

Article

A Few-Shot Fault Diagnosis Algorithm for Rotating Machinery Based on Time-Frequency Dual-Stream Mamba and Joint Metric Learning

Yuan Xu, Boning Wu, Yingnan Gao  and Zhenzhen Jin * 

Guangxi Key Laboratory of Manufacturing System & Advanced Manufacturing Technology, School of Mechanical Engineering, Guangxi University, Nanning 530004, China; 2311391134@st.gxu.edu.cn (Y.X.); 2511390213@st.gxu.edu.cn (B.W.); 2511300015@st.gxu.edu.cn (Y.G.)

* Correspondence: gxdxjz@gxu.edu.cn

Abstract

Rotating machinery plays a critical role in transmission systems, while the scarcity of fault samples and labeled data limits the performance of existing diagnostic methods under few-shot conditions. This paper proposes a few-shot fault diagnosis method for rotating machinery based on a time-frequency dual-stream Mamba architecture and joint metric learning. Raw vibration signals and multi-scale short-time Fourier transform time-frequency views are constructed to characterize transient impacts and frequency-band energy distributions from complementary perspectives. A dual-stream feature extraction network is developed, where the time-domain branch captures impulsive fault patterns, and the time-frequency spatial Mamba and channel attention branches extract spatial dependencies and fault-sensitive responses. Moreover, a hierarchical selective fusion mechanism is introduced to adaptively integrate complementary features across different domains. Finally, a covariance-based joint metric learning module is designed to model class distributions using second-order statistics of support samples and classify query samples through local similarity aggregation. Experimental results show that the proposed method consistently outperforms comparison methods under different rotational speeds, training sample scales, and 8-way 1-shot/5-shot tasks, validating its effectiveness and robustness for few-shot rotating machinery fault diagnosis.

Keywords: rotating machinery; few-shot; joint metric learning; feature extraction; time-frequency views



Academic Editor: Francesc Pozo

Received: 8 July 2026

Revised: 17 July 2026

Accepted: 21 July 2026

Published: 23 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Rotating machinery (RM) plays a crucial role in power transmission and actuation in industrial production, transportation, and other engineering fields, and its operating condition directly affects production efficiency, operational safety, and the stable operation of industrial equipment [1,2]. During long-term operation, critical components such as bearings and gears are continuously exposed to alternating loads, impact-induced vibration, and varying operating conditions, making them susceptible to wear, cracking, spalling, looseness, and compound damage. Once abnormalities occur in rotating components, equipment efficiency may deteriorate, and unplanned shutdowns or even cascading damage may follow [3,4]. Therefore, developing effective fault diagnosis (FD) methods for RM, enabling timely health-state identification and accurate fault classification, is essential for

reducing maintenance costs, preventing unexpected downtime, and ensuring the stable operation of industrial equipment.

Traditional FD methods for RM generally follow a diagnostic pipeline consisting of signal processing, feature extraction, and shallow classifiers. These methods typically extract time-domain, frequency-domain, time-frequency (TF), or nonlinear dynamic features from monitoring signals such as vibration and current signals, and then employ classifiers such as SVM and nearest-neighbor methods for condition identification [5–7]. Wang et al. [8] proposed refined compound multiscale slope entropy to address the nonlinear and nonstationary characteristics of rolling bearing vibration signals and combined it with an SVM optimized by the grey wolf optimizer to identify bearing faults of different types and severity levels. From the perspective of periodic component identification, Cheng et al. [9] developed an adaptive periodic mode decomposition method that extracts fault-induced periodic impulsive components through Ramanujan subspaces, thereby improving feature extraction for rolling bearing faults. To address the difficulty of adaptively determining the center frequency and penalty factor in variational mode extraction, Pang et al. [10] proposed a recursive variational mode extraction method that progressively separates fault-related band-limited components. These methods offer a certain degree of interpretability and can achieve satisfactory diagnostic performance under specific operating conditions. Nevertheless, their effectiveness depends heavily on handcrafted feature design, parameter selection, and expert knowledge. Under complex operating environments, increased noise interference, or varying operating conditions, the robustness and generalizability of manually engineered features are often difficult to maintain, which may ultimately degrade diagnostic performance.

With advances in sensing technologies, industrial big data, and artificial intelligence, deep learning methods have been increasingly applied to intelligent FD of RM. Compared with traditional approaches, deep models can automatically learn deep nonlinear representations from raw signals or TF maps, thereby reducing reliance on handcrafted features and providing stronger representation capability for complex FD tasks [11–14]. Sinitsin et al. [15] proposed an FD method that combines hybrid inputs with a CNN-MLP architecture, using information from different sensing locations and input forms to improve fault detection and localization. Li et al. [16] integrated convolutional networks with Transformer architectures and introduced a joint learning strategy to enhance signal denoising and FD simultaneously, thereby improving diagnostic robustness under complex noise conditions. Hou et al. [17] developed a fault recognition framework that captured both local and global characteristics by integrating frequency-domain information, a parallel feature fusion encoder, and a cross-flipping decoding structure, thereby increasing the accuracy of FD. Chen et al. [18] combined channel attention, spatial attention, residual connections, and a soft-thresholding mechanism to improve feature selection under noisy conditions. Overall, deep learning enables end-to-end feature learning and has substantially advanced the level of automation in RM FD. However, most existing deep diagnostic models are developed under the assumptions of sufficient training data and relatively stable data distributions. In practical engineering applications, RM operates in healthy states for most of its service life, resulting in a limited quantity of available fault samples. In addition, manual annotation of collected fault data is time-consuming and labor-intensive. When labeled samples are scarce, deep models are prone to overfitting, unstable decision boundaries, and degraded generalization performance. These limitations are particularly evident in the diagnosis of compound faults, incipient faults, and cross-condition scenarios, where conventional supervised learning methods often fail to learn sufficiently stable and discriminative fault representations. Therefore, improving the feature extraction and

class discrimination capabilities of RM FD models under limited labeled data remains an important challenge in intelligent operation and maintenance.

To alleviate the limitations caused by insufficient labeled data, recent studies have explored the utilization of self-supervised learning (SSL) for FD. By designing pretext tasks, SSL can learn general representations from unlabeled data and thereby reduce, to some extent, the dependence on manually annotated samples [19–21]. Cui et al. [22] employed the Signal Transformer as the backbone network, learning fault representations from unlabeled data before fine-tuning the model with a limited number of labeled samples. To address the difficulty of training Transformers with scarce labeled data, Cen et al. [23] proposed MSFormer, a masked SSL framework in which the underlying structure of vibration signals is learned by reconstructing or predicting masked segments. Xu et al. [24] developed a self-supervised transfer learning framework for cross-machine FD by exploiting multi-scale TF semantic consistency, and designed a semantic-guided soft contrastive loss to alleviate false negative samples arising in traditional contrastive learning paradigms. These studies demonstrate the potential of SSL for reducing reliance on labeled data. However, such methods generally still require a substantial amount of unlabeled data for pretraining, and discrepancies may exist between the pretext objective and the downstream few-shot classification task. When only limited unlabeled data are available for the target machine, or when considerable distribution discrepancies exist between the source and target domains, the transfer performance of self-supervised pretrained models may remain limited.

Different from SSL, few-shot learning focuses on allowing models to generalize to unseen categories or tasks with limited labeled data, offering a promising solution for rotating machinery FD under data scarcity conditions. Among existing approaches, meta-learning- and metric-learning-based methods have become prominent directions in few-shot fault diagnosis (FSFD) research [25–27]. Wang et al. [28] showed a metric-driven meta-learning framework in the feature space, which integrates supervised representation learning with episodic training strategies for metric optimization. The method jointly exploits single-sample attribute information and group-wise similarity information to accommodate bearing and gear FD under various limited-data conditions. Wen et al. [29] developed a Siamese neural network framework that integrates few-shot learning with multi-state training, using a multi-stage training strategy and a multi-source feature fusion network to improve few-shot bearing fault recognition. Su et al. [30] proposed a data-reconstruction hierarchical recurrent meta-learning method, in which noise is suppressed and informative components are extracted during the reconstruction stage, while a recurrent training strategy is employed in the meta-learning stage to diagnose bearing faults. Luo et al. [31] introduced an elastic prototypical network that incorporates squeeze-and-excitation attention into an enhanced feature encoder and designs a metric module with an elasticity factor, thereby improving few-shot transfer diagnosis under nonstationary rotational speeds. Lin et al. [32] proposed a generalized MAML framework for heterogeneous-signal-driven cross-domain FD, where a channel-interaction feature encoder and a weight-guided factor are used to enhance meta-knowledge extraction and generalization across diverse tasks. Li et al. [33] developed a meta-learning method for FSFD under complex operating conditions. The raw signals are transformed into TF maps and organized into meta-learning tasks, allowing the model to learn suitable initialization parameters from tasks constructed under known conditions and rapidly adapt to few-shot fault recognition under previously unseen conditions. Despite the progress achieved by these few-shot and metric learning methods in reducing dependence on labeled samples, several limitations remain. First, some approaches rely primarily on a single time-domain, frequency-domain, or TF representation, making it difficult to simultaneously preserve transient fault-induced impacts and energy distribution patterns across different frequency bands. Second, existing feature

extractors often emphasize local convolutional features, recurrent temporal dependencies, or global dependency modeling with Transformers, while insufficiently exploiting the complementarity among raw time-domain characteristics, multi-scale time-frequency (MSTF) spatial structures, and channel responses. In addition, many metric learning methods use class prototypes or Euclidean distance as the primary basis for classification and therefore focus mainly on the mean representation of support features. This may overlook within-class distribution structures and local feature correlations, limiting robustness in few-shot diagnostic scenarios.

To address the above limitations, this study presents an FSFD method for RM, termed time–frequency dual-stream mamba with joint metric learning (TFDS-MJML). Taking vibration signals as input, the proposed method jointly characterizes fault states from two complementary perspectives: raw time-domain responses and MSTF distributions, thereby overcoming the limitations of a single signal representation in describing complex fault characteristics. A time-domain impulsive feature branch, an MSTF Mamba feature extraction branch, and a hierarchical selective fusion mechanism are developed to adaptively integrate transient time-domain information, spatial dependencies in the TF plane, and channel response characteristics according to the specific few-shot diagnostic task. In addition, a joint metric learning strategy combining class-specific covariance distribution modeling with local similarity aggregation is introduced to improve the representation of class distributions from limited support samples and enhance the stability of query-to-class matching. Through these designs, TFDS-MJML improves the discriminability of fault representations and the stability of class boundaries under data-scarce conditions, providing an effective framework for intelligent FSFD of RM.

The main contributions of this paper are as follows:

- (1) A TF dual-stream Mamba feature extraction framework is proposed for FSFD of RM. By jointly exploiting raw time-domain signals and multi-scale short-time Fourier transform (STFT) based TF views, the framework extracts complementary fault information from three perspectives: time-domain impulsive responses, spatial structures in the TF plane, and channel-wise correlations. This design alleviates the limitations of a single input representation in characterizing complex fault patterns.
- (2) A hierarchical selective fusion mechanism is developed to dynamically integrate MSTF spatial features, TF channel-attention features, and raw time-domain impulsive features. Spatial and channel features are first fused within the TF domain, after which the resulting TF representation is further integrated with the time-domain representation. This hierarchical scheme allows the model to dynamically adjust the contribution of each branch according to sample-specific fault characteristics, thereby reducing information redundancy and the representational bias introduced by fixed fusion strategies.
- (3) A joint metric learning classification strategy based on covariance distribution modeling and local similarity aggregation is constructed. Second-order statistical information derived from support samples is used to characterize within-class distribution structures, while few-shot classification is performed by measuring the similarity between query features and class-specific covariance matrices. This strategy improves the stability of class discrimination when only limited labeled samples are available.

The remainder of this paper is organized as follows. Section 2 presents the proposed FSFD framework, including signal representation, TF dual-stream feature extraction, hierarchical selective fusion, and covariance-based metric classification. Section 3 describes the bearing platform, dataset construction, preprocessing procedures, comparison methods, and implementation details. Section 4 provides comprehensive experimental evaluations, including comparative experiments, ablation studies, and detailed analysis of the diagnos-

tic performance under different operating conditions. Finally, Section 5 summarizes the main conclusions.

