applied sciences

MDPI

Article

A Fault Diagnosis Method for Plunger Pumps Based on Multi-Scale Convolution and Attention

Linlin Liu, Shuhui Hao *, Ruonan Yin, Kewen Li and Liechong Wang

Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China; z23070140@s.upc.edu.cn (L.L.); b22070022@s.upc.edu.cn (R.Y.); s23070008@s.upc.edu.cn (L.W.)

* Correspondence: haoshuhui2021@163.com

Abstract

Plunger pumps serve as core power equipment in oilfield water injection systems, where their reliable operation directly affects crude oil recovery efficiency and production safety. Failures such as mechanical wear and seal leakage can cause injection pressure fluctuations, increased energy consumption, and even pipeline burst accidents. This study addresses the challenges in plunger pump fault diagnosis, including the difficulty in capturing multi-scale fault features, interference from redundant information in high-dimensional feature spaces, and high model computational complexity. We propose a lightweight fault diagnosis approach called Multi-scale Attention Neural Network (MSLAN), which combines multi-scale convolution and attention mechanisms. In this model, a Separable Multi-scale Fusion Module (SMSF) employs parallel multi-branch convolutional kernels to acquire fault signatures across multiple scales, while computational overhead is reduced through depthwise separable convolution and shared pointwise convolution. Additionally, a Multi-Branch Parallel Attention Module (MBPA) is introduced to finely model complex inter-channel dependencies through a four-branch parallel structure, enhancing the perception of key features and suppressing redundant information. Experimental results on a self-constructed plunger pump dataset, the Case Western Reserve University bearing dataset, and the Southeast University gearbox dataset demonstrate that MSLAN achieves F1-scores of 88.95%, 98.89%, and 99.90%, respectively. While maintaining high diagnostic accuracy, the model exhibits significantly lower parameter count and computational cost compared to baseline models, effectively balancing diagnostic precision and computational efficiency. Ablation studies and visualization analyses further validate the effectiveness of each module. This study establishes an accurate and efficient intelligent fault diagnosis solution for plunger pumps, which is also readily applicable to a broader range of rotating machinery.

Keywords: time series; fault diagnosis; deep neural network; multi-scale convolution

Check for updates

Academic Editor: Thomas Lindner

Received: 26 March 2026
Revised: 12 April 2026
Accepted: 15 April 2026
Published: 12 June 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) license.

1. Introduction

The water injection pump serves as a core power device in oilfield development, maintaining reservoir pressure through high-pressure water injection while undertaking the dual functions of energy conversion and fluid transportation. Sustained production system operation and safety hinge directly on the equipment's operating condition, underscoring its pivotal role in crude oil recovery and oilfield development performance [1]. Due to prolonged exposure to harsh operating conditions such as high pressure and highly corrosive media, water injection pumps are prone to mechanical wear, seal failure, and

other faults. These issues may lead to fluctuations in injection pressure, reduced injection efficiency, increased energy consumption, and, in severe cases, production accidents such as pipeline bursts. In deep oilfield development scenarios, where bottom-hole pressure gradients are high and medium compositions are more complex, water injection pump failures can lead to paralysis of the water injection system, reduce crude oil recovery rates, and severely compromise the economic benefits of oilfield development. Figure 1 shows a photograph of a water injection pump at an oilfield site.



Figure 1. Field photograph of a water injection pump at an oilfield site.

Among various types of water injection pumps, the plunger pump has become the mainstream model in modern oilfield water injection systems due to its high working pressure, high volumetric efficiency, and flexible flow rate adjustment capability. It pressurizes the medium through the reciprocating motion of plungers within cylinders, which alters the volume of the working chamber. However, its complex mechanical structure leads to diverse fault modes. Moreover, these faults involve high-dimensional feature spaces containing substantial redundant and ineffective features with coupling relationships. Different faults may exhibit similar patterns in pressure, flow rate, or vibration anomalies, further interfering with the diagnostic process and reducing the accuracy of fault identification.

In recent years, researchers have applied both traditional machine learning and deep learning to plunger pump fault diagnosis [2–8]. However, existing methods still suffer from several limitations. First, different fault types exhibit scale variations in feature representation, and fixed-size convolution kernels struggle to comprehensively capture multi-scale features, limiting the model's adaptability to complex operating conditions. Second, high-dimensional feature spaces often contain substantial redundant and ineffective information, which interferes with the accuracy of fault identification. Third, existing diagnostic models generally suffer from high computational complexity and large memory consumption, making them difficult to deploy on industrial edge devices and unable to meet the requirements for real-time fault diagnosis and early warning in industrial settings. Therefore, a diagnostic model must not only possess strong feature extraction and representation capabilities but also address the requirements of lightweight design and real-time performance.

To address the above issues, this paper proposes a fault diagnosis method for plunger pumps based on multi-scale convolution and attention, termed MSLAN. The method captures fault features at different scales through a multi-scale fusion convolution module, introduces an attention mechanism to suppress redundant information, and adopts lightweight strategies such as depthwise separable convolutions to reduce model complexity. Experiments on a self-constructed plunger pump dataset and two public datasets validate the effectiveness and generalizability of the proposed method.

2. Related Work

2.1. Fault Diagnosis Methods Based on Signal Processing and Traditional Machine Learning

In recent years, researchers have extensively investigated strategies combining signal processing with traditional machine learning for plunger pump fault diagnosis. Such methods typically first employ advanced signal processing techniques to decompose and enhance features from raw signals, followed by classification using traditional machine learning models.

For instance, Gao et al. [2] used pump discharge pressure signals as the analysis object and adopted Empirical Wavelet Transform (EWT) for signal decomposition and reconstruction to extract features, ultimately validating the effectiveness of this feature extraction method through experiments. Zhao et al. [3] proposed a method combining Local Mean Decomposition (LMD) with Support Vector Machine (SVM). They reconstructed vibration signals using LMD and extracted sample entropy and standard deviation to construct feature vectors, achieving high-precision diagnosis of various fault states in plunger pumps. Li et al. [4] focused on the pulsating pressure signals of plunger pumps, employing wavelet packet transform for denoising and feature frequency band extraction, and constructed a BP neural network diagnostic model to identify fault states based on time-domain features, verifying the effectiveness of this method in plunger pump fault diagnosis.

Although these methods have achieved certain results, their inherent limitations constrain practical application effectiveness. On one hand, such methods heavily rely on signal processing techniques for manual feature design. This process not only requires deep domain expertise but also makes the quality of feature extraction directly dependent on expert experience, rendering the diagnostic process time-consuming and labor-intensive. On the other hand, shallow network models such as SVM and BP neural networks, limited by their structural constraints, possess insufficient capacity for representing fault features and struggle to comprehensively capture multi-scale characteristics of signals under complex operating conditions, resulting in inadequate generalization ability when facing actual working conditions such as variable loads and strong noise.

2.2. Fault Diagnosis Methods Based on Deep Learning

Convolutional neural networks (CNNs) have demonstrated marked advantages in mechanical fault diagnosis, largely due to their hierarchical feature learning mechanism—a key strength enabled by the rapid progress of deep learning.

Tang et al. [5] proposed an axial plunger pump fault diagnosis method combining CNN-SE-LSTM with multi-sensor data. Experimental results showed high accuracy and robustness even under noise interference. Xu et al. [6] proposed an axial plunger pump fault diagnosis method based on CNNs, achieving a diagnostic accuracy of 92.17% by training on time-frequency images. Huang et al. [7] introduced a D-1DCNN-based framework for identifying faults in axial plunger pumps, which leverages deep learning to autonomously extract features from vibration signals. The results indicated that this method achieves a fault diagnosis accuracy of 95%. Wang et al. [8] proposed a rolling bearing fault diagnosis

method based on a multi-feature parallel fusion encoder and long short-term memory network (MPFE-LSTM). This method extracts frequency-domain features and employs a parallel fusion encoder to capture both local and global features, maintaining high diagnostic accuracy even under strong noise interference.

3. Related Theory

3.1. Multi-Scale Convolution Block

The structure of the Multi-scale Convolution (MC) block [9] is illustrated in Figure 2. Unlike conventional single-scale convolution kernels, which can only capture features from a single local receptive field, this module employs parallel convolution kernels of different sizes to simultaneously extract feature information from multiple local receptive fields, effectively enhancing the capability to capture features at different scales. Specifically, features extracted by different convolution kernels are first concatenated along the channel dimension. Subsequently, a Batch Normalization (BN) layer is applied to maintain the stability of information distribution along the channel dimension, preventing data distribution shifts caused by feature concatenation. Finally, the GELU (Gaussian Error Linear Units) activation function is introduced for nonlinear mapping. Compared with traditional activation functions such as ReLU, GELU preserves partial gradients of negative features, alleviates the gradient vanishing problem, and enhances the model's ability to represent complex features.

