF-Net [15]	94.13	95.61	97.51	98.78	99.51	99.56	99.63	99.73
QS-Former [39]	92.14	94.11	96.13	96.25	98.21	98.14	99.63	99.53
Ensembling Transformer [30]	96.82	98.23	98.17	98.98	99.61	99.14	99.89	99.76
Proposed MixMamba-Fewshot	**99.33**	**99.53**	**99.52**	**99.68**	**99.77**	**99.83**	**99.90**	**99.93**

The bold values in the tables indicate the best performance for each evaluation metric

The paired t-test is performed on the predicted probabilities for the corresponding labels obtained from the softmax outputs of the proposed and compared models. A significance level of $\alpha = 0.05$ was used to indicate a statistically significant difference between the proposed method and the compared models.

5.3 Results on the CWRU bearing dataset

In this section, we compared the classification performance of the proposed model with other advanced methods on the CWRU Bearing Dataset. All models were re-implemented based on the open-source code provided by the authors to

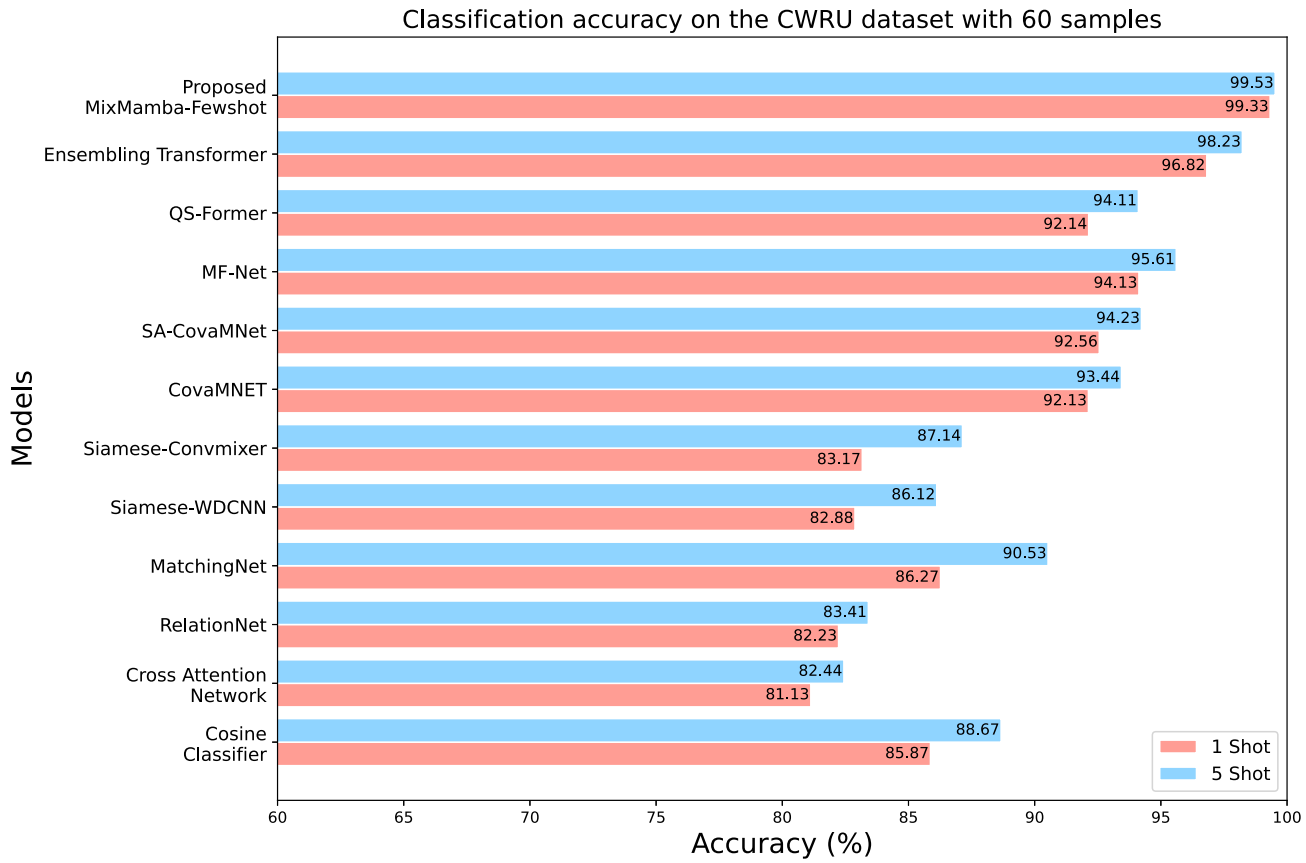


Fig. 5 Bar chart illustrating the classification accuracy on the CWRU dataset with 60 training samples

ensure transparency and fairness. The experiments were conducted in two scenarios: a 10-way 1-shot task and a 10-way 5-shot task, with varying amounts of training data, ranging from very limited (only 60 training samples) to more abundant data with 19,800 samples. The results presented in Table 3 show that the proposed model outperforms current state-of-the-art few-shot methods. In the 10-way 1-shot task with only 60 training samples, few-shot models using cosine distance, such as Cosine Classifier [33] and MatchingNet [34], or models using cross-attention mechanisms like the Cross-Attention Network [35]. As shown in Table 3, along with models leveraging non-linear relationships like the Relation Network [36] and models based on Siamese networks such as Siamese WDCNN [6] and Siamese Convmixer [37] achieved reasonably good performance with accuracy over 80%. MatchingNet, in particular, achieved the highest accuracy in this group with 86.27%. As shown in Fig. 5, the bar chart highlights the superior classification accuracy of the proposed approach compared to other models in both the 1-shot and 5-shot cases, even with as few as 60 training samples.

