Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AES-FusionNet: A ResNet-GRU Multi-Modal Fusion Framework for AE Bearing Fault Diagnosis

Yi Chen
Guangdong University of Petrochemical Technology
Wenmin Wen
Guangdong University of Petrochemical Technology
Wanying Wu
Guangdong University of Petrochemical Technology
Junqi Liu
Guangdong University of Petrochemical Technology
Yi Pan
Guangdong University of Petrochemical Technology
Yanzuo Chang
158762072280163.com

Guangdong University of Petrochemical Technology

Article

Keywords: Rolling bearing, Acoustic emission technique, Fault diagnosis, Deep learning, Multimodal fusion

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11676883/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AES-FusionNet: A ResNet-GRU Multi-Modal Fusion Framework for AE Bearing Fault Diagnosis

Yi Chen[1], Wenmin Wen[1], Wanying Wu[1], Junqi Liu[1], Yi Pan[2], Yanzuo Chang[1, *]

[1] Guangdong University of Petrochemical Technology, School of Energy and Power Engineering, Maoming, China
[2] Guangdong University of Petrochemical Technology, School of Computer Science, Maoming, China
*e-mail: 15876207228@163.com

ABSTRACT

Rolling bearings are among the core components of rotating machinery, and acoustic emission (AE) technology is notably sensitive to incipient faults, thus attracting increasing attention in the field of bearing fault diagnosis in recent years. However, AE signals are characterized by high frequency, non-stationarity, and susceptibility to noise interference, rendering single-modality diagnostic methods inadequate for fully exploiting the multidimensional information embedded in the signals, which leads to insufficient fault feature extraction and limited diagnostic accuracy. To address these issues, this paper proposes a deep learning-based multimodal fusion network that takes one-dimensional raw AE signals and two-dimensional Markov Transition Field (MTF) images as dual-channel inputs, wherein the image branch employs an improved ResNet34 network to extract spatial textural features, while the signal branch adopts a two-layer Gated Recurrent Unit (GRU) to capture temporal features. An experimental platform for AE-based bearing diagnosis was established, and data covering six bearing fault conditions were collected to validate the proposed model. Experimental results show that the model achieves an overall accuracy of 95.74% and an average correct classification rate of 98.05% across the six fault types. Comparative experiments demonstrate that the proposed model outperforms nine mainstream methods, including 1D-CNN and 2D-CNN, while ablation studies validate the effectiveness of the multimodal fusion strategy and the SK-Net attention mechanism. These results confirm the strong applicability of the proposed model to AE fault signals and provide a viable technical solution for this field.

Keywords: Rolling bearing; Acoustic emission technique; Fault diagnosis; Deep learning; Multimodal fusion

1. INTRODUCTION

With the continuous advancement of the Industry 4.0 initiative, the global manufacturing sector is undergoing a profound transformation toward digitalization [1], and rotating machinery, as a core asset of modern industrial systems, is subject to increasingly stringent requirements regarding operational

safety and reliability [2][3]. As a critical component of rotating machinery, rolling bearings are frequently exposed to complex loads and harsh operating conditions, rendering them highly susceptible to wear, fatigue, and other forms of damage [4]. Reliance solely on conventional periodic maintenance or post-hoc repair strategies is inadequate to meet the demands of modern production for long-term, stable equipment operation. Consequently, real-time monitoring and fault diagnosis of rolling bearings not only contribute to the effective reduction of unforeseen accidents but also significantly lower operation and maintenance costs. Therefore, fault diagnosis of rolling bearings constitutes a pivotal element in ensuring the stable operation of mechanical equipment and, in turn, serves as a key driver for promoting high-quality development in the manufacturing sector.

Traditional vibration analysis methods suffer from low sensitivity to incipient faults, susceptibility to noise interference, and strong dependence on operating conditions, which render them inadequate for meeting the increasingly stringent maintenance requirements of modern industrial systems. Consequently, acoustic emission (AE) technology, as a dynamic non-destructive testing method, has attracted growing attention in the field of bearing fault diagnosis [5][6]. Compared with vibration signals, AE signals offer distinct advantages, including higher sensitivity, a broader frequency bandwidth, and greater robustness against environmental noise, thereby enabling superior detection capability for early-stage faults and facilitating real-time tracking of fault evolution processes [7]. Furthermore, AE technology is applicable for monitoring under harsh conditions such as high temperature, flammable, and explosive environments [8]. Owing to these merits, acoustic emission technology provides a promising alternative technical pathway for bearing fault diagnosis.

The advancement of deep learning techniques has opened up new avenues for acoustic emission-based bearing fault diagnosis, as these methods are capable of learning hierarchical feature representations directly from raw data [9], thereby eliminating the need for labor-intensive manual signal feature analysis [10]. To date, a number of mainstream deep learning architectures, including convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and Transformers, have achieved notable success in the field of mechanical fault diagnosis, as exemplified by the works of Prosvirin [11], Jiang [12], Li [13], and Mao [14], among others, which collectively provide powerful methodological support for acoustic emission diagnostic techniques [15].