2. Methodology

2.1. Overall Framework of the Proposed TFDS-MJML

During the operation of RM, critical components remain in healthy conditions for most of their service life. Consequently, fault samples are often costly to acquire, sparsely labeled, and unevenly distributed across classes. To enable stable class discrimination with only a limited quantity of labeled samples, this study formulates diagnostic tasks using an episodic few-shot learning strategy. Unlike conventional supervised training, which directly learns a fixed mapping from inputs to predefined classes, episodic learning simulates the N-way K-shot setting encountered at test time during the training process. Within each episode, the model learns the relationships between query and support samples, thereby improving its generalization capability under data-scarce conditions. The episodic task can be formulated as follows:

$$\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N \times K}, \quad \mathcal{Q} = \{(x_j^q, y_j^q)\}_{j=1}^{N \times Q} \quad (1)$$

Here, $\mathcal{S}$ and $\mathcal{Q}$ denote the support set and query set, respectively. N is the quantity of fault classes included in each episode, K is the quantity of support samples per class, and Q is the quantity of query samples per class. x_i^s and x_j^q represent support and query samples, respectively, while y_i^s and y_j^q denote their corresponding class labels. Given the support set, the objective of the model is to determine the fault class to which each query sample belongs.

The overall workflow of TFDS-MJML is shown in Figure 1. The diagnostic process of the proposed method mainly consists of the following steps:

Step 1: The raw vibration signals collected from the RM are first segmented using a sliding window and subjected to amplitude normalization to obtain fixed-length sample segments. Two input representations, corresponding to the time domain and frequency domain, are then constructed for each sample. An episodic N-way K-shot training scheme is adopted, in which the limited labeled samples are divided into a support set and a query set. The support set provides a small quantity of reference samples for each fault class, whereas the query set simulates the diagnosis of previously unseen samples. By repeatedly constructing support-query episodes during training, the model learns inter-class relationships under a task formulation consistent with that used at test time, thereby improving its adaptability to few-shot bearing FD scenarios.

Step 2: During the feature extraction stage, a TF dual-stream Mamba feature extractor is developed. For the MSTF views, a spatial Mamba branch and a channel attention branch are constructed. The former captures local textures and long-range spatial dependencies in the TF maps, whereas the latter models the relationships among responses associated with different scales, frequency bands, and feature channels. For the raw time-domain view, a time-domain impulsive feature branch is employed to preserve transient fault information that may be attenuated during the TF transformation process.

Step 3: After obtaining the multi-source feature representations, a hierarchical selective fusion mechanism is employed to adaptively integrate the features extracted by diverse branches. Within the TF domain, the spatial Mamba features are first fused with the channel-attention features to form an integrated TF representation. This representation is then fused with the raw time-domain impulsive features at the second fusion stage to obtain the final fault representation. Through this hierarchical process, the contribution of each branch can be dynamically adjusted according to the fault characteristics of individual samples.

Step 4: During the metric-based classification stage, the fused features of support samples belonging to the same class are used to construct a class-specific covariance matrix, which characterizes the distribution structure of each fault class in the feature space. The second-order statistical similarity between the query features and each class-specific covariance matrix is then computed. The resulting local similarity responses for all classes are fed into a 1D convolutional classifier to produce the fault-class scores of the query sample. In this way, TFDS-MJML can effectively identify RM fault states from only a limited quantity of labeled samples.

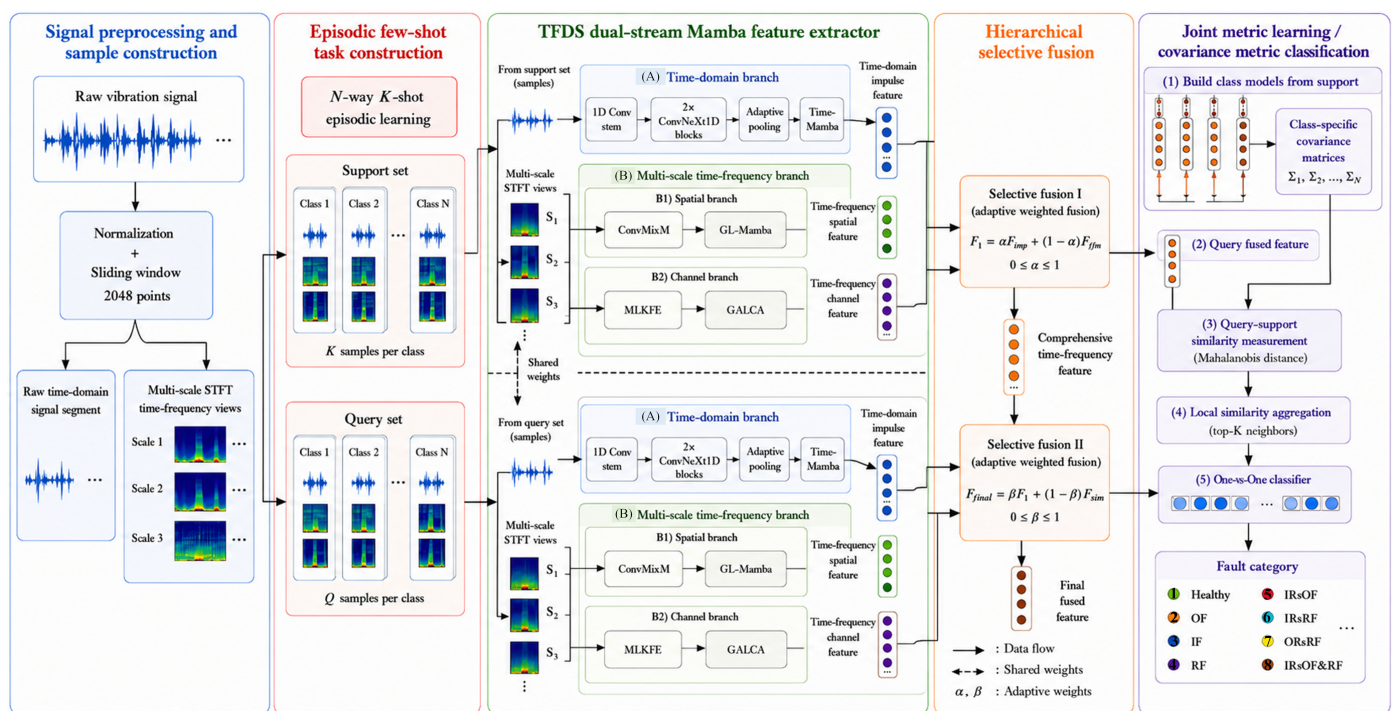


Figure 1. The brief workflow of TFDS-MJML.

2.2. Signal Preprocessing and MSTF View Construction

RM vibration signals generally contain fault-induced impacts, background noise, and nonstationary components caused by variations in rotational speed. When only a single time-domain sequence is used, the model cannot explicitly characterize the distribution of fault-related energy along the frequency axis. Conversely, reliance on a single TF representation may attenuate local transient impact information. Therefore, as shown in Figure 2, both raw time-domain inputs and MSTF inputs are constructed during the preprocessing stage. The acquired continuous vibration signals are first segmented into samples of length L using a sliding window. Amplitude normalization is then performed for each sample along the channel dimension to reduce the effects of differences in acquisition batches and sensor sensitivities. It should be noted that the raw time-domain signal and the multi-scale STFT maps originate from the same sample and therefore do not require additional labeled data. Their difference lies in the form in which fault information is represented. The raw waveform preserves the exact temporal location and amplitude variation of transient impacts, whereas the STFT maps explicitly reveal the temporal evolution of spectral energy. Using several STFT resolutions further reduces dependence on a single TF resolution. Thus, the dual-view construction enriches the fault representation available from each limited labeled sample without increasing the annotation requirement.

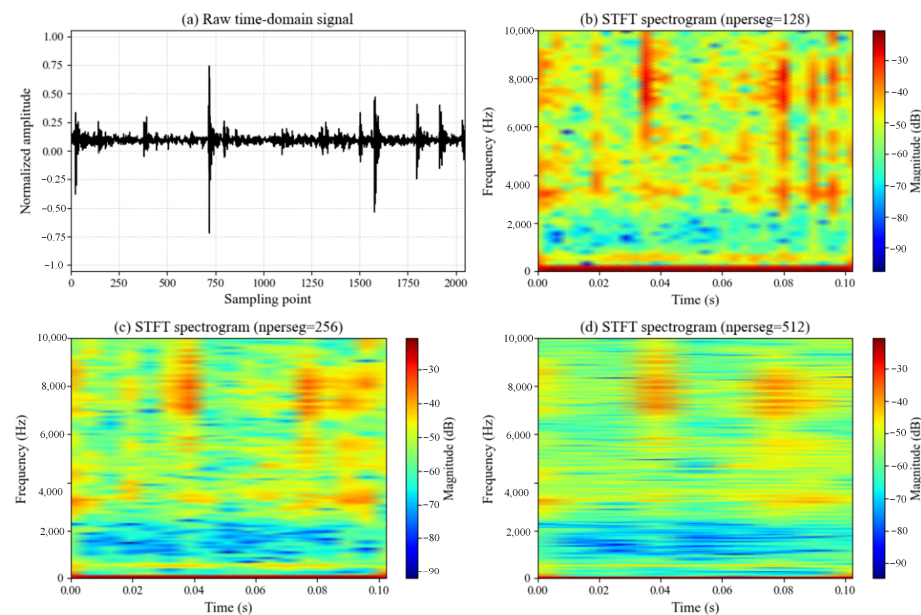


Figure 2. Original time-domain input and MSTF input display.

$$\tilde{x}_{i,c}(n) = 2 \frac{x_{i,c}(n) - \min(x_{i,c})}{\max(x_{i,c}) - \min(x_{i,c}) + \varepsilon} - 1 \quad (2)$$

Here, $x_{i,c}(n)$ represents the n -th sampling point of the i -th sample on the c -th sensor channel. $\tilde{x}_{i,c}(n)$ is the normalized signal value. $\max(x_{i,c})$ and $\min(x_{i,c})$ denote the maximum and minimum values within that sample channel, respectively. And ε is a tiny constant used to prevent division by zero. The normalized samples serve directly as input for the time-domain branch and are also used to generate MSTF representations.

To characterize fault features across varying TF resolutions, a multi-scale STFT is employed to construct a frequency-domain representation. Short windows are more sensitive to local impulses and transient changes, whereas long windows are better suited for capturing stable frequency components and fault modulation patterns. Consequently, concatenating TF maps generated using multiple window lengths along the channel dimension provides a more comprehensive spectral representation for the subsequent network. For a normalized signal $\tilde{x}(n)$, the STFT at the m -th scale can be expressed as follows.

$$X_m(\tau, \omega) = \sum_{n=0}^{L-1} \tilde{x}(n) w_m(n - \tau) e^{-j\omega n} \quad (3)$$

Here, $X_m(\tau, \omega)$ represents the short-time spectrum at time position τ and angular frequency ω for the m -th scale. $w_m(\cdot)$ is the window function corresponding to the m -th window length. L is the sample length. and j is the imaginary unit. In the actual implementation, different nfft values correspond to different time and frequency resolutions, and the transformed power spectrum is further mapped to a logarithmic spectrum and resized to a uniform dimension.

$$\mathcal{T}(x) = \text{Concat}(\Phi_{m_1}(x), \Phi_{m_2}(x), \dots, \Phi_{m_M}(x)) \quad (4)$$

In the equation, $\Phi_m(x)$ represents the 2D TF map obtained at the m -th STFT scale, M is the number of scales, and $\text{Concat}(\cdot)$ denotes concatenation along the channel dimension. For a single-channel vibration signal, using three STFT scales results in an MSTF input represented as $x_{tf} \in \mathbb{R}\{3 \times H_0 \times W_0\}$. For multi-channel signals, the number of input

channels is expanded accordingly to the product of the number of sensor channels and the number of scales.

In the experiment, the STFT window lengths were set to {128, 256, 512}. The sampling frequency was 20 kHz, and the FFT length was equal to the corresponding window length. Thus, the three windows corresponded to time spans of 6.4, 12.8, and 25.6 milliseconds, with nominal frequency intervals of 156.25, 78.13, and 39.06 Hz respectively. They occupied 1/16, 1/8, and 1/4 of each 2048-point sample segment, forming a two-fold multiplicative multi-resolution sequence. The 128-point window emphasizes local transient energy changes, the 256-point window achieves a balance between time and frequency resolution, and the 512-point window can more precisely characterize stable frequency components and modulation structures. At rotational speeds of 1500 and 2100 revolutions per minute, the corresponding rotation periods were 40 and 28.57 milliseconds respectively. Therefore, the selected scales can characterize the fracture responses within different sub-period time ranges without over-averaging multiple rotation periods.