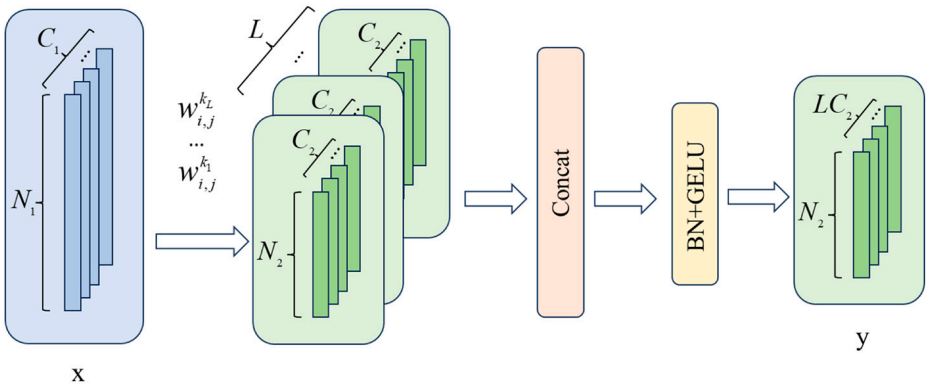

Figure 2. MC block.

The above multi-scale convolution process can be expressed by Equations (1) and (2):

$$y^{k_l} = Concat_{j=1}^{C_2}\left(\sum_{i=1}^{C_1} w_{i,j}^{k_l} * x_i\right) \tag{1}$$

$$y = GELU\left(BN\left(Concat\left(y^{k_1}, y^{k_2}, \cdots, y^{k_L}\right)\right)\right) \tag{2}$$

where $X \in R^{C_1 \times N_1}$ denotes the input feature matrix, C_1 indicates the input channel count, N_1 corresponds to the input dimension, $y \in R^{LC_2 \times N_2}$ represents the output feature matrix, C_2 specifies the number of output channels per convolution kernel, N_2 gives the output dimension, L denotes the quantity of parallel convolution kernel sizes, $w^{k_l} \in R^{C_2 \times C_1 \times k_l}$ stands for the weight parameters of the l-th convolution kernel, * signifies the convolution operation, and $Concat(\cdot)$ indicates the feature channel concatenation operation.

Using the convolution operation as a foundation, the parameter count (Params) and floating-point operations (FLOPs) for the multi-scale convolution block are derived. These two metrics correspond to the module's spatial complexity and time complexity, respectively, and are given by:

$$Params_{MC} = \sum_{l=1}^{L} k_l \times C_1 \times C_2 \tag{3}$$

$$FLOPs_{MC} = \sum_{l=1}^{L} k_l \times C_1 \times C_2 \times N_2 \tag{4}$$

3.2. Attention Mechanism

Attention mechanisms have represented a notable development in deep learning over the past several years. The central concept lies in empowering models to selectively concentrate on pertinent input features while filtering out extraneous content, consequently improving discriminative power and resilience. Initially applied to natural language processing, this approach has subsequently been extended to other areas, including computer vision and audio processing. The Squeeze-and-Excitation (SE) module, introduced by Hu et al. [10], exemplifies a classic channel attention mechanism designed to augment the representational capability of convolutional neural networks. Its structure is depicted in Figure 3.

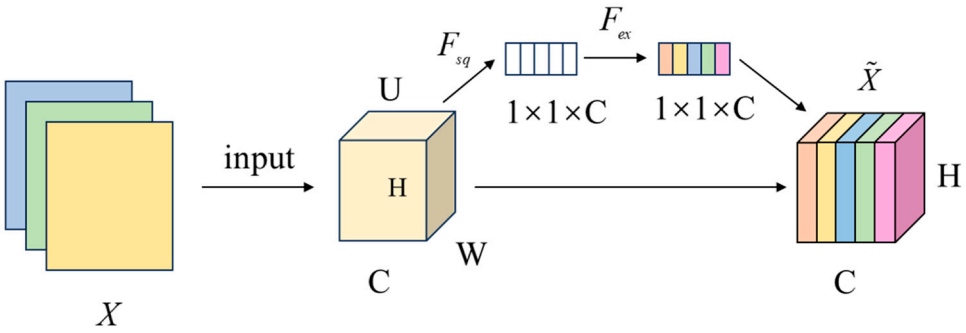

Figure 3. SE block.

Adaptively weighting individual channels constitutes the core function of the SE block, which encourages the model to prioritize useful features. The structural procedure is carried out in the following manner:

1. Squeeze Stage

Global average pooling (GAP) is employed to collapse the spatial dimensions of each channel in the feature map x into a single scalar. Assuming the input feature dimensions are $X \in R^{H \times W \times C}$, after global pooling, a vector C is generated:

$$z_C = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j,c} \tag{5}$$

where z_c denotes the global feature corresponding to channel c, with H and W representing the height and width of the feature map, and C indicating the channel count.

2. Excitation Stage

The Squeeze-stage output z is fed into two fully connected layers, which generate adaptive per-channel weights. These layers incorporate a ReLU activation function, succeeded by a Sigmoid function that scales the activations to the $[0, 1]$ range. The weight for each channel is output as:

$$S = \sigma(W_2 \delta(W_1 z)) \tag{6}$$

where W_1 and W_2 correspond to the weights of the two fully connected layers, δ represents ReLU activation, and σ denotes the Sigmoid function. Through this process, adaptive weights are learned for each channel.

3. Recalibration Stage

The channel weights s are applied to the input feature map through element-wise multiplication, yielding weighted channels where critical information is emphasized and irrelevant content is suppressed:

$$\hat{X}_C = S_C \cdot X_C \tag{7}$$

where S_C is the weight for channel c.

4. Method

To address the limitations of fixed convolution kernels in capturing multi-scale fault features and the adverse effects of redundant information within high-dimensional feature spaces on diagnostic accuracy, a fault diagnosis framework named Multi-scale Attention Neural Network (MSLAN) is proposed for plunger pump applications. This framework integrates multi-scale convolution with attention mechanisms to enhance fault recognition performance through multi-level feature fusion. The architecture comprises several key components. First, a Separable Multi-scale Fusion (SMSF) module is designed to extract multi-dimensional fault features. Next, a Multi-Branch Parallel Attention (MBPA) module applies channel-level optimization, strengthening critical fault indicators while attenuating redundant information. Global average pooling then aggregates and compresses the feature information across output channels, followed by a fully connected layer that maps the resulting features to a fault category weight vector, where the vector dimension corresponds to the number of fault types. Finally, the Softmax function converts this vector into a probability distribution, outputting the probability of the sample belonging to each fault category and thereby achieving accurate fault state identification. The structure of the proposed model is depicted in Figure 4.

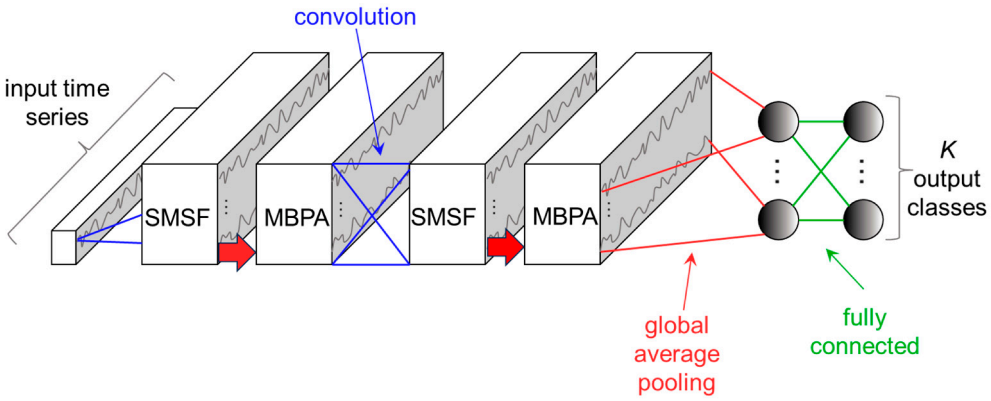

Figure 4. Structure of the proposed MSLAN.

4.1. Separable Multi-Scale Fusion Module (SMSF)

In multi-scale convolution modules, cross-channel convolution can integrate information from various channel dimensions. Traditional cross-channel convolution requires independently computing convolution kernel weights for all input channels for each output channel, which leads to a significant increase in parameter count and computational complexity, failing to meet the requirements of real-time plunger pump fault monitoring in industrial settings. To address this issue, this paper proposes a Separable Multi-scale Fusion Module (SMSF), which decouples cross-channel and spatial convolution to achieve efficient multi-scale feature extraction. Its core structure is illustrated in Figure 5 and consists of three stages:

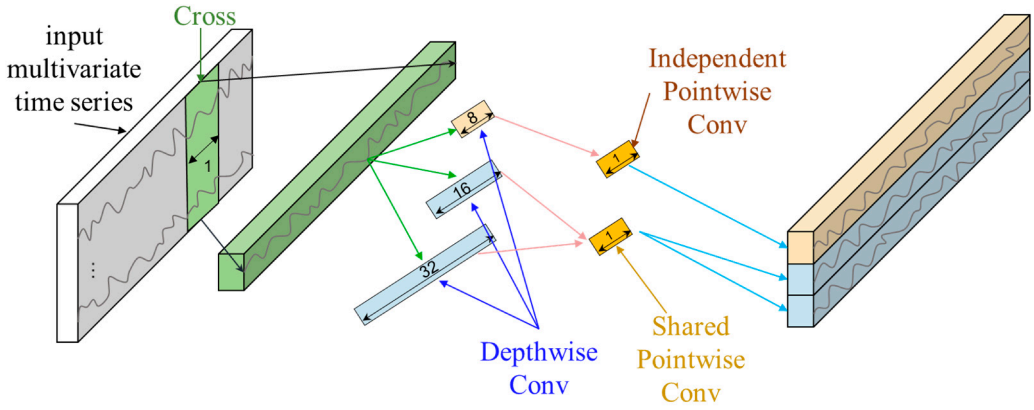

Figure 5. Structure of the SMSF module.