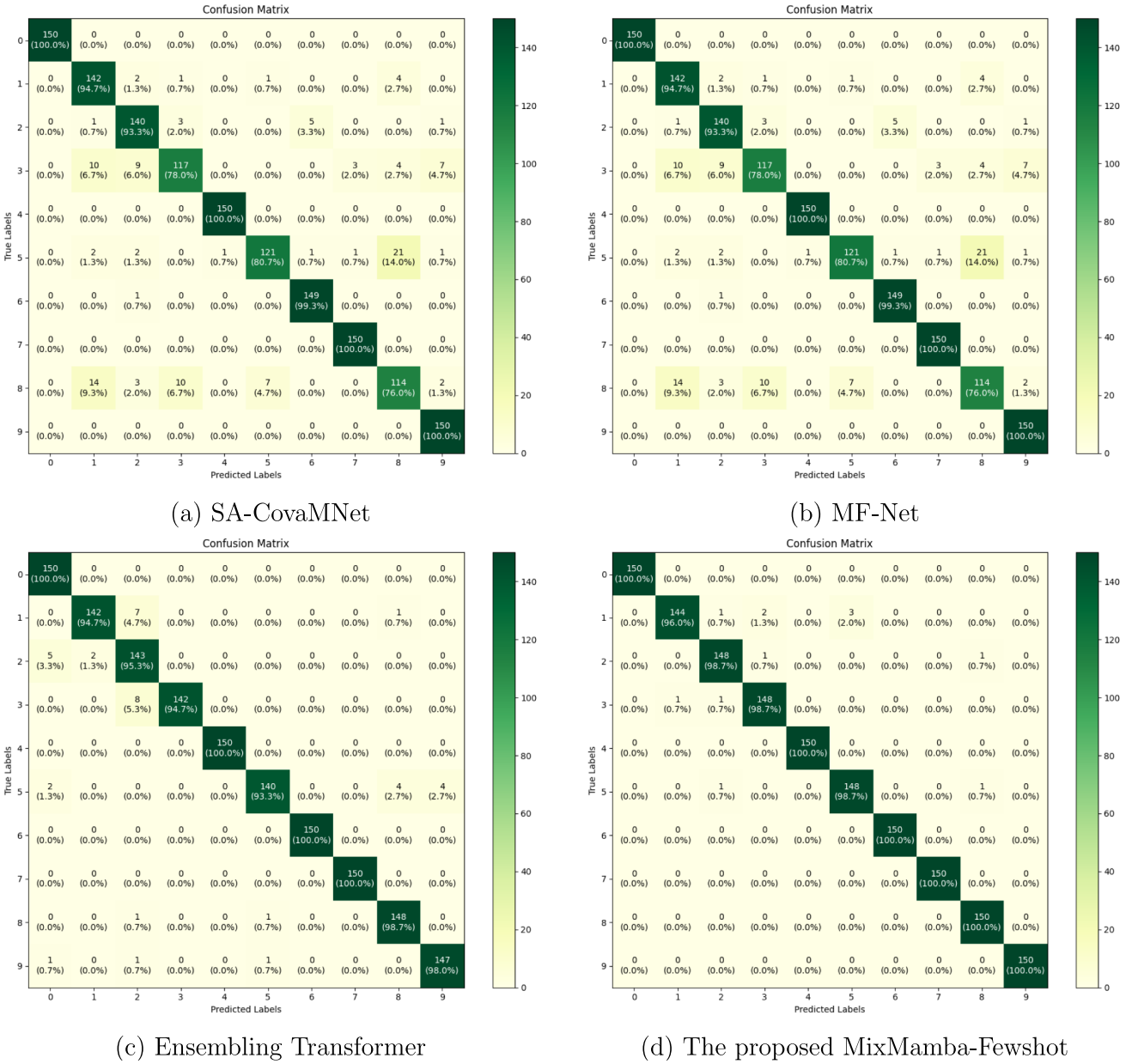
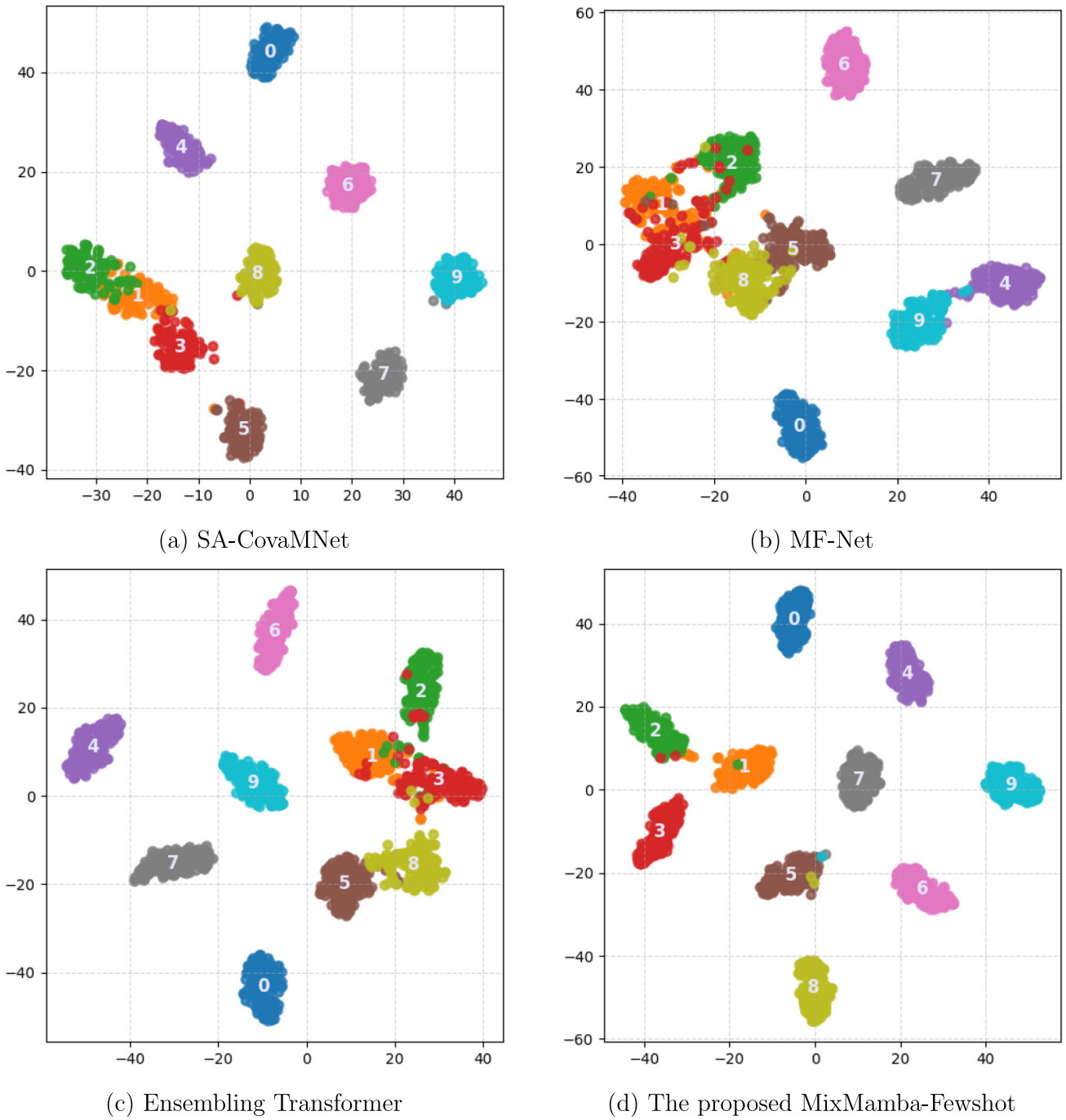


Fig. 6 Visualization of the confusion matrix for the top-performing methods in Table 3 in the 10-way 1-shot scenario, using 60 training samples from the CWRU Bearing Dataset

Significantly, models like CovaMNet [18] and its variants, such as SA-CovaMNet [38] and MF-Net [15], leveraged the ability to capture relationships between features through covariance matrices, attaining accuracy exceeding 90%, with MF-Net [15] achieving the highest at 94.13%. The application of the self-attention mechanism from Transformers in the few-shot model QS-Former [39] was also a breakthrough, yielding impressive results with 92.14% accuracy. Recognizing the strengths of these approaches, the proposed Ensembling Transformer model [30] skillfully combines the self-attention mechanism of the Transformer with the use of covariance matrices through Mahalanobis distance, achieving an accuracy of up to 96.82%. Notably, when applying Mamba's new attention mechanism to extract features and model their relationships through CovaMNet [18], our proposed model achieved an outstanding accuracy of 99.33%.



(a) SA-CovaMNet (b) MF-Net

(c) Ensembling Transformer (d) The proposed MixMamba-Fewshot

Fig. 7 Visualization of the t-SNE for the top-performing methods in Table 3 in the 10-way 1-shot scenario, using 60 training samples from the CWRU Bearing Dataset

To give a more thorough overview of the models' performance and classification abilities, we presented confusion matrices in Fig. 6 and t-SNE visualizations in Fig. 7, which are detailed in Table 3.

To assess the statistical significance of the improvements achieved by the proposed method, we conducted a paired t-test on the predicted probabilities for the corresponding labels obtained from the proposed MixMamba-Fewshot method and compared models. The analysis was performed for both 1-shot and 5-shot scenarios on the CWRU dataset under varying training sample sizes. Table 4 presents the p-values and significance status for pairwise comparisons between the proposed model and other state-of-the-art methods. A significance difference is confirmed (Yes) if the p-value is smaller than significance level $\alpha = 0.05$.

