machines

Article

MSA-Mamba: A Frequency-Enhanced Multi-Scale Adaptive State Space Model for Motor Fault Diagnosis

Xutao Lu [1], Xingpeng An [2], Jing Li [3], Minghuan He [1] and Hao Liu [4,*]

[1] School of Mechatronic Engineering, North University of China, Taiyuan 030051, China; luxutao@nuc.edu.cn (X.L.); ty123456he@163.com (M.H.)
[2] Hangzhou Zhiyuan Research Institute Co., Ltd., Hangzhou 310024, China; anxingpeng@zy-cs.com.cn
[3] School of Electrical and Control Engineering, North University of China, Taiyuan 030051, China; lj@nuc.edu.cn
[4] School of Mechanical Engineering, North University of China, Taiyuan 030051, China
* Correspondence: 20250180@nuc.edu.cn

Abstract

In complex industrial environments, vibration and acoustic signals are often corrupted by strong noise, resulting in non-stationary and low signal-to-noise ratio (SNR) characteristics that hinder effective feature extraction and degrade model generalization under limited samples. This study proposes a frequency-enhanced multi-scale adaptive state space Mamba (MSA-Mamba) model for motor fault diagnosis. A CNN Stem is first introduced for local feature extraction and sequence compression, followed by an adaptive frequency-domain filtering module to suppress noise and enhance fault-related frequency components. An MSA-Mamba Block combining multi-scale convolutional attention and parallel state-space modeling is then developed for adaptive feature fusion. In addition, Mixup and label smoothing strategies are employed to improve small-sample generalization. Experimental results show that MSA-Mamba achieves accuracies of 88.28%, 98.44%, and 99.48% on single-, three-, and four-channel datasets, respectively, and maintains 91.41% accuracy under strong-noise conditions. The proposed method demonstrates superior robustness and generalization for motor fault diagnosis.

Keywords: motor fault diagnosis; Mamba; state space model; multi-scale feature fusion; small-sample learning

1. Introduction

In modern industrial systems, motors serve as critical power units, while bearings and transmission systems play essential roles in load support and power transmission, respectively. The operational conditions of these components directly affect the safety and reliability of industrial production processes [1]. With the development of industrial equipment toward higher loads, higher rotational speeds, and more complex operating conditions, vibration and acoustic signals collected during operation are inevitably contaminated by strong background noise and exhibit significant non-stationarity and multi-scale coupling characteristics.

Consequently, weak fault-related features in the time–frequency domain are easily submerged by noise [2], making traditional diagnostic methods that rely on manually designed features ineffective in capturing sensitive fault information. Data-driven intelligent fault diagnosis approaches, benefiting from automatic representation learning and end-to-end modeling capabilities, provide an effective solution to reduce dependence on

Academic Editor: Ahmed Abu-Siada

Received: 2 August 2026
Revised: 7 September 2026
Accepted: 14 September 2026
Published: 18 September 2026

expert-designed features and have become a major research focus in the field of intelligent industrial maintenance [3,4].

1.1. Related Work

Based on existing studies, current fault diagnosis approaches can be generally categorized into three groups: local feature modeling methods based on convolutional neural networks (CNNs), sequence dependency modeling methods based on temporal models, and robust diagnosis methods based on multimodal fusion and data augmentation [5,6].

(1) CNN-based local feature modeling methods.

CNN-based methods have been widely adopted in vibration-based fault diagnosis owing to their strong local feature extraction capability and computational efficiency. A common strategy is to combine CNNs with temporal or attention modules to jointly capture local patterns and global context. Hybrid architectures integrating CNN with LSTM have demonstrated improved bearing fault diagnosis by combining local feature learning with sequential modeling [7]. To address small-sample and noisy conditions, lightweight designs based on dilated causal convolution and local binary convolution have been proposed to reduce model complexity while maintaining robustness [8]. Multi-scale feature extraction strategies with global context enhancement have further improved motor fault diagnosis under variable-speed conditions [9]. Meanwhile, CNN-based methods have enhanced early fault feature extraction and temporal dependency learning through identity mapping and adaptive optimization strategies [10]. Residual attention mechanisms have also been incorporated to suppress noise interference and enhance diagnostic performance in complex environments [11]. Despite these advances, purely CNN-based methods remain limited in capturing long-range temporal dependencies, which restricts their effectiveness in characterizing complex fault evolution processes.

(2) Temporal model-based sequence dependency modeling methods.

To overcome this limitation, temporal models—including Transformers, LSTMs, and their variants—have been introduced to capture long-range sequential dependencies that CNN-based methods tend to overlook. A representative approach combines convolutional local feature extraction with Transformer-based global dependency modeling, achieving over 99% accuracy across multiple benchmark datasets [12]. Memory mechanisms and dual Transformer architectures have further been proposed to capture long-term dependencies in temporal and spatiotemporal prediction tasks [13]. Encoder–decoder structures with enhanced multi-head attention mechanisms have been explored to jointly model short- and long-term temporal dynamics, improving accuracy and stability in multi-step forecasting [14]. Graph-based temporal models integrating spatial–temporal feature extraction with residual structures have additionally demonstrated robust and generalizable fault detection in complex industrial systems [15]. However, these methods are predominantly designed for single-modality inputs and may underperform when heterogeneous sensing signals are available, limiting their applicability in multimodal diagnostic scenarios.

(3) Robust diagnosis methods based on multimodal fusion and data augmentation.

To address the dual challenges of limited labeled data and strong noise interference in practical deployments, multimodal fusion and data augmentation have emerged as complementary strategies for improving diagnostic robustness and generalization. On the augmentation side, diffusion-based approaches integrating multimodal non-Gaussian modeling with conditional GANs have been proposed to generate high-fidelity fault samples and improve small-sample performance [16]. Multimodal collaborative frameworks combining heterogeneous sensing modalities—such as infrared thermal images and operational parameters—with self-attention networks and decision-level fusion strategies have demonstrated superior cross-condition generalization for chiller fault diagnosis [17].

Physics-informed diffusion mechanisms have further been adopted to replace adversarial training, effectively mitigating data imbalance while improving sample quality and diagnostic accuracy [18]. To alleviate mode collapse and gradient vanishing in GAN-based generation, auxiliary classifier GANs incorporating Wasserstein loss and spectral normalization have been developed for bearing and gear fault diagnosis [19]. Ensemble augmentation frameworks employing multi-source generation and parallel training strategies, coupled with data selection mechanisms, have additionally enhanced sample diversity and industrial diagnostic performance [20]. Nevertheless, existing multimodal and augmentation methods often rely on complex pipelines or domain-specific designs, limiting their generalizability to broader fault diagnosis tasks.

Based on the above analysis, representative methods are systematically summarized from three perspectives—method type, technical approach, and key innovation—as listed in Table 1.

Table 1. Summary of Representative Existing Fault Diagnosis Methods.

Category	Ref	Method	Innovations
CNN-based Local Feature Modeling	[7]	CNN + LSTM	Constructs LFLB for local features and introduces LSTM for global dependency modeling, enabling joint local–global feature learning.
	[8]	LBTCNN	Employs dilated causal convolution to increase depth with limited parameters and uses LBC to reduce computations, alleviating overfitting and enhancing noise robustness.
	[9]	GC-ResCNN	Uses hierarchical structure for multiscale features, global context module for discriminative enhancement, and multi-feature fusion for adaptive integration.
	[10]	CNN + Identity Mapping + Adam	Introduces quadruple identity mapping to alleviate gradient vanishing and proposes power-exponential Adam to avoid local optima and oscillations.
	[11]	MRA-CNN	Designs MLMod for multiscale features, RAMod for noise suppression via attention, and residual attention learning to reduce information loss.
Temporal Sequence Modeling	[12]	CNN + Transformer	Integrates CNN local feature extraction and Transformer long-range temporal dependency modeling for joint local–global feature learning.
	[13]	ALTDC	Adopts intra/inter-sequence memory mechanisms and dual spatio-temporal Transformers for joint temporal–spatial modeling.
	[14]	DTDM	Models multi-step temporal dependencies via encoder–decoder structure and captures long/short-term temporal dynamics with extended multi-head attention.
	[15]	DMTL-GCNGRU	Extracts spatio-temporal features via GCN-GRU and enhances multi-task representation robustness with attention and residual structure.
Multi-modal Fusion for Robust Diagnosis	[16]	DDPM + cGAN	Improves fault sample generation quality and sampling efficiency via multi-modal non-Gaussian diffusion and conditional GAN.
	[17]	CWGAN + SA-BAGAN-GP	Enhances variable-condition diagnosis robustness through dual-modal data augmentation and D-S evidence fusion.
	[18]	Diffusion Model + Data Augmentation	Replaces GAN with diffusion model to improve sample quality and diversity and build a unified evaluation framework.
	[19]	MACGAN	Resolves mode collapse and gradient vanishing via Wasserstein loss and spectral normalization with optimized classifier compatibility.
	[20]	Ensemble Data Augmentation	Proposes MSDA combining multiple GANs, with cross-training and parallel training, plus data filtering (df-ct-MSDA) to select high-quality samples.

1.2. Problem Statement

Existing studies have demonstrated that deep learning-based approaches have achieved remarkable progress in the field of industrial fault diagnosis. CNN-based models exhibit strong capability in local feature extraction, temporal models show advantages in long-range dependency modeling, and multimodal fusion combined with data augmentation strategies has great potential for improving model robustness and alleviating data scarcity issues.

However, further analysis reveals that the applicability and stability of existing methods in complex industrial scenarios still require improvement, mainly in the following aspects:

(1) Limited local feature modeling capability.

Existing CNN-based models mainly focus on extracting local time–frequency features, while their ability to characterize long-term temporal dependencies embedded in vibration signals remains insufficient. Consequently, diagnostic performance tends to degrade significantly under non-stationary and complex operating conditions, such as variable speed and variable load scenarios.

(2) Insufficient multimodal fusion mechanisms.

Most existing methods employ feature concatenation or decision-level fusion strategies, which provide limited exploration of the intrinsic semantic correlations among multi-source data. The collaborative modeling capability across different modalities remains inadequate, preventing the full exploitation of complementary information provided by multimodal signals.

(3) Limited model generalization capability.

Current fault diagnosis models are prone to overfitting under small-sample conditions, resulting in severe degradation of generalization performance. Although generative data augmentation approaches represented by generative adversarial networks (GANs) and diffusion models have been introduced to alleviate this issue, they still suffer from inherent limitations in small-sample industrial scenarios, including difficult training convergence, mode collapse, low inference efficiency, and insufficient fidelity in preserving multi-scale coupled signal characteristics [21,22].