Although deep learning-based acoustic emission bearing fault diagnosis has achieved considerable progress, several challenges remain to be addressed. For instance, most existing approaches rely on single-modality data inputs [16][17], such as one-dimensional raw AE waveforms or two-dimensional time-frequency representations; however, AE signals inherently contain rich information, and a single modality is often insufficient to comprehensively reveal fault characteristics [18]. In contrast to unimodal approaches, multimodal analysis enables the construction of more comprehensive fault feature representations by integrating complementary information from multiple sources, thereby broadening the applicability of deep learning models in practical engineering scenarios [19]. To date, multimodal fusion methods have been extensively adopted across various engineering domains. For example, Liu et al. [20] combined one-dimensional time-series signals with two-dimensional time-frequency images through a dual-channel architecture comprising Swin Transformer and CNN-BiGRU-Attention to

achieve feature fusion, thereby complementarily extracting temporal and time-frequency domain features and consequently improving both accuracy and robustness. Li et al. [21] employed raw time-series data and two types of time-frequency representations as three distinct modalities, utilizing a large language model (LLM) for cross-modal feature reprogramming and incorporating time- or frequency-domain statistical indicators as prompts, so as to overcome the limitation of unimodal deep learning methods in simultaneously exploiting both temporal and spectral information. Zhang et al. [22] implemented cross-modal bidirectional semantic alignment between time-domain and time-frequency-domain features via a bidirectional cross-attention mechanism, and introduced a dynamic gating mechanism to adaptively adjust the fusion weights of each modality, thereby effectively decoupling multiple feature components in compound faults and enabling precise identification thereof. Hu et al. [23] simultaneously utilized one-dimensional time-series data and two-dimensional time-frequency images, embedded Laplacian wavelet convolution to incorporate physical prior knowledge, and employed a Mahalanobis-prototype joint metric to enhance diagnostic reliability and interpretability under few-sample conditions.

Although multimodal fault diagnosis has achieved promising results in engineering applications, a number of issues remain to be resolved. First, most existing methods directly adopt raw waveforms or time-frequency representations as network inputs without performing effective feature enhancement or representation optimization on the input data, which impedes the model's ability to fully extract discriminative features from the original signals. Second, in terms of image feature extraction, conventional convolutional neural networks employ convolution kernels with fixed receptive fields that are incapable of adaptively adjusting according to the input content, thereby restricting the model's capacity to capture fault characteristics across varying scales.

To address the aforementioned issues, this paper proposes a deep learning-based network termed AES-FusionNet(ResNet-GRU Multi-Modal Fusion Framework), which constructs a dual-channel parallel deep learning framework to fully leverage both the one-dimensional temporal information and the two-dimensional image data embedded in acoustic emission signals. In the signal branch, a Gated Recurrent Unit (GRU) is employed to extract one-dimensional temporal features, thereby capturing the dynamic evolution patterns throughout the fault development process. In the image branch, the one-dimensional signals are first transformed into Markov Transition Field (MTF) images to preserve the state transition information across different time intervals, following which a ResNet architecture integrated with a Selective Kernel attention (SK-attention) mechanism is utilized to extract deep spatial features from the images. The features derived from the two branches are subsequently integrated through a fusion layer, enabling in-depth mining and complementary utilization of multimodal information. This design effectively overcomes the limitations associated with unimodal inputs, namely the incompleteness of information representation and the insufficient exploitation of complementary relationships across different feature dimensions, thereby significantly enhancing both the accuracy and robustness of bearing fault diagnosis. The main contributions of this paper are as follows:

(1) This paper proposes a deep learning-based multimodal diagnostic model that constructs a dual-channel feature input by fusing the one-dimensional temporal information and two-dimensional image data of acoustic emission signals, thereby effectively overcoming the

limitation of incomplete information representation inherent in single-modality approaches.

(2) An image feature extraction module based on ResNet34 embedded with SK-attention is introduced, which strengthens the capability of extracting critical frequency-domain features from MTF images and addresses the insufficiency of existing methods in utilizing frequency-domain and time-frequency-domain information.

(3) A temporal feature extraction module based on Gated Recurrent Unit (GRU) is constructed, in which a two-layer network stacking configuration is adopted to enhance the model's capability of capturing long-term temporal dependencies in one-dimensional acoustic emission signals, thereby remedying the deficiency of existing methods in adequately modeling dynamic temporal evolution patterns.

2. RELATED THEORETICAL BACKGROUND

2.1 Residual Neural Network

The residual neural network (ResNet) was proposed by He et al. in 2015 [24] to address the degradation problem encountered by deep neural networks. The key innovation of ResNet lies in the design of the residual block, which effectively alleviates the issues of vanishing and exploding gradients while substantially reducing computational complexity.

The residual block serves as the core component of the residual neural network [25]; a standard residual block consists of a residual path and a shortcut connection, and its structure is illustrated in Fig.1.