For all sample sizes in the 1-shot and 5-shot scenarios, the p-values were consistently below the significance level $\alpha = 0.05$, indicating that the observed improvements of the proposed model are statistically significant. For instance, in the 1-shot case with 60 samples, the proposed model outperformed the Cosine Classifier [33] with a p-value of 0.0019, while the comparison with the MatchingNet [34] yielded a p-value of 0.0033. Similarly, in the 5-shot case with 19800 samples, the comparison with the Ensembling Transformer [30] resulted in a p-value of 0.0228. This analysis confirms that the performance gains achieved by the proposed MixMamba-Fewshot model are not due to random variations but are statistically significant. The findings further validate the robustness and effectiveness of the proposed approach in bearing fault diagnosis tasks under limited data conditions.

5.4 Results on the PU bearing dataset

In addition to evaluating performance on the CWRU Bearing Dataset, we tested the effectiveness of our proposed model on a more complex dataset, the PU Bearing Dataset. This dataset includes both natural faults that occur during motor operation and artificial faults caused by various reasons. We conducted experiments in two scenarios: 13-way 1-shot and 13-way 5-shot, with the number of training samples gradually

Table 4 The p-value indicates the statistical significance of the results on the CWRU dataset between the proposed method and compared models across various training sample cases

Pairwise Comparison	60 samples		90 samples		300 samples		19800 samples	
	p-value	$p < 0.05$	p-value	$p < 0.05$	p-value	$p < 0.05$	p-value	$p < 0.05$
(a) For 1-shot case								
Proposed vs. Cosine Classifier [33]	0.0019	Yes	0.0039	Yes	0.0021	Yes	0.0028	Yes
Proposed vs. Cross Attention Network [35]	0.0006	Yes	0.0084	Yes	0.0082	Yes	0.0012	Yes
Proposed vs. RelationNet [36]	0.0007	Yes	0.0096	Yes	0.0054	Yes	0.0077	Yes
Proposed vs. MatchingNet [34]	0.0033	Yes	0.0114	Yes	0.0097	Yes	0.0302	Yes
Proposed vs. Siamese-WDCNN [6]	0.0061	Yes	0.0055	Yes	0.0087	Yes	0.0215	Yes
Proposed vs. Siamese-Convmixer [37]	0.0060	Yes	0.0159	Yes	0.0092	Yes	0.0030	Yes
Proposed vs. CovaMNET [18]	0.0097	Yes	0.0187	Yes	0.0175	Yes	0.0181	Yes
Proposed vs. SA-CovaMNet [38]	0.0048	Yes	0.0142	Yes	0.0144	Yes	0.0283	Yes
Proposed vs. MF-Net [15]	0.0256	Yes	0.0204	Yes	0.0211	Yes	0.0245	Yes
Proposed vs. QS-Former [39]	0.0164	Yes	0.0180	Yes	0.0122	Yes	0.0256	Yes
Proposed vs. Ensembling Transformer [30]	0.0289	Yes	0.0236	Yes	0.0251	Yes	0.0352	Yes
(b) For 5-shot case								
Proposed vs. Cosine Classifier [33]	0.0129	Yes	0.0025	Yes	0.0047	Yes	0.0096	Yes
Proposed vs. Cross Attention Network [35]	0.0005	Yes	0.0135	Yes	0.0028	Yes	0.0099	Yes
Proposed vs. RelationNet [36]	0.0008	Yes	0.0007	Yes	0.0002	Yes	0.0047	Yes
Proposed vs. MatchingNet [34]	0.0166	Yes	0.0077	Yes	0.0087	Yes	0.0167	Yes
Proposed vs. Siamese-WDCNN [6]	0.0077	Yes	0.0061	Yes	0.0071	Yes	0.0079	Yes
Proposed vs. Siamese-Convmixer [37]	0.0089	Yes	0.0178	Yes	0.0052	Yes	0.0046	Yes
Proposed vs. CovaMNET [18]	0.0154	Yes	0.0182	Yes	0.0104	Yes	0.0335	Yes
Proposed vs. SA-CovaMNet [38]	0.0199	Yes	0.0175	Yes	0.0096	Yes	0.0144	Yes
Proposed vs. MF-Net [15]	0.0151	Yes	0.0335	Yes	0.0107	Yes	0.0184	Yes
Proposed vs. QS-Former [39]	0.0174	Yes	0.0088	Yes	0.0056	Yes	0.0206	Yes
Proposed vs. Ensembling Transformer [30]	0.0207	Yes	0.0198	Yes	0.0118	Yes	0.0228	Yes

Table 5 Comparison of accuracy (in %) of the proposed model with other models under different data conditions in the PU Bearing Dataset

Few-shot models	195 samples		260 samples		650 samples		1300 samples		25844 samples	
	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
Cosine Classifier [33]	61.24	69.53	73.14	78.02	85.12	85.25	93.07	93.81	97.24	98.01
Cross Attention Network [35]	75.17	77.72	79.06	79.63	83.14	84.32	90.07	91.26	97.23	97.46
RelationNet [36]	63.14	71.28	72.28	76.05	79.11	79.68	87.74	89.06	91.16	97.22
MatchingNet [34]	56.72	64.14	67.12	67.06	78.07	78.13	77.25	77.28	87.06	90.92
Siamese-WDCNN [6]	66.72	74.59	79.81	79.13	82.06	84.62	90.01	93.42	98.07	98.76
Siamese-Convmixer [37]	70.28	77.72	79.06	79.63	83.14	84.32	90.07	91.26	97.23	97.46
CovaMNET [18]	76.38	80.23	84.25	88.46	89.92	93.67	97.13	97.21	99.06	99.35
SA-CovaMNet [38]	72.75	79.92	83.12	89.14	90.06	90.17	92.83	93.14	98.07	99.15
MF-Net [15]	77.13	83.25	87.06	89.17	91.23	92.14	92.06	95.87	99.23	99.37
QS-Former [39]	80.17	82.28	88.13	88.61	94.02	95.17	96.28	96.66	98.13	99.52
Ensembling Transformer [30]	80.08	86.15	87.13	89.46	94.68	96.72	98.14	99.23	**99.62**	**99.76**
Proposed MixMamba-Fewshot	**83.79**	**90.21**	**89.28**	**91.95**	**96.05**	**97.85**	**98.26**	**99.33**	99.59	99.69

The bold values in the tables indicate the best performance for each evaluation metric

increasing from 195 to 25,844 samples. The results of our proposed model, along with those of the compared models, are presented in Table 5.

The results, as provided in Table 5, show that for a small number of samples, such as 195, due to the complexity of the data, models like Cosine Classifier [33], Cross Attention Network [35], Relation Net [36], Matching Net [34],