2.3. Time-Domain and MSTF Dual-Stream Mamba Feature Extractor

After obtaining the raw time-domain signals and MSTF maps, a TF dual-stream Mamba feature extractor is constructed, as illustrated in Figure 3, to jointly exploit these two complementary sources of information. The time-domain branch operates directly on the normalized raw vibration sequences and is primarily intended to extract fault-induced impulsive components and short-term fluctuations. The TF branch takes the multi-scale STFT maps as input and learns TF representations from two perspectives, namely spatial dependency modeling and channel-response characterization. To ensure consistency in the subsequent fusion and metric-based classification stages, all branches are designed to produce feature maps with the same quantity of channels and identical spatial dimensions.

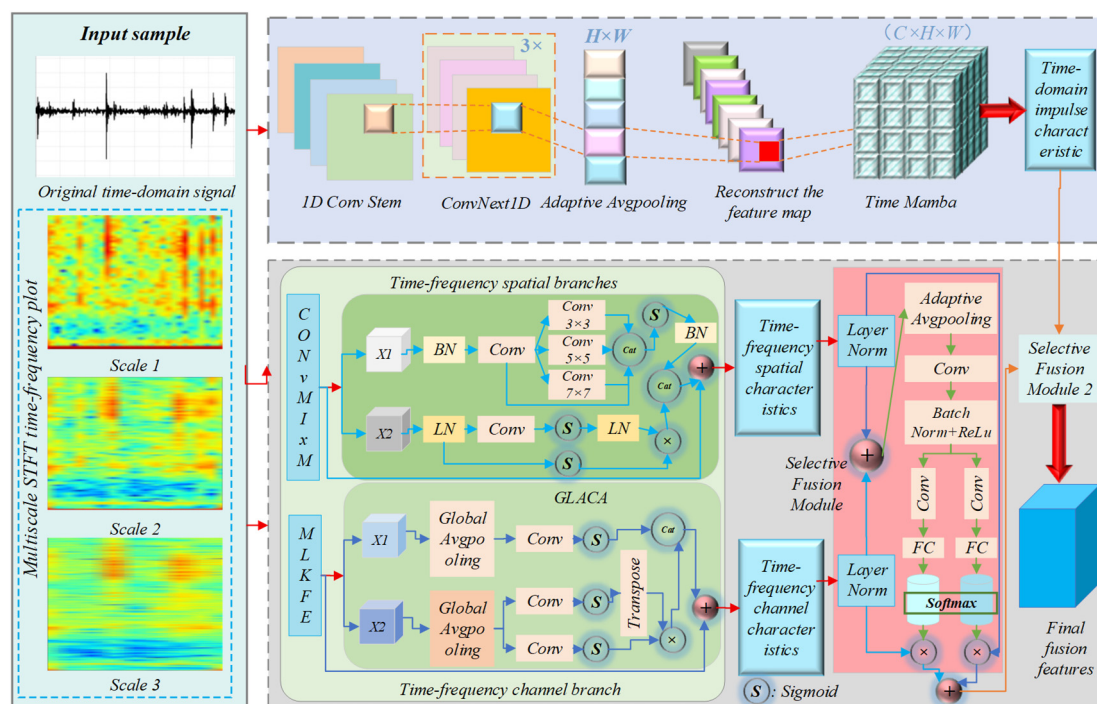


Figure 3. The working process of the TF dual-stream Mamba feature extractor.

2.3.1. Raw Time-Domain Impulse Feature Branch

The raw time-domain branch is designed to compensate for the potential attenuation of transient information during TF transformation. For localized damage in RM, fault

signals often exhibit periodic impacts, localized spikes, and abrupt short-term amplitude variations, which are more directly represented in the raw time-domain sequence. The branch first employs a 1D convolutional stem to model local receptive fields in the input sequence, followed by normalization and nonlinear activation to improve the stability of feature representation. Three stacked ConvNeXt1D blocks are then used to extract deeper temporal patterns. Within each block, depthwise convolution captures local temporal structures, pointwise convolution performs channel mixing, and residual connections help maintain training stability in the deep network.

To align the time-domain features with the 2D feature maps produced by the TF branches, adaptive pooling is applied after 1D feature extraction to compress the sequence length to $H \times W$. The resulting features are then reshaped into a 2D feature map and passed to the Time-Mamba module, which models long-range dependencies among local temporal responses. This process can be formulated as follows:

$$F_t = \mathcal{M}_t(\mathcal{R}(\mathcal{P}(\mathcal{E}_t(x_{raw})))) \quad (5)$$

Here, x_{raw} denotes the normalized raw time-domain signal. $\mathcal{E}_t(\cdot)$ represents the time-domain encoding function composed of a 1D convolutional stem and ConvNeXt1D blocks. $\mathcal{P}(\cdot)$ is an adaptive pooling operation. $\mathcal{R}(\cdot)$ denotes the reconstruction operation transforming 1D features into a 2D feature map. $\mathcal{M}_t(\cdot)$ is the Time-Mamba mapping, and $F_t \in \mathbb{R}^{C \times H \times W}$ represents the final time-domain features.

2.3.2. MSTF Spatial Mamba Branch

The MSTF spatial branch is primarily designed to capture local textures and long-range spatial dependencies in TF maps. First, the multi-scale STFT representation x_{tf} is processed by a convolutional mixing module (ConvMixM) for patch embedding, which maps the input into a feature map with reduced spatial resolution. Compared with conventional stacked convolutions, ConvMixM performs local mixing across neighboring TF regions while retaining a compact structure, facilitating the extraction of spatial patterns around fault-related frequency bands. The resulting feature map is then fed into a Global–Local Mamba (GL-Mamba) module, where local spatial convolution and two-dimensional selective scanning are jointly employed to model both local interactions and global dependencies across the TF plane.

$$F_s = \mathcal{M}_s(\mathcal{E}_s(x_{tf})) \quad (6)$$

In Equation (6), $\mathcal{E}_s(\cdot)$ denotes the ConvMixM spatial encoding function, $\mathcal{M}_s(\cdot)$ denotes the GL-Mamba spatial modeling function, and $F_s \in \mathbb{R}^{C \times H \times W}$ represents the multi-scale spatio-temporal-frequency features.

In GL-Mamba, 2D selective scanning converts feature maps into sequences along multiple directions for state-space modeling, and subsequently aggregates the responses from these different directions into a 2D output. This process can be expressed as:

$$\hat{Z} = \text{SS2D}(Z) = \sum_{r=1}^4 \text{S6}(\text{Scan}_r(Z)) \quad (7)$$

Here, Z denotes the input feature map, $\text{Scan}_r(\cdot)$ represents the scanning operation in the r -th direction with the four directions typically corresponding to horizontal, vertical, and their respective reverse scans, $\text{S6}(\cdot)$ signifies the selective state-space mapping, and $\hat{Z}$ represents the aggregated features capturing spatial dependencies. This architecture enables the establishment of a large receptive field at a low computational cost, making it suitable for processing TF maps characterized by non-stationary energy distributions.

2.3.3. MSTF Channel Attention Branch

Unlike the spatial branch, the MSTF channel attention branch focuses on differences in the contributions of channel responses. Since multi-scale STFT representations organize information associated with different window lengths, sensors, or frequency bands along the channel dimension, individual channels may contribute differently to fault identification. As shown in Figure 4, the branch first employs a multi-scale large-kernel feature extraction (MLKFE) module to extract TF texture features under different receptive fields. A global and local channel attention (GALCA) module is then applied to model both local channel dependencies and global inter-channel relationships. This process can be expressed as follows:

$$F_c = \mathcal{A}_c(\mathcal{E}_c(x_{tf})) \quad (8)$$

In the equation, $\mathcal{E}_c(\cdot)$ denotes the channel-branch encoding function composed of convolution, pooling, and multi-scale large-kernel feature extraction modules. $\mathcal{A}_c(\cdot)$ represents the global-local channel attention map. and $F_c \in \mathbb{R}^{[C \times H \times W]}$ is the channel-enhanced TF feature. MLKFE enhances local texture responses through multi-scale convolutional attention, while GALCA further recalibrates features based on inter-channel relationships, enabling the model to highlight scales, frequency bands, and sensor responses relevant to faults.

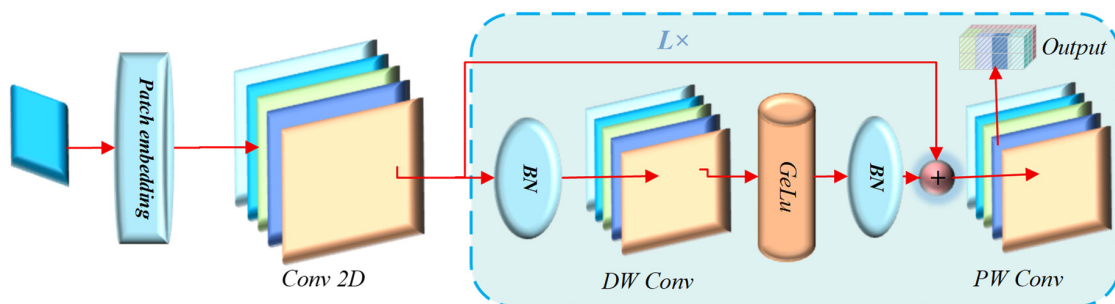


Figure 4. ConvMixM workflow diagram.

The spatial Mamba branch and the channel attention branch are complementary: the former focuses on the spatial correlations within TF spectrograms, while the latter emphasizes modeling the importance of different feature channels. Integrating both branches for MSTF spectrograms enhances the model's ability to represent the complex fault characteristics of bearings.

Although the spatial Mamba branch and the channel-attention branch are designed to extract complementary information from the same MSTF input, their responses are not necessarily consistent for every sample. The spatial branch may strongly activate a localized TF region because of its distinctive texture or long-range dependency, whereas the channel branch may assign a relatively low importance to the corresponding response if it is regarded as redundant or noise-sensitive. Conversely, a feature channel emphasized by the channel branch may not exhibit a sufficiently stable spatial structure. Directly adding or concatenating such inconsistent representations may cause feature-scale imbalance, redundant activation, or interference between fault-related and noise-related responses. These effects may increase intra-class feature variation and reduce the stability of few-shot class discrimination. Therefore, the two parallel features are not directly used as the final TF representation but are first coordinated through the first-stage selective fusion module.

2.3.4. Hierarchical Selective Fusion Module

After multi-branch feature extraction, direct fusion by concatenation or element-wise addition may fail to account for the varying contributions of different branches across fault samples. To address this issue, a hierarchical selective fusion strategy is adopted. The

strategy consists of two Selective Branch modules. At the first stage, the spatial Mamba features F_s and channel-attention features F_c are selectively fused within the TF domain to produce an integrated TF representation F_{tf} . At the second stage, F_{tf} is further fused with the raw time-domain feature F_t , yielding the final fault representation F .

For any two features F_1 and F_2 to be fused, they are first normalized and summed. then, a channel descriptor vector is obtained via global average pooling, followed by the generation of branch weights using a lightweight mapping function and Softmax.

$$[\alpha_1, \alpha_2] = \text{Softmax}(g(\text{GAP}(F_1 + F_2))) \quad (9)$$

$$F_{fuse} = \alpha_1 \odot F_1 + \alpha_2 \odot F_2 \quad (10)$$

In the above equation, $\text{GAP}(\cdot)$ denotes global average pooling, $g(\cdot)$ represents a weight generation function composed of pointwise convolution, batch normalization, and non-linear activation, α_1 and α_2 denote the adaptive weights for the two branches respectively, $\odot$ represents element-wise multiplication, and F_{fuse} represents the fused features. Since the weights are dynamically generated based on the sample's own features, this module can adjust the source of information according to the feature characteristics of different fault samples.

$$F_{tf} = \text{SB}_1(F_s, F_c), \quad F = \text{SB}_2(F_{tf}, F_t) \quad (11)$$

In the equation, $\text{SB}_1(\cdot)$ denotes the first-stage selective fusion module, $\text{SB}_2(\cdot)$ denotes the second-stage selective fusion module, F_{tf} represents the integrated TF features, and $F \in \mathbb{R}^{[C \times H \times W]}$ is the fault representation ultimately used for classification. Through this hierarchical design, the model first performs information filtering within the TF domain before incorporating raw time-domain impulse information, thereby preventing mutual interference between features of different modalities during early fusion.

2.4. Covariance Metric Classification Module

Following dual-stream feature extraction and selective fusion, the model obtains the final feature representations for the query set and the support set. To address the issue of limited sample sizes per class in few-shot scenarios, a covariance-based metric classification strategy is adopted. As illustrated in Figure 5, this strategy does not rely on a fixed classifier to memorize training classes. Instead, it constructs class distribution descriptors using support set samples from the current episode and performs classification based on the consistency between query samples and these class distributions.