1.3. Contributions

To address these challenges, this study proposes a Frequency-Enhanced Multi-Scale Adaptive State Space Mamba (MSA-Mamba) for reliable motor fault diagnosis. Three practical problems are targeted: fault features masked by noise in raw vibration–acoustic signals, insufficient single-scale representations for multi-frequency fault signatures, and overfitting under limited labeled data. The main contributions are summarized as follows:

(1) A convolutional prior enhancement module and a multi-scale convolutional attention mechanism are introduced to strengthen local feature extraction and adaptive multi-scale fault representation across different frequency resolutions.

(2) A frequency-domain adaptive filtering module is developed to dynamically suppress noise interference by weighting fault-relevant spectral components and fusing complementary time–frequency features, improving robustness under strong noise conditions.

(3) Mixup augmentation and label smoothing are incorporated to smooth decision boundaries and improve inter-class discrimination, enhancing model generalization under limited training data without additional data collection cost.

(4) Extensive experiments on a multimodal motor fault dataset validate the proposed method under different input modalities, training sample sizes (5–100%), and composite noise levels (SNR from -10 dB to 10 dB), demonstrating consistent superiority over representative baselines.

The remainder of this paper is organized as follows. Section 2 reviews the theoretical background. Section 3 presents the proposed MSA-Mamba framework. Section 4 reports the experimental results and analysis. Section 5 concludes the paper.

2. Theoretical Background

Rotating motors are typical electromechanical systems whose operating conditions can be characterized by vibration and acoustic signals. Healthy signals generally exhibit stable amplitudes and concentrated spectral energy, whereas faults may cause periodic impacts, modulation, sidebands, and spectral broadening. Therefore, effective fault diagnosis requires capturing both local fault patterns and long-range temporal dependencies [23].

CNNs extract local fault features through convolutional operations, such as impacts, waveform distortions, and short-term oscillations. Although deeper layers enlarge the receptive field and capture higher-level features, modeling long-range dependencies requires deeper networks or larger kernels, increasing computational cost. Repeated convolution and down-sampling may also weaken fine-grained information associated with weak or early-stage faults [24].

Mamba is a selective state-space model (SSM)-based architecture that adaptively updates state-space parameters through input-driven dynamic parameter mapping. It leverages a hardware-accelerated parallel-scan kernel to speed up its inherent sequential recurrent computation, achieving long-range temporal dependency modeling with linear computational complexity and enabling efficient extraction of global sequential features, as illustrated in Figure 1 [25].

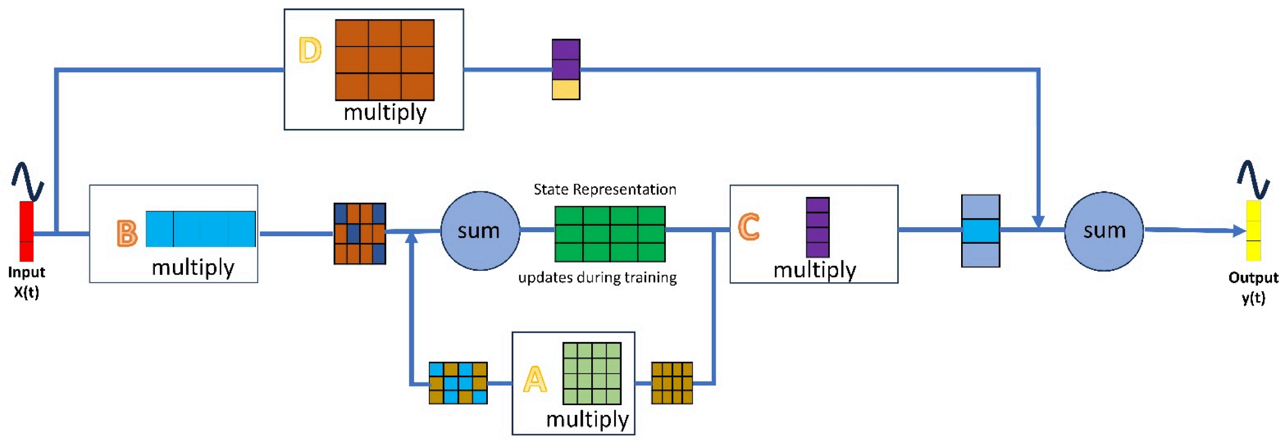


Figure 1. Illustration of the Mamba architecture.

Compared with CNNs, it captures long-range temporal dependencies without deep stacking, reducing computational overhead. However, the original Mamba has three inherent limitations: (1) unidirectional state propagation fails to fully capture bidirectional temporal correlations; (2) lack of multi-scale feature extraction and time–frequency fusion limits sensitivity to weak fault characteristics in acoustic signals; (3) insufficient fusion of complementary multi-channel information in vibration signals and constrained cross-modal interaction in vibration–acoustic scenarios.

Existing studies on Mamba enhancement mainly focus on three directions: bidirectional scanning for enhanced context awareness, multi-scale hierarchical architectures for richer feature representations, and multi-modal fusion modules for improved robustness. Nevertheless, breakthroughs are still needed in feature representation, generalization, and robustness, particularly in weak fault feature extraction, multi-channel information complementarity, and deep cross-modal collaboration [26].

3. Frequency-Enhanced Multi-Scale Adaptive State Space Mamba

3.1. Framework Design of MSA-Mamba

To address the limitations of the original Mamba model discussed in Section 2, including limited multi-scale feature extraction capability and inadequate time–frequency information fusion under complex operating conditions, this section develops a MSA-Mamba model based on the time–frequency characteristics of different fault types to improve motor fault diagnosis performance in complex scenarios.

The overall framework follows a processing pipeline consisting of local feature extraction, frequency-domain enhancement, global dependency modeling, and discriminative prediction. While preserving the advantages of Mamba in long-sequence modeling, the proposed model further improves robustness and generalization capability. The overall architecture is illustrated in Figure 2.

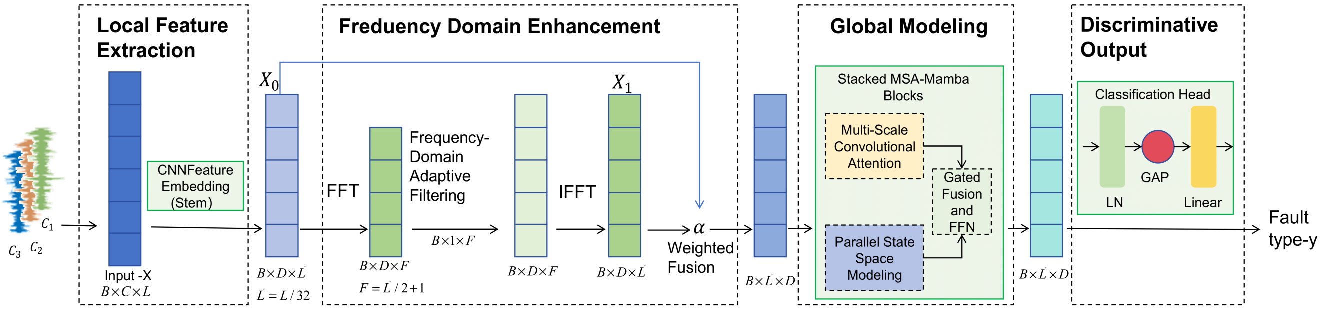


Figure 2. Illustration of the MSA-Mamba architecture.

Specifically, at the input stage, the model receives multi-channel temporal signals $\mathbf{X} \in \mathbb{R}^{B\times C\times L}$, where B denotes the batch size, C represents the number of input channels (dynamically supporting 1/3/4-channel configurations), and L indicates the sequence length.

First, the convolutional prior enhancement module extracts hierarchical features from raw signals through progressive down-sampling, generating stable high-level representations. Then, the frequency-domain adaptive filtering module applies FFT to transform temporal features into the frequency domain, where learnable frequency gains predicted by an MLP are used to suppress noise components and enhance fault-related spectral information. A residual fusion mechanism is further introduced to adaptively regulate filtering strength and preserve informative features.

Subsequently, multiple MSA-Mamba Blocks are stacked for global feature modeling. Each block integrates a parallel state space module and a multi-scale convolutional attention module to capture long-range temporal dependencies and multi-scale local patterns. A learnable gating mechanism is employed to fuse dual-branch features, followed by residual connections and feed-forward networks to improve representation capability. Finally, Layer Normalization and Global Average Pooling generate global feature vectors for fault classification through a fully connected layer.

During training, Mixup augmentation and label smoothing loss are adopted to alleviate overfitting and improve generalization under limited-sample conditions.

The detailed training procedure is described in Algorithm 1.

Algorithm 1: Framework of MSA-Mamba

Input: Training batch $\{(\mathbf{X}, \mathbf{Y})\}, \mathbf{X} \in \mathbb{R}^{B \times C \times L}$
Output: Predicted label $\hat{\mathbf{y}}$, loss $\mathcal{L}$