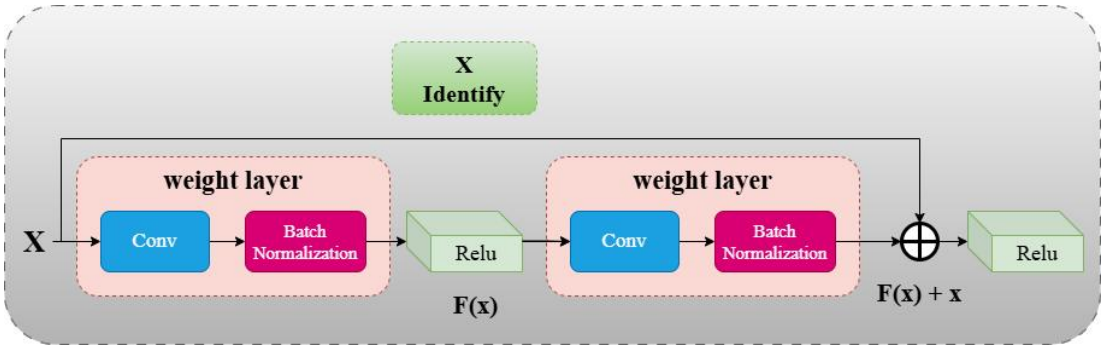

Figure. 1. Structure of the residual block in ResNet

The residual neural network is constructed by stacking multiple residual modules. Commonly used variants with different depths include ResNet18, ResNet34, ResNet50, and ResNet101, where the numbers denote the number of network layers. Deeper networks are capable of learning more complex features, albeit at the cost of increased computational load. The commonly adopted network architecture is illustrated in Fig. 2.

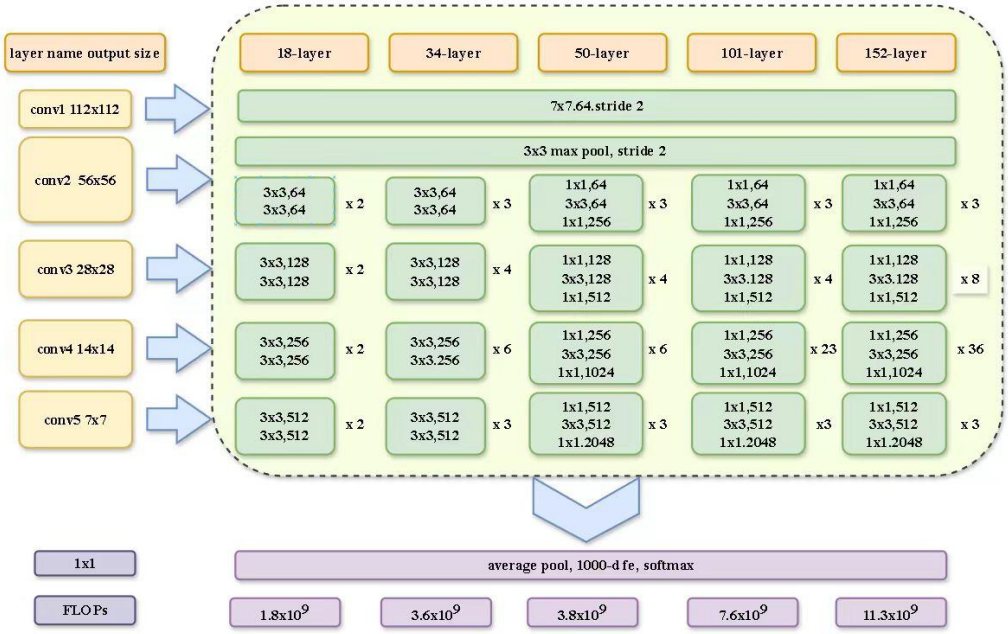

Figure. 2. Common architectures of ResNet

2.2 Gated Recurrent Unit

The Gated Recurrent Unit (GRU) was proposed by Cho et al. [26] and, like the Long Short-Term Memory (LSTM) network, is a variant of the Recurrent Neural Network (RNN). In traditional RNNs, the gradient weights accumulate over time during backpropagation, leading to exponential decay or explosion, which prevents the model from effectively capturing long-range temporal dependencies.

The GRU eliminates the long-term memory transmission branch of LSTM [27], and its structure is illustrated in Fig. 3.