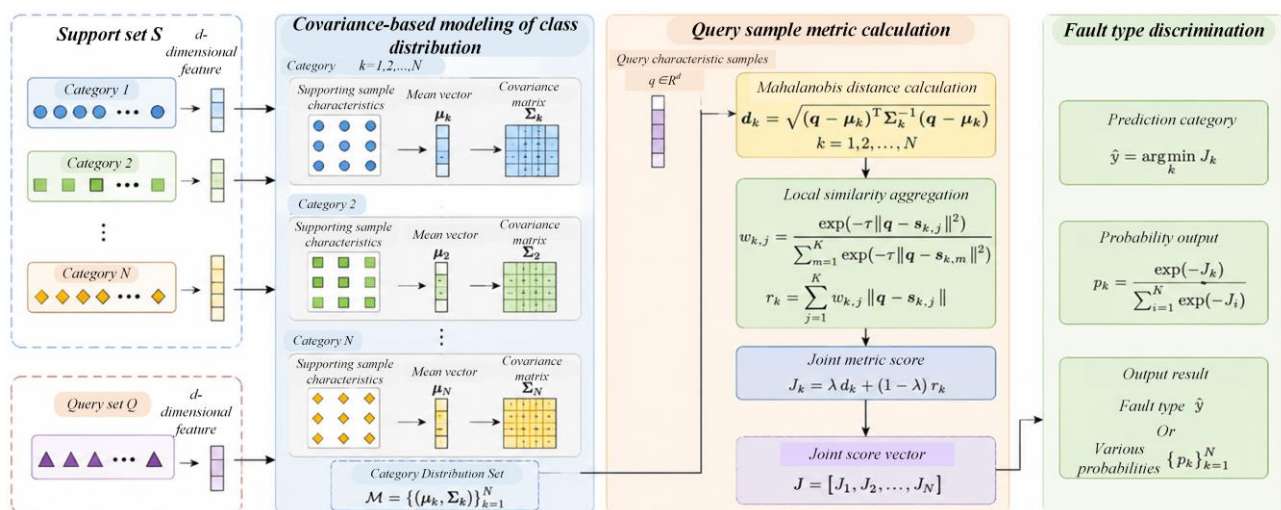


Figure 5. Flowchart of the metric classification strategy based on covariance.

2.4.1. Class-Specific Covariance Matrix Construction

Let C be the quantity of channels and $H \times W$ be the spatial dimensions of the final feature map, and let $d = H \times W$. For the i -th class, there are K samples in the support set, and their fused features can be represented as $F_i^s \in R\{K \times C \times H \times W\}$. By treating the channel response at each spatial location as a local deep descriptor, one obtains K_d local descriptors. Denoting the j -th local descriptor as $z_{i,j}$, $j \in R^C$, the mean vector and covariance matrix for the i -th class are given by:

$$\mu_i = \frac{1}{Kd} \sum_{j=1}^{Kd} z_{i,j} \quad (12)$$

$$\Sigma_i = \frac{1}{Kd-1} \sum_{j=1}^{Kd} (z_{i,j} - \mu_i)(z_{i,j} - \mu_i)^T \quad (13)$$

In Equations (12) and (13), μ_i denotes the mean vector of the local descriptors for support samples of the i -th class, and $\Sigma_i \in R^{(C \times C)}$ represents the covariance matrix for that class. This matrix captures the second-order statistical relationships between features across different channels within the class. Compared to metrics relying solely on class centers, the covariance representation preserves a richer structure of the intra-class distribution.

2.4.2. Query-Support Similarity Measurement and Conv1D Classification

For the query sample, its final feature map is likewise flattened into $Z_q \in R^{(C \times d)}$, where each column corresponds to a local descriptor at a spatial location. To evaluate the consistency between the query sample and the distribution of the support set for class i , a quadratic form is employed to compute the response of the query's local descriptors with respect to the class covariance matrix:

$$r_i = \text{diag}(Z_q^T \Sigma_i Z_q) \quad (14)$$

Here, $r_i \in R^d$ represents the local similarity vector of the query sample with respect to the i -th class, and $\text{diag}(\cdot)$ denotes the operation of extracting the diagonal elements of a matrix. This operation enables the calculation of the matching degree between query features and the second-order statistical distribution of each class at every spatial location, thereby yielding a fine-grained description of local similarity.

Subsequently, the local similarity vectors between the query sample and all N classes are concatenated in the order of the classes, and a 1D convolution is applied to aggregate the d local responses for each class, resulting in the final class scores:

$$R = \text{Concat}(r_1, r_2, \dots, r_N) \quad (15)$$

$$\hat{y} = \text{Conv1D}(R, \text{kernel} = d, \text{stride} = d) \quad (16)$$

In the equation, $R \in R^{[Nd]}$ represents the concatenated similarity vector, and $\hat{y} \in R^N$ in Equation (16) denotes the prediction scores for the query sample across the N fault categories. Since both the convolutional kernel size and the stride are set to d , the 1D convolution compresses the local similarity response corresponding to each category into a single statistical score. This classification approach utilizes the distributional structure of the support set to perform category discrimination. Compared to simple Euclidean distance or cosine similarity, it is better suited for characterizing complex fault feature relationships under few-shot conditions.

2.5. Training Objective and Diagnostic Procedure

TFDS-MJML employs an end-to-end training approach. In each training episode, the model first extracts fused features for each class based on the support set and constructs covariance matrices, then calculates similarity scores between query samples and each class. For a training batch containing B query samples, cross-entropy loss is used to supervise and optimize the class prediction results:

$$\mathcal{L}_{ce} = -\frac{1}{B} \sum_{b=1}^B \log \frac{\exp(\hat{y}_{b,y_b})}{\sum_{n=1}^N \exp(\hat{y}_{b,n})} \quad (17)$$

Here, B denotes the quantity of query samples, N represents the quantity of fault categories, $\hat{y}_{b,n}$ is the predicted score indicating the likelihood that the b -th query sample belongs to the n -th category, and y_b is the true label of that query sample. By minimizing $\mathcal{L}_{ce}$, the feature extractor, the selective fusion module, and the covariance classifier can be jointly optimized under a unified objective.

The diagnostic procedure of the model can be summarized as follows. First, the bearing vibration signals are segmented using a sliding window and normalized. Raw time-domain inputs and MSTF inputs are then constructed for each sample. During training, N -way K -shot tasks are randomly generated, and the shared feature extraction network is used to extract representations from both the support and query sets. The resulting multi-branch features are integrated through the hierarchical selective fusion mechanism to obtain the final fault representation. Query samples are subsequently classified based on the class-specific covariance matrices constructed from the support set. This procedure allows the model to continuously simulate few-shot diagnostic scenarios during training and rapidly identify previously unseen query samples at test time using only a limited number of support samples.

3. Introduction to Datasets and Experiments

3.1. Data Sources and Processing

3.1.1. Bearing Fault Simulation Test Rig and Data Acquisition

To evaluate the reliability of the proposed TFDS-MJML method, data acquisition experiments were conducted using a self-developed bearing fault simulation test rig. As shown in Figure 6, the experimental platform mainly consists of a drive motor, test bearing, braking device, dynamic torque sensor, vibration sensor, control system, cooling system, and signal acquisition system. The drive motor provides a stable rotational speed for the bearing test system, while the braking device is used to adjust the load condition. The dynamic torque sensor monitors torque variations during power transmission, and the vibration sensor, mounted near the test bearing, records vibration responses during bearing operation. The signal acquisition system synchronously records the vibration data throughout the experiments, providing the basis for subsequent fault sample construction and model training. During the experiments, test bearings under different health conditions were installed on the test rig in sequence, and the operating state of the rig was adjusted through the control system. Vibration signals for each health condition were collected after the bearing speed stabilized at 1500 r/min and 2100 r/min, respectively. The sampling frequency was set to 20 kHz, and data were continuously recorded for 60 s under each health condition. Data collected at the two rotational speeds were used to further assess the diagnostic adaptability and stability of the proposed model under different operating conditions.

The relationship between the present experimental procedure and relevant international standards should be clarified. The ISO 13373 [34] series provides general guidance

for machinery vibration data acquisition, processing, analysis, and vibration-based diagnosis, while ISO 5348 [35] describes the mechanical mounting of accelerometers and the influence of mounting conditions on measurement fidelity. These standards provide an engineering reference for vibration monitoring and data interpretation. However, the present study focuses on algorithmic classification using vibration samples acquired from a controlled laboratory test rig and does not establish ISO-based vibration severity thresholds or alarm limits.

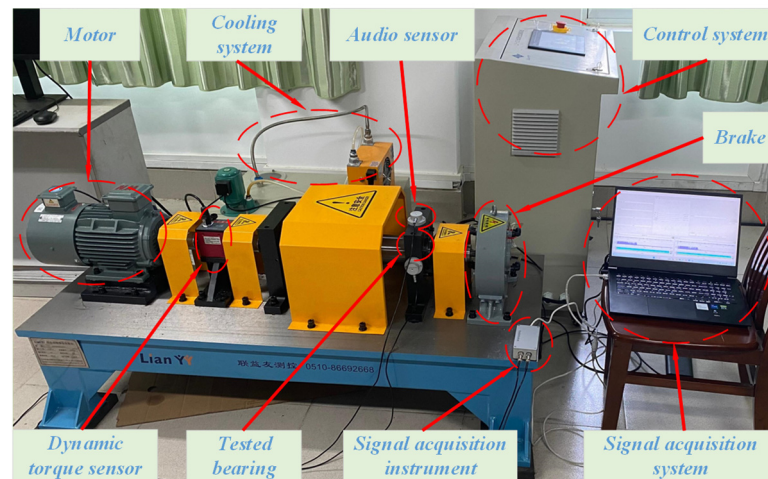


Figure 6. The experimental bench and its components.

To simulate typical localized damage that may occur in bearings during the long-term operation of RM, wire-cut electrical discharge machining was used to introduce localized cracks at various locations on the bearings. These locations included critical components such as the outer race, inner race, and rolling elements. Additionally, composite damage states involving multiple locations were created to more comprehensively capture the variations in vibration characteristics under complex damage conditions. The faulty bearings constructed for the experiment are shown in Figure 7. The dataset comprises eight health states: normal condition, outer race fault, inner race fault, rolling element fault, composite outer race and rolling element fault, composite inner race and rolling element fault, composite inner and outer race fault, and composite inner race, outer race, and rolling element fault. To facilitate model training and experimental analysis, these eight states were assigned numerical labels ranging from 0 to 7.

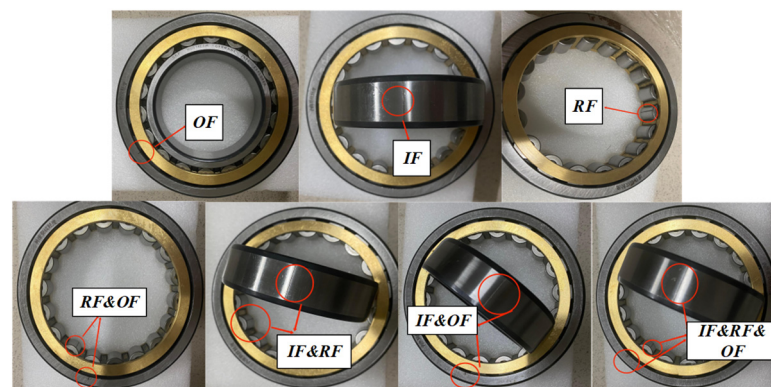


Figure 7. Schematic diagrams of seven different faulty bearings.

ISO 15243 [36] provides standardized terminology and classification principles for rolling-bearing damage and failure modes according to their characteristics and possible

causes. In the present experiment, localized defects were artificially introduced at the inner race, outer race, and rolling elements to construct controlled diagnostic categories. These artificial defects represent predefined laboratory fault locations and should not be regarded as complete reproductions of all in-service damage mechanisms defined in ISO 15243. Therefore, the outputs of TFDS-MJML indicate the predicted experimental health states rather than ISO-certified bearing failure categories.