S1: CNN Feature Embedding (Stem)
$\mathbf{X}_0 \leftarrow \mathrm{CNNStem}(\mathbf{X})$;
S2: Frequency-Domain Adaptive Filtering
$\boldsymbol{F} \leftarrow \mathrm{FFT}(\boldsymbol{X}_0)$
$\boldsymbol{A} \leftarrow |\boldsymbol{F}|$
$\boldsymbol{A}_c \leftarrow \mathrm{AdaptiveAvgPool}(\boldsymbol{A})$
$\boldsymbol{W}_f \leftarrow \sigma(\mathrm{Conv}(\mathrm{GELU}(\mathrm{Conv}(\boldsymbol{A}_c))))$
$\boldsymbol{F}' \leftarrow \boldsymbol{F} \odot \mathrm{Interp}\left(\boldsymbol{W}_f\right)$
$\boldsymbol{X}_1 \leftarrow \mathrm{IFFT}(\boldsymbol{F}')$
$\boldsymbol{X}_1 \leftarrow \alpha \boldsymbol{X}_1 + (1-\alpha)\boldsymbol{X}_0$
S3: Sequence Preparation
$\boldsymbol{Z} \leftarrow \mathrm{Transpose}(\boldsymbol{X}_1)$
$\boldsymbol{Z} \leftarrow \boldsymbol{Z} + \mathrm{PE}(\boldsymbol{Z})$
S4: Stacked MSA-Mamba Blocks
For $i = 1$ **to** N **do**
    Multi-Scale Convolutional Attention
    for $s = 1$ **to** S **do**
        $\boldsymbol{F}_s \leftarrow \mathrm{DWConv}_{k_s}(\boldsymbol{Z})$
        $\boldsymbol{z}_s \leftarrow \mathrm{GAP}(\boldsymbol{F}_s)$
        $\alpha_s \leftarrow \mathrm{Softmax}(\mathrm{MLP}(\boldsymbol{z}_s))$
        $\boldsymbol{Z}_{msa} \leftarrow \sum_{s=1}^{S} \alpha_s \boldsymbol{F}_s$
        Parallel State Space Modeling
        $(\boldsymbol{U}, \boldsymbol{V}) \leftarrow \mathrm{Linear}(\boldsymbol{Z})$
        $\boldsymbol{U} \leftarrow \mathrm{DWConv}(\boldsymbol{U})$
        $(\Delta_t, \boldsymbol{B}, \boldsymbol{C}) \leftarrow \mathrm{Linear}(\boldsymbol{U})$
        $\boldsymbol{H} \leftarrow \mathrm{ParallelScan}(\Delta_t, \boldsymbol{B}, \boldsymbol{U})$
        $\boldsymbol{Z}_{ssm} \leftarrow (\boldsymbol{C} \cdot \boldsymbol{H}) \odot \mathrm{SiLU}(\boldsymbol{V})$
        Gated Fusion and FFN
        $\boldsymbol{Z} \leftarrow \boldsymbol{Z} + \sigma(g_1)\boldsymbol{Z}_{msa} + \sigma(g_2)\boldsymbol{Z}_{ssm}$
        $\boldsymbol{Z} \leftarrow \boldsymbol{Z} + \mathrm{FFN}(\mathrm{LN}(\boldsymbol{Z}))$
S5: Classification Head
$\boldsymbol{Z} \leftarrow \mathrm{LN}(\boldsymbol{Z})$
$\boldsymbol{z} \leftarrow \mathrm{GAP}(\boldsymbol{Z})$
$\hat{\boldsymbol{y}} \leftarrow \mathrm{Linear}(\boldsymbol{z})$
S6: Mixup Training Strategy
Sample $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$ and permutation π
$\boldsymbol{X}_{mix} \leftarrow \lambda \boldsymbol{X} + (1-\lambda)\boldsymbol{X}_\pi$
$\boldsymbol{Y}_{mix} \leftarrow \lambda \boldsymbol{Y} + (1-\lambda)\boldsymbol{Y}_\pi$
$\boldsymbol{Y}_{mix} \leftarrow (1-\varepsilon)\boldsymbol{Y}_{mix} + \varepsilon / K$
$\hat{\boldsymbol{y}} \leftarrow f(\boldsymbol{X}_{mix})$
$\mathcal{L} \leftarrow \mathrm{CE}\left(\hat{\boldsymbol{y}}, \boldsymbol{Y}_{mix}\right)$
return $\hat{\boldsymbol{y}}, \mathcal{L}$

3.2. Convolutional Prior Enhancement

To improve the feature representation capability of the proposed model under limited-sample conditions, a deep one-dimensional convolutional embedding structure is introduced at the input stage to enhance the model's inductive bias. Compared with sequence models that solely rely on state space modeling, convolutional networks possess local connectivity and parameter-sharing properties, enabling effective extraction of stable local patterns with limited training data and reducing the dependence on large-scale datasets.

Given the input temporal signal $\boldsymbol{X} \in \mathbb{R}^{B\times C\times L}$, the convolutional embedding module progressively extracts local features through multiple one-dimensional convolutional layers and compresses the sequence length through stride operations. The operation of the (l)-th convolutional layer can be formulated as:

$$\mathbf{X}^{(l)} = \phi\Big(\mathrm{BN}\Big(\mathrm{Conv}_{k_l,s_l}(\mathbf{X}^{(l-1)})\Big)\Big), \tag{1}$$

where $\mathrm{Conv}_{k_l,s_l}(\cdot)$ denotes a one-dimensional convolution operation with kernel size k_l and stride s_l, $\mathrm{BN}(\cdot)$ represents batch normalization, and $\phi(\cdot)$ denotes the GELU activation function.

Through hierarchical stacking of convolutional layers, the receptive field is gradually expanded, enabling hierarchical feature modeling from fine-grained local patterns to high-level semantic representations. After L_c convolutional layers, the embedded feature representation is obtained as $\mathbf{X}_{\mathrm{emb}} = \mathbf{X}^{(L_c)}$. During this process, the sequence length is reduced from the original (L) to

$$L' = L/\prod_{l=1}^{L_c} s_l, \tag{2}$$

This compression process preserves essential fault-related information while reducing the computational complexity of subsequent modules.

To further improve model generalization, a dropout regularization operation is introduced at the end of the convolutional module.

$$\mathbf{X}_{\mathrm{out}} = \mathrm{Dropout}(X_{\mathrm{emb}}), \tag{3}$$

3.3. Frequency Domain Enhancement

In practical industrial scenarios, vibration and acoustic signals collected from motors are usually contaminated by complex noise interference. In particular, periodic noises such as power-frequency interference are often concentrated within specific frequency bands. Most existing studies mainly focus on time-domain sequence modeling while lacking explicit characterization of frequency-domain structural information, resulting in degraded diagnostic performance under low signal-to-noise ratio (SNR) conditions [27].

To address this issue, a frequency-domain adaptive modeling mechanism is introduced in this study. By explicitly modeling spectral information during feature extraction, the proposed method dynamically separates noise components from effective signal characteristics, thereby improving model robustness.

Given the input feature representation $\boldsymbol{X} \in \mathbb{R}^{B\times C\times L}$, FFT is applied along the temporal dimension to obtain the frequency-domain representation $\mathbf{F} = \mathcal{F}(X) \in \mathbb{C}^{B\times C\times L}$, where $\mathcal{F}(\cdot)$ represents the Fourier transform operation.

To reduce the influence of phase instability, the amplitude spectrum is extracted as the frequency-domain feature representation $\mathbf{A} = |\mathbf{F}|$.

Subsequently, adaptive pooling is applied to compress the spectrum representation, obtaining global frequency distribution information while reducing computational complexity $\mathbf{A}_c = \mathrm{Pool}(\mathbf{A})$.

Based on this representation, a lightweight convolutional network is introduced to learn the importance weights of different frequency components $\mathbf{W}_f = \sigma(\text{Conv}(\mathbf{A}_c))$, where $\sigma(\cdot)$ denotes the Sigmoid activation function. To match the original spectral resolution, interpolation is employed to recover the weight dimension $\widetilde{\mathbf{W}}_f = \text{Interp}\left(\mathbf{W}_f\right)$.

The original spectrum is then adaptively modulated as $\mathbf{F}' = \mathbf{F} \odot \widetilde{\mathbf{W}}_f$, where $\odot$ denotes element-wise multiplication. This operation suppresses noise-dominant frequency bands while enhancing discriminative fault-related frequency components.

Afterward, IFFT is performed to convert the enhanced frequency-domain features back into the time domain $\mathbf{X}' = \mathcal{F}^{-1}(\mathbf{F}')$.

To prevent excessive attenuation of useful information during frequency filtering, a residual adaptive fusion mechanism is introduced $\mathbf{Y} = \alpha\mathbf{X}' + (1-\alpha)\mathbf{X}$, where $\alpha \in [0,1]$ is a learnable fusion coefficient. This design enables effective noise suppression while preserving the original signal structure, thereby improving model stability and generalization capability.

3.4. Global Feature Modeling

Global feature modeling is achieved by stacking multiple MSA-Mamba Blocks. Each MSA-Mamba Block consists of a parallel state space modeling module and a multi-scale convolutional attention module, which are designed to capture long-range temporal dependencies and multi-scale local feature representations, respectively.

The overall architecture of the proposed block is illustrated in Figure 3. Note that the $\mathbf{D}$ matrix (feedthrough/skip-connection term) shown in Figure 1 is omitted in Figure 3 for visual clarity; its function is realized by the residual connection $Z_k = y'_k + X_k$ in the implementation.

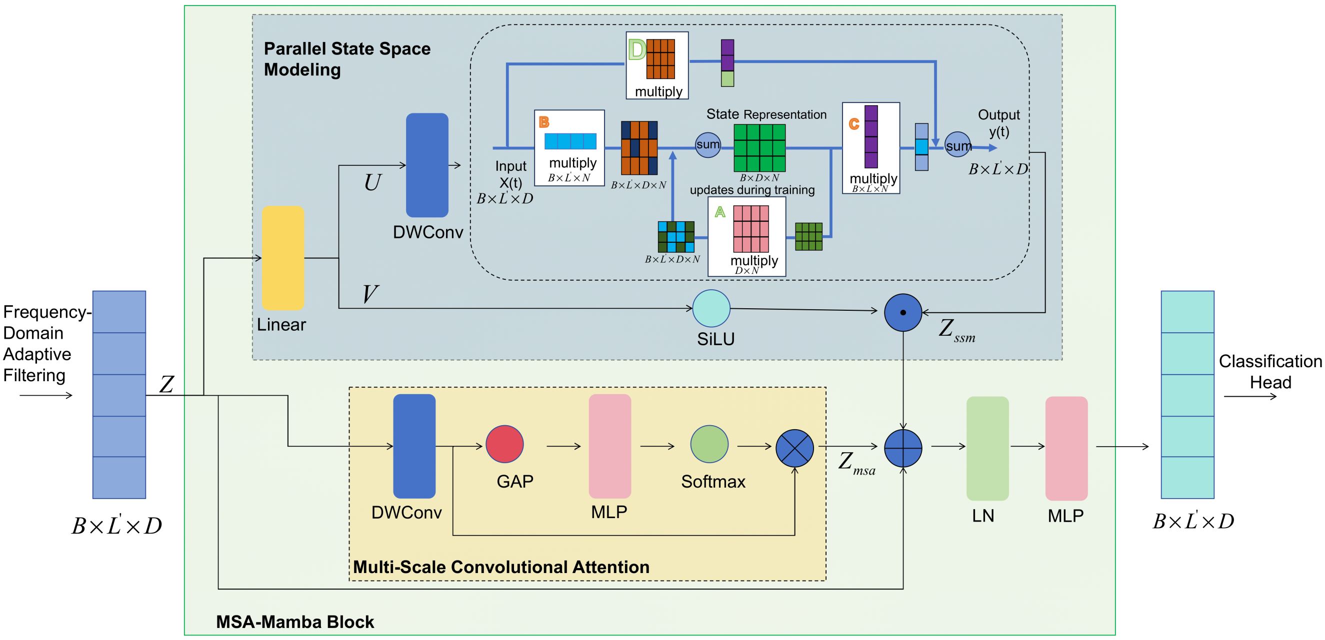


Figure 3. Illustration of the MSA-Mamba Block.

3.4.1. Parallel State Space Model

Motor signal fault evolution exhibits long-span temporal dependencies. The SSM, the core module of Mamba, is well-suited to capture such long-range sequential relationships. Nevertheless, when stacked into multiple parallel multi-scale branches, repeated step-by-step recurrent computation reduces overall GPU-parallel efficiency. To tackle this problem, this work adopts a branch-parallel state-space computation strategy to improve efficiency while retaining long-sequence modeling capability.