1. Dynamic Channel Compression and Cross-Channel Fusion

Cross-channel convolution with a kernel size of 1 is employed to integrate multi-channel information and adjust the channel dimension. The expression can be formulated as:

$$z = Concat_{j=1}^{C_2}\left(\sum_{i=1}^{C_1} w_{ij}^{1} * x_i\right) \tag{8}$$

where $X \in R^{C_1 \times N_1}$ is the input, $z \in R^{C_2 \times N_1}$ represents the compressed feature, C_2 is the number of compressed channels, $w^1 \in R^{C_2 \times C_1}$ denotes the weight of the 1 convolution, $Concat(\cdot)$ denotes the convolution operation, and N_1 is the input time dimension.

2. Multi-scale Depthwise Convolution with Shared Weights

Following this step, parallel separable convolutions with different kernel sizes are performed on z. Separable convolution links each input channel to a corresponding output channel on a one-to-one basis, yielding substantial savings in learnable parameters and FLOPs. At the same time, varying kernel sizes in parallel separable convolutions ensure multi-scale local receptive field feature extraction remains intact. For convolution kernels where $k > 8$, shared pointwise convolution is introduced to eliminate redundancy. This procedure is formulated as:

$$y^{k_l} = Concat_{j=1}^{C_2}\left(W^{k_l} * \left(v_j^{k_l} * z_j\right)\right) \tag{9}$$

where $v^{k_1} \in R^{C_2 \times k_l}$ represents the weight of the depthwise convolution with kernel size k_l and W^{k_l} denotes the pointwise convolution weight. For branches where $k_l \leq 8$, each branch independently learns W^{k_l}; for branches where $k_l > 8$, the same W^{k_l} is shared.

3. Feature Fusion and Nonlinear Mapping

The features from multiple local receptive fields are concatenated and mapped through a batch normalization (BN) layer and Gaussian Error Linear Unit (GELU) activation function:

$$y = GELU\left(BN\left(Concat\left(y^{k_1}, \ldots, y^{k_L}\right)\right)\right) \tag{10}$$

The optimized parameter count and computational complexity are as follows:

$$Params_{SMCF} = C_1 \times C_2 + C_2 \times \sum_{l=1}^{L} k_l + (L - m) \times C_2^2 + C_2^2 \tag{11}$$

$$FLOPS_{SMCF} = C_1 \times C_2 \times N_1 + C_2 \times N_2 \times \sum_{l=1}^{L} k_l + L \times N_2 \times C_2^2 \tag{12}$$

where m denotes the number of branches with shared pointwise convolution (in this paper, $m = 2$), and $L - m$ denotes the number of branches with independent pointwise convolution. This design employs shared pointwise convolution for large kernel branches to streamline parameters and reduce computational redundancy, while retaining independent learning capabilities for small kernel branches to ensure fine-grained feature extraction accuracy. Through the differentiated configuration of shared and independent mechanisms, this approach balances the requirements for multi-scale fault feature representation in plunger pumps while enhancing the computational efficiency of the module in industrial scenarios.

4.2. Multi-Branch Parallel Attention Module (MBPA)

In multi-scale convolutional feature fusion, the contribution of different channels to the final output varies significantly as the number of channels increases. Inspired by the multi-scale convolution module, this paper proposes a Multi-Branch Parallel Attention Module (MBPA). Compared with the classic SE module, MBPA employs a four-branch parallel structure instead of a single-branch design, enhancing channel dependency modeling capability to achieve adaptive reinforcement of critical features and precise suppression of redundancy. Additionally, a residual connection is introduced to alleviate the gradient vanishing problem in deep networks and preserve original feature information, further improving adaptability to multi-branch convolutional features. The core structure of the module is illustrated in Figure 6, and its specific workflow is as follows:

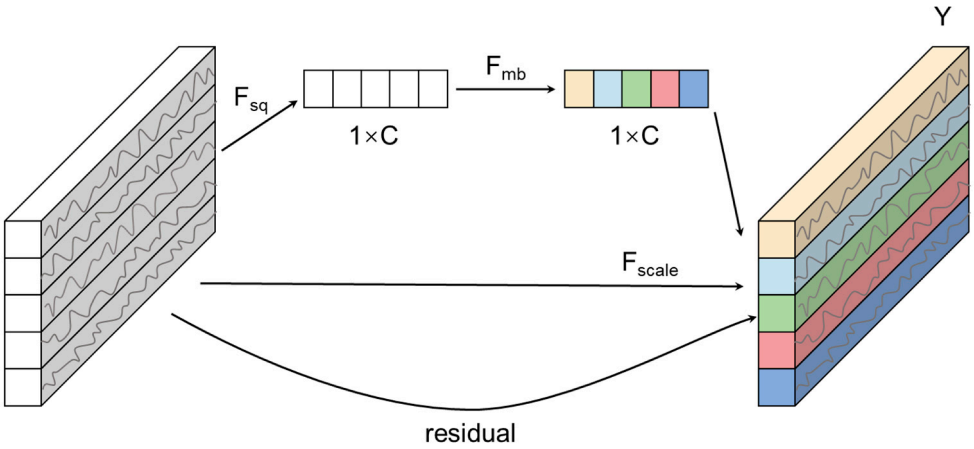

Figure 6. Structure of the MBPA module.

1. Squeeze Stage

By applying adaptive one-dimensional global average pooling to the input feature map, the sequence dimension is compressed and channel statistics are generated. Assuming the input feature dimensions are $X \in R^{C \times L}$, with C indicating the channel count and L the sequence length, the resulting channel descriptor from global average pooling is:

$$z_C = \frac{1}{L} \sum_{i=1}^{L} X_{C,i} \tag{13}$$

where z_c represents the global feature of the c-th channel, $z \in R^C$.

2. Multi-Branch Excitation Stage

The vector z obtained from the Squeeze stage is processed through parallel fully connected layers in multiple branches. Each branch performs linear transformation followed by the ReLU activation function:

$$w_k = \delta(W_{1k} z) \tag{14}$$

where $W_{1k} \in R^{C/r \times C}$ is the weight matrix of the k-th branch, r is the channel reduction factor, and $w_k \in R^{C/r}$. The outputs of the four branches are then concatenated along the channel dimension:

$$w = Concat(w_1, w_2, w_3, w_4) \in R^{4C/r} \tag{15}$$

The feature dimension is then restored to the original number of channels C through a second fully connected layer, and the channel attention weights are generated using the Sigmoid function:

$$s = \sigma(W_2 w) \tag{16}$$

where $W_2 \in R^{4C/r \times C}$ is the weight matrix, σ denotes the Sigmoid function, and the output $s \in R^C$.

3. Scale Recalibration and Residual Connection

The channel attention weights s learned through the process are used to perform channel-wise recalibration on the input features X_C, achieving enhancement of important feature channels and suppression of redundant channels. Specifically, element-wise multiplication is performed between the channel weights s and the input features X_C:

$$\hat{X}_C = s \cdot X_C \tag{17}$$

where $\hat{X}_C$ denotes the recalibrated feature map. To further maintain smooth gradient flow and enhance feature representation capability, a residual connection mechanism is introduced:

$$Y = X + \hat{X}_C \tag{18}$$

where Y is the final output of the module. The introduction of the residual connection not only preserves the original feature information but also reinforces the representation of important features through attention weighting.

The MBPA module, through its multi-branch parallel design, overcomes the limitation of traditional single-branch attention mechanisms in modeling complex channel dependencies. This module can more effectively capture intricate inter-channel dependencies, enhancing the model's sensitivity to weak fault features in plunger pumps and the specificity of key feature selection, thereby providing more discriminative feature support for subsequent fault classification tasks.

5. Experiment

5.1. Dataset

To comprehensively validate the effectiveness and generalization capability of the MSLAN model presented herein, three datasets were selected for experimental validation. Real plunger pump data collected from an oilfield in Shandong Province constitute a self-constructed dataset used to verify the method's diagnostic performance in real-world engineering settings. Generalization across rotating machinery fault diagnosis tasks is assessed using the CWRU rolling bearing dataset [11] and the Southeast University gearbox dataset [12], demonstrating that the proposed approach is applicable beyond plunger pumps and can be extended to other mechanical components.