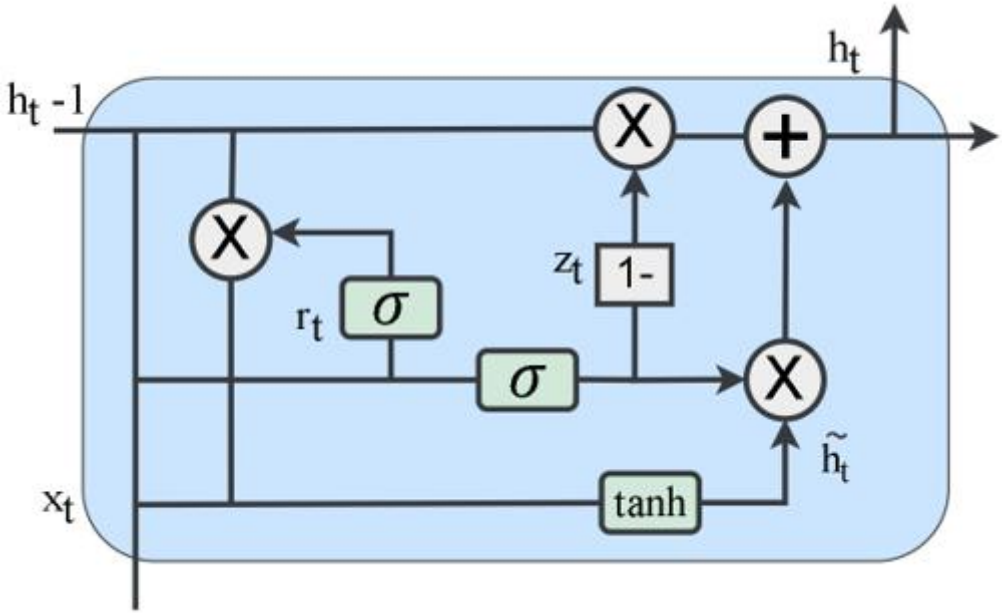

Figure. 3. Structure of the GRU

The computation of the GRU at time step t is as follows. As shown in Figure 3, let the input be X_t, the hidden state from the previous time step be H_{t-1}, and the current hidden state be H_t. First, the reset gate R_t and the update gate Z_t are computed based on X_t and H_{t-1}, with the sigmoid function compressing the values to the range between 0 and 1, thereby serving as a gating mechanism:

$$R_t = \sigma(W_{xr}X_t + W_{hr}H_{t-1} + b_r) \tag{1}$$

$$Z_t = \sigma(W_{xz}X_t + W_{hz}H_{t-1} + b_z) \tag{2}$$

The next step is to compute the candidate hidden state $\tilde{h}_t$, The reset gate R_t is multiplied element-wise with H_{t-1} to determine which historical information should be retained; after this gating process, the result is combined with the input X_t and passed through the tanh activation function to generate the candidate hidden state:

$$\tilde{h}_t = tanh(W_{xh}X_t + W_{hh}(R_t eH_{t-1}) + b_h) \tag{3}$$

Finally, the hidden state H_t is computed by performing a linear interpolation between the previous hidden state H_{t-1} and the current candidate hidden state $\tilde{h}_t$. The update gate Z_t determines the balance between the old information to be retained and the new information to be incorporated, thereby enabling the network to smoothly preserve long-term dependencies.

$$h_t = (1 - Z_t)eH_{t-1} + Z_t e\tilde{h}_t \tag{4}$$

2.3 Selective Kernel Attention Mechanism

The Selective Kernel Attention mechanism (SK-attention) was proposed by Li et al. in 2019 [28]. The workflow of SK-attention consists of three core steps, namely Split, Fuse, and Select, and its structure is illustrated in Fig. 4.

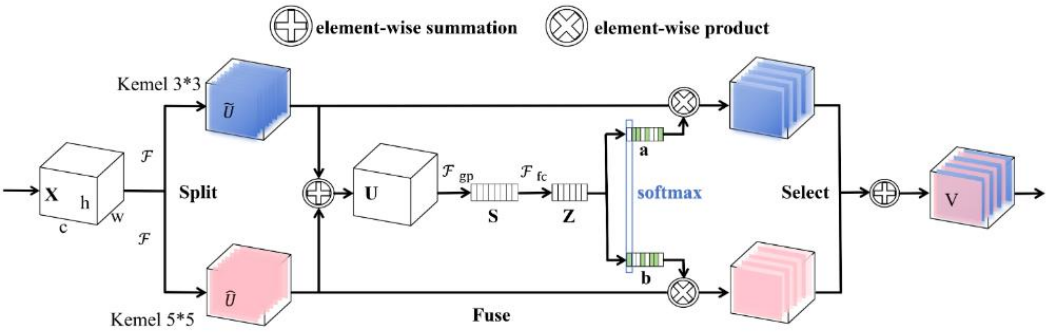

Figure. 4. Structure of SK-attention

The Split step constructs two parallel convolutional branches. Branch 1 employs a 3×3 dilated convolution to extract detailed features, while Branch 2 also adopts a 3×3 dilated convolution to capture global structural features, with the output feature map denoted as $U_2 \in R^{H \times W \times C}$.The convolution operations

in both branches are identical in configuration, ensuring that the output feature maps are strictly consistent in spatial dimensions and channel numbers.

The Fuse step constitutes the core of SK-attention, which aggregates global information and employs gating encoding to generate attention weight vectors. This step consists of two sub-steps, namely global information squeezing and gating vector excitation.

The global information squeezing operation first performs element-wise summation over the multi-branch feature maps to obtain the global feature map U_3. Subsequently, global average pooling is applied to U_3, compressing the feature map of dimensions $H \times W \times C$ into a global descriptor $S \in R_C$, thereby achieving the squeezing of global spatial information. The value of the c-th channel is computed as follows:

$$S_C = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i,j), C = 1,2,\cdots,C \tag{5}$$

In the gating vector excitation step, the descriptor s is first passed through the first fully connected layer for dimensionality reduction to obtain the hidden layer feature z, which is then upscaled through the second fully connected layer to generate the attention weight vectors $a \in R_C$ and $b \in R_C$ corresponding to each branch. These weights are subsequently normalized via the Softmax function to ensure that the sum of the branch weights for each channel equals unity, where a_c denotes the attention weight for the c-th channel of the first branch and b_c denotes that of the second branch. The complete computation is formulated as follows:

$$\begin{cases} a = Softmax(W_a z) \\ b = Softmax(W_b z) \end{cases} \tag{6}$$

The final Select step corresponds to the output stage, in which the dynamic attention weights generated by the Fuse step are applied to perform channel-wise weighted fusion across the different branches, yielding the final output feature map denoted as $V \in R^{H \times W \times C}$. The output for the c-th channel is computed as follows:

$$V_c = a_c \square U_{1,c} + b_c \square U_{2,c} \tag{7}$$

U1,c and $U_{2,c}$ denote the c-th channel outputs of the two branches, respectively. Through this dynamic weighting process, the network is able to assign adaptive weights to features across different channels and scales, thereby achieving dynamic adjustment of the receptive field in response to the input content.