Specifically, given the input feature sequence $\mathbf{X} \in \mathbb{R}^{B\times L\times D}$, the input-driven state representation and gating branch are generated through linear projections $\mathbf{U} = \boldsymbol{W}_u\mathbf{X}$, $\mathbf{G} = \sigma(\mathbf{W}_g\boldsymbol{X})$, where $\mathbf{U}$ denotes the input-dependent state driving term and $\mathbf{G}$ represents the gating coefficient.

The continuous-time state space model is formulated as:

$$\frac{d\mathbf{h}(t)}{dt} = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{u}(t),\ y(t) = \mathbf{C}\mathbf{h}(t), \tag{4}$$

After discretization with a time interval Δt, the state transition process can be expressed as:

$$\mathbf{h}_k = \overline{\mathbf{A}}\mathbf{h}_{k-1} + \overline{\mathbf{B}}\mathbf{u}_k, \tag{5}$$

where the discretized parameters are defined as:

$$\overline{\mathbf{A}} = \exp(\Delta t\mathbf{A}), \tag{6}$$

$$\overline{\mathbf{B}} = \left(\int_0^{\Delta t} \exp(\tau\mathbf{A})d\tau\right)\mathbf{B}, \tag{7}$$

To enable efficient parallel computation, the state transition matrix is approximated through diagonalization:

$$\mathbf{A} = \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_D), \tag{8}$$

rather than approximating a full dense matrix. Under this decoupling assumption, cross-channel interactions among different hidden dimensions are eliminated. Thus, the state update can be simplified into an element-wise operation:

$$\mathbf{h}_k = \exp(\Delta t\mathbf{\Lambda}) \odot \mathbf{h}_{k-1} + \widetilde{\mathbf{B}} \odot \mathbf{u}_k, \tag{9}$$

By expanding the recursive formulation, the parallel representation of the state propagation is obtained as:

$$\mathbf{h}_k = \textstyle\sum_{i=1}^{k}\left(\prod_{j=i+1}^{k}\overline{\mathbf{A}}\right) \odot \overline{\mathbf{B}}\mathbf{u}_i, \tag{10}$$

This formulation constitutes the core of the Parallel scan operation shown in Figure 3, which implements the linear recurrence in Equation (9) via a binary-tree reduction that exploits the associativity of the state update operator. At each level of the tree, adjacent state pairs are merged as: $(\overline{\mathbf{A}}_j, \mathbf{s}_j) \oplus (\overline{\mathbf{A}}_i, \mathbf{s}_i) = (\overline{\mathbf{A}}_j \odot \overline{\mathbf{A}}_i, \overline{\mathbf{A}}_j \odot \mathbf{s}_i + \mathbf{s}_j)$. Thereby, the original sequential recurrent computation is transformed into parallel tensor operations, reducing the complexity from $\mathcal{O}(L)$ to $\mathcal{O}(\log L)$ across L time steps and significantly reducing sequential dependencies along the temporal dimension.

The state output is first obtained as $\boldsymbol{y}_k = \mathbf{C}\mathbf{h}_k$. Subsequently, a gating mechanism is introduced to perform nonlinear modulation on the output $\boldsymbol{y'}_k = \mathbf{G}_k \odot \boldsymbol{y}_k$, where G_k denotes the adaptive gating coefficient and $\odot$ represents element-wise multiplication. Finally, the residual connection is incorporated to obtain the module output $\mathbf{Z}_k = \boldsymbol{y'}_k + \mathbf{X}_k$.

3.4.2. Multi-Scale Convolutional Attention Mechanism

Motor mechanical fault signals typically exhibit prominent multi-scale characteristics, where different fault types correspond to unique temporal patterns and frequency distributions. The original Mamba architecture primarily implements global dependency modeling via state space recursion but lacks the ability to capture local multi-scale fault features [28].

To address this deficiency, a multi-scale convolutional attention mechanism is introduced to enhance the extraction and adaptive fusion of multi-scale temporal fault features.

Given the input feature representation $\mathbf{X} \in \mathbb{R}^{B \times L \times D}$, multiple one-dimensional convolutions with different kernel sizes are first applied in parallel:

$$\mathbf{F}_s = \mathrm{Conv}_{k_s}(\mathbf{X}), s = 1, \ldots, S, \tag{11}$$

where k_s denotes the kernel size of the s-th convolution branch. Different kernel sizes provide diverse receptive fields, enabling the extraction of multi-scale temporal representations.

Subsequently, global average pooling (GAP) is applied to each scale feature to obtain compact scale descriptors $\mathbf{z}_s = \mathrm{GAP}(\mathbf{F}_s)$. Based on these descriptors, a shared multilayer perceptron is utilized to estimate the importance of each scale $\mathbf{w}_s = \mathrm{MLP}(\mathbf{z}_s)$. The scale attention weights are normalized using the SoftMax function:

$$\alpha_s = \frac{\exp(\mathbf{w}_s)}{\sum_{i=1}^{S} \exp(\mathbf{w}_i)}, \tag{12}$$

The multi-scale features are then adaptively fused according to the learned attention weights:

$$\mathbf{F}_{\mathrm{fusion}} = \sum\nolimits_{s=1}^{S} \alpha_s \cdot \mathbf{F}_s, \tag{13}$$

Finally, a point-wise convolution (1×1 convolution) is employed for channel-wise feature integration $\mathbf{Y} = \mathrm{Conv}_{1\times 1}(\mathbf{F}_{\mathrm{fusion}})$ and the residual connection is introduced to obtain the final output $\mathbf{Z} = \mathbf{Y} + \mathbf{X}$.

3.5. Mixup-Based Training Optimization Strategy

In practical industrial motor diagnosis scenarios, labeled fault samples are generally scarce, easily leading to model overfitting and poor generalization. To address this issue, Mixup data augmentation and label smoothing are integrated into the training process. Mixup generates virtual samples by linearly interpolating paired inputs and labels, effectively expanding the training data distribution and mitigating overfitting. Meanwhile, label smoothing optimizes output probability distributions to smooth classification decision boundaries, further improving model generalization toward unseen fault conditions under limited-sample scenarios.

Let the training samples and their corresponding labels be denoted as $(\mathbf{x}_i, \mathbf{y}_i)$ and $\left(\mathbf{x}_j, \mathbf{y}_j\right)$, respectively. First, the interpolation coefficient is sampled from a Beta distribution $\lambda \sim \mathrm{Beta}(\alpha, \alpha), \lambda \in [0, 1]$, where α is a hyperparameter that controls the mixing degree between samples.

Then, the mixed samples and labels are generated according to the interpolation coefficient:

$$\begin{cases} \boldsymbol{x}_{\mathrm{mix}} = \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_j \\ \boldsymbol{y}_{\mathrm{mix}} = \lambda \boldsymbol{y}_i + (1-\lambda)\boldsymbol{y}_j \end{cases}, \tag{14}$$

During the training process, the model takes the mixed samples as inputs and is optimized using a weighted cross-entropy loss function:

$$\mathcal{L}_{\mathrm{mix}} = \lambda \cdot \mathcal{L}_{\mathrm{CE}}(f, (\boldsymbol{x}_{\mathrm{mix}})\boldsymbol{y}_i) + (1-\lambda) \cdot \mathcal{L}_{\mathrm{CE}}\left(f, (\boldsymbol{x}_{\mathrm{mix}})\boldsymbol{y}_j\right), \tag{15}$$

where $f(\cdot)$ represents the model prediction function and $\mathcal{L}_{\mathrm{CE}}$ denotes the cross-entropy loss function.

To further enhance the capability of the model in characterizing uncertainty near class boundaries, the label smoothing strategy is adopted. The original one-hot labels are transformed into smoothed labels:

$$\boldsymbol{y}^{\mathrm{LS}} = (1-\varepsilon)\boldsymbol{y} + \frac{\varepsilon}{K}, \tag{16}$$

where ε represents the smoothing coefficient and K denotes the number of categories.

The final training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{mix}}\left(y^{\mathrm{LS}}\right) \tag{17}$$

By introducing smoothing operations in both the sample space and label space, the proposed strategy effectively improves the generalization performance and robustness of the model under limited-sample conditions.

4. Experimental Results and Analysis

4.1. Experimental Setup

The experiments were conducted under a unified hardware and software environment. An NVIDIA RTX 4090 GPU was used for accelerated training, with an Intel Core series CPU. The model was implemented based on Python 3.9 and the PyTorch 2.1 framework, with NumPy, Pandas, and Scikit-learn utilized for data processing and evaluation.

The HUST motor multi-modal fault dataset was adopted for validation, which was collected using the experimental platform shown in Figure 4 [29].

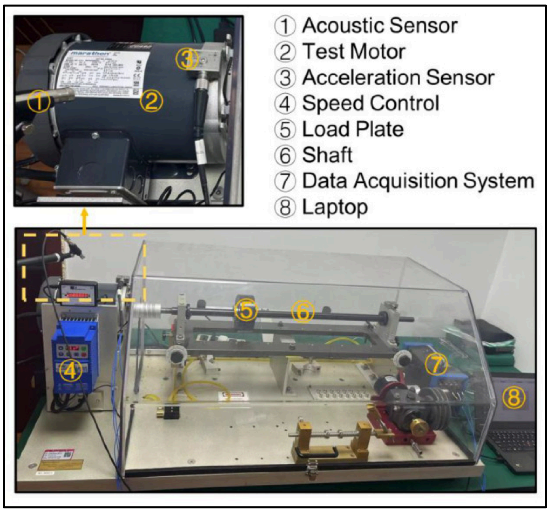


Figure 4. HUST motor fault dataset data acquisition platform [29].

The dataset contains six health states: one healthy condition and five single-fault conditions, including bearing fault, bowed rotor, broken rotor bars, rotor misalignment, and voltage unbalance. These conditions cover representative mechanical faults (bowed rotor, broken rotor bars, rotor misalignment), electrical faults (voltage unbalance), and component-level faults (bearing fault), providing a comprehensive benchmark for motor fault-diagnosis research. Each faulty motor sample is independent, and all faults were artificially preset by the manufacturer.

To comprehensively evaluate the proposed method across three claimed contributions—local feature modelling, multimodal fusion, and model generalization—three experimental settings were designed in conjunction with three input modality configurations.

To evaluate the feature extraction and fusion capabilities of the proposed model under different information conditions, three input modes were considered: single-channel acoustic signals, three-channel vibration signals, and four-channel vibration–acoustic fusion signals, as detailed in Table 2. For the four-channel vibration–acoustic group, multimodal fusion is achieved by stacking the synchronized raw vibration and acoustic signals along the channel dimension before they are fed into the CNN stem.

Table 2. Experiment Configuration.