Self-Built Dataset: This dataset originates from real operating condition data of a plunger pump in an oilfield in Shandong Province, reflecting equipment operating states in actual industrial environments. The experimental object is a triplex single-acting plunger pump used in the oilfield water injection system. The monitoring system deploys multiple types of sensors at key positions of the pump body, including a pump discharge pressure sensor, a housing vibration acceleration sensor, a crosshead temperature sensor, and a

drive motor current sensor. The data acquisition system synchronously records signals from each channel, forming a multi-source time-series dataset. The dataset contains three fault types of the plunger pump, as well as a healthy state. The fault types are connecting flange misalignment, stuffing box leakage, and dynamic seal failure. The dataset contains a total of 2412 samples, including 954 samples of healthy status, 504 samples of connecting flange misalignment, 504 samples of stuffing box leakage, and 450 samples of dynamic seal failure. Figure 7 illustrates the sensor placement of the field data acquisition platform, where markers 1–7 indicate the installation positions of vibration sensors for collecting signals in the X, Y, and Z directions.



Figure 7. Vibration sensor placement on the field data acquisition platform.

CWRU Dataset: The rolling bearing dataset from Case Western Reserve University (CWRU) is among the most authoritative and extensively employed public datasets in fault diagnosis research, originating from the university's Electrical Engineering Laboratory. To assess the proposed method's performance in bearing fault identification, this dataset was selected for supplementary experiments. The experimental setup comprises a motor, a torque sensor, and a dynamometer, with the test bearings supporting the motor shaft. During the experiments, three categories of single-point defects—namely inner race, outer race, and ball faults—were artificially introduced into the bearings via electric discharge machining, with fault diameters of 0.007, 0.014, and 0.021 inches [11]. Vibration signals were acquired using an accelerometer positioned on the drive-end motor housing, operating at a sampling frequency of 12 kHz. For this study, the operating condition under 0 hp load was adopted, with a sampling length of 512 and an overlap rate of 0.75. Specifically, ten distinct fault types were selected for this study, with each type containing 117 samples.

Southeast University Dataset: Originating from Southeast University, the university's gearbox dataset is recognized as one of the commonly used public datasets within the fault diagnosis community. This dataset was selected for supplementary experiments to validate the proposed method's capability for gear fault diagnosis. The experimental platform comprises a motor, motor controller, planetary gearbox, parallel-axis gearbox, brake, and controller, with the test gears housed inside the gearbox. During the experiments, four categories of faults—tooth surface wear, gear root crack, tooth defect, and tooth breakage—were artificially created in the gears through machining, alongside a healthy state serving as a control [12]. Vibration measurements were obtained via accelerometers installed at strategic locations on the test rig, collecting signals from eight channels at

a sampling frequency of 5120 Hz. The operating condition with a rotational speed of 1200 r/min and a load of 0 N·m was selected for this study, utilizing a sampling length of 1024 and an overlap rate of 0.75. Specifically, five distinct health conditions were selected, with each condition containing 4092 samples.

Each dataset was randomly divided into training, validation, and test sets. Table 1 summarizes the sample sizes for each dataset.

Table 1. Statistics of datasets.

Dataset	Train	Val	Test	Total
Self-Built Dataset	1928	242	242	2412
CWRU Dataset	640	170	360	1170
Southeast University Dataset	11,455	2865	6140	20,460

5.2. Experimental Setup

All experiments were conducted on a hardware platform consisting of an Intel Core i5-12400 processor, an NVIDIA GeForce RTX 5060 8 GB graphics card, and 16 GB RAM. The programs were written in Python 3.9, with the deep learning framework PyTorch 2.7.1 and CUDA version 12.8.

To comprehensively evaluate the performance of MSLAN, all model parameters were initialized according to a preset random seed. During the training process, parameters were updated through the backpropagation algorithm and optimized using the Adam optimizer. The learning rate was set to 0.001, and the batch size was set to 32. To ensure convergence to the optimal state, the maximum number of training epochs was set to 200, and an early stopping mechanism was introduced. The F1-Score on the validation set was used as the monitoring metric. If this metric showed no improvement for ten consecutive training epochs, training was terminated early, and the model weights with the best performance were saved.

To evaluate the robustness of the proposed model in complex industrial environments, Gaussian white noise was added to the raw signals of the CWRU rolling bearing dataset and the Southeast University gearbox dataset to simulate background noise interference under actual operating conditions. A single noise level was employed in this experiment, with a uniform signal-to-noise ratio (SNR) of 10 dB. Specifically, for the CWRU dataset, the standard deviation of the raw signal is 0.36, the standard deviation of the added noise is 0.11, and the noise variance is 0.10 times the signal variance. For the Southeast University gearbox dataset, the standard deviation of the raw signal is 0.05, the standard deviation of the added noise is 0.02, and the noise variance is 0.10 times the signal variance. Noise addition was performed prior to data input, and all compared methods were trained and tested under the same noise conditions to ensure fairness of the experimental results.

The key architectural parameters of the proposed MSLAN model are presented in Table 2.

Table 2. Key architectural parameters of the MSLAN model.

Parameter	Value
Number of stacked SMSF modules	2
Number of stacked MBPA modules	2
Parallel kernel sizes	[8, 16, 32]
Channel reduction factor r in MBPA	16
Bottleneck channels	8
Base number of filters n_filters	16

5.3. Comparison Experiments

To verify the advanced performance of MSLAN in plunger pump fault diagnosis, a comparative experiment was designed to comprehensively compare it with other deep learning models. The selected models encompass commonly used time series classification models (FCN [13], ResNet [13], InceptionTime [14], TimesNet [15]), models frequently employed in fault diagnosis (WDD-CNN [16], CNN-BiLSTM [17], WDCNN [18]), and lightweight models (MobileNet [19], ShuffleNet [20], GhostNet [21]). Through comparison with the above three categories, the comprehensive performance of MSLAN in terms of diagnostic accuracy, model complexity, and computational efficiency was thoroughly evaluated. To ensure fairness, the learning rate was adjusted to the optimal value specified in the original papers or kept consistent with MSLAN training. The batch size was uniformly set to 32, the number of epochs to 100, and early stopping was adopted. The loss function was uniformly set as cross-entropy loss, and the Adam optimizer was used. Model performance was evaluated using F1-score, FLOPs, and Params, with experimental results presented in Tables 3–5.

Table 3. Comparative experimental results on the self-built dataset.

Model	F1-Score/%	FLOPs/M	Params/M
MSLAN	88.95	8.791	0.006
FCN	87.39	552.338	0.270
ResNet	83.51	353.668	3.848
InceptionTime	88.90	480.988	0.234
TimesNet	87.22	459.145	0.075
WDD-CNN	83.62	22.272	0.410
CNN-BiLSTM	88.75	72.218	0.550
WDCNN	81.18	2.866	0.091
MobileNet	72.94	11.323	0.070
ShuffleNet	83.25	46.584	0.486
GhostNet	84.40	4.898	0.009

Table 4. Comparative experimental results on the CWRU dataset.

Model	F1-Score/%	FLOPs/M	Params/M
MSLAN	98.89	4.351	0.006
FCN	98.62	271.583	0.266
ResNet	97.79	175.691	3.849
InceptionTime	99.87	168.411	0.164
TimesNet	98.88	229.081	0.075
WDD-CNN	98.88	8.520	0.406
CNN-BiLSTM	90.40	34.901	0.549
WDCNN	96.63	0.778	0.061
MobileNet	87.48	5.602	0.072
ShuffleNet	92.67	23.112	0.489
GhostNet	97.20	2.326	0.009

The comparative experimental results demonstrate that MSLAN achieves excellent and stable diagnostic performance across all three datasets, with F1-scores of 88.95%, 98.89%, and 99.90% on the self-constructed dataset, CWRU dataset, and gearbox dataset, respectively. While maintaining high accuracy, both the computational cost (FLOPs) and the number of parameters (Params) of MSLAN are significantly lower than those of the other compared models. Compared with time-series classification models such as FCN, ResNet, InceptionTime, and TimesNet, MSLAN reduces computational overhead

by one to two orders of magnitude while achieving comparable or even superior accuracy. In comparison with commonly used fault diagnosis models such as WDD-CNN, CNN-BiLSTM, and WDCNN, MSLAN exhibits a notable improvement in F1-Score on the self-constructed dataset while maintaining a more lightweight architecture. Furthermore, relative to lightweight models such as MobileNet, ShuffleNet, and GhostNet, MSLAN achieves substantially higher diagnostic accuracy across all datasets while retaining a clear advantage in model complexity. Overall, MSLAN achieves an optimal balance among diagnostic accuracy, model lightweightness, and computational efficiency, fully validating its effectiveness and superiority in the task of plunger pump fault diagnosis.

Table 5. Comparative experimental results on the Southeast University dataset.

Model	F1-Score/%	FLOPs/M	Params/M
MSLAN	99.90	4.416	0.006
FCN	99.06	278.004	0.271
ResNet	98.67	177.294	3.850
InceptionTime	99.60	240.995	0.235
TimesNet	99.60	229.769	0.075
WDD-CNN	99.60	12.454	0.413
CNN-BiLSTM	98.40	36.620	0.551
WDCNN	99.60	1.695	0.067
MobileNet	95.72	5.687	0.070
ShuffleNet	98.80	23.368	0.486
GhostNet	98.40	2.498	0.009

It is worth noting that the F1-scores of all comparative methods on the self-constructed dataset are significantly lower than those on the two public datasets, indicating that the self-constructed dataset itself presents a higher level of diagnostic difficulty. This is primarily attributed to the stronger environmental noise, more complex operational fluctuations, and inter-class overlap among fault features inherent in the field-collected plunger pump data, which faithfully reflects the intrinsic challenges of real-world industrial fault diagnosis.