2.4 Markov Transition Field

The Markov Transition Field (MTF) is fundamentally rooted in the modeling of dynamic transition characteristics of time series. This method discretizes the amplitude information into a finite number of states [29], computes the transition probabilities between different states, and ultimately maps the resulting probability information into pixel values, thereby achieving the transformation from time-series signals to spatial images.

The computational procedure for the two-dimensional data representation of the Markov Transition Field is as follows. For a one-dimensional acoustic emission time series $X = \{X_1, X_2, \dots, X_N\}$ of length N, the first step involves data discretization and state partitioning, in which the amplitude range is divided into Q discrete state intervals. This quantization process converts the continuous-valued sequence into a discrete form, yielding a state sequence denoted as $S = \{S_1, S_2, \dots, S_N\}$, where each $S_i \in \{1, 2, \dots, Q\}$ represents the state at the corresponding time instant. In fault diagnosis applications, the number of states Q is determined according to the specific circumstances, with a common value ranging between 8 and 64. Specifically, a $Q \times Q$ transition count matrix C is constructed, in which each element C_{ij} denotes the number of transitions from state i to state j. This matrix is then normalized row-wise to obtain the first-order Markov transition probability matrix W, as follows:

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1Q} \\ w_{21} & w_{22} & \cdots & w_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ w_{Q1} & w_{Q2} & \cdots & w_{QQ} \end{bmatrix} \tag{8}$$

Where W_{ij} denotes the probability that the system transitions from state i at time $t-1$ to state j at time t, and is computed as follows:

$$w_{ij} = \frac{C_{ij}}{\sum_{k=1}^{Q} C_{ik}} \tag{9}$$

The probability matrix is extended along the time axis to obtain a transition field image of dimension $N \times N$, where the pixel value at the i-th row and j-th column is defined as the transition probability from the state s_i at data point i to the state s_j at data point j:

$$M_{ij} = w_{s_i s_j} \tag{10}$$

In this manner, the MTF encodes the dynamic characteristics of the time series into spatial textural information within the image. Each pixel in the image inherently contains temporal distance information. Unlike approaches that simply truncate and reorganize signals into two-dimensional matrices, the MTF characterizes the transition properties of the time series among different amplitude states, thereby preserving the temporal dependencies of the signal.

3. PROPOSED METHOD

This paper proposes a multimodal fusion model that fully leverages both the one-dimensional temporal information and the two-dimensional image data of acoustic emission signals. Through a dual-branch parallel extraction architecture, the model achieves in-depth mining of multimodal features. The overall framework of the proposed model is illustrated in Fig. 5, which mainly consists of two core components, namely the dual-channel feature extraction module and the feature fusion and classification module.

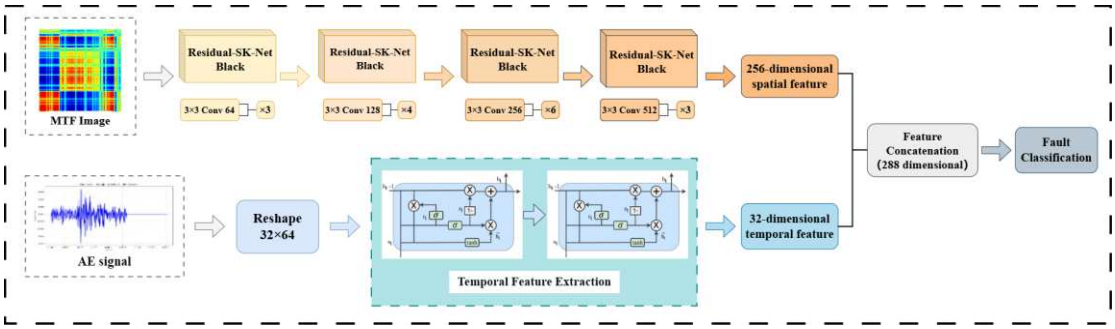

Figure. 5. Design of the multimodal dual-channel fusion network

The dual-channel feature extraction module consists of an image branch and a signal branch. The image branch employs ResNet34 as the backbone network, with SK-attention embedded in the residual blocks to extract deep spatial features from the images. The one-dimensional signal branch adopts a two-layer GRU network architecture to capture the temporal dependencies of the one-dimensional acoustic emission signals. Finally, through feature fusion, the feature vectors extracted from the two branches are concatenated. The fused features are then fed into a three-layer fully connected network, combined with a decreasing Dropout strategy to prevent overfitting, and ultimately output the classification results for six types of bearing conditions.