Experiment Name	Number of Channels	Signal Type	Objective
Single-channel Experiment	Single channel	Acoustic signal	To verify the model's ability to extract fault features of single-modal signals.
Multi-channel Experiment	Three channels	Vibration signal	To investigate the model's capability to model spatial multi-dimensional vibration information.
Multi-channel Experiment	Four channels	Vibration–acoustic fused signal	To validate the comprehensive diagnosis performance under multi-modal information fusion scenarios.

The single-channel experiment validates the model's ability to extract discriminative fault features from unimodal acoustic signals. The three-channel experiment investigates the model's capability to capture spatial multi-dimensional vibration information. The four-channel experiment validates comprehensive diagnosis performance under multimodal information fusion scenarios, directly supporting the multimodal fusion claim.

Beyond the standard clean-data evaluation, model generalization is further assessed through two additional experiments. A noise robustness experiment evaluates all models under composite noise at five SNR levels ranging from -10 dB to 10 dB, and a few-shot experiment examines performance under varying training set sizes from 5% to 100% of the full dataset. This multi-experiment design ensures that each claimed contribution is supported by appropriate and targeted evaluation evidence, with accuracy and macro-averaged F1-score reported as the primary metrics. The macro-averaged F1-score is particularly critical for ensuring fair evaluation across all six fault categories, including minority classes.

All raw recordings were first partitioned into training, validation, and test subsets at a 7:1:2 ratio before signal segmentation. Each recording was then uniformly segmented into non-overlapping windows of 2048 points. All windows sliced from the same raw recording were retained within a single subset to prevent data leakage. A total of 1920 samples were finally constructed. The random seed was fixed to 42, and this identical group-level split was adopted for all baseline models and the proposed MSA-Mamba to ensure a fair comparison.

Four representative time-series models were selected as baselines, covering convolutional, recurrent, attention-based, and state-space architectures: CNN-1D, BiLSTM, Transformer, and the original Mamba. CNN-1D extracts local fault patterns through convolution operations; BiLSTM captures bidirectional temporal dependencies; Transformer models long-range dependencies via multi-head self-attention; Mamba achieves efficient sequence modeling with linear complexity based on selective state space modeling [30–32].

To ensure fair comparison, all models were trained under identical settings with a fixed input length of 2048 points. The AdamW optimizer was used with an initial learning rate of 1×10^{-3}, weight decay of 1×10^{-4}, and batch size of 32. Cosine annealing learning rate scheduling, gradient clipping, and a fixed random seed of 42 were uniformly applied. Standard, noise robustness, and few-shot experiments used max epochs/early stopping patience of 80/20, 50/15, and 60/15, respectively. The proposed MSA-Mamba adopts $d_{\text{model}} = 64$, $d_{\text{state}} = 16$, with 3 SSM layers and 3 multi-scale branches.

4.2. Fault Diagnosis Results with Single-Channel Acoustic Signals

To evaluate the fault diagnosis capability of the proposed model under single-sensor conditions, acoustic signals with a single channel were adopted for the six-class fault classification task. The diagnostic results of different models on the clean single-channel dataset are presented in Figure 5.

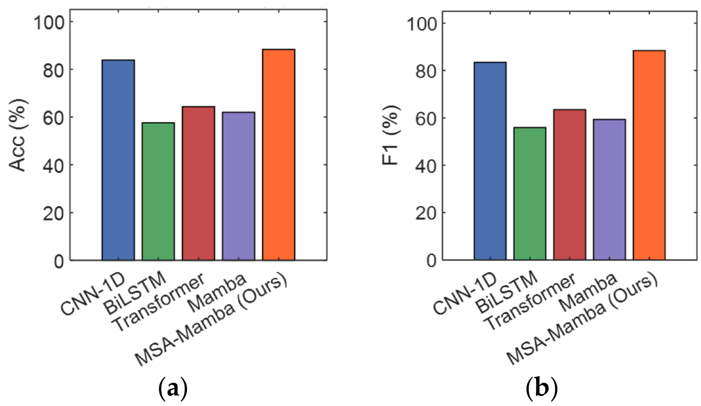


Figure 5. Diagnostic results on clean single-channel data: (**a**) Accuracy; (**b**) Weighted F1-score.

The corresponding quantitative results are summarized in Table 3.

Table 3. Diagnostic results on clean single-channel data (%).

Model	Accuracy	Weighted F1-Score
CNN-1D	83.85	83.46
BiLSTM	57.55	55.93
Transformer	64.32	63.48
Mamba	61.98	59.34
MSA-Mamba	88.28	88.42

As shown in Figure 5 and Table 3, single-channel acoustic signals contain limited feature information, resulting in insufficient performance for conventional sequence models. BiLSTM achieves the lowest accuracy of 57.55%, while Transformer and original Mamba show limited capability in extracting weak acoustic fault patterns. CNN-1D obtains the best baseline performance due to its ability to capture local discriminative features.

In contrast, the proposed MSA-Mamba achieves an accuracy of 88.28% and a weighted F1-score of 88.42%. This improvement benefits from the combination of CNN-based prior enhancement and multi-scale convolutional attention, which enables adaptive extraction of local fault patterns at different temporal scales. Meanwhile, the parallel SSM effectively captures long-range temporal dependencies.

These results demonstrate that the proposed model maintains superior fault recognition performance under single acoustic sensing conditions, providing an effective solution for low-cost and deployable industrial online fault monitoring.

4.3. Multi-Channel Experimental Results

In practical industrial scenarios, multi-sensor systems are commonly deployed for collaborative monitoring. To evaluate the multi-modal fusion capability of the proposed model, two multi-channel experiments were conducted, including three-channel vibration signals and four-channel vibration–acoustic fusion signals.

(1) Three-Channel Vibration Signal Experiment.

Vibration signals contain rich mechanical fault-related impact features, and multi-channel inputs further enhance feature diversity. The diagnostic results on the clean three-channel vibration dataset are presented in Figure 6.

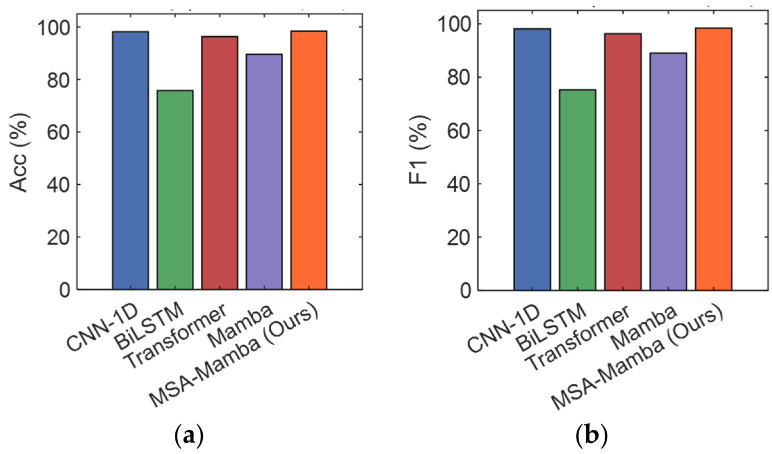


Figure 6. Diagnostic results on clean three-channel vibration data: (**a**) Accuracy; (**b**) Weighted F1-score.

The corresponding quantitative results are summarized in Table 4.

Table 4. Diagnostic results on three-channel vibration data (%).

Model	Accuracy	Weighted F1-Score
CNN-1D	98.18	98.15
BiLSTM	75.78	75.24
Transformer	96.35	96.28
Mamba	89.58	88.98
MSA-Mamba	98.44	98.40

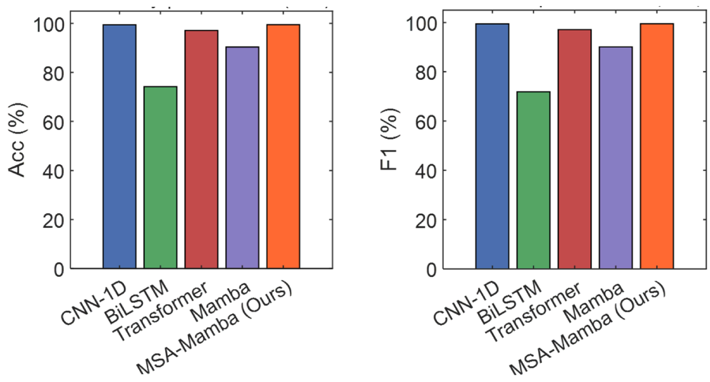
The results indicate that CNN-1D achieves competitive performance by exploiting local feature extraction capability, while Transformer benefits from global dependency modeling. However, the original Mamba exhibits relatively limited representation ability, resulting in lower diagnostic accuracy.

The proposed MSA-Mamba achieves an accuracy of 98.44% and a weighted F1-score of 98.40%, outperforming all baseline models. This improvement demonstrates that the multi-scale convolutional attention mechanism effectively integrates local features from multi-channel vibration signals, while the enhanced temporal–frequency representation further strengthens fault feature extraction.

(2) Four-Channel Vibration–Acoustic Fusion Experiment.

The diagnostic results using four-channel vibration–acoustic fusion signals are shown in Figure 7.

As shown in Figure 7 and Table 5, the diagnostic performance of all models is further improved after multi-modal fusion. The proposed frequency-domain adaptive filtering module adaptively recalibrates the mixed spectral components from multiple channels, suppresses redundant information, and enhances the complementary characteristics between vibration and acoustic signals.



(**a**) (**b**)

Figure 7. Diagnostic results on clean four-channel fusion data: (**a**) Accuracy; (**b**) Weighted F1-score.

The corresponding statistical results are presented in Table 5.

Table 5. Diagnostic results on four-channel vibration–acoustic fusion data (%).

Model	Accuracy	Weighted F1-Score
CNN-1D	99.42	99.42
BiLSTM	74.22	71.88
Transformer	97.14	97.12
Mamba	90.36	90.10
MSA-Mamba	99.48	99.48

The proposed MSA-Mamba achieves the highest accuracy and weighted F1-score of 99.48%, demonstrating superior capability in multi-modal feature fusion and complex fault pattern recognition. These results verify the effectiveness of the proposed architecture for industrial multi-sensor fault diagnosis applications.

4.4. Generalization Performance Evaluation

In industrial fault diagnosis, the acquisition and annotation of fault samples are costly, resulting in limited available training data. To evaluate the generalization capability of the proposed model under small-sample conditions, six training sample scales (96, 192, 384, 768, and 1344 samples) were constructed. Small-sample experiments were conducted under composite noise conditions.

The diagnostic accuracy results for single-channel, three-channel, and four-channel fusion inputs are presented in Figure 8, Figure 9, and Figure 10, respectively.