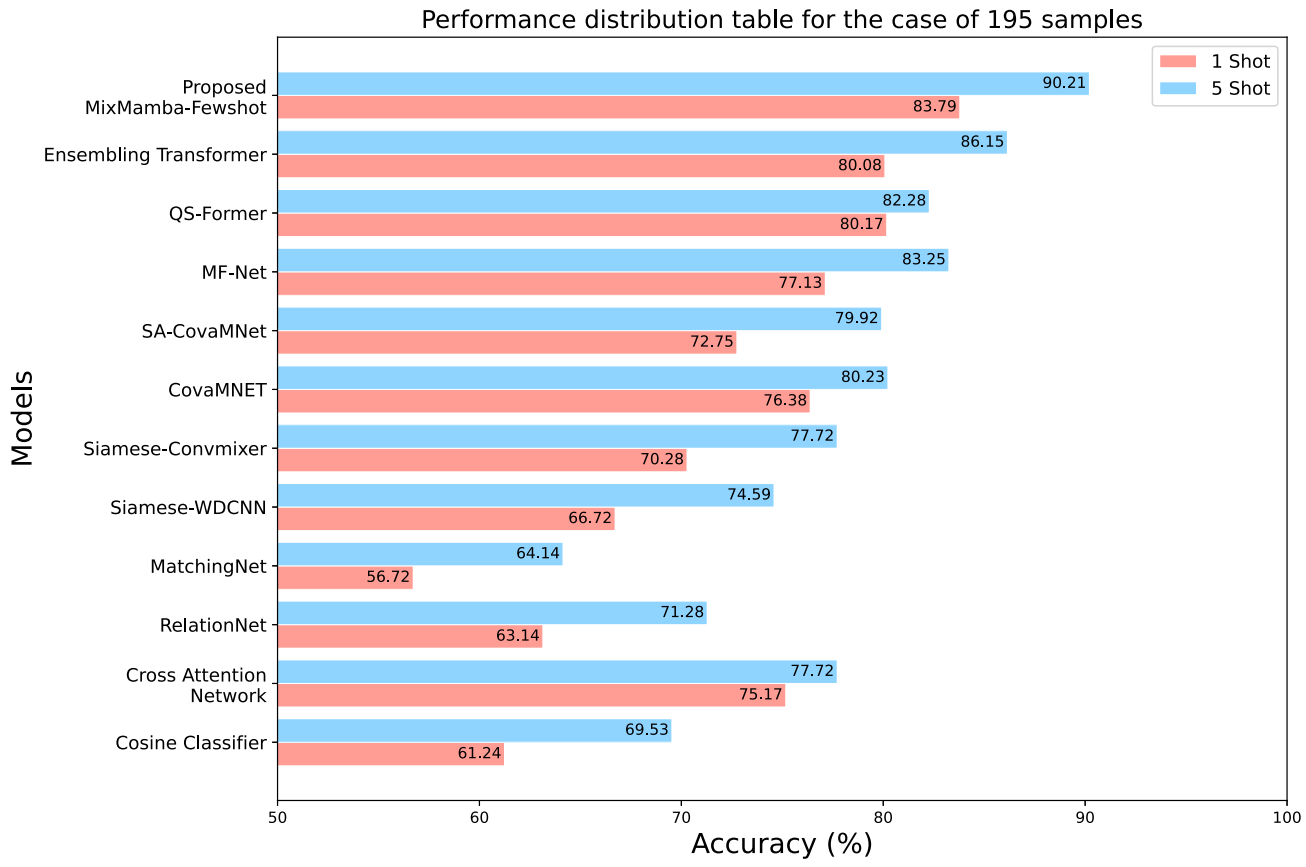


Fig. 8 Bar chart showing the classification accuracy on the PU dataset with 195 samples

Siamese-WDCNN [6], and Siamese-Convmixer [37] did not perform well, with accuracy ranging from 56.72% to 77.72%. When using the CovaMNet network [18] and its variants, such as SA-CovaMNet [38] and MF-Net [15], the results improved, with MF-Net achieving the highest accuracy of 83.25% in the 13-way 5-shot scenario.

The Ensembling Transformer model [30], which combines local and global feature extraction through Mahalanobis distance and the Transformer's self-attention mechanism, achieved an impressive result of 86.15% in the 13-way 5-shot case. Our proposed model outperformed others, achieving 83.79% and 90.21% accuracy for 1-shot and 5-shot, respectively. As the number of training samples increased, the accuracy of all models improved, but our model consistently maintained its leading performance under various data conditions. In the case of diverse data with 25,844 samples, our model achieved 99.59% for 1-shot and 99.69% for 5-shot. Although this was slightly lower than the Ensembling Transformer [30], based on the results in Table 5, we are confident in our model's ability to learn effectively under limited data conditions - a common challenge in real-world applications. This is clearly illustrated in the bar chart, in Fig. 8, with as few as 195 training samples, the proposed approach achieved an accuracy of 83.79% for the 1-shot case

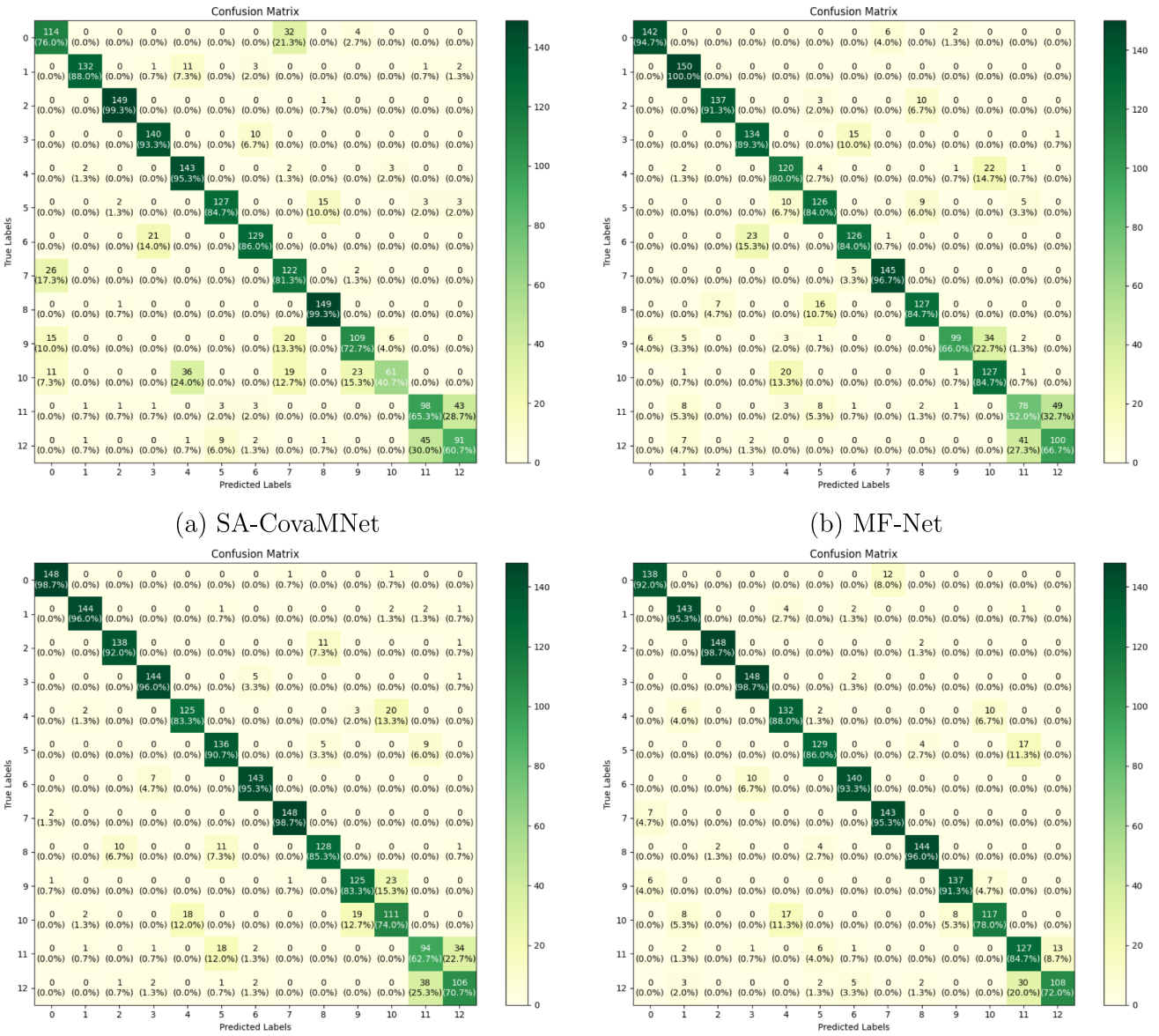
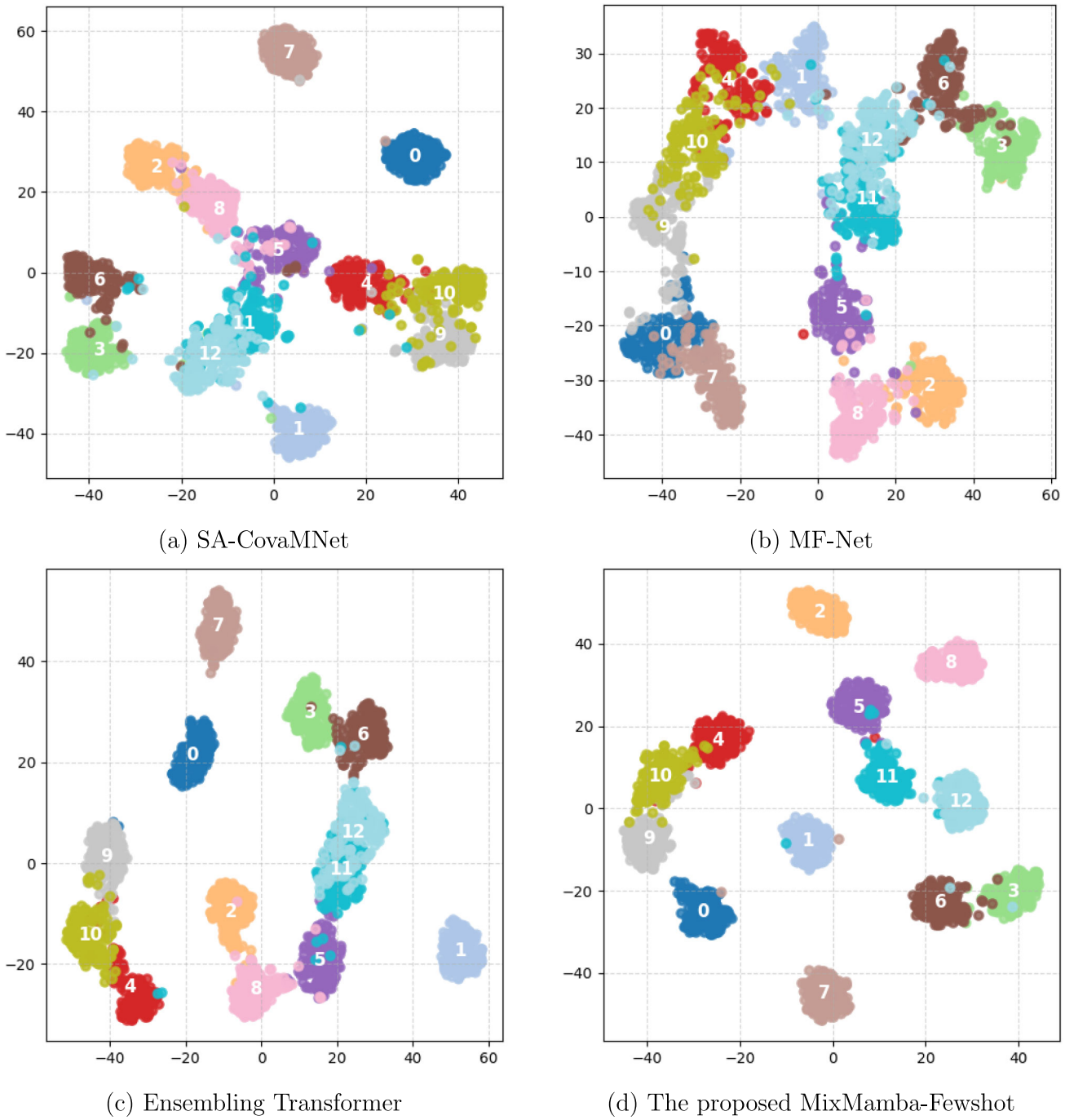