5.4. Ablation Experiments

To achieve higher diagnostic accuracy in plunger pump fault diagnosis tasks, MSLAN introduces cross-channel convolution in the SMSF module to effectively integrate multi-source information, and employs depthwise separable convolution along with shared pointwise convolution strategies to reduce computational redundancy while efficiently extracting multi-dimensional fault features. Additionally, a multi-branch parallel attention mechanism is introduced in the MBPA module to enhance the perception capability of critical fault features. To validate the effectiveness of the above components in plunger pump fault diagnosis tasks, multiple ablation experiments are conducted on the self-constructed dataset, where one or a combination of modules is selected for each experiment. The final ablation experiment results are shown in Table 6. It is worth noting that, since multi-scale convolution serves as the core structure of the network, it is not subjected to ablation analysis in these experiments.

From the experimental results, it can be observed that the complete model, integrating cross-channel convolution, depthwise separable convolution with shared pointwise convolution strategies, and the MBPA attention mechanism, achieves the optimal diagnostic performance, with an F1-Score of 88.95%, which is 2.08 percentage points higher than that of the baseline model. Meanwhile, the computational complexity is significantly reduced to 8.79 MFLOPs, representing a reduction of over 90%, and the parameter count is reduced by an order of magnitude. Regarding the contribution of each component, the introduc-

tion of cross-channel convolution alone reduces computational complexity but leads to a decrease in accuracy due to information loss. The depthwise separable convolution with shared pointwise convolution strategy significantly improves computational efficiency while maintaining accuracy. In the comparison of attention mechanisms, MBPA achieves a 1.34 percentage point improvement in accuracy compared to the standard SE module under similar computational complexity, validating the effectiveness of the multi-branch parallel structure in modeling channel dependencies. In summary, the synergistic effect of SMSF and MBPA enables the model to achieve optimal diagnostic accuracy while maintaining extremely high computational efficiency, fully demonstrating the effectiveness and practicality of the proposed method in plunger pump fault diagnosis tasks.

Table 6. Ablation experiment results.

No.	CC *	DSConv + SPW *	SE	MBPA	F1-Score (%)	FLOPs (M)	Params (M)
1					86.87	103.216	0.050
2	√				84.65	34.300	0.017
3		√			87.20	12.398	0.006
4			√		87.19	103.413	0.051
5				√	88.53	103.415	0.053
6	√	√			87.14	8.592	0.004
7	√	√		√	88.95	8.791	0.006

* "CC" denotes dynamic channel compression; "DSConv + SPW" denotes depthwise convolution with shared pointwise weights; "√" indicates that the module is used in the corresponding experiment.

5.5. Visualization Analysis

To further demonstrate the effectiveness of the MBPA module, fault samples are randomly selected from the self-built plunger pump dataset, the CWRU dataset, and the Southeast University dataset in this section to visualize the feature maps of the two MBPA modules in the MSLAN network. The feature maps before and after the operation of the two MBPA modules are shown in Figures 8–10, where the left side of each figure shows the original feature map and the recalibrated feature map of the first MBPA layer, and the right side shows those of the second MBPA layer.

The intra-module attention mechanism assigns channel-wise attention weights at each layer, revealing that channels with higher weights have greater influence on downstream computations. This channel prioritization intensifies as the network advances toward the output layer. By this stage, the model has nearly completed its processing of the original data, feature variations show little further change, and the most informative channels have been identified. From the self-built dataset results presented in Figure 8, the feature map changes observed in the first MBPA layer remain modest, while only a subset of channels retain large values within the second MBPA layer. In Figure 9 on the CWRU dataset, the attention weight changes in both MBPA modules are very distinct, with different channels being activated and suppressed after each layer. Figure 10, obtained from the Southeast University dataset, illustrates that with increasing network depth, the feature maps evolve from a uniformly distributed pattern in the first layer toward alternating light-dark stripe patterns in the second layer—a shift that demonstrates how the attention mechanism enables the network to discern channel significance.

To intuitively evaluate the classification performance and feature learning capability of the MSLAN model, ResNet, WDCNN, and MobileNet were selected for visual comparison with MSLAN. The t-SNE [22] dimensionality reduction visualization and confusion matrix analysis were conducted on three datasets, and the results are presented in Figures 11–18.

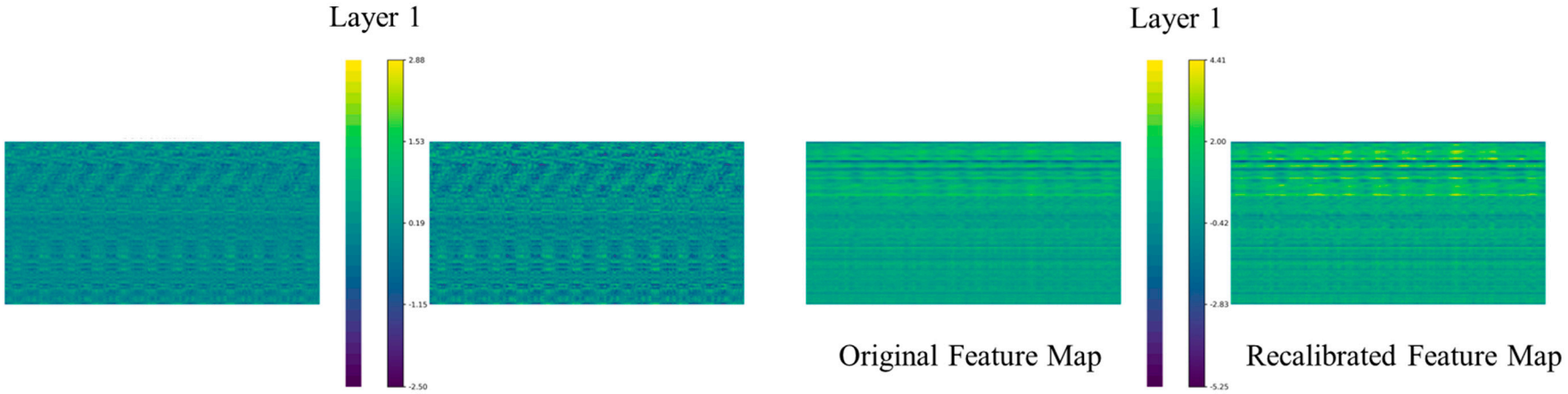

Figure 8. Visualization of intra-module attention on self-built dataset.

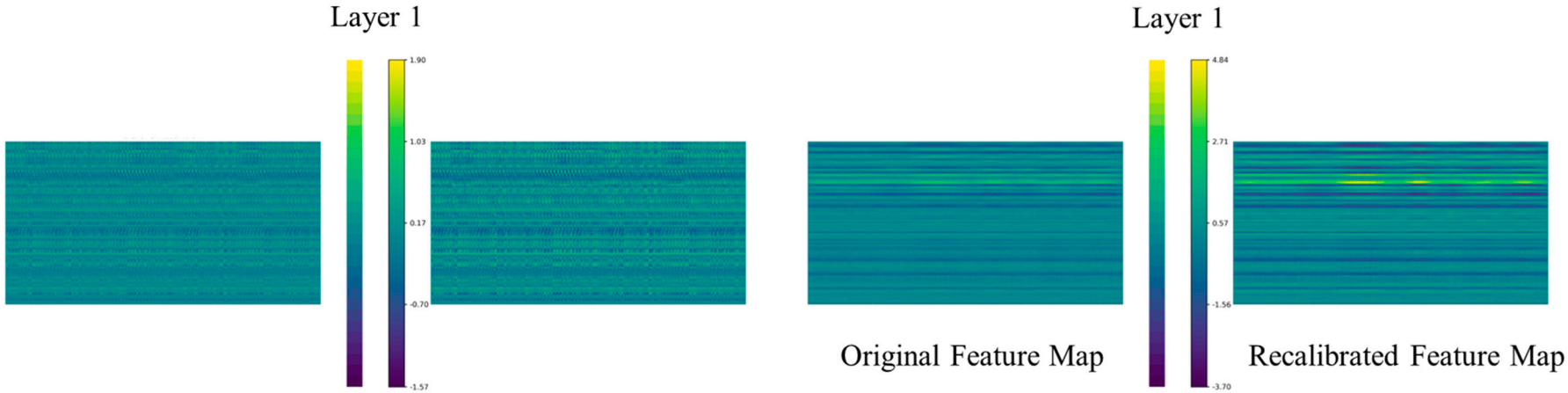

Figure 9. Visualization of intra-module attention on the CWRU dataset.