3.1.2. Data Standardization and Partitioning

After the experimental data acquisition, the raw vibration signals corresponding to various health states are first normalized to the range of $(-1, 1)$ to mitigate amplitude scale discrepancies caused by differing operating conditions and data acquisition batches. Subsequently, a sliding window approach is employed to segment the continuous vibration signals. The window length is set to 2048 sampling points with a step size of 512 points, thereby generating fixed-length bearing vibration samples. The continuous vibration signal is segmented using a sliding window of 2048 points and a step size of 512 points. At a sampling frequency of 20 kHz, each sampling point corresponds to a time interval of 102.4 milliseconds. Since the rotation periods at 1500 revolutions per minute and 2100 revolutions per minute are 40 milliseconds and 28.57 milliseconds respectively, each segmented sample contains approximately 2.56 and 3.58 axial rotations. The selected window length can provide sufficient temporal context to retain the repetitive pulses and modulation patterns caused by faults, and can effectively maintain this even under various operating conditions. A step size of 512 points corresponds to 25.6 milliseconds, with 75% overlap between adjacent sampling points. This overlap strategy can generate samples with different initial phases, reduce the possibility of losing transient pulses near the segmentation boundaries, and provide a sufficiently large candidate sample pool for constructing event-based few-sample tasks. During dataset partitioning, a class-balanced strategy is adopted to construct the training candidate set and the test set. The test set is reserved exclusively for model performance evaluation and is excluded from the training process. The training candidate set is utilized to formulate few-shot diagnostic tasks. Specifically, N-way K-shot tasks—each comprising a support set and a query set—are dynamically constructed during the training phase. The support set provides a small quantity of labeled samples for each fault category, while the query set is used to compute classification loss and update model parameters. Given that the bearing dataset encompasses eight health states, the experiments are configured for two types of few-shot tasks: 8-way 1-shot and 8-way 5-shot. To further analyze the model's diagnostic performance across varying labeled sample sizes, multiple subsets with limited training sample quantities are derived from the training candidate set and evaluated against the same test set.

3.2. Experimental Setup and Implementation

First, the key structural parameters and details of the TFDS-MJML model are outlined to enhance the transparency of the experimental process. Subsequently, ablation studies are conducted on the primary modules of TFDS-MJML to verify the contribution of each strategy to model performance. Finally, to further evaluate the effectiveness of TFDS-MJML, five state-of-the-art FSFD algorithms are selected as benchmarks, and data collected from a bearing test rig under various operating conditions are used to validate the diagnostic advantages and stability of TFDS-MJML.

3.2.1. Model Implementation Details

In the TFDS-MJML model, the normalized bearing vibration samples are simultaneously represented as raw time-domain inputs and MSTF inputs. The TF map at each scale is resized to 64×64 . For a single-channel vibration signal, the resulting MSTF input has a

dimension of $3 \times 64 \times 64$. During feature extraction, the time-domain branch consists of a two-layer 1D convolutional stem, three ConvNeXt1D blocks, adaptive average pooling, and a Time-Mamba module. This branch is used to extract impulsive components and short-term fluctuation patterns from the raw vibration signals, after which the output is reshaped into a two-dimensional feature map of size $64 \times 16 \times 16$. The MSTF stream contains a spatial modeling branch and a channel enhancement branch. The former is composed of ConvMixM and GL-Mamba modules, whereas the latter consists of MLKFE and GALCA modules. Both branches produce feature representations of size $64 \times 16 \times 16$. The model subsequently employs a two-stage Selective Branch fusion mechanism. TF spatial features are first fused with TF channel features, after which the integrated TF representation is further fused with the raw time-domain features to obtain the final fault representation. During classification, the covariance metric module constructs class-specific covariance matrices from support samples of the same class and evaluates the similarity between each query sample and the corresponding class distributions. A 1D convolutional classifier is then used to aggregate the local similarity responses associated with each class, with both the kernel size and stride set to $16 \times 16 = 256$. The experiments are conducted under two few-shot task settings, namely 8-way 1-shot and 8-way 5-shot. Cross-entropy loss is used for end-to-end optimization during training. Model parameters are updated using the AdamW optimizer, with an initial learning rate of 1×10^{-3} and a weight decay coefficient of 1×10^{-4} . The learning rate is reduced every 10 epochs by a decay factor of 0.5, and the total quantity of training epochs is set to 100. To ensure a fair comparison, all experiments are conducted using identical data partitions and parameter settings, and the final reported results are obtained by averaging multiple independent runs.

3.2.2. Ablation Study Design Plan

To evaluate the contribution of the key components in TFDS-MJML to few-shot bearing FD, two groups of ablation studies were designed. First, the effectiveness of the TF dual-stream representation strategy was investigated by comparing different combinations of input representations and feature extraction branches. The evaluated configurations included the raw time-domain branch only, the single-scale TF branch only, the MSTF branch only, the model without the raw time-domain impulsive feature branch, and the complete TF dual-stream model. This group of experiments was designed to examine the role of raw time-domain signals in preserving fault-induced impulsive information, the contribution of multi-scale STFT representations to characterizing energy distributions across different frequency bands, and the benefit of complementary fusion between time-domain and TF features for few-shot diagnostic performance. Secondly, ablation experiments were conducted on the hierarchical selective fusion mechanism and the covariance metric classification module. In these experiments, the two-stage selective branch fusion module was replaced with channel concatenation, adaptive gated fusion, and adaptive interactive fusion. The spatial-channel selective fusion within the TF domain was removed, and the second-stage selective fusion between the time domain and the TF domain was cancelled. The covariance metric classification module was replaced with a prototype classifier based on Euclidean distance. Comparisons with the complete model were used to assess the role of selective fusion in dynamically adjusting the contributions of different branches and the effectiveness of the covariance metric module in exploiting second-order statistical information from the support set to improve few-shot class discrimination. All ablation studies were conducted using the same data partitions, number of training samples, and few-shot task settings. Subsequently, a sensitivity analysis was conducted on the configuration of the STFT window length to quantitatively evaluate its impact on diagnostic performance using a small sample size. Four combinations of three scale window lengths were compared:

64/128/256, 128/256/512, 256/512/1024, and 64/256/1024. The number of STFT scales was fixed at three to avoid interference between the window length and the quantity of input scales. Finally, a controlled backbone network replacement experiment was conducted to compare the feature modeling module based on Mamba with the mainstream Transformer encoder. By combining Time-Mamba or 1D Transformer in the time-domain branch, and GL-Mamba or 2D Transformer in the TF space branch, four configurations were constructed. For the Transformer variants, the input and output feature dimensions and the number of encoding blocks, as well as the downstream fusion and metric classification modules, were all consistent with the corresponding Mamba variants. To ensure fair comparisons, each ablation model modified only the component under investigation, while the remaining network architecture, training parameters, and testing procedure were kept unchanged. The 8-way 5-shot task under the 1500 r/min operating condition was selected as the primary experimental setting for analysis. By comparing the variations in diagnostic accuracy across different ablation configurations, the effectiveness of the proposed dual-domain feature representation, hierarchical selective fusion, and covariance metric classification strategies can be systematically evaluated.

3.2.3. Introduction to Few-Shot Contrastive Models

To demonstrate the superiority of TFDS-MJML over existing few-shot models, this paper selects five few-shot algorithms for comparison. All the comparative algorithms utilize the same datasets as the TFDS-MJML model. The selected algorithms and brief descriptions of them are presented in Table 1.

Table 1. Comparison models and their descriptions.

Methods	Description
Cross Attention Net [37]	A few-shot classification framework that introduces cross-attention interactions between support and query feature maps to enhance target-related discriminative regions and improve metric-based recognition under limited labeled samples.
MWPN [38]	A multiscale wavelet prototypical network that combines wavelet convolution and prototype-based metric learning to extract scale-sensitive vibration features and improve FSFD across operating conditions and components.
MDTFS [39]	A TF dual-branch few-shot diagnostic network that integrates time-domain and TF feature extraction with ensemble metric learning to jointly exploit local temporal variations and spectral-temporal fault patterns.
FS-TFD-MEN [40]	A multi-source multi-view task-guided FSFD method that constructs complementary views and adaptively fuses branch features according to task-level discriminative information.
DCA-BiGRU [41]	A dual-channel attention-enhanced BiGRU network that combines feature attention with bidirectional temporal modeling to capture fault-sensitive sequential dependencies for FSFD.

4. Ablation Experiment and Comparative Test Analysis

4.1. Ablation Experiment

4.1.1. Experiment 1: The Impact of TF Dual-Stream Feature Characterization Strategy

To demonstrate the impact of the proposed TF dual-stream feature representation strategy on few-shot bearing FD performance, an ablation study with different input configurations and feature extraction branch combinations was conducted. In these experiments, only the input views and their corresponding feature extraction branches were modified, while other settings, including network training parameters, data partition strategy, episodic task construction scheme, and covariance metric classification module, were kept unchanged. Specifically, A1 retains only the raw time-domain branch to evaluate

the diagnostic contribution of impulse responses and short-term fluctuations contained in raw vibration signals. A2 employs only a single-scale STFT-based TF branch to verify the fault representation capability under a single TF resolution. A3 adopts a multi-scale STFT-based TF branch to analyze the enhancement effect of different TF resolutions on fault feature representation. A4 introduces a single-scale TF branch in addition to the raw time-domain branch to investigate the complementary effect between time-domain features and single-scale TF features. A5 represents the complete model, which simultaneously utilizes the raw time-domain branch and the MSTF branch, and obtains the final fault representation through the hierarchical selective fusion mechanism.

As shown in Table 2, the A5 configuration achieves the best diagnostic performance with an accuracy of 98.75%. In contrast, the A1 model using only the raw time-domain branch obtains a relatively lower accuracy of 82.25%. This indicates that although raw vibration signals preserve fault-induced impulses and local transient responses, relying solely on time-domain information is insufficient to fully characterize the differences among complex fault states under few-shot conditions. Compared with A1, A2 achieves improved performance, demonstrating that STFT-based TF representations can provide additional fault-related information, including frequency-band energy distributions and time-varying spectral patterns. However, the diagnostic capability of A2 remains limited because a single STFT scale cannot simultaneously capture transient impulse characteristics and stable frequency components. After introducing multi-scale STFT representations, A3 further improves the diagnostic accuracy to 89.38%, indicating that TF views generated with different window lengths can provide complementary fault information from multiple TF resolutions and enhance the discriminative capability for different fault categories. Furthermore, by comparing A3 and A5, it can be observed that incorporating the raw time-domain impulse feature branch in addition to the multi-scale TF branch enables the complete model to achieve the best diagnostic performance under different training sample scales. This demonstrates that raw time-domain information and multi-scale TF information exhibit strong complementarity. Their combination allows the model to simultaneously preserve fault-induced impulse responses and frequency-domain energy distribution characteristics, thereby improving the performance of few-shot bearing FD.

Table 2. The experimental results of the ablation experiment 1.

	Original Time- Domain Branch	Single- Scale STFT	Multiscale STFT	Time- Frequency Spatial Branches	Time- Frequency Channel Branch	Time- Frequency Fusion	Accuracy (%)
A1	✓						82.50
A2		✓		✓	✓		85.63
A3			✓	✓	✓		89.38
A4	✓	✓		✓	✓	✓	92.25
A5	✓		✓	✓	✓	✓	98.75

4.1.2. Experiment 2: The Impact of Hierarchical Fusion and Covariance Measurement Module

To further verify the contribution of the hierarchical selective fusion mechanism and covariance metric classification module to the diagnostic performance of TFDS-MJML, several structural variants were constructed and compared with the complete model while keeping the input configuration, feature extraction backbone, data partition strategy, and training parameters unchanged. Specifically, B1 replaces the two-stage selective branch fusion modules with channel concatenation to evaluate the effectiveness of the adaptive fusion strategy compared with fixed fusion approaches. B2 employs Sigmoid-gated fusion,

generating learnable gated mappings from the two input features to dynamically control their relative contributions. B3, on the other hand, adopts cross-attention fusion, explicitly modeling the interaction between the two feature branches through the query-key-value attention mechanism. B4 removes the spatial-channel selective fusion within the TF domain and only performs direct integration of spatial Mamba features and channel attention features. B5 eliminates the second-stage selective fusion between the TF domains to analyze the importance of dynamically integrating raw time-domain impulse features and multi-scale TF features. B6 replaces the covariance metric classification module with a Euclidean distance-based prototypical classifier to investigate the role of second-order statistical distribution modeling in few-shot classification. B7 represents the complete TFDS-MJML model.