3.1 Design Rationale of the Model Framework

Most existing deep learning-based fault diagnosis approaches rely on unimodal inputs or straightforward cascades of off-the-shelf modules, without dedicated synergistic consideration of acoustic emission signal characteristics. The AES-FusionNet proposed herein is not a trivial assembly of ResNet, GRU, and SK-attention components; instead, the essence of the architectural design resides in the complementarity and synergistic interaction among these modules.

First, the combination of the signal branch and the image branch constitutes the most fundamental synergistic logic of the model. Since the fault characteristics of acoustic emission signals are simultaneously embedded in temporal dynamic evolution and spatial textural distribution, the signal branch captures the periodic patterns of faults along the time axis, while the image branch extracts the spatial patterns of amplitude state transitions in the MTF images. The two branches extract features from the temporal sequence and spatial texture dimensions, respectively, and after concatenation at the fusion layer, the model is able to jointly exploit both types of information for comprehensive judgment. This combination endows the model with the capability of synthesizing information from both dynamic and static perspectives, achieving a more holistic diagnostic performance compared with single-branch approaches.

Second, the integration of ResNet and SK-attention within the image branch further enhances the adaptability of spatial feature extraction. The textural features of MTF images exhibit significant scale variations; for instance, outer-race cracks correspond to large-scale periodic textures, whereas ball wear corresponds to localized fine textures. ResNet provides sufficient network depth to establish a foundation for feature extraction, while SK-attention introduces a multi-branch dynamic selection mechanism on this basis, enabling the network to adaptively select convolution kernels of different sizes according to

the specific textural characteristics of the input image. The synergy between the two components ensures that the network possesses adequate capacity to handle complex textures while remaining flexible in accommodating scale variations.

Finally, regarding the selection of the feature fusion strategy, this paper adopts feature-level concatenation rather than a complex cross-modal attention mechanism. Since the inputs to both the image branch and the signal branch originate from the same acoustic emission signal, they exhibit strong physical and semantic relevance, thereby obviating the need for sophisticated cross-modal alignment.

3.2 Two-dimensional image data branch

The image branch of the model is built upon the ResNet34 framework, with the number of convolutional blocks stacked in each stage being 3, 4, 6, and 3, respectively, and an SK-attention module is incorporated into each residual block. The input image size is 64×64, which falls into the category of small-sized images. The depth of ResNet34 is sufficient to provide an adequate receptive field, while its parameter count remains moderate, making it less prone to overfitting on small-scale datasets.

The improved residual block of ResNet34 is shown in Fig. 6. For MTF images, the textural features corresponding to different fault types exhibit varying scales. SK-attention is capable of adaptively selecting appropriate receptive fields, thereby extracting discriminative features more effectively. The final output of the image branch is a 256-dimensional feature vector, which serves as a compact representation of the spatial structural information and textural information embedded in the MTF images.

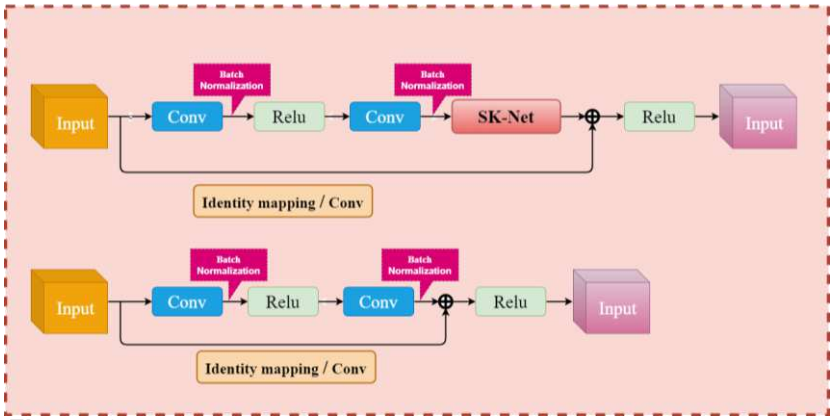

Figure. 6. Specific position of SK-attention embedded in the residual block

3.3 One-dimensional signal branch

The signal branch adopts a two-layer GRU network for temporal feature extraction from the one-dimensional acoustic emission signals, as shown in Fig. 7. The first GRU layer takes the reshaped 32×64 matrix as input, with each time step having an input dimension of 64 across a total of 32 time steps, and the hidden state dimension is set to 64. The second GRU layer takes the output of the first layer as its input, with the hidden state dimension reduced to 32. This design enables the network to extract temporal features at different levels of abstraction. The signal branch ultimately outputs a 32-dimensional feature vector, which serves as an encoding of the temporal dependencies. The dimensionality of this 32-dimensional vector is substantially smaller than the 256-dimensional output of the image branch. This intentionally unbalanced design allows the image branch to provide rich spatial textural information,

while the signal branch supplies lightweight temporal information, and the two modalities complement each other.