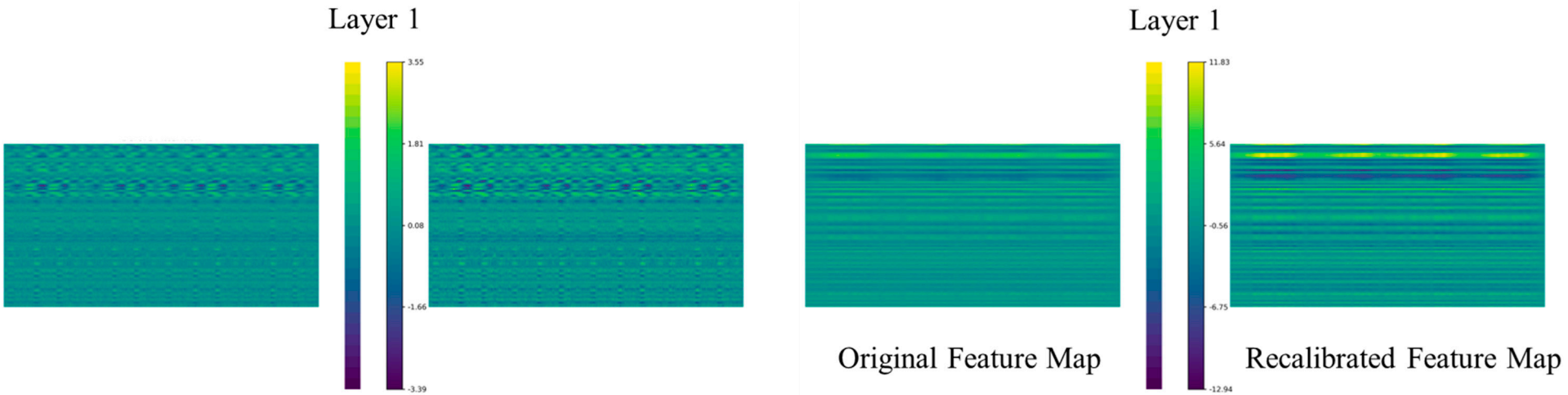

Figure 10. Visualization of intra-module attention on the Southeast University dataset.

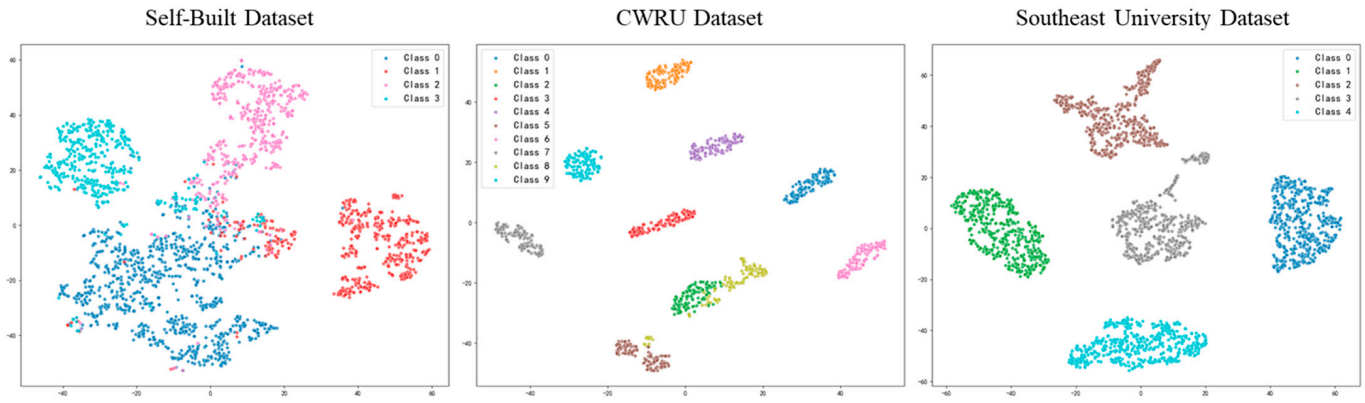

Figure 11. T-SNE visualization of MSLAN.

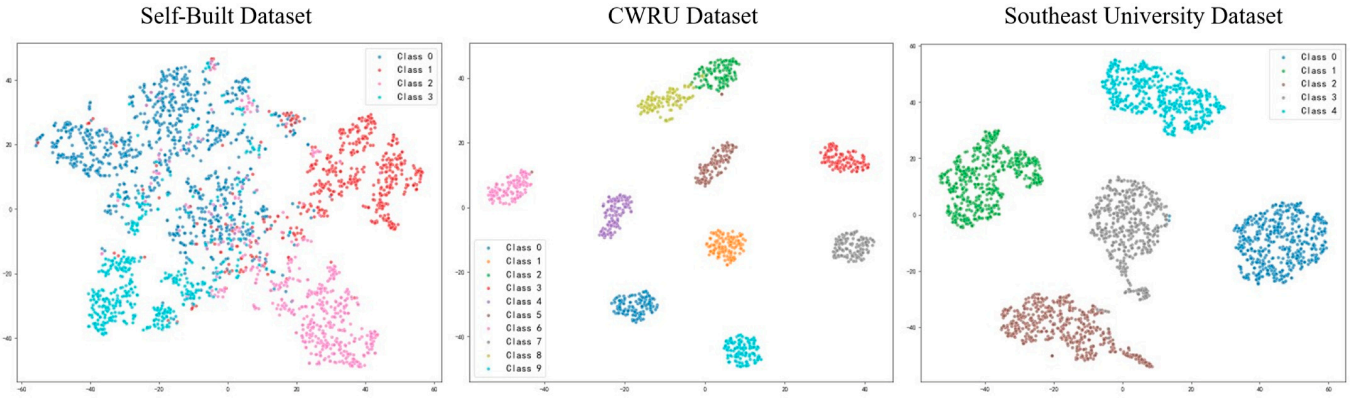

Figure 12. T-SNE visualization of ResNet.

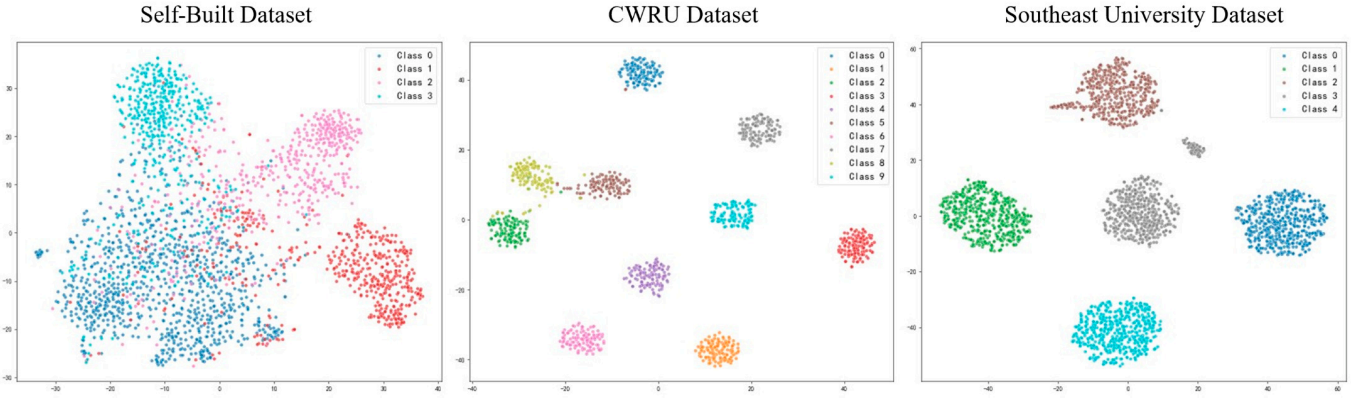

Figure 13. T-SNE visualization of WDCNN.

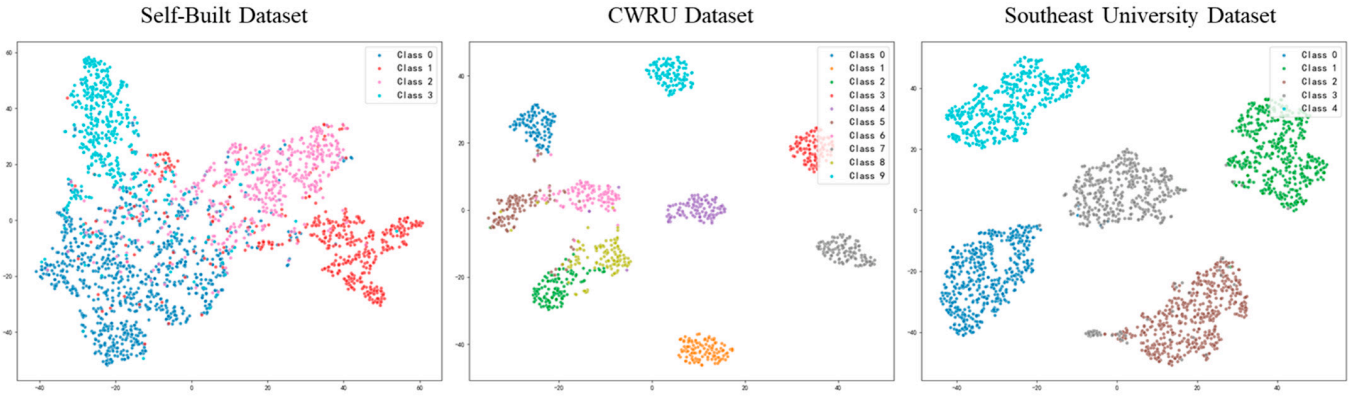

Figure 14. T-SNE visualization of MobileNet.