Figure 8 presents the experimental results of different ablation configurations. It can be observed that when the selective fusion mechanism is replaced by other fusion strategies, the accuracy rates of B1-B3 all decrease, significantly lower than that of the complete model. This indicates that the concatenation method of B1-B3 cannot dynamically adjust the contribution of different feature branches based on the features of individual fault samples, resulting in information redundancy and a poor feature fusion effect. Compared with the complete model, both B4 and B5 exhibit performance degradation, achieving accuracies of only 91.25% and 90.50%, respectively. These results demonstrate that both the spatial-channel complementary relationship within the TF domain and the cross-domain complementarity between time-domain and frequency-domain representations play important roles in fault feature learning. Among them, B5 shows a more significant performance decline after removing the second-stage selective fusion, indicating that raw time-domain impulse features and MSTF features cannot be fully exploited through simple combination. Instead, an adaptive fusion strategy is required to effectively coordinate the complementary information between different modalities. Furthermore, B6 exhibits the most significant performance degradation after replacing the covariance metric module with the Euclidean distance-based prototype classifier. This result suggests that under few-shot conditions, relying only on first-order class mean representations is insufficient to characterize the intra-class distribution variations of support samples. In contrast, the covariance metric module can exploit second-order statistical relationships among support-set features, thereby improving the matching capability between query samples and class distributions. Overall, the complete TFDS-MJML model consistently achieves the best performance under different training sample scales, validating the rationality and effectiveness of the proposed hierarchical selective fusion mechanism and covariance metric classification module.

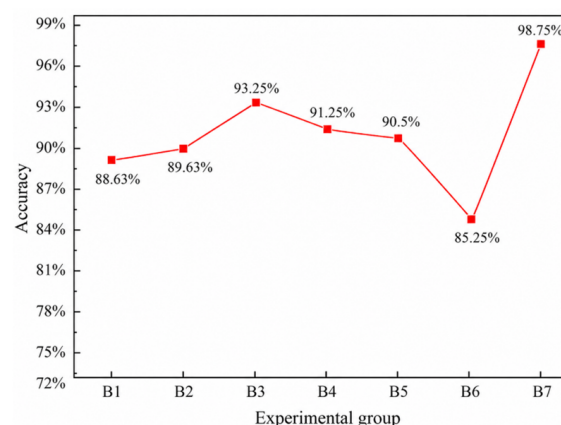


Figure 8. Bar chart of the results of the ablation experiment 2.

4.1.3. Sensitivity Analysis of STFT Window-Length Combinations

To quantitatively evaluate the impact of STFT window length on the diagnostic performance with a small sample size, four tri-scale configurations were compared: 64/128/256, 128/256/512, 256/512/1024, and 64/256/1024. To avoid interference between the window length and the number of scales, the number of scales in the STFT was fixed at three. Except for the window length configuration, all other settings remained unchanged. As shown in Table 3, different combinations of STFT window lengths can lead to measurable changes in the diagnostic accuracy of a small sample. The 64/128/256 configuration performs better in terms of time positioning, but has relatively weaker frequency resolution. Compared with the selected configuration, the accuracy decreased by 8.37% percentage points. The 256/512/1024 configuration increases the frequency component density, but reduces the retention ability for local transient changes, resulting in an accuracy decrease of 6.62% percentage points. The 64/256/1024 configuration covers a wider resolution range, but the intervals between adjacent scales are too large, weakening the continuity of the multi-scale representation. Among the various candidate schemes evaluated, the 128/256/512 configuration shows the highest overall performance. These results indicate that the selected window lengths achieve an appropriate balance between time positioning, frequency resolution, and multi-scale complementarity.

Table 3. The experimental results of the ablation experiment 3.

Window Lengths	64	128	256	512	1024	Accuracy (%)
C1	✓	✓	✓			90.38
C2		✓	✓	✓		98.75
C3			✓	✓	✓	92.13
C4	✓		✓		✓	88.50

4.1.4. Comparison Between Mamba and Transformer Backbones

In order to investigate whether the performance improvement of TFDS-MJML is related to the feature modeling strategy based on Mamba, we conducted a controlled backbone network replacement experiment. The complete model was compared with three variants, which respectively replaced Time-Mamba and/or GL-Mamba with standard Transformer encoders.

As shown in Table 4, replacing either Time-Mamba or GL-Mamba with the corresponding Transformer encoder leads to measurable changes in diagnostic performance and computational cost. The comparison between D2 and D1 reflects the contribution of Time-Mamba, while the comparison between D3 and D1 demonstrates the effect of GL-Mamba. The complete D4 configuration achieves the best overall trade-off between few-shot accuracy and computational efficiency. These results suggest that the performance gain does not merely originate from the dual-stream structure but is also associated with the selective state-space modeling mechanisms used in the temporal and TF spatial branches.

Table 4. The experimental results of the ablation experiment 4.

Configuration	Time Branch	TF Spatial Branch	Training Time per Epoch (s)	Accuracy (%)
D1	1D Transformer	2D Transformer	16.12	85.25
D2	Time-Mamba	2D Transformer	14.89	89.88
D3	1D Transformer	GL-Mamba	13.68	92.38
D4	Time-Mamba	GL-Mamba	10.25	98.75

4.2. Comparative Test

To further validate the effectiveness of the proposed TFDS-MJML model for few-shot RM FD, five representative FSFD models were selected as comparison methods, and comparative experiments were conducted using the bearing fault simulation test rig dataset. To fully exploit the collected data, experiments were carried out under two rotational speed conditions, where Experiment A represents the dataset collected at 1500 r/min and Experiment B represents the dataset collected at 2100 r/min. Meanwhile, experiments were performed under both 8-way 1-shot and 8-way 5-shot scenarios, with three different training sample scales of 48, 240, and 8000 samples, respectively, to investigate the diagnostic performance of different models under varying numbers of labeled samples. To ensure the fairness of the comparative experiments, all models were evaluated using the same data partition strategy, training sample scales, and few-shot task configurations. Each experiment was repeated five times, and the average accuracy, recall, and F1-score were adopted as the final evaluation metrics.

As shown in Tables 5 and 6, the proposed method consistently achieves superior diagnostic performance compared with the comparison models under different rotational speed conditions and few-shot task settings. This demonstrates that the proposed TF dual-stream feature extraction architecture and covariance metric-based classification strategy can effectively improve the fault recognition capability under limited labeled samples. From the perspective of training sample variations, the accuracy, recall, and F1-score of all models exhibit different degrees of improvement as the total number of training samples increases from 48 to 240 and 8000, indicating that additional labeled samples can enhance the learning of fault category boundaries. Specifically, under the 1500 r/min condition, the accuracy of TFDS-MJML increases from 93.88% to 96.75% and 99.23% in the 8-way 1-shot scenario, while that in the 8-way 5-shot scenario improves from 96.75% to 98.38% and 99.58%. Under the 2100 r/min condition, the accuracy increases from 92.63% to 95.88% and 98.88% in the 1-shot scenario, and from 95.88% to 97.75% and 99.38% in the 5-shot scenario. These results indicate that the proposed method not only maintains high diagnostic accuracy under extremely limited sample conditions but also benefits from additional training samples for further performance improvement. Furthermore, the overall performance of all models under the 5-shot scenario is higher than that under the 1-shot scenario. This is mainly because more support samples per class provide more reliable class representations and improve the estimation accuracy of class distributions. However, as the number of training samples increases, the performance gap between the 1-shot and 5-shot scenarios gradually decreases. This phenomenon suggests that TFDS-MJML can effectively exploit additional support samples to enhance category distribution estimation under extremely limited data conditions, while the learned fault representations become increasingly stable when sufficient training samples are available. A comparison between the two rotational speed conditions shows that the diagnostic performance of all models under 2100 r/min is slightly lower than that under 1500 r/min. This is mainly attributed to the fact that higher rotational speeds introduce more complex dynamic responses in bearing vibration signals, increasing the coupling among fault impulses, background vibrations, and frequency-band energy distributions, thereby making fault identification more challenging. Nevertheless, TFDS-MJML still maintains competitive diagnostic performance under the 2100 r/min condition, demonstrating its strong adaptability and robustness to different operating speeds. Among the comparison methods, FS-TFD-MEN and Cross-AttentionNet achieve relatively better performance in most experimental settings. Specifically, FS-TFD-MEN improves few-shot feature representation capability through multi-view feature construction and task-related feature fusion, achieving an accuracy of 92.13% under the 8-way 5-shot scenario with 48 training samples at 1500 r/min. Cross-AttentionNet enhances target-

related feature representation by exploiting cross-attention interactions between support and query samples, showing relatively stable performance under different sample scales. However, compared with TFDS-MJML, these methods still exhibit noticeable performance degradation under extremely limited sample conditions. This further demonstrates that TFDS-MJML possesses stronger fault feature extraction and category matching capabilities when only a small number of labeled samples are available.

The results in Tables 5 and 6 demonstrate that the proposed method maintains significant diagnostic advantages under different training sample scales, shot configurations, and rotational speed conditions. This superiority can be mainly attributed to the following aspects. First, the raw time-domain branch preserves the impulse responses and short-term fluctuation information contained in bearing fault signals. Second, the multi-scale STFT branch characterizes fault-related frequency-band energy distributions from multiple TF resolutions. Furthermore, the hierarchical selective fusion mechanism dynamically adjusts the contributions of time-domain information, TF spatial features, and channel responses according to sample characteristics, thereby enhancing the complementarity among different feature sources. Finally, the covariance metric module exploits the second-order statistical distribution information of support-set features to characterize class structures. Compared with prototype-based metric approaches that rely only on class mean representations, the proposed module can more comprehensively capture intra-class variations under few-shot conditions. Therefore, the proposed method is capable of constructing a more stable and discriminative fault feature space with limited labeled samples, thereby improving the diagnostic accuracy and generalization capability of few-shot bearing FD.

Table 5. The accuracy rate, recall rate and F1 index of each model under a rotational speed of 1500 rpm in different task scenarios.

(A) 48 TRAINING SAMPLES						
METHOD	Accuracy		Recall		F1	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
CROSS ATTENTION NET	87.38	90.75	86.94	90.31	87.12	90.46
MWPN	84.63	88.25	84.11	87.80	84.32	88.01
MDTFS	86.50	89.63	85.96	89.18	86.21	89.37
FS-TFD-MEN	89.25	92.13	88.87	91.75	89.03	91.94
DCA-BIGRU	82.75	86.38	82.16	85.91	82.43	86.13
PROPOSED	93.88	96.75	93.52	96.41	93.69	96.58
(B) 240 TRAINING SAMPLES						
METHOD	Accuracy		Recall		F1	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
CROSS ATTENTION NET	91.50	94.13	91.06	93.72	91.27	93.91
MWPN	89.38	92.00	88.84	91.57	89.12	91.76
MDTFS	90.75	93.25	90.32	92.88	90.51	93.03
FS-TFD-MEN	93.25	95.63	92.91	95.24	93.07	95.42
DCA-BIGRU	87.25	90.38	86.71	89.92	86.95	90.14
PROPOSED	96.75	98.38	96.43	98.07	96.58	98.22
(C) 8000 TRAINING SAMPLES						
METHOD	Accuracy		Recall		F1	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
CROSS ATTENTION NET	96.63	97.88	96.32	97.61	96.46	97.73
MWPN	95.25	96.75	94.88	96.44	95.05	96.58
MDTFS	96.00	97.25	95.67	96.93	95.82	97.08
FS-TFD-MEN	97.38	98.50	97.12	98.23	97.25	98.36
DCA-BIGRU	93.88	95.63	93.41	95.20	93.63	95.39
PROPOSED	99.23	99.58	99.05	99.43	99.14	99.50

Table 6. The accuracy rate, recall rate and F1 index of each model under a rotational speed of 2100 rpm in different task scenarios.