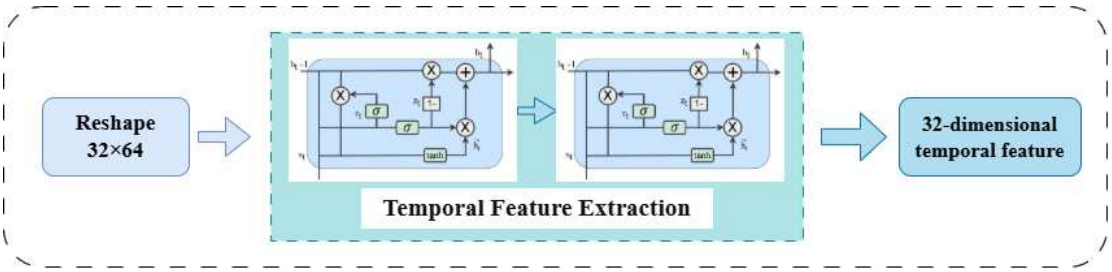

Figure. 7. Structure of the signal branch

The spatial structural and textural features extracted by the image branch exhibit strong discriminative capability for distinguishing different fault types, while the temporal evolution features of the fault signals extracted by the GRU network are beneficial for capturing the periodic patterns of fault impulses.

After the two branches, feature-level fusion is performed, leveraging both the image structural information and the original temporal information of the one-dimensional signals, thereby avoiding the problem of incomplete information inherent in single-modality approaches. The parallel design of the two branches ensures that the feature extraction processes of the two modalities are independent of each other, allowing the network parameters of each branch to be optimized separately and reducing the complexity of model training. The model adopts a feature concatenation strategy, in which the 256-dimensional feature vector output by the image branch and the 32-dimensional feature vector output by the signal branch are concatenated to form a 288-dimensional fused feature vector.

Let the output of the image branch be denoted as f_{img} and that of the signal branch as f_{sig}. The fused feature F_{fuse} is then expressed as:

$$F_{fuse} = Concat\left(f_{img}, f_{sig}\right) \in R^{288} \tag{11}$$

The enhanced classifier is designed to process the 288-dimensional fused feature vector. It consists of a three-layer fully connected network, incorporating a decreasing Dropout strategy to prevent overfitting.

The first fully connected layer of the classifier has an input dimension of 288 and an output dimension of 128, followed by a ReLU activation function and a Dropout layer. The second fully connected layer has an input dimension of 128 and an output dimension of 64, followed by a ReLU activation function and a Dropout layer. The third fully connected layer serves as the output layer, with an input dimension of 64 and an output dimension of 6, corresponding to the six types of bearing conditions, followed by a Softmax activation function.

The decreasing Dropout strategy adopted in each layer of the network is designed with certain distinctions. In the shallower layers of the network, the Dropout rate is set to 0.5 to mitigate the impact of relatively high feature dimensions and parameter counts. In the deeper layers, the Dropout rate is reduced to 0.3 to accommodate the lower feature dimensions and parameter scales. This design

effectively preserves the model's generalization capability while maximizing the representational power of the deep-layer features.

The classifier adopts cross-entropy loss as the loss function, which is expressed as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\left(\hat{y}_i, c\right) \tag{12}$$

Where N denotes the batch size, C denotes the number of classes, $y_{i,c}$ is the ground-truth label of the i-th sample, and $y_{i,c}(\hat{y}_i, c)$ is the predicted probability that the i-th sample belongs to class c.

4. EXPERIMENTAL VALIDATION AND ANALYSIS

4.1 Description of the Experimental Setup

The data used in this experiment were acquired from a dedicated test rig, which consists of a motor, a probe, a wired AE system, an AE signal sensor, a preamplifier, and a host control system, as shown in Fig. 8(a). The motor employed in the experiment is a variable-frequency speed-regulating three-phase asynchronous motor of model YBP90L-4. The sensor used for acquiring acoustic emission signals needs to be used in conjunction with the AE probe, which is of model WD FS15, with a maximum amplitude of 63.21 dB and a resonant frequency of 273.44 kHz. The core sensing element is a physical acoustic sensor of model WD, which belongs to the broadband sensor type and is capable of preserving complete frequency information, thereby facilitating feature extraction and fault classification. The relevant parameters of the sensor are listed in Table 1. As shown in Fig. 8(c), the preamplifier serves to amplify weak signals and improve the signal-to-noise ratio, as the signals acquired by the sensor are relatively weak and require amplification to reach a level suitable for transmission and processing. The AE mainframe is responsible for converting the electrical signals received from the preamplifier, as shown in Fig. 8(b), into digital signals via analog-to-digital (AD) conversion, and performing preliminary data organization. Both before and after AD conversion, the mainframe further processes the signals to enhance signal quality.

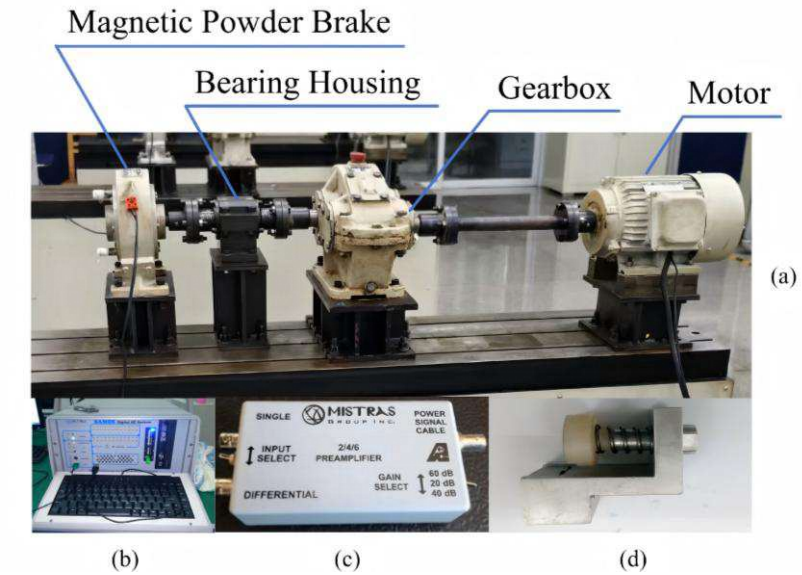

Figure. 8. (a) Structure of the test rig (b) AE mainframe (c) Preamplifier (d) Magnetic mounting base