Fig. 9 Visualization of the confusion matrix for the top-performing methods in Table 5 in the 13-way 5-shot scenario, using 195 training samples from the PU Bearing Dataset

and 90.21% for the 5-shot case, significantly outperforming CovaMNet, which achieved 76.38% and 80.23% for the 1-shot and 5-shot cases, respectively.

To further evaluate the performance of each fault class, we used a confusion matrix and t-SNE visualization technique in Figs. 9 and 10. Based on the confusion matrix data, it is evident that the models perform well on artificial errors due to their clear structure and easier identification. However, when dealing with natural errors in classes 10, 11, and 12 (with severity levels of 2 or 3 according to Table 2), the models face greater challenges. This is due to the complexity of natural error data, which is often unpredictable and non-linear, making accurate classification more difficult. The t-SNE visualization further highlights this issue, as some



(a) SA-CovaMNet (b) MF-Net

(c) Ensembling Transformer (d) The proposed MixMamba-Fewshot

Fig. 10 Visualization of the t-SNE for the top-performing methods in Table 3 in the 13-way 5-shot scenario, using 195 training samples from the PU Bearing Dataset

Table 6 The p-value indicates the statistical significance of the results on the PU dataset between the proposed method and compared models across various training sample cases

Pairwise Comparison	195 samples		260 samples		650 samples		1300 samples		25844 samples	
	p-value	$p < 0.05$	p-value	$p < 0.05$	p-value	$p < 0.05$	p-value	$p < 0.05$	p-value	$p < 0.05$
(a) For 1-shot case										
Proposed vs. Cosine Classifier [33]	0.0024	Yes	0.0033	Yes	0.0054	Yes	0.0162	Yes	0.0135	Yes
Proposed vs. Cross Attention Network [35]	0.0050	Yes	0.0107	Yes	0.0050	Yes	0.0088	Yes	0.0118	Yes
Proposed vs. RelationNet [36]	0.0063	Yes	0.0049	Yes	0.0009	Yes	0.0041	Yes	0.0071	Yes
Proposed vs. MatchingNet [34]	0.0003	Yes	0.0047	Yes	0.0009	Yes	0.0036	Yes	0.0046	Yes
Proposed vs. Siamese-WDCNN [6]	0.0034	Yes	0.0116	Yes	0.0085	Yes	0.0130	Yes	0.0156	Yes
Proposed vs. Siamese-Convmixer [37]	0.0058	Yes	0.0157	Yes	0.0096	Yes	0.0119	Yes	0.0150	Yes
Proposed vs. CovaMNET [18]	0.0088	Yes	0.0176	Yes	0.0143	Yes	0.0262	Yes	0.0274	Yes
Proposed vs. SA-CovaMNet [38]	0.0074	Yes	0.0182	Yes	0.0152	Yes	0.0157	Yes	0.0194	Yes
Proposed vs. MF-Net [15]	0.0085	Yes	0.0199	Yes	0.0193	Yes	0.0151	Yes	0.0537	No
Proposed vs. QS-Former [39]	0.0104	Yes	0.0206	Yes	0.0212	Yes	0.0246	Yes	0.0449	Yes
Proposed vs. Ensembling Transformer [30]	0.0098	Yes	0.0216	Yes	0.0239	Yes	0.0396	Yes	0.1758	No
(b) For 5-shot case										
Proposed vs. Cosine Classifier [33]	0.0049	Yes	0.0071	Yes	0.0098	Yes	0.0155	Yes	0.0289	Yes
Proposed vs. Cross Attention Network [35]	0.0084	Yes	0.0098	Yes	0.0080	Yes	0.0076	Yes	0.0176	Yes
Proposed vs. RelationNet [36]	0.0068	Yes	0.0050	Yes	0.0009	Yes	0.0110	Yes	0.0133	Yes
Proposed vs. MatchingNet [34]	0.0089	Yes	0.0038	Yes	0.0008	Yes	0.0100	Yes	0.0009	Yes
Proposed vs. Siamese-WDCNN [6]	0.0110	Yes	0.0099	Yes	0.0085	Yes	0.0318	Yes	0.0335	Yes
Proposed vs. Siamese-Convmixer [37]	0.0059	Yes	0.0102	Yes	0.0089	Yes	0.0253	Yes	0.0361	Yes
Proposed vs. CovaMNET [18]	0.0094	Yes	0.0153	Yes	0.0214	Yes	0.0367	Yes	0.0424	Yes
Proposed vs. SA-CovaMNet [38]	0.0128	Yes	0.0365	Yes	0.0201	Yes	0.0325	Yes	0.0399	Yes
Proposed vs. MF-Net [15]	0.0159	Yes	0.0370	Yes	0.0253	Yes	0.0330	Yes	0.0424	Yes
Proposed vs. QS-Former [39]	0.0129	Yes	0.0271	Yes	0.0388	Yes	0.0315	Yes	0.0530	No
Proposed vs. Ensembling Transformer [30]	0.0254	Yes	0.0332	Yes	0.0437	Yes	0.0716	No	0.0472	Yes