(A) 48 TRAINING SAMPLES						
METHOD	Accuracy		Recall		F1	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
CROSS ATTENTION NET	85.75	89.38	85.22	88.91	85.48	89.12
MWPN	83.25	86.88	82.70	86.43	82.96	86.64
MDTFS	84.63	88.25	84.13	87.76	84.36	88.00
FS-TFD-MEN	87.50	91.00	87.08	90.62	87.26	90.80
DCA-BIGRU	80.88	84.75	80.31	84.23	80.56	84.47
PROPOSED	92.63	95.88	92.25	95.55	92.43	95.70
(B) 240 TRAINING SAMPLES						
METHOD	Accuracy		Recall		F1	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
CROSS ATTENTION NET	90.13	93.25	89.67	92.84	89.88	93.03
MWPN	87.88	91.13	87.32	90.68	87.59	90.88
MDTFS	89.25	92.38	88.80	91.95	89.01	92.16
FS-TFD-MEN	91.88	94.75	91.43	94.38	91.64	94.56
DCA-BIGRU	85.75	89.25	85.14	88.72	85.42	88.96
PROPOSED	95.88	97.75	95.50	97.42	95.68	97.58
(C) 8000 TRAINING SAMPLES						
METHOD	Accuracy		Recall		F1	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
CROSS ATTENTION NET	95.38	97.13	95.02	96.86	95.19	96.99
MWPN	94.00	95.88	93.57	95.52	93.78	95.69
MDTFS	94.88	96.50	94.46	96.16	94.66	96.31
FS-TFD-MEN	96.25	97.88	95.91	97.57	96.08	97.72
DCA-BIGRU	92.38	94.50	91.86	94.05	92.10	94.26
PROPOSED	98.88	99.38	98.70	99.21	98.79	99.29

To provide a more intuitive comparison of the diagnostic performance of different models under various training sample scales, Figures 9 and 10 present the three-dimensional bar charts of the best diagnostic accuracies achieved by each model under the 8-way 1-shot and 8-way 5-shot scenarios at 1500 r/min and 2100 r/min, respectively. As shown in the figures, the diagnostic accuracies of all comparison models exhibit an overall increasing trend as the total number of training samples increases from 48 to 240 and 8000. This indicates that increasing the number of training samples can enhance the models' ability to learn bearing fault characteristics and improve the stability of class decision boundaries. Meanwhile, the accuracies of most models under the 5-shot scenario are higher than those under the 1-shot scenario, demonstrating that increasing the number of support samples per class enables more reliable class representations and improves the matching accuracy of query samples.

Under the 1500 r/min operating condition, TFDS-MJML achieves the highest diagnostic accuracy across all training sample scales. When the total number of training samples is only 48, TFDS-MJML achieves accuracies of 95.13% and 97.38% under the 1-shot and 5-shot scenarios, respectively, which are significantly higher than those of other comparison models. In contrast, although the comparison models gradually improve with increasing training sample size, their overall performance remains inferior to TFDS-MJML. Under the 2100 r/min operating condition, all models maintain relatively competitive diagnostic performance, while TFDS-MJML still achieves the best recognition results. These results demonstrate that the proposed model not only possesses strong fault feature extraction capability under extremely limited sample conditions but also maintains favorable generalization performance under different rotational speed conditions.

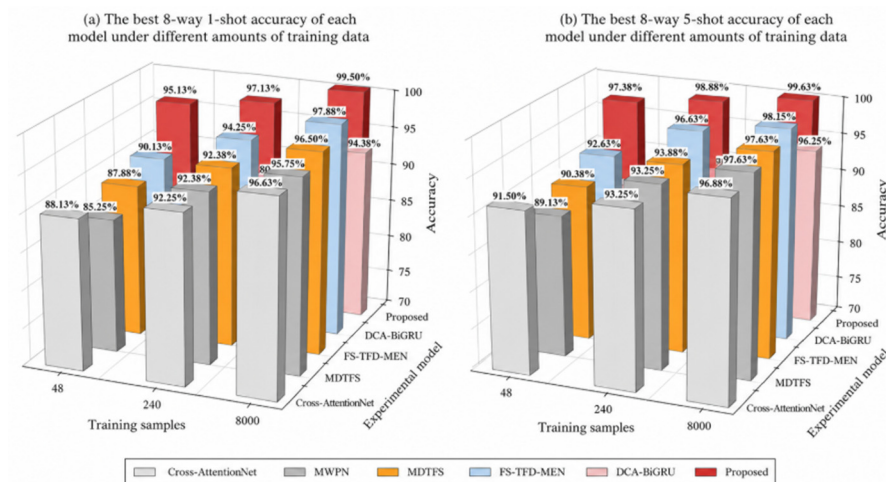


Figure 9. The optimal diagnostic accuracy rates of each model under 1500 rpm for different numbers of training samples.

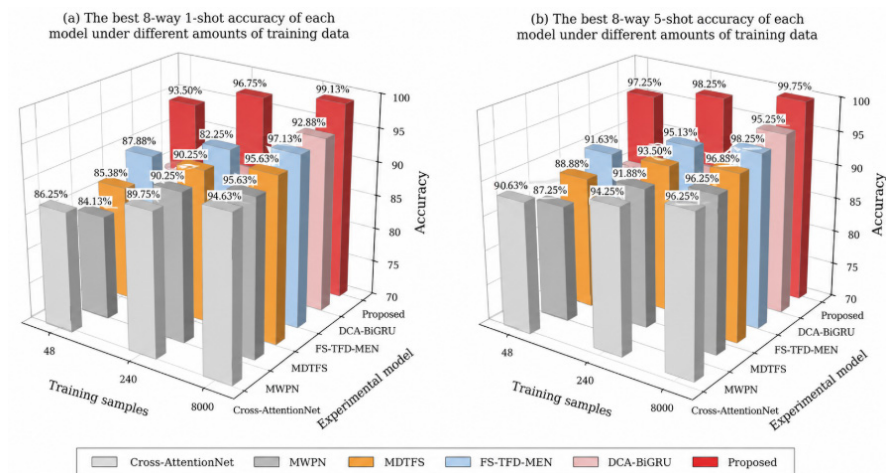


Figure 10. The optimal diagnostic accuracy rates of each model under 2100 rpm for different numbers of training samples.

To further investigate the differences in recognition performance among various models across different bearing health states, Figure 11 presents the radar charts of classification accuracies achieved by each model under the 8-way 1-shot and 8-way 5-shot scenarios at 1500 r/min. The vertices of the radar charts correspond to the eight health states, namely Normal, OF, IF, RF, RF&OF, IF&RF, IF&OF, and IF&RF&OF. A radar curve closer to the outer boundary indicates higher recognition accuracy for the corresponding health state. As illustrated in Figure 11, the radar curves of TFDS-MJML consistently remain closer to the outer boundary under both 1-shot and 5-shot tasks, while maintaining a relatively balanced distribution across different fault categories. In contrast, other comparison methods exhibit varying degrees of contraction in certain categories, indicating that their recognition capability for complex fault conditions remains limited. These results further demonstrate that TFDS-MJML not only improves the overall diagnostic accuracy but also enhances the balance of recognition performance among different bearing health states.

In summary, the advantages of TFDS-MJML are reflected not only in the overall average diagnostic accuracy but also in its more balanced recognition capability across different health states. This can be attributed to the fact that TFDS-MJML simultaneously exploits the raw time-domain branch to preserve fault-induced impulse responses, utilizes the MSTF branch to extract frequency-band energy distribution characteristics, and enhances the complementarity among different feature sources through the hierarchical selective

fusion mechanism. Meanwhile, the covariance metric module leverages the second-order statistical distribution information of support-set samples to characterize class structures, thereby improving the matching accuracy between query samples and different fault categories. Consequently, the proposed method exhibits strong discriminative capability and diagnostic stability for both single-component faults and compound fault categories.

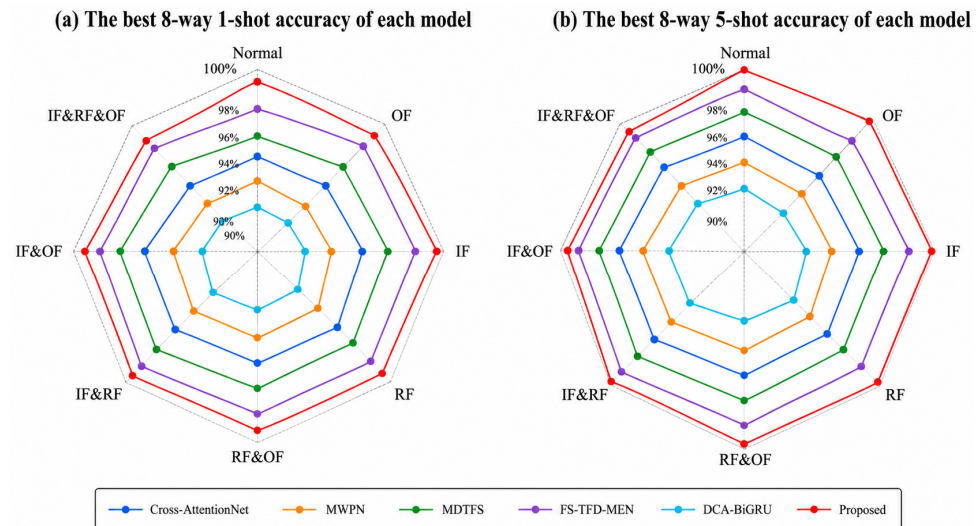


Figure 11. The diagnostic accuracy rates of each model for different health conditions of the bearings at 1500 rpm.

5. Conclusions

To address the challenges of limited fault sample acquisition, insufficient feature representation of single-domain signals, and the limited capability of conventional metric-based classifiers in characterizing class distributions for few-shot RM FD, this paper proposes a FSFD method for RM based on a TF dual-stream Mamba network and joint metric learning. Taking vibration signals of RM as inputs, the proposed method jointly exploits raw time-domain signals and MSTF views to characterize fault states. By integrating time-domain impulse feature extraction, MSTF Mamba modeling, hierarchical selective fusion, and joint metric-based classification, the proposed method achieves multi-class fault identification under limited labeled sample conditions. The main contributions and conclusions of this study are summarized as follows:

- (1) A TF dual-stream representation method is proposed for RM FD. While preserving transient impulses and periodic fluctuation information contained in the raw vibration signals, as well as frequency-band energy distribution characteristics embedded in MSTF representations, the proposed method provides complementary descriptions of different fault states from both temporal responses and TF structures. This strategy effectively alleviates the limitation of single-domain representations in comprehensively characterizing complex fault features.
- (2) A feature extraction architecture consisting of a time-domain impulse feature branch, an MSTF spatial Mamba branch, and a TF channel attention branch is constructed. Specifically, the time-domain branch enhances the representation capability of fault-related impulses and local fluctuations; the TF spatial Mamba branch captures local structural patterns and long-range dependencies within TF representations; and the channel attention branch emphasizes fault-sensitive frequency bands and critical response channels. This architecture improves the feature representation capability of the proposed model under FSFD scenarios.

- (3) A hierarchical selective fusion mechanism and a joint metric learning-based classification strategy are developed. The hierarchical selective fusion mechanism first adaptively integrates spatial and channel features in the TF domain, and subsequently combines time-domain features with comprehensive TF features, enabling the model to dynamically adjust the contributions of different feature branches according to specific few-shot diagnostic tasks. The proposed joint metric learning strategy introduces class-specific covariance distributions from the support set and local similarity aggregation, which not only considers class-center information but also exploits second-order statistical structures within limited support samples to characterize class distributions. Consequently, the stability of matching between query samples and support-set categories is improved.
- (4) Extensive experiments, including ablation studies, comparative evaluations, and few-shot diagnosis experiments under different operating conditions, are conducted on a self-built RM bearing fault test platform. Experimental results demonstrate that the proposed method achieves competitive diagnostic performance under both 8-way 1-shot and 8-way 5-shot scenarios, and consistently outperforms the selected comparison methods under different training sample sizes and rotational speed conditions. As the number of training samples increases, the diagnostic accuracy of all models gradually improves, while the proposed method maintains superior recognition performance. Under the 8-way 5-shot setting, the proposed model achieves a maximum diagnostic accuracy of 99.75%, demonstrating its effectiveness, stability, and generalization capability for few-shot RM FD.

In summary, the proposed method combines the original time-domain signal to retain the transient effects caused by faults, and utilizes multi-scale STFT representation to characterize the energy distribution, modulation patterns, and frequency bands sensitive to faults over time. Through hierarchical selective fusion and covariance-based joint metric learning, these complementary features are adaptively integrated and matched under limited support samples, enabling accurate identification of the health status of various rotating machinery under a small number of samples. Future work will further investigate challenging engineering conditions, including strong noise, load variations, cross-condition generalization, and the impact of multi-sensor configurations, and enhance lightweight deployment and online diagnostic capabilities to improve its applicability in actual rotating machinery systems.

Author Contributions: Conceptualization, Y.X. and B.W.; methodology, Y.X.; software, Y.X.; validation, Y.X., Z.J. and B.W.; formal analysis, Y.X.; investigation, B.W.; resources, Y.G.; data curation, Y.X.; writing—original draft preparation, Y.X.; writing—review and editing, Z.J.; visualization, Y.G.; supervision, Z.J.; project administration, Z.J.; funding acquisition, Z.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Guangxi Natural Science Foundation under (Grant No. 2026GXNSFAA00640916, 2025GXNSFBA069131), Guangxi Science and Technology Program (Grant Number Guike AD25069080), 2025 Guangxi key laboratory of auto parts and vehicle technology open research topic (Grant No. 2025GKLACVTKF05), Foundational Research Capacity Enhancement Program for Young and Middle-Aged Teachers in Guangxi Higher Education Institutions (2025KY0060) and Innovation Project of Guangxi Graduate Education (YCBZ2026064).