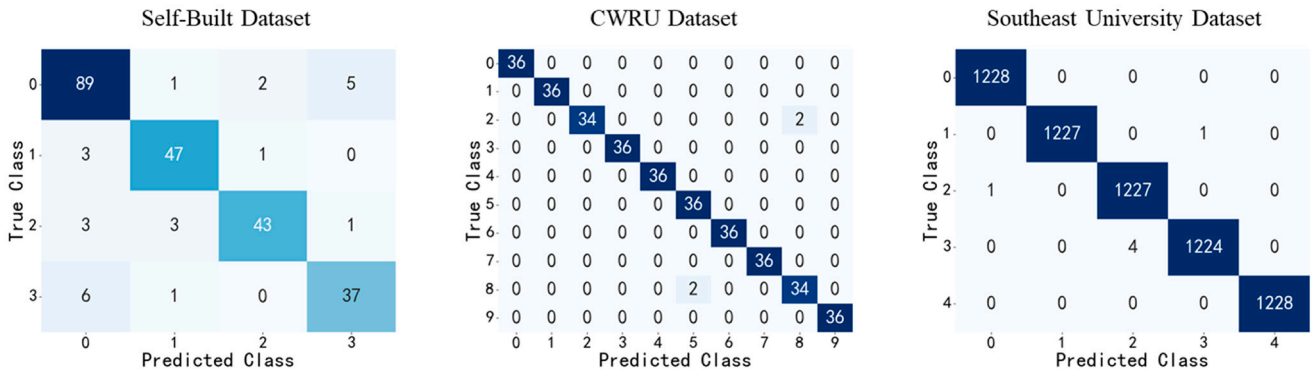

Figure 15. Confusion matrix of MSLAN.

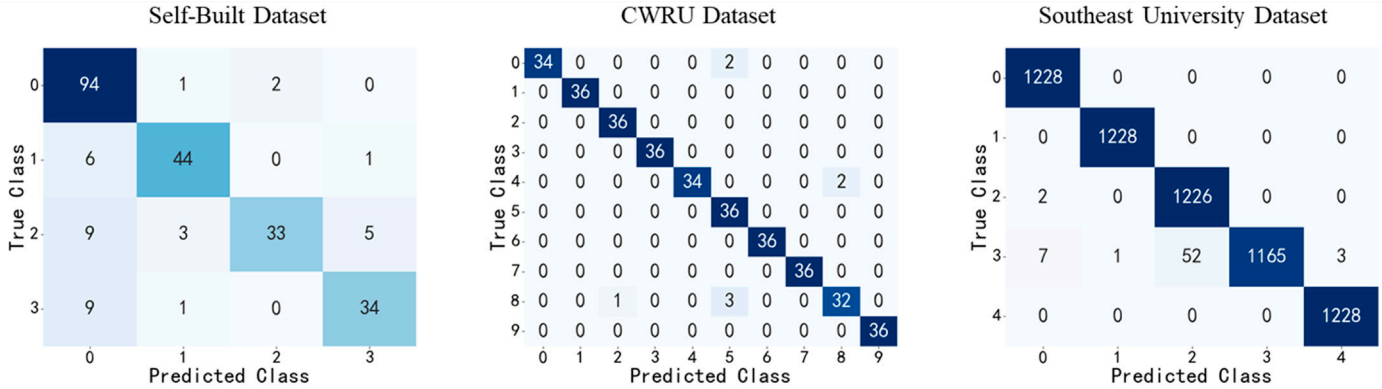

Figure 16. Confusion matrix of ResNet.

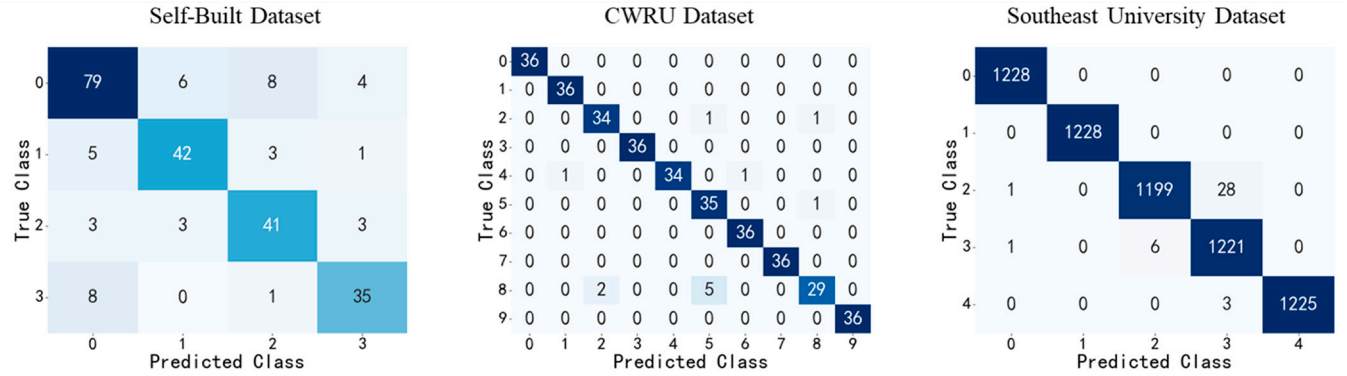

Figure 17. Confusion matrix of WDCNN.

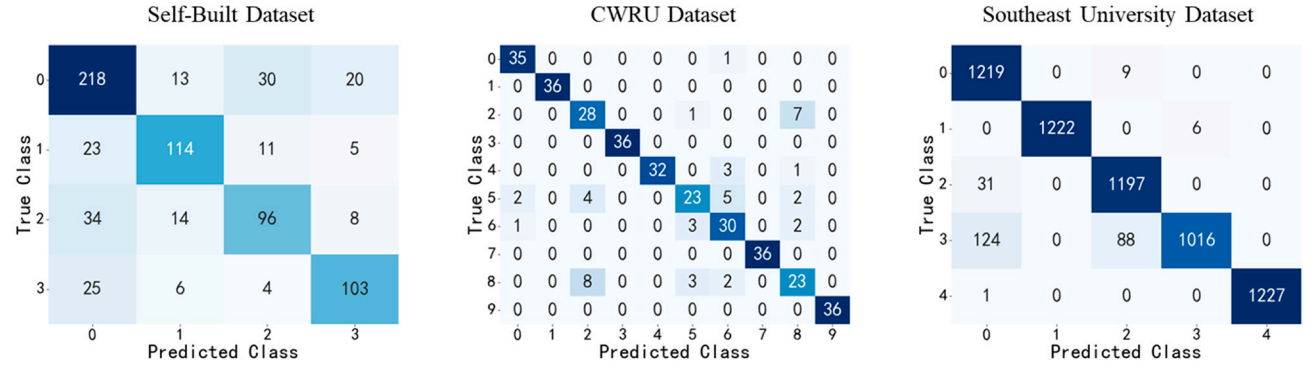

Figure 18. Confusion matrix of MobileNet.

The t-SNE visualization results show that the features extracted by MSLAN exhibit the clearest clustering structure in two-dimensional space. Different fault categories are

well-separated with distinct boundaries, and samples within the same class are tightly clustered, indicating that MSLAN effectively learns highly discriminative fault features. In contrast, the feature distributions of ResNet and WDCNN show some degree of inter-class overlap, with blurred boundaries between certain fault categories. MobileNet exhibits the poorest clustering performance, with multiple categories intermixed and difficult to distinguish. Compared to the public CWRU and Southeast University datasets, the self-constructed dataset exhibits more noticeable inter-class overlap in the t-SNE visualization, with samples of different fault types not being clearly separated and the clustering structure being relatively ambiguous. The self-constructed dataset originates from real-world oilfield operating conditions, where the signals inevitably contain complex environmental noise, operational fluctuations, and sensor measurement errors, resulting in more intricate fault feature distributions and relatively ambiguous inter-class boundaries. In contrast, the CWRU and Southeast University datasets are typically collected under controlled laboratory conditions, where fault patterns are more distinct and noise levels are lower, leading to better feature clustering performance. Although the clustering performance of the self-constructed dataset is inferior to that of the public datasets, MSLAN still achieves a better clustering structure compared to other methods, further demonstrating its resilience to complex industrial data.

From the confusion matrix analysis, MSLAN achieves the highest classification accuracy across all datasets, with significantly deeper diagonal colors compared to the other models and very few misclassified samples. ResNet and WDCNN exhibit some misclassifications between easily confusable fault categories, while MobileNet shows the most severe misclassification, with notably low recognition accuracy for several categories.

Overall, the visualization results confirm that MSLAN consistently outperforms ResNet, WDCNN, and MobileNet across both feature extraction and classification tasks, attesting to the proposed method's robustness and efficacy in plunger pump fault diagnosis.