error classes do not show clear separation. Nevertheless, the proposed model, PAM-Mamba, not only achieves high performance in classifying artificial errors but also excels compared to other models in handling challenging natural errors like classes 10, 11, and 12. The t-SNE plot demonstrates that PAM-Mamba has better class separation, leading to improved overall accuracy. These results underscore PAM-Mamba's strong potential for practical applications in engine monitoring and fault diagnosis systems, where complex and diverse data are frequently encountered.

To evaluate the statistical significance of the results achieved by the proposed MixMamba-Fewshot framework on the PU dataset, we performed a paired t-test. Similar to the case of CWRU dataset, the test was conducted on the predicted probabilities for the corresponding labels obtained from the softmax outputs of the proposed model and the compared state-of-the-art methods. Table 6 provides the p-values and significance levels ($\alpha < 0.05$) for pairwise comparisons in both the 1-shot and 5-shot scenarios under varying sample sizes.

For the 1-shot case, the p-values were consistently below the significance level $\alpha = 0.05$, except in a few instances involving methods such as MF-Net [15] and QS-Former [39] when evaluated with larger sample sizes. For example, the proposed model outperformed the Cosine Classifier [33] with a p-value of 0.0024 for 195 samples and achieved a p-value of 0.0003 when compared to MatchingNet [34] for the same sample size. In the 5-shot case, similar trends were observed. The proposed model demonstrated significant improvements over most of the baseline methods across all sample sizes, with p-values less than 0.05. For instance, the proposed model yielded a p-value of 0.0084 when compared to the Cross Attention Network [35] for 195 samples, and a p-value of 0.0068 when compared to RelationNet [36] for the same dataset. However, a few methods, such as the Ensembling Transformer [30], did not show statistically significant differences for larger sample sizes (e.g., 1300 and 25844 samples), with p-values exceeding $\alpha = 0.05$.

The significance analysis confirms that the observed performance improvements of the proposed MixMamba-Fewshot framework are statistically significant across most scenarios, especially for smaller sample sizes. This underscores the robustness and effectiveness of the proposed method in addressing the challenges of bearing fault diagnosis with limited data.

5.5 Ablation study

5.5.1 Ablation study on the proposed feature extractor module

In this subsection, we conduct ablation experiments to analyze the impact of each component on the performance of the proposed model. Specifically, we use each individual branch of the model to extract features from the query set **Q** and the support set **S**. Then, we employ the baseline CovaMNet [18] to capture second-order statistical information from the features of **Q** and **S**. Our model introduces a significant improvement over the conventional CovaMNet [18] by utilizing two parallel branches, each tasked with extracting different features from the samples. The proposed PCA-PSA branch captures the variations occurring after convolutional layers, helping the model maintain stability during training. Meanwhile, the VSS branch incorporates the selective scanning mechanism of SS2D, allowing it to focus on the most informative regions of the data, which is crucial under conditions of limited training data.

In Tables 7 and 8, we present the results of ablation experiments in the context of a 5-shot learning scenario with three different training sample conditions for each dataset. In the first experiment, to clarify the impact of spatial and channel-wise feature extraction, we utilized only the PCA-PSA branch from our model. In the CWRU Bearing Dataset, with 60 training samples, the model achieved an accuracy of 98.80%. For the PU Bearing Dataset, with 195 training samples, the accuracy reached 89.54%. Overall, the classification

Table 7 Ablation Study for different branches of the proposed model on the CWRU Bearing Dataset under various data conditions

(a) 1-shot case

PCA-PSA	VSS	60 samples	90 samples	300 samples	19800 samples
✓	-	98.33	98.87	98.93	99.13
-	✓	95.93	96.33	97.47	97.87
✓	✓	99.33	99.52	99.77	99.90

(b) 5-shot case

PCA-PSA	**VSS**	**60 samples**	**90 samples**	**300 samples**	**19800 samples**
✓	-	98.80	99.27	99.67	99.73
-	✓	96.40	96.80	96.73	98.13
✓	✓	**99.53**	99.68	**99.83**	**99.93**

The bold values in the tables indicate the best performance for each evaluation metric

Table 8 Ablation Study for different branches of the proposed model on the PU Bearing Dataset under various data conditions

(a) 1-shot case

PCA-PSA	VSS	195 samples	260 samples	650 samples	1300 samples	25844 samples
✓	-	81.74	87.18	93.33	96.81	98.44
-	✓	79.38	82.08	90.41	94.15	97.86
✓	✓	83.79	89.28	96.05	98.26	99.59

(b) 5-shot case

PCA-PSA	VSS	**195 samples**	**260 samples**	**650 samples**	**1300 samples**	**25844 samples**
✓	-	87.54	88.15	96.92	97.94	99.18
-	✓	85.64	86.26	95.95	96.05	97.69
✓	✓	**90.21**	91.95	**97.85**	99.33	**99.69**

The bold values in the tables indicate the best performance for each evaluation metric

accuracy of the PCA-PSA branch was only approximately 1% lower than that of the complete model. These impressive results demonstrate that spatial and channel features play a crucial role in the performance of our model. In the next experiment, where we used only the VSS branch, the results were lower compared to the PCA-PSA branch.

Specifically, under different data conditions, the accuracy using only the VSS branch was 1% - 4% lower than that of the PCA-PSA branch. This indicates that the selective scanning mechanism, when used in isolation, is less effective. It also suggests that separating the branches for individual training leads to suboptimal model learning, as each branch is responsible for extracting distinct features from the data. Recognizing this, we combined both the PCA-PSA and VSS branches to achieve optimal performance. This combination allowed the model to leverage the strengths of both feature extraction techniques, leading to significant improvements in results. The performance of the proposed model after combining the two branches showed a notable increase, reinforcing the comprehensive and effective nature of the model in classification tasks under limited training data conditions.

5.5.2 Ablation study on classification and feature extraction strategies

In this section, we conducted an ablation study to analyze the contributions of different classification modules and feature extraction strategies. Our study compared the PAM-Mamba feature extractor with classification modules such as CovaMNET [18], as proposed in our methodology, and the recent Ensembling Transformer (Transformer-based) module introduced by Vu et al. [30]. Additionally, we evaluated alternative feature extraction strategies such as the MLKFE (Multiscale Large Kernel Feature Extraction) from Vu et al., ViT from Vision Transformer [40], VSS (Mamba), PCA-PSA, and the proposed PAM-Mamba modules. We performed the ablation studies for both the CWRU and PU datasets in the 1-shot and 5-shot learning scenarios.