Data Availability Statement: The datasets generated during the current study are not publicly available due to the project confidentiality requirements but are available from the corresponding author on reasonable request.

Acknowledgments: The authors would like to thank all colleagues and collaborators who provided valuable guidance and support during this study. All individuals mentioned have given their consent to be acknowledged.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RM	Rotating machinery
FD	fault diagnosis
TF	time-frequency
FSFD	few-shot fault diagnosis
SSL	Self-supervised learning
TFDS-MJML	Time-Frequency Dual-Stream Mamba with Joint Metric Learning
MSTF	multi-scale time-frequency
STFT	short-time Fourier transform
ConvMixM	Convolutional mixing module
GL-Mamba	Global-local Mamba
MLKFE	Multi-scale large kernel feature extraction
GALCA	Global and local channel attention

References

- Dai, C.; He, D.; Jin, Z.; Zhang, X.; Chen, G.; Wu, J.; Zhao, J.; Xu, Y. Digital twin-assisted graph contrastive domain adaptation for small-sample bearing fault diagnosis. *Struct. Health Monit.* **2026**, 14759217261440798. [\[CrossRef\]](#)
- Bao, Z.; Liu, C.; Yang, H.; Zhang, J.; Li, Y. From theory to industry: A survey of deep learning-enabled bearing fault diagnosis in complex environments. *Eng. Appl. Artif. Intell.* **2026**, 163, 113068. [\[CrossRef\]](#)
- Xu, Y.; He, D.; Sun, H.; Jin, Z.; Zhao, M. Self-supervised learning for train bearing fault diagnosis based on time-frequency dual domain prediction. *Struct. Health Monit.* **2026**, 14759217251405584. [\[CrossRef\]](#)
- Mohammad, M.; Ibryaeva, O.; Sinitsin, V.; Ereemeeva, V. A Computationally Efficient Method for the Diagnosis of Defects in Rolling Bearings Based on Linear Predictive Coding. *Algorithms* **2025**, 18, 58. [\[CrossRef\]](#)
- Wang, Z.; Yao, L.; Chen, G.; Ding, J. Modified multiscale weighted permutation entropy and optimized support vector machine method for rolling bearing fault diagnosis with complex signals. *ISA Trans.* **2021**, 114, 470–484. [\[CrossRef\]](#) [\[PubMed\]](#)
- Chen, Z.; Yang, Y.; He, C.; Liu, Y.; Liu, X.; Cao, Z. Feature extraction based on hierarchical improved envelope spectrum entropy for rolling bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **2023**, 72, 3518912. [\[CrossRef\]](#)
- Li, T.; Yu, M.; Ma, T.; Du, Y.; Dou, S. A Variable-Speed and Multi-Condition Bearing Fault Diagnosis Method Based on Adaptive Signal Decomposition and Deep Feature Fusion. *Algorithms* **2025**, 18, 753. [\[CrossRef\]](#)
- Wang, J.; Zheng, J.; Pan, H.; Tong, J.; Liu, Q. Refined composite multiscale slope entropy and its application in rolling bearing fault diagnosis. *ISA Trans.* **2024**, 152, 371–384. [\[CrossRef\]](#) [\[PubMed\]](#)
- Cheng, J.; Yang, Y.; Li, X.; Cheng, J. Adaptive periodic mode decomposition and its application in rolling bearing fault diagnosis. *Mech. Syst. Signal Process.* **2021**, 161, 107943. [\[CrossRef\]](#)
- Pang, B.; Nazari, M.; Tang, G. Recursive variational mode extraction and its application in rolling bearing fault diagnosis. *Mech. Syst. Signal Process.* **2022**, 165, 108321. [\[CrossRef\]](#)
- Luo, R.; Xie, H.; Wen, H.; He, H.; Li, Y.; Wang, K. Cross-Domain Bearing Fault Diagnosis Under Class Imbalance: A Dynamic Maximum Triple-View Classifier Discrepancy Network. *Algorithms* **2026**, 19, 228. [\[CrossRef\]](#)
- Xu, Y.; He, D.; Luo, X.; Jin, Z.; Sun, H. CSSCL: A cross-sensor self-contrastive learning for nondestructive fault detection of heavy locomotive motors in label-scarce scenarios. *Nondestruct. Test. Eval.* **2026**, 1–34. [\[CrossRef\]](#)
- Zhao, B.; Zhang, X.; Li, H.; Yang, Z. Intelligent fault diagnosis of rolling bearings based on normalized CNN considering data imbalance and variable working conditions. *Knowl.-Based Syst.* **2020**, 199, 105971. [\[CrossRef\]](#)
- Chen, Z.; Mauricio, A.; Li, W.; Gryllias, K. A deep learning method for bearing fault diagnosis based on Cyclic Spectral Coherence and Convolutional Neural Networks. *Mech. Syst. Signal Process.* **2020**, 140, 106683. [\[CrossRef\]](#)
- Sinitsin, V.; Ibryaeva, O.; Sakovskaya, V.; Ereemeeva, V. Intelligent bearing fault diagnosis method combining mixed input and hybrid CNN-MLP model. *Mech. Syst. Signal Process.* **2022**, 180, 109454. [\[CrossRef\]](#)
- Li, S.; Ji, J.C.; Xu, Y.; Feng, K.; Zhang, K.; Feng, J.; Beer, M.; Ni, Q.; Wang, Y. Dconformer: A denoising convolutional transformer with joint learning strategy for intelligent diagnosis of bearing faults. *Mech. Syst. Signal Process.* **2024**, 210, 111142. [\[CrossRef\]](#)

17. Hou, Y.; Wang, J.; Chen, Z.; Ma, J.; Li, T. Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved Transformer. *Eng. Appl. Artif. Intell.* **2023**, *124*, 106507. [\[CrossRef\]](#)
18. Chen, Y.; Zhang, D.; Zhang, H.; Wang, Q.-G. Dual-path mixed-domain residual threshold networks for bearing fault diagnosis. *IEEE Trans. Ind. Electron.* **2022**, *69*, 13462. [\[CrossRef\]](#)
19. Ding, Y.; Zhuang, J.; Ding, P.; Jia, M. Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings. *Reliab. Eng. Syst. Saf.* **2022**, *218*, 108126. [\[CrossRef\]](#)
20. He, D.; Xu, Y.; Sun, H.; Jin, Z.; Zhao, M. Self-supervised learning for vehicle bearing fault diagnosis based on time–frequency dual-domain contrast and fusion. *Nonlinear Dyn.* **2025**, *113*, 17385–17412. [\[CrossRef\]](#)
21. Wu, K.; Nie, Y.; Wu, J.; Wang, Y. Prior knowledge-based self-supervised learning for intelligent bearing fault diagnosis with few fault samples. *Meas. Sci. Technol.* **2023**, *34*, 105104. [\[CrossRef\]](#)
22. Cui, L.; Tian, X.; Wei, Q.; Liu, Y. A self-attention based contrastive learning method for bearing fault diagnosis. *Expert Syst. Appl.* **2024**, *238*, 121645. [\[CrossRef\]](#)
23. Cen, J.; Yang, Z.; Wu, Y.; Hu, X.; Jiang, L.; Chen, H.; Si, W. A mask self-supervised learning-based transformer for bearing fault diagnosis with limited labeled samples. *IEEE Sens. J.* **2023**, *23*, 10359. [\[CrossRef\]](#)
24. Xu, Y.; Xu, E.; Gao, Y.; Jin, Z. MTFSC: A Self-Supervised Transferable Representation Learning Algorithm for Diagnosing Cross-Machine Faults in Rotating Machinery. *Algorithms* **2026**, *19*, 507. [\[CrossRef\]](#)
25. Huang, Q. Few-Shot Class-Incremental Learning with Prompt Alignment and Subspace Prototype Aggregation. *Algorithms* **2026**, *19*, 407. [\[CrossRef\]](#)
26. Jiang, Y.; Qiu, Z.; Zheng, L.; Dong, Z.; Jiao, W.; Tang, C.; Sun, J.; Xuan, Z. Recursive prototypical network with coordinate attention: A model for few-shot cross-condition bearing fault diagnosis. *Appl. Acoust.* **2025**, *231*, 110442. [\[CrossRef\]](#)
27. Feng, Y.; Chen, J.; Zhang, T.; He, S.; Xu, E.; Zhou, Z. Semi-supervised meta-learning networks with squeeze-and-excitation attention for few-shot fault diagnosis. *ISA Trans.* **2022**, *120*, 383–401. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Wang, D.; Zhang, M.; Xu, Y.; Lu, W.; Yang, J.; Zhang, T. Metric-based meta-learning model for few-shot fault diagnosis under multiple limited data conditions. *Mech. Syst. Signal Process.* **2021**, *155*, 107510. [\[CrossRef\]](#)
29. Wen, C.; Xue, Y.; Liu, W.; Chen, G.; Liu, X. Bearing fault diagnosis via fusing small samples and training multi-state Siamese neural networks. *Neurocomputing* **2024**, *576*, 127355. [\[CrossRef\]](#)
30. Su, H.; Xiang, L.; Hu, A.; Xu, Y.; Yang, X. A novel method based on meta-learning for bearing fault diagnosis with small sample learning under different working conditions. *Mech. Syst. Signal Process.* **2022**, *169*, 108765. [\[CrossRef\]](#)
31. Luo, J.; Shao, H.; Lin, J.; Liu, B. Meta-learning with elastic prototypical network for fault transfer diagnosis of bearings under unstable speeds. *Reliab. Eng. Syst. Saf.* **2024**, *245*, 110001. [\[CrossRef\]](#)
32. Lin, J.; Shao, H.; Zhou, X.; Cai, B.; Liu, B. Generalized MAML for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals. *Expert Syst. Appl.* **2023**, *230*, 120696. [\[CrossRef\]](#)
33. Li, C.; Li, S.; Zhang, A.; He, Q.; Liao, Z.; Hu, J. Meta-learning for few-shot bearing fault diagnosis under complex working conditions. *Neurocomputing* **2021**, *439*, 197–211. [\[CrossRef\]](#)
34. Kuemle, H.; Gross, T.; Kolerus, J. Machine vibrations and diagnostics the world of ISO. In *Proceedings of the Industry Applications Society 60th Annual Petroleum and Chemical Industry Conference*; IEEE: Piscataway, NJ, USA, 2013; pp. 1–13. [\[CrossRef\]](#)
35. Hummel, K.; Hix, J.; Cárdenas, E. Mounted Accelerometer Frequency Response of Adhesive Products and Aluminum Frame Quick Mounts. *Vibration* **2025**, *8*, 61. [\[CrossRef\]](#)
36. Vencl, A.; Gašić, V.; Stojanović, B. Fault tree analysis of most common rolling bearing tribological failures. *IOP Conf. Ser. Mater. Sci. Eng.* **2017**, *174*, 012048. [\[CrossRef\]](#)
37. Hou, R.; Chang, H.; Ma, B.; Shan, S.; Chen, X. Cross attention network for few-shot classification. In *Proceedings of the 33rd international Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019*; Curran Associates Inc.: Red Hook, NY, USA, 2019; Volume 32.
38. Yue, K.; Li, J.; Chen, J.; Huang, R.; Li, W. Multiscale wavelet prototypical network for cross-component few-shot intelligent fault diagnosis. *IEEE Trans. Instrum. Meas.* **2022**, *72*, 3502411. [\[CrossRef\]](#)
39. Xu, Y.; He, D.; Dai, C.; Jin, Z.; Chen, Y.; Fu, Y. A few-shot fault diagnosis method for ship motors based on multi-source data and multi-view task guidance. *Ocean Eng.* **2026**, *362*, 126532. [\[CrossRef\]](#)
40. Xu, Y.; Wu, B.; Gao, Y.; Jin, Z.; Li, K. FS-TFD-MEN: An asynchronous motor few-shot fault diagnosis framework based on integrated metrics. *Meas. Sci. Technol.* **2026**, *37*, 175901. [\[CrossRef\]](#)
41. Zhang, X.; He, C.; Lu, Y.; Chen, B.; Zhu, L.; Zhang, L. Fault diagnosis for small samples based on attention mechanism. *Measurement* **2022**, *187*, 110242. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.