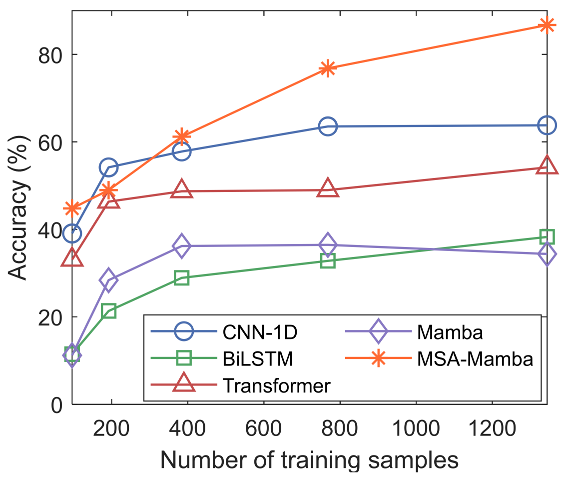


Figure 8. Diagnostic accuracy under single-channel small-sample conditions.

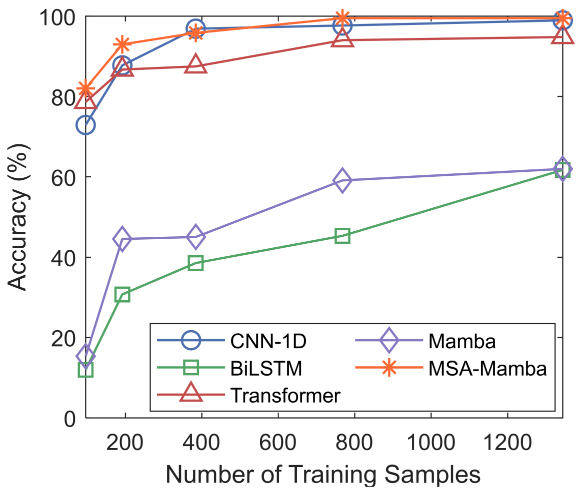


Figure 9. Diagnostic accuracy under three-channel small-sample conditions.

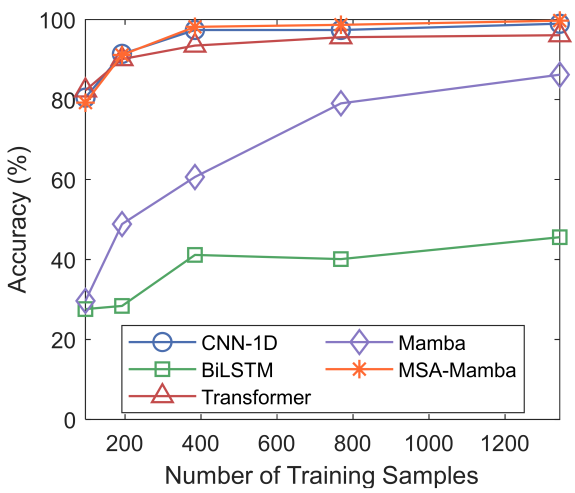


Figure 10. Diagnostic accuracy under four-channel fusion small-sample conditions.

The corresponding quantitative results are summarized in Tables 6–8.

Table 6. Diagnostic accuracy under single-channel small-sample conditions (%).

Model	96	192	384	768	1344
CNN-1D	39.06	54.17	57.81	63.54	63.80
BiLSTM	11.46	21.35	28.91	32.81	38.28
Transformer	33.07	46.35	48.70	48.96	54.17
Mamba	11.20	28.39	36.20	36.46	34.38
MSA-Mamba	44.79	48.96	61.20	76.82	86.72

Table 7. Diagnostic accuracy under three-channel small-sample conditions (%).

Model	96	192	384	768	1344
CNN-1D	72.92	87.76	96.88	97.66	98.96
BiLSTM	11.98	30.73	38.54	45.31	61.72
Transformer	78.65	86.72	87.50	94.01	94.79
Mamba	15.36	44.53	45.05	59.11	61.98
MSA-Mamba	82.03	92.97	95.83	99.48	99.48

Table 8. Diagnostic accuracy under four-channel fusion small-sample conditions (%).

Model	96	192	384	768	1344
CNN-1D	80.47	91.41	97.40	97.40	98.96
BiLSTM	27.60	28.39	41.15	40.10	45.57
Transformer	82.29	90.10	93.49	95.57	96.09
Mamba	29.64	48.85	60.62	79.06	86.20
MSA-Mamba	79.22	91.15	98.18	98.70	99.74

A comparative analysis of the three input modalities demonstrates that conventional sequence models generally suffer from limited generalization under small-sample conditions. BiLSTM is constrained by its recurrent structure and requires sufficient sequential data for effective feature learning, resulting in consistently lower accuracy across different input modes. Mamba lacks feature enhancement and regularization mechanisms, leading to unstable convergence and poor generalization with limited samples. CNN-1D maintains relatively stable performance due to its local feature extraction capability, but its feature representation ability is limited under single-channel conditions. Transformer provides strong global modeling capability; however, its large parameter scale increases the risk of overfitting when training data are insufficient.

In contrast, MSA-Mamba shows prominent advantages in the majority of small-sample scenarios. The CNN-Stem convolutional prior enhancement reduces dependence on large-scale data, while Mixup augmentation and label smoothing effectively alleviate overfitting. For most sample-size configurations and input modalities, MSA-Mamba yields better diagnostic performance than the compared baselines, and tends to converge more rapidly with growing training sample size.

These results indicate that the proposed MSA-Mamba effectively mitigates performance degradation caused by insufficient industrial fault samples and maintains strong generalization capability across different data scales and multi-modal input conditions.

4.5. Robustness Evaluation Under Noise Interference

Industrial environments are often affected by various types of environmental noise and electromagnetic interference. To evaluate the anti-noise robustness of the proposed model, different levels of composite Gaussian noise and periodic interference were added to the test signals, and the diagnostic performance of different models under noisy conditions was compared.

Five SNR levels, including 10 dB, 5 dB, 0 dB, −5 dB, and −10 dB, were selected to simulate operating conditions ranging from mild interference to severe noise contamination. The signal degradation process under different SNR levels is illustrated in Figure 11.

As the SNR decreases, fault-related components are gradually masked by noise, resulting in increased waveform distortion and higher difficulty in fault identification.

Based on the previous experimental settings, three input modalities, including single-channel acoustic signals, three-channel vibration signals, and four-channel vibration–acoustic fusion signals, were further evaluated. The diagnostic accuracy of different models under various noise levels is presented in Figures 12–14.

The corresponding quantitative results are summarized in Tables 9–11.

The experimental results demonstrate that conventional models generally exhibit insufficient robustness against noise interference. BiLSTM and Mamba are highly sensitive to noise, and their feature extraction capability significantly degrades under low-SNR conditions. Although CNN-1D and Transformer possess certain feature selection capabilities, their classification performance decreases substantially under severe noise environments.

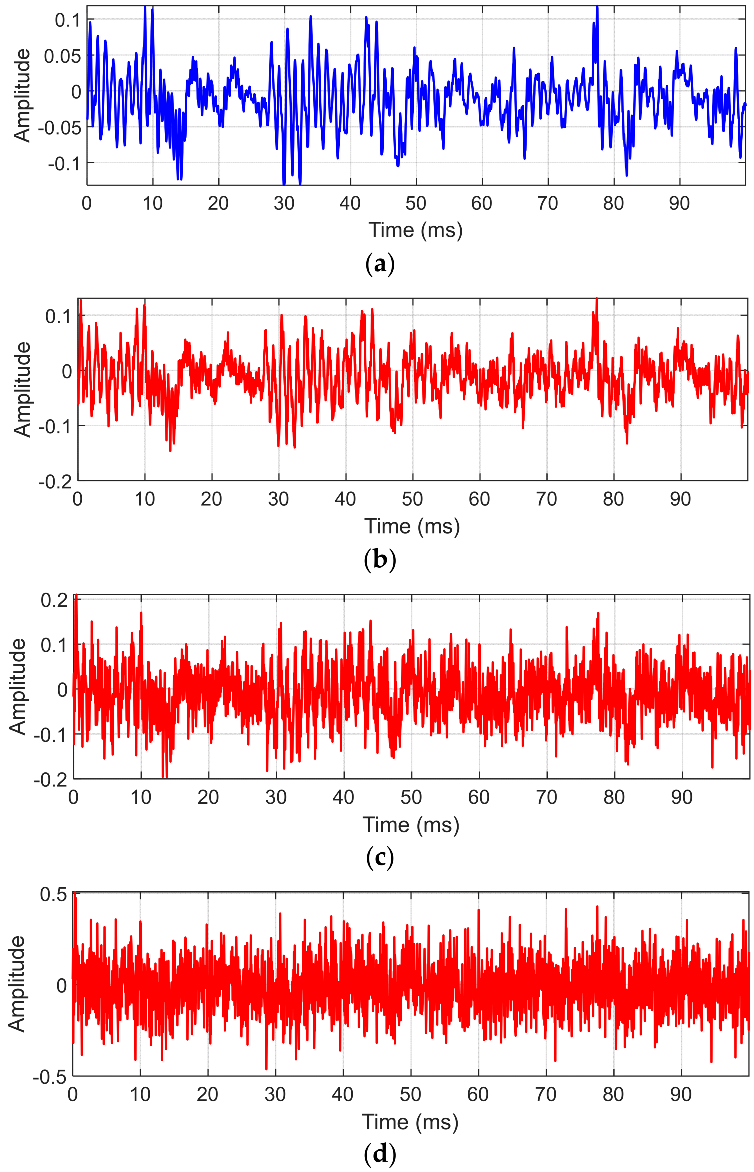


Figure 11. Examples of noisy signals (30 Hz X-direction vibration signal, Broken-Tooth fault): (**a**) Original signal; (**b**) 10 dB; (**c**) 0 dB; (**d**) −10 dB.

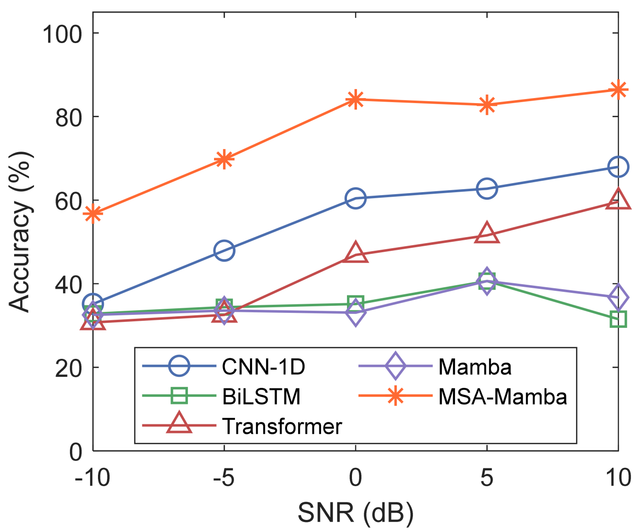


Figure 12. Diagnostic accuracy of single-channel acoustic signals under different SNR levels.