As this ablation study utilizes both the MLKFE feature extraction strategy and the Transformer-based classification module from the recent work of Vu et al., we outline the relationship between the Transformer-based approach and our proposed study. The work by Vu et al. [30] employs

Table 9 Comparison of different feature extraction and classification modules across varying training samples on the CWRU dataset

Cases	Feature Extraction	Classification Module	60 samples		90 samples		300 samples		19800 samples	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Δ_1	MLKFE	Transformer-based	96.82	98.23	98.17	98.98	99.61	99.14	98.14	99.76
Δ_2	ViT	Transformer-based	93.32	94.06	95.15	95.83	96.68	97.22	98.36	98.97
Δ_3	VSS	Transformer-based	95.88	96.52	96.94	97.15	97.41	97.28	98.38	99.15
Δ_4	PCA-PSA	Transformer-based	97.13	98.18	98.39	98.85	99.38	99.23	99.65	99.72
Δ_5	PAM-Mamba	Transformer-based	98.11	98.65	98.87	99.02	99.31	99.60	99.48	99.67
Δ_6	MLKFE	CovaMNET	96.21	96.85	97.29	97.64	98.01	98.55	99.28	99.50
Δ_7	ViT	CovaMNET	92.47	93.87	95.06	95.52	96.33	96.89	97.04	98.38
Δ_8	VSS	CovaMNET	95.93	96.40	96.33	96.80	97.47	96.73	97.87	98.13
Δ_9	PCA-PSA	CovaMNET	98.33	98.80	98.87	99.27	98.93	99.67	99.13	99.73
Δ_{10}	PAM-Mamba	CovaMNET	**99.33**	**99.53**	**99.52**	**99.68**	**99.77**	**99.83**	**99.90**	**99.93**

The bold values in the tables indicate the best performance for each evaluation metric

Table 10 Comparison of different feature extraction and classification modules across varying training samples on the PU dataset

Cases	Feature Extraction	Classification Module	195 samples		260 samples		650 samples		1300 samples		25844 samples	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Δ_1	MLKFE	Transformer-based	80.08	86.15	87.13	89.46	94.68	96.72	98.14	99.23	**99.62**	**99.76**
Δ_2	ViT	Transformer-based	76.64	78.22	79.85	81.23	86.54	88.28	92.36	93.57	97.51	98.49
Δ_3	VSS	Transformer-based	78.27	82.91	82.37	85.45	90.28	94.60	96.33	97.42	98.57	98.90
Δ_4	PCA-PSA	Transformer-based	80.42	86.20	87.50	89.72	95.18	97.04	97.92	98.76	99.51	99.64
Δ_5	PAM-Mamba	Transformer-based	81.15	86.81	87.32	89.52	94.33	96.17	98.22	98.66	99.54	99.67
Δ_6	MLKFE	CovaMNET	79.17	83.59	85.70	87.88	92.18	94.39	97.43	97.80	98.32	98.56
Δ_7	ViT	CovaMNET	76.92	79.62	81.36	81.97	87.22	87.81	92.35	93.11	98.51	98.94
Δ_8	VSS	CovaMNET	79.38	85.64	82.08	86.26	90.41	95.95	94.15	96.05	97.86	97.69
Δ_9	PCA-PSA	CovaMNET	81.74	87.54	87.18	88.15	93.33	96.92	96.81	97.94	98.44	99.18
Δ_{10}	PAM-Mamba	CovaMNET	**83.79**	**90.21**	**89.28**	**91.95**	**96.05**	**97.85**	**98.26**	**99.33**	99.59	99.69

The bold values in the tables indicate the best performance for each evaluation metric

a Transformer architecture for classification, utilizing cross-attention mechanisms to model correlations between support and query sets. The referenced work incorporates a Multi-scale Large Kernel Feature Extraction module to enhance representation. In contrast, our approach does not utilize large kernel feature extraction or Transformer-based components. We employ the Mamba architecture combined with a lightweight ConvMixer-inspired module for feature extraction. For classification module, we adopt the Covariance Metric Networks (CovaMNET) framework for efficient feature correlation. Additionally, the use of the Mahalanobis distance also differs. While the work by Vu et al. combines Mahalanobis metrics with cross-attention outputs in its local branch for classification, our method directly applies CovaMNET for metric learning, ensuring computational simplicity.

The results from our ablation study are provided in Tables 9, and 10 for the CWRU and PU datasets respectively. The results confirm that while the Transformer-based classification strategy demonstrates competitive performance, the CovaMNET structure combined with the PAM-Mamba feature extractor, Δ_{10} case, consistently outperforms Transformer-based approaches (Δ_1 to Δ_5 cases). Specifically, on the CWRU dataset, PAM-Mamba with CovaMNET achieves 99.33% and 99.53% accuracy in the 1-shot and 5-shot cases, respectively, with only 60 training samples.

Similarly, on the PU dataset, this combination achieves 83.79% and 90.21% accuracy in the 1-shot and 5-shot cases, respectively, with only 195 training samples. These results demonstrate the robustness of our lightweight feature extraction strategy and metric learning. Furthermore, the ablation study highlights the effectiveness of the combination of lightweight feature extraction and effective covariance metric learning in CovaMNet. The integration of the Mamba architecture with the channel-spatial attention mechanism of PCA-PSA, provides a robust and efficient framework for few-shot bearing fault diagnosis.

6 Conclusion

In this study, we propose a new model called MixMamba-Fewshot, based on the attention ConvMixer and Mamba architecture with few-shot learning method CovaMNET, to tackle the problem of bearing fault classification. This method leverages the selective attention mechanism of the state-space model, combined in parallel with extracting complex features along both the channel and spatial dimensions from the spectrogram of bearing fault signals, using the Priority Attention Mixer. This approach enables our model to integrate diverse local and global information across both spatial and channel dimensions. Under limited data conditions, experimental results on two datasets, CWRU and PU Bearing Datasets, demonstrate that the proposed model achieves higher accuracy compared to previous methods, highlighting its high practical applicability. In the future, we will extend our research and experiments to other types of industrial fault signal, such as motor shaft faults and gear faults, also under limited data conditions, to fully exploit the potential of the model. Furthermore, we will continue to refine the model by building upon existing frameworks, aiming to further enhance its performance and efficiency.

Acknowledgements This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05−2021.34.

Data Availability The data are sourced from publicly available datasets and properly cited.

References

1. Smith DM (1969) Bearing development and bearing theory. J Bearings Turbomach 3–5. Boston, MA: Springer US
2. Samanta P, Murmu N, Khonsari M (2019) The evolution of foil bearing technology. Tribol Int 135:305–323. ISSN: 0301-679X