6. Discussion

This paper proposes a lightweight fault diagnosis method based on multi-scale convolution and attention mechanisms, termed MSLAN, to address the challenges in plunger pump fault diagnosis, including the difficulty in capturing multi-scale fault features, interference from redundant information in high-dimensional feature spaces, and high model computational complexity. By leveraging the SMSF module, the method efficiently captures fault features across different scales with minimal computational expense, successfully addressing the receptive field limitation of a single convolution kernel. Additionally, the MBPA module is introduced, which significantly enhances the model's ability to perceive critical fault features and effectively suppresses redundant information through refined modeling of complex inter-channel dependencies. Experiments conducted on three datasets—a self-built plunger pump dataset, the CWRU bearing dataset, and the Southeast University gearbox dataset—confirm that the proposed method delivers outstanding performance for plunger pump fault diagnosis while demonstrating strong generalization across other rotating machinery fault diagnosis tasks. More importantly, the incorporation of lightweight techniques—namely depthwise separable convolution and shared pointwise convolution—enables the model to considerably reduce parameter count and computational burden while sustaining high diagnostic accuracy, meeting the requirements for deployment on edge devices and real-time diagnosis in industrial contexts.

Although the proposed method achieves remarkable results in terms of diagnostic accuracy and computational efficiency, certain limitations remain.

First, in terms of multi-source feature fusion, while the proposed method can process input data from multiple channels including vibration, pressure, temperature, and current

sensors, it primarily relies on channel concatenation for feature fusion, failing to fully exploit the complementarity and correlation among different sensor signals at the physical mechanism level. Specifically, pressure signals reflect the pump's working load and fluid pulsation characteristics, vibration signals reflect impact and wear conditions of mechanical components, temperature signals reflect thermodynamic changes in friction pairs, and current signals reflect the motor's driving state and load fluctuations. These signals often exhibit different temporal scale response characteristics and coupling relationships when faults occur. Simple channel stacking struggles to effectively model the deep synergistic mechanisms among heterogeneous sensors, potentially limiting the model's comprehensive perception of complex fault patterns.

Second, validation of the method was performed on a self-built dataset alongside two public datasets, all gathered under relatively controlled experimental environments or specific operating conditions. In actual oilfield production, equipment operating conditions are complex and variable, with dynamic changes in load, rotational speed, environmental noise, and other factors imposing higher demands on the model's generalization capability. Although experiments simulated some interference by adding Gaussian white noise, this still deviates from the complex and non-stationary noise present in real industrial environments.

Third, the lower diagnostic performance of the proposed method on the self-constructed plunger pump dataset, compared with the two public datasets, is not a model flaw but an objective reflection of the inherent challenges posed by real-world industrial data. The self-constructed dataset, collected from actual oilfield operating conditions, contains stronger environmental noise, more complex operational fluctuations, and inter-class overlap of fault features, truly reflecting the greater difficulty of industrial fault diagnosis tasks. Future improvements for this challenging scenario may include the introduction of robust denoising modules, the design of imbalanced-data loss functions, and the exploration of multi-sensor fusion strategies to further improve diagnostic performance in complex industrial environments.

To address the above limitations, future work will focus on the following directions.

First, a physics-informed multi-source feature fusion mechanism will be constructed. We plan to deeply analyze the physical response patterns of sensor signals such as vibration, pressure, temperature, and current during fault occurrence, and design feature alignment and fusion strategies based on physical priors. For instance, temporal alignment mechanisms can handle differences in time scales among different signals, or attention mechanisms can be introduced to adaptively learn the contribution weights of each sensor under different fault modes. Leveraging synergistic information from heterogeneous sensors through deep mining, the model's diagnostic capability for compound faults and incipient weak faults can be further augmented.

Second, the cross-domain generalization capability of the model will be enhanced. To cope with the variability of operating conditions, domain adaptation or meta-learning strategies will be adopted, empowering the model to extract transferable fault features from known conditions and achieve fast adaptation to unseen conditions using only a few samples. Additionally, we plan to construct industrial datasets covering a wider range of operating conditions, including more fault types and more complex environmental interference, to provide more rigorous test benchmarks for evaluating model robustness.

Third, efforts will be directed toward enhancing model interpretability. Deep learning models often operate as black boxes, rendering their decision-making processes challenging to interpret. To strengthen confidence in diagnostic results, future investigations will integrate visualization approaches, including class activation mapping, to conduct a granular analysis of the model's attention regions and salient features. Uncovering the foundations

of diagnostic decisions will provide field engineers with clearer fault insights, paving the way for human–machine collaborative intelligent diagnosis frameworks.

In summary, the MSLAN method proposed in this paper provides an effective solution for intelligent fault diagnosis of plunger pumps that achieves both high accuracy and high efficiency. Through continuous research in areas such as multi-source feature fusion, cross-condition generalization, and interpretability, this technology is expected to evolve toward greater intelligence, reliability, and practicality, providing strong support for intelligent operation and maintenance of oilfield equipment.

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

Funding: This work was supported by the major project of the National Natural Science Foundation of China (51991365) and the Natural Science Foundation of Shandong Province of China (ZR2021MF082).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Han, W.T. Failure Analysis and Reconstruction Treatment of Oil Injection Pump. *Petrochem. Ind. Technol.* **2024**, *31*, 102–104.
2. Gao, J.W.; Zhang, Z.K. Research on Fault Feature Extraction Method of Plunger Pump Based on Empirical Wavelet Transform. *Coal Mine Mach.* **2023**, *44*, 177–179. [CrossRef]
3. Zhao, L.H.; Cheng, H.; Li, W.Y.; Guan, C. Research on Fault Diagnosis Method of Plunger Pump Based on LMD and Support Vector Machine. *Mach. Des. Manuf.* **2022**, *373*, 238–241. [CrossRef]
4. Li, Y.M.; Li, M. BP Network Fault Diagnosis of Piston Pump Pulsation Pressure Signal Based on Wavelet Packet Transform. *Hydraul. Pneum. Seals* **2023**, *43*, 123–126.
5. Tang, H.B.; Gong, Y.C.; Dong, J.Y.; Chen, S. Fault Diagnosis of Axial Piston Pump Based on CNN-SE-LSTM and Multi-sensor Data. *Mach. Tool Hydraul.* **2024**, *52*, 224–232.
6. Xu, C.L.; Lan, Y.; Ba, G.L.; Wu, B. Application of Convolutional Neural Networks in Fault Diagnosis of Axial Piston Pump. *Mach. Des. Manuf.* **2024**, 1–7. [CrossRef]
7. Xu, C.L.; Huang, J.H.; Lan, Y.; Wu, B.; Niu, C.; Ma, X.; Li, B. Fault Diagnosis of Axial Piston Pump Based on D-1DCNN. *J. Mech. Electr. Eng.* **2021**, *38*, 1494–1500.
8. Wang, M.B.; Gao, Q.; Feng, Y.L.; Zhao, N.; Chen, J.; Lv, C.X. Research on Bearing Fault Diagnosis via Multi-Feature Fusion Transformer-LSTM. *Mech. Sci. Technol. Aerosp. Eng.* **2025**, 1–11. [CrossRef]
9. Fang, H.; Deng, J.; Bai, Y.; Feng, B.; Li, S.; Shao, S.; Chen, D. CLFormer: A Lightweight Transformer Based on Convolutional Embedding and Linear Self-Attention with Strong Robustness for Bearing Fault Diagnosis under Limited Sample Conditions. *IEEE Trans. Instrum. Meas.* **2021**, *71*, 3504608. [CrossRef]
10. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018*; IEEE: New York, NY, USA, 2018; pp. 7132–7141.
11. Smith, W.A.; Randall, R.B. Rolling Element Bearing Diagnostics Using the Case Western Reserve University Data: A Benchmark Study. *Mech. Syst. Signal Process.* **2015**, *64*, 100–131. [CrossRef]
12. Chen, C. Methodologies for Fault Diagnosis of Rotary Machine Based on Transfer Learning. Ph.D. Thesis, Southeast University, Nanjing, China, 2020. [CrossRef]
13. Wang, Z.; Yan, W.; Oates, T. Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline. In Proceedings of the 2017 International Joint Conference on Neural Networks, Anchorage, AK, USA, 14–19 May 2017; pp. 1578–1585.
14. Ismail Fawaz, H.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D.F.; Weber, J.; Webb, G.I.; Idoumghar, L.; Muller, P.A.; Petitjean, F. InceptionTime: Finding AlexNet for Time Series Classification. *Data Min. Knowl. Discov.* **2020**, *34*, 1936–1962. [CrossRef]

15. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. *arXiv* **2022**, arXiv:2210.02186.
16. Sonmez, E.; Kacar, S.; Uzun, S. A New Deep Learning Model Combining CNN for Engine Fault Diagnosis. *J. Braz. Soc. Mech. Sci. Eng.* **2023**, *45*, 644. [CrossRef]
17. Xu, L.; Zhao, G.; Zhao, S.; Wu, Y.; Chen, X. Fault Diagnosis Method for Tractor Transmission System Based on Improved Convolutional Neural Network–Bidirectional Long Short-Term Memory. *Machines* **2024**, *12*, 492. [CrossRef]
18. Zhang, W.; Peng, G.; Li, C.; Chen, Y.; Zhang, Z. A New Deep Learning Model for Fault Diagnosis with Good Anti-Noise and Domain Adaptation Ability on Raw Vibration Signals. *Sensors* **2017**, *17*, 425. [CrossRef] [PubMed]
19. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [CrossRef]
20. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018*; IEEE: New York, NY, USA, 2018; pp. 6848–6856.
21. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 1580–1589.
22. van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

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.