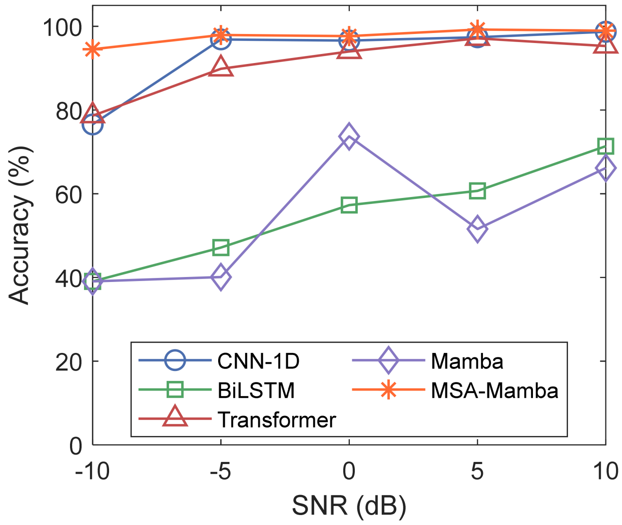


Figure 13. Diagnostic accuracy of three-channel vibration signals under different SNR levels.

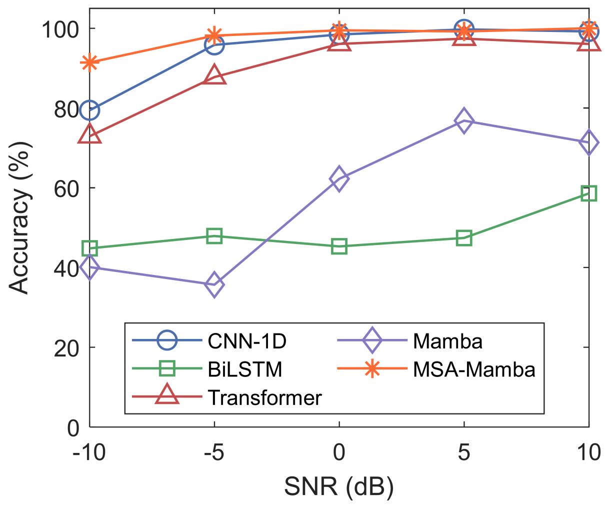


Figure 14. Diagnostic accuracy of four-channel vibration–acoustic fusion signals under different SNR levels.

Table 9. Diagnostic accuracy of single-channel acoustic signals under different SNR levels (%).

Model	10 dB	5 dB	0 dB	−5 dB	−10 dB
CNN-1D	67.97	62.76	60.42	47.92	35.16
BiLSTM	31.51	40.62	35.16	34.38	32.81
Transformer	59.64	51.56	46.88	32.55	30.73
Mamba	36.72	40.62	33.07	33.59	32.55
MSA-Mamba	86.46	82.81	84.11	69.79	56.77

Table 10. Diagnostic accuracy of three-channel vibration signals under different SNR levels (%).

Model	10 dB	5 dB	0 dB	−5 dB	−10 dB
CNN-1D	98.70	97.40	96.61	96.88	76.56
BiLSTM	71.35	60.68	57.29	47.14	39.06
Transformer	95.31	97.14	94.01	89.84	78.65
Mamba	66.15	51.56	73.70	40.10	39.06
MSA-Mamba	98.96	99.22	97.66	97.92	94.53

Table 11. Diagnostic accuracy of four-channel vibration–acoustic fusion signals under different SNR levels (%).

Model	10 dB	5 dB	0 dB	−5 dB	−10 dB
CNN-1D	99.22	99.74	98.44	95.83	79.43
BiLSTM	58.59	47.40	45.31	47.92	44.79
Transformer	96.09	97.40	96.09	87.76	72.92
Mamba	71.35	76.82	62.24	35.68	40.10
MSA-Mamba	99.46	99.22	99.48	98.18	91.41

In contrast, the proposed MSA-Mamba maintains superior robustness across all noise levels and input modalities. Benefiting from the frequency-domain adaptive filtering module, the model can dynamically suppress irrelevant spectral interference while preserving fault-related frequency components. Meanwhile, the multi-scale convolutional attention mechanism further enhances discriminative feature extraction under noisy conditions.

Particularly, under the severe noise condition of −10 dB, the four-channel vibration–acoustic fusion model still achieves a diagnostic accuracy of 91.41%, improving by 15.08%, 104%, 25.35%, and 126% compared with CNN-1D, BiLSTM, Transformer, and Mamba, respectively. These results verify that the proposed MSA-Mamba can effectively adapt to complex industrial environments with strong noise interference.

4.6. Impact of Signal Characteristics on MSA-Mamba Performance

Based on the foregoing experimental results, the performance gain of MSA-Mamba over the baseline models is not constant but is closely associated with the feature richness and degradation level of the input signals. Under relatively ideal conditions with rich information, such as multi-channel inputs, high signal-to-noise ratios, and sufficient training samples, 1D-CNN can effectively extract local impact features, leaving only a limited performance margin for MSA-Mamba. In contrast, under information-scarce or severely degraded conditions, including single-channel acoustic signals, heavy noise interference, and extremely limited training samples, MSA-Mamba exhibits substantially greater advantages. This improvement can be attributed to the complementary capabilities of multi-scale convolutional attention in enhancing weak local features, parallel SSM blocks in modeling long-range dependencies, and frequency-domain adaptive filtering in suppressing noise. These findings indicate that the primary advantage of MSA-Mamba lies not in pursuing marginal performance gains under ideal conditions, but in achieving greater robustness and generalization capability in complex industrial scenarios characterized by weak features, heavy noise, and limited training data.

4.7. Model Complexity and Ablation Analysis

To comprehensively evaluate the computational cost and deployment efficiency of the different models, their parameter counts, MACs, FLOPs, CPU inference latency, and memory consumption were compared. All tests were conducted under the same conditions, using four-channel vibration–acoustic fused inputs with a batch size of 1, thereby ensuring a consistent and fair comparison across models. The detailed results are presented in Table 12.

In terms of theoretical computational cost, MSA-Mamba achieves the lowest computational complexity among all compared models, with only 0.010 G MACs and 0.019 G FLOPs. This indicates that the proposed multi-scale adaptive architecture can effectively exploit temporal redundancy while reducing unnecessary computational operations. In terms of deployment efficiency, MSA-Mamba maintains a low peak memory consumption of 11.14 MB, which is only slightly higher than that of the lightweight CNN-1D and substantially lower than those of Transformer and Mamba. Under the same hardware and testing backend, MSA-Mamba also achieves a considerably lower CPU inference latency than Mamba, reducing the latency from 18.071 ms to 8.164 ms, corresponding to a reduction of approximately 54.8%.

Table 12. Complexity and Deployment Performance Comparison of Different Models.

Model	Params (M)	MACs (G)	FLOPs (G)	Latency (ms)	Memory (MB)
CNN-1D	0.178	0.096	0.192	1.003	10.22
BiLSTM	0.566	0.135	0.269	3.562	48.58
Transformer	0.642	0.202	0.405	2.399	14.10
Mamba	0.176	0.039	0.077	18.071	17.25
MSA-Mamba	0.255	0.010	0.019	8.164	11.14

To evaluate the effectiveness of the key components and training strategies in MSA-Mamba, an ablation study was conducted to investigate the contributions of Frequency Filtering, Multi-Scale Attention, Mixup, and Label Smoothing to the diagnostic performance. Specifically, while keeping the backbone architecture, training settings, and data partitioning unchanged, individual components were removed separately, and Frequency Filtering and Multi-Scale Attention were further removed simultaneously to examine both their individual contributions and potential complementary effects. To further assess the model's adaptability under data-limited and noisy conditions, the experiments were conducted with 96, 192, and 384 training samples under two representative SNR levels of 10 dB and −10 dB. These experiments provide a systematic evaluation of the contributions of the individual components to few-shot learning and robustness under severe noise. The experimental results are presented in Figure 15.

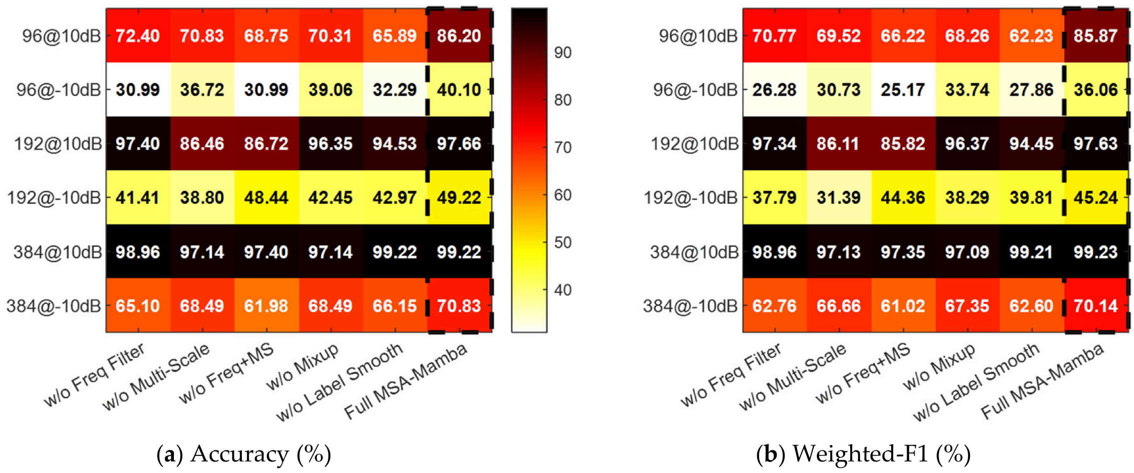


Figure 15. Ablation Performance under Different Data and Noise Conditions.

The ablation results demonstrate that full MSA-Mamba achieves consistently strong performance under different data and noise conditions. At 10 dB SNR, its Accuracy increases from 86.20% to 97.66% and 99.22% as the training samples increase from 96 to 192 and 384, with corresponding Weighted-F1 values of 85.87%, 97.63%, and 99.23%. With only 96 samples, removing Label Smoothing causes the largest degradation, reducing Accuracy to 65.89%, or 20.31 percentage points below the full model, highlighting the importance of regularization in limited-data settings.

Under −10 dB SNR, all variants experience substantial performance degradation, while the full model maintains the best or near-best performance. Its Accuracy increases from 40.10% to 49.22% and 70.83% as the training samples increase, with Weighted-F1 reaching 36.06%, 45.24%, and 70.14%, respectively. With 384 samples, removing both Frequency Filtering and Multi-Scale Attention reduces Accuracy to 61.98%, which is 8.85 percentage points lower than the full model. These results confirm the complementary contributions of frequency enhancement and multi-scale feature modeling and demonstrate the robustness of MSA-Mamba under limited-data and severe-noise conditions.