3. Yadav E, Chawla V (2022) An explicit literature review on bearing materials and their defect detection techniques. In: Materials Today: Proceedings 50. 2nd International Conference on Functional Material, Manufacturing and Performances (ICFMMP-2021), pp 1637–1643
4. Goudarzi MM, Jahromi SJ, Nazarboland A (2009) Investigation of characteristics of tin-based white metals as a bearing material. Mater Des 30(6):2283–2288
5. Wu K, Yu K, Chen C, Wu J, Liu Y (2024) Optimal transport strategy-based meta-attention network for fault diagnosis of rotating machinery with zero sample. Appl Intell 1–17
6. Zhang A, Li S, Cui Y, Yang W, Dong R, Hu J (2019) Limited data rolling bearing fault diagnosis with few-shot learning. Ieee Access 7:110895–110904
7. Zhang Y, Zhao X, Liang H, Chen P (2024) Multiscale dilated convolution and swin-transformer for small sample gearbox fault diagnosis. Appl Intell 1–17
8. Kumar H, Upadhyaya G (2023) Fault diagnosis of rolling element bearing using continuous wavelet transform and k-nearest neighbour. In: Materials today: proceedings 92, pp 56–60
9. Sawaqed LS, Alrayes AM (2020) Bearing fault diagnostic using machine learning algorithms. Prog Artif Intell 9(4):341–350
10. Zhang J, Yi S, Liang G, Hongli G, Xin H, Hongliang S (2020) A new bearing fault diagnosis method based on modified convolutional neural networks. Chin J Aeronaut 33(2):439–447
11. Chen X, Zhang B, Gao D (2021) Bearing fault diagnosis base on multi-scale cnn and lstm model. J Intell Manuf 32(4):971–987
12. Yang Z, Cen J, Liu X, Xiong J, Chen H (2022) Research on bearing fault diagnosis method based on transformer neural network. Meas Sci Technol 33(8):085111
13. Koch G, Zemel R, Salakhutdinov R et al (2015) Siamese neural networks for one-shot image recognition. In: Icml deep learning workshop, vol 2. 1. Lille. pp 1–30
14. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
15. Vu MH, Pham VT (2023) Mixerformer-covariance metric neural network: A new few-shot learning model for bearing fault diagnosis. In: 2023 12th international conference on control, automation and information sciences (ICCAIS), pp 639–644
16. Shen H, Zhao D, Wang L, Liu Q (2023) Bearing fault diagnosis based on prototypical network. In: International conference on mechatronics engineering and artificial intelligence (MEAI 2022), vol 12596, SPIE. pp 79–84
17. Li C, Li S, Zhang A, He Q, Liao Z, Hu J (2021) Meta-learning for few-shot bearing fault diagnosis under complex working conditions. Neurocomputing 439:197–211
18. Li W, Xu J, Huo J, Wang L, Gao Y, Luo J (2019) Distribution consistency based covariance metric networks for few-shot learning. In: AAAI Conference on artificial intelligence. https://api.semanticscholar.org/CorpusID:69672216
19. Zheng X, Yue C, Wei J, Xue A, Ge M, Kong Y (2023) Few-shot intelligent fault diagnosis based on an improved meta-relation network. Appl Intell 53(24):30080–30096
20. Gu A, Dao T (2023) Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752
21. Zhu L, Liao B, Zhang Q, Wang X, Liu W, Wang X (2024) Vision mamba: Efficient visual representation learning with bidirectional state space model. ArXiv arXiv:2401.09417 https://api.semanticscholar.org/CorpusID:267028142
22. Liu Y, Tian Y, Zhao Y, Yu H, Xie L, Wang Y, Ye Q, Liu Y (2024) Vmamba: Visual state space model. CoRR arXiv:2401.10166 https://doi.org/10.48550/arXiv.2401.10166
23. Trockman A, Kolter JZ (2022) Patches are all you need? arXiv preprint arXiv:2201.09792
24. Park Y, Azaña J (2010) Optical signal processors based on a time-spectrum convolution. Opt Lett 35(6):796–798
25. Yue Y, Li Z (2024) Medmamba: Vision mamba for medical image classification. arXiv:2403.03849
26. Le TV, Le H-M-Q, Vu VY, Tran T-T, Pham V-T (2023) Attention convmixer model and application for fish species classification. EAI Endorsed Trans Ind Netw Intell Syst 10:e2
27. Smith WA, Randall RB (2015) Rolling element bearing diagnostics using the case western reserve university data: A benchmark study. Mech Syst Signal Process 64:100–131
28. Paderborn University bearing data center (2014). Online. https://mb.uni-paderborn.de/kat/forschung/datacenter/bearing-datacenter Accessed Jan 2021
29. Alonso-González M, Díaz VG, Pérez BL, G-Bustelo BCP, Anzola JP, (2023) Bearing fault diagnosis with envelope analysis and machine learning approaches using cwru dataset. IEEE Access 11:57796–57805
30. Vu M-H, Nguyen V-Q, Tran T-T, Pham V-T, Lo M-T (2024) Few-shot bearing fault diagnosis via ensembling transformer-based model with mahalanobis distance metric learning from multiscale features. IEEE Transactions on Instrumentation and Measurement
31. Nguyen V-Q, Vu M-H, Pham V-T, Tran T-T (2023) A deep learning approach based on mlp-mixer models for bearing fault diagnosis. In: 2023 International conference on system science and engineering (ICSSE), pp 16–21. IEEE
32. Neupane D, Seok J (2020) Bearing fault detection and diagnosis using case western reserve university dataset with deep learning approaches: A review. IEEE Access 8:93155–93178
33. Park K, Hong JS, Kim W (2020) A methodology combining cosine similarity with classifier for text classification. Appl Artif Intell 34(5):396–411
34. Vinyals O, Blundell C, Lillicrap T, Wierstra D et al (2016) Matching networks for one shot learning. Adv Neural Inf Process Syst 29
35. Hou R, Chang H, Ma B, Shan S, Chen X (2019) Cross attention network for few-shot classification. Adv Neural Inf Process Syst 32
36. Sung F, Yang Y, Zhang L, Xiang T, Torr PH, Hospedales TM (2018) Learning to compare: Relation network for few-shot learning. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp 1199–1208
37. Vu M-H, Nguyen V-Q, Tran T-T, Pham V-T (2023) A new convmixer-based approach for diagnosis of fault bearing using signal spectrum. In: Conference on information technology and its applications. Springer, pp 3–14
38. Zhai J, Han L, Xiao Y, Yan M, Wang Y, Wang X (2023) Few-shot fine-grained fish species classification via sandwich attention CovaMNet. Front Mar Sci 10:1149186
39. Wang X, Wang X, Jiang B, Luo B (2023) Few-shot learning meets transformer: Unified query-support transformers for few-shot classification. IEEE Trans Circ Syst Vid Technol 33(12):7789–7802
40. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30

Nhu-Linh Than is currently a fourth-year student majoring in Control Engineering and Automation at Hanoi University of Science and Technology. His research interests focus on Computer Vision, Deep Learning, and Few-shot Learning. He is passionate about deep learning algorithms, particularly models that enhance image recognition and machine learning with limited data.

Van Quang Nguyen is a fourth-year student at Hanoi University of Science and Technology, majoring in Control Engineering and Automation. His academic and research interests focus on Computer Vision, Deep Learning, Segmentation, and Few-shot Learning. Throughout his studies, he has actively participated in research projects on Few-shot Learning for fault diagnosis in bearings and medical image segmentation, aiming to develop innovative solutions in artificial intelligence.

Gia-Bao Truong is a fourth-year student at Hanoi University of Science and Technology majoring in Control Engineering and Automation, is currently working as an AI Engineer at VinBigdata in Hanoi. His research interests include computer vision, speech recognition, and natural language understanding. Throughout his academic journey, he has focused on applying few-shot learning techniques to the diagnosis of bearing faults, addressing the challenges posed by limited data availability in industrial environments.

Van-Truong Pham received B.S. and M.S. degrees in Electrical Engineering from Hanoi University of Science and Technology, Vietnam, and the Ph.D. degree in Electrical Engineering from National Central University, Taiwan. From 2013 to 2016, he held a post-doctoral position at the National Central University, Taoyuan, Taiwan. He is currently an Associate Professor with the Department of Automation Engineering, School of Electrical and Electronic, Hanoi University of Science and Technology, Hanoi, Vietnam. His research interests include computer vision, image processing, deep learning, signal processing, and applied control theory.

Thi-Thao Tran received B.S. and M.S. degrees in Electrical Engineering from Hanoi University of Science and Technogy, Vietnam, in 2003 and 2005, respectively, and the Ph.D. degree in Electrical Engineering from National Central University, Taiwan. She is currently an Associate Professor with the Department of Automation Engineering, School of Electrical and Electronic, Hanoi University of Science and Technology, Hanoi, Vietnam. Her research interests include computer vision, image processing, deep learning, and signal processing.