Notably, the Weighted-F1 and Macro-F1 values are nearly identical across the experimental settings. This is mainly attributed to the relatively balanced distribution of the fault classes in the evaluation dataset, such that the class-frequency weighting used in Weighted-F1 has only a limited effect on the overall score. Therefore, both metrics provide highly consistent performance trends, and Macro-F1 is not separately reported to avoid redundant information.

5. Conclusions

This study addresses the challenges of strong noise interference, prominent multi-scale characteristics, and limited sample availability in motor fault signals under complex industrial environments. A frequency-enhanced multi-scale adaptive state space Mamba (MSA-Mamba) is proposed for robust intelligent fault diagnosis. The proposed model introduces convolutional embedding to incorporate local structural priors and enhance feature representation. A frequency-domain adaptive filtering module is developed to suppress noise interference while emphasizing critical frequency components. Furthermore, multi-scale convolutional attention is integrated with a parallel state space model to achieve joint modeling of local fine-grained features and long-range temporal dependencies. In addition, Mixup and label smoothing strategies are employed to further improve the model generalization capability.

Experimental results demonstrate that MSA-Mamba achieves the best or second-best performance across single-channel, three-channel, and four-channel multimodal inputs. In particular, the proposed method achieves a diagnostic accuracy of 99.48% under the four-channel fusion condition and maintains stable performance under limited-sample and high-noise scenarios, outperforming the compared approaches overall. The single-channel experiments verify the capability of the proposed model in extracting weak fault-related features, while the three-channel and four-channel experiments demonstrate its effectiveness in multi-scale and multimodal feature fusion. The few-shot experiments further confirm its strong generalization ability, and the noise robustness experiments validate its superior resistance to interference.

In summary, the proposed MSA-Mamba achieves superior feature extraction performance and strong robustness against synthetic noise perturbations on the publicly available HUST motor multimodal dataset under the considered operating conditions. The experimental results demonstrate its effectiveness and reliability for motor fault diagnosis across the evaluated scenarios. Future work will investigate lightweight deployment and online adaptive updating mechanisms to further improve the real-time applicability of MSA-Mamba on resource-constrained industrial edge devices.

Author Contributions: Conceptualization, X.L. and M.H.; methodology, X.L.; validation, X.A.; investigation, X.A.; data curation, X.A.; writing—original draft preparation, M.H.; writing—review and editing, J.L. and H.L.; visualization, X.L.; supervision, H.L.; funding acquisition, none. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: Regarding the potential use of AI writing tools, we would like to clarify that AI tools (Deepseek-V3.1) were only used for language polishing, grammar checking, and improving academic expression during manuscript preparation. The AI tools were not involved in the development of the proposed MSA-Mamba model, experimental design, data processing, result analysis, or scientific conclusions.

Conflicts of Interest: Xingpeng An was employed by Hangzhou Zhiyuan Research Institute Co., Ltd. Other authors declare no conflict of interest.

References

1. Romanssini, M.; de Aguirre, P.C.C.; Compassi-Severo, L.; Girardi, A.G. A review on vibration monitoring techniques for predictive maintenance of rotating machinery. *Eng* **2023**, *4*, 1797–1817. [CrossRef]
2. Wang, H.; Xie, J. Fault diagnosis of rolling bearings based on acoustic signals in strong noise environments. *Appl. Sci.* **2025**, *15*, 1389. [CrossRef]
3. Li, G.; Yang, F. Intelligent fault diagnosis technology for mechanical equipment: Current status and development. *Sustain. Dev.* **2025**, *1*, 40–50. [CrossRef]
4. Chan, F.T.S.; Ding, K. Industrial intelligence-driven production and operations management. *Int. J. Prod. Res.* **2023**, *61*, 4215–4219. [CrossRef]
5. Xia, L.; Han, Q.; Yu, S. Industrial intelligence and industrial structure change: Effect and mechanism. *Int. Rev. Econ. Financ.* **2024**, *93*, 1494–1506. [CrossRef]
6. Cen, J.; Yang, Z.; Liu, X.; Xiong, J.; Chen, H. A review of data-driven machinery fault diagnosis using machine learning algorithms. *J. Vib. Eng. Technol.* **2022**, *10*, 2481–2507. [CrossRef]
7. Zhang, C.; Wang, W.; Zhang, C.; Fan, B.; Wang, J.; Gu, F.; Yu, X. Extraction of local and global features by a convolutional neural network–long short-term memory network for diagnosing bearing faults. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2022**, *236*, 1877–1887. [CrossRef]
8. Xue, Y.; Yang, R.; Chen, X.; Tian, Z.; Wang, Z. A novel local binary temporal convolutional neural network for bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 3525013. [CrossRef]
9. Xu, Y.; Yan, X.; Sun, B.; Liu, Z. Global contextual residual convolutional neural networks for motor fault diagnosis under variable-speed conditions. *Reliab. Eng. Syst. Saf.* **2022**, *225*, 108618. [CrossRef]
10. Wang, Y.; Xiao, Z.; Cao, G. A convolutional neural network method based on Adam optimizer with power-exponential learning rate for bearing fault diagnosis. *J. Vibroeng.* **2022**, *24*, 666–678. [CrossRef]
11. Jia, L.; Chow, T.W.S.; Wang, Y.; Yuan, Y. Multiscale residual attention convolutional neural network for bearing fault diagnosis. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 3519413. [CrossRef]
12. Ahmed, H.O.A.; Nandi, A.K. Convolutional-transformer model with long-range temporal dependencies for bearing fault diagnosis using vibration signals. *Machines* **2023**, *11*, 746. [CrossRef]
13. Huang, L.; Huang, J.; Chen, P.; Li, H.; Cui, J. Long-term sequence dependency capture for spatiotemporal graph modeling. *Knowl.-Based Syst.* **2023**, *278*, 110818. [CrossRef]
14. Jiang, B.; Yang, H.; Wang, Y.; Liu, Y.; Geng, H.; Zeng, H.; Ding, J. Dynamic temporal dependency model for multiple steps ahead short-term load forecasting of power system. *IEEE Trans. Ind. Appl.* **2024**, *60*, 5244–5254. [CrossRef]
15. Shahid, M.A.; Zhang, X.; Qin, X.; Peng, K. A novel deep multi-task learning model for spatial–temporal fault detection and diagnosis in industrial systems. *Can. J. Chem. Eng.* **2025**, *103*, 5910–5934. [CrossRef]
16. Li, X.; Wu, X.; Yu, J.; Wang, T.; Shi, C.; Choi, J.H.; Chu, F. A fault diagnosis data augmentation method integrating multimodal non-Gaussian denoising diffusion generative adversarial network. *Adv. Eng. Inform.* **2025**, *68*, 103776. [CrossRef]
17. You, Y.; Zhao, Y.; Ke, Y.; Tang, J.; Yang, B. A novel solution for data uncertainty and insufficient in data-driven chiller fault diagnosis based on multi-modal data fusion. *Energy Build.* **2025**, *328*, 115197. [CrossRef]
18. Yang, X.; Ye, T.; Yuan, X.; Zhu, W.; Mei, X.; Zhou, F. A novel data augmentation method based on denoising diffusion probabilistic model for fault diagnosis under imbalanced data. *IEEE Trans. Ind. Inform.* **2024**, *20*, 7820–7831. [CrossRef]
19. Li, W.; Zhong, X.; Shao, H.; Cai, B.; Yang, X. Multi-mode data augmentation and fault diagnosis of rotating machinery using modified ACGAN designed with new framework. *Adv. Eng. Inform.* **2022**, *52*, 101552. [CrossRef]
20. Jiang, X.; Zheng, J.; Zhuang, X.; Ge, Z. Ensemble data augmentation for imbalanced fault diagnosis. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 3528312. [CrossRef]
21. Xing, J.; Li, Y.; Zhuang, Y.; Wang, D.; Zhang, H. IEVAEGAN: An input enhancement VAEGAN for rotating component fault diagnosis with extremely limited data. *IEEE Sens. J.* **2024**, *24*, 39958–39966. [CrossRef]
22. Yang, H.; Song, Y.; Wang, D.; Xing, J.; Li, Y. A global information-guided denoising diffusion probabilistic model for fault diagnosis with imbalanced data. *Eng. Appl. Artif. Intell.* **2025**, *147*, 110312. [CrossRef]

23. Zhou, P.; Chen, S.; He, Q.; Wang, D.; Peng, Z. Rotating machinery fault-induced vibration signal modulation effects: A review with mechanisms, extraction methods and applications for diagnosis. *Mech. Syst. Signal Process.* **2023**, *200*, 110489. [CrossRef]
24. Ruan, D.; Wang, J.; Yan, J.; Gühmann, C. CNN parameter design based on fault signal analysis and its application in bearing fault diagnosis. *Adv. Eng. Inform.* **2023**, *55*, 101877. [CrossRef]
25. Qu, H.; Ning, L.; An, R.; Fan, W.; Derr, T.; Liu, H.; Xu, X.; Li, Q. A survey of mamba. *ACM Trans. Intell. Syst. Technol.* **2026**, *17*, 116. [CrossRef]
26. Miao, Q.; Jia, L.; Xie, K.; Fu, K.; Yang, Z. A comprehensive survey and taxonomy of mamba: Applications, Challenges, and Future Directions. *Inf. Fusion* **2025**, *130*, 104094. [CrossRef]
27. Jiao, W.; Han, X.; Yu, C.; Xu, X.; Zhang, Q.; Zhang, X.; Wang, B.; Li, L.; Xu, F. Power load forecasting based on time-frequency domain feature fusion. *Eng. Appl. Artif. Intell.* **2026**, *173*, 114438. [CrossRef]
28. Xiao, P.; Dong, Y.; Zhao, J.; Peng, T.; Geiß, C.; Zhong, Y.; Taubenböck, H. MF-Mamba: Multi-scale convolution and mamba fusion model for semantic segmentation of remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5405916. [CrossRef]
29. Zhao, C.; Shen, W.; Zio, E.; Ma, H. Multimodal unified generalization and translation network for intelligent fault diagnosis under dynamic environments. *Eng. Appl. Artif. Intell.* **2025**, *162*, 112559. [CrossRef]
30. Chen, C.C.; Liu, Z.; Yang, G.; Wu, C.C.; Ye, Q. An improved fault diagnosis using 1D-convolutional neural network model. *Electronics* **2021**, *10*, 59. [CrossRef]
31. Wu, J.; Wang, Z.; Hu, Y.; Tao, S.; Dong, J. Runoff forecasting using convolutional neural networks and optimized bi-directional long short-term memory. *Water Resour. Manag.* **2023**, *37*, 937–953. [CrossRef]
32. Irani, H.; Metsis, V. Positional encoding in transformer-based time series models: A survey. *Int. J. Mach. Learn. Cybern.* **2026**, *17*, 339. [CrossRef]