Table 1. Sensor parameters

Sensor Information	Parameter
AE sensor model	WD broadband acoustic emission sensor
Sensor dimensions (mm)	18.17
Operating temperature (°C)	-6–175
Center frequency (kHz)	55
Frequency range (kHz)	100–1000
Interface type	DUAL BNC

During the experiment, the sampling frequency of the acoustic emission signal test rig was set to 55 kHz. The total operating time of the test rig accumulated to 1 hour, with each single recording session lasting 2 minutes. Six types of bearing conditions were simulated on the fault test rig, namely Normal, No Shaft, Outer Crack, Inner Crack, Ball Wear, and Missing Ball (one ball missing). The length of each one-dimensional raw acoustic emission data sample was 2,048 data points.

Through visualization analysis, the time-frequency diagrams of the six types of fault signals collected in the experiment are shown in Fig. 9.

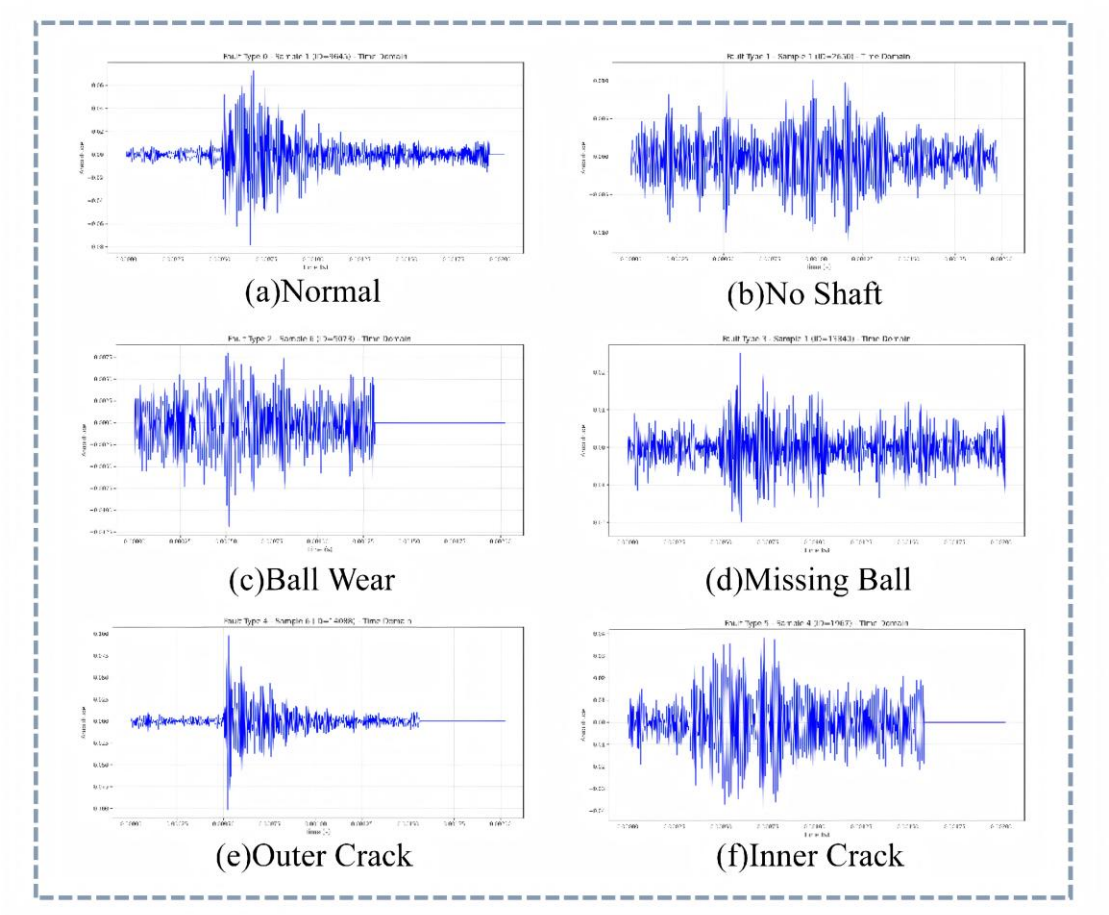

Figure. 9. Acoustic emission time-domain waveforms under different fault conditions. (a)-(f) correspond to different fault types

The experiments were conducted in a Linux operating system environment, and the relevant environmental parameters are listed in Table 2. The initial learning rate of the model was set to 0.008, the training batch size was set to 16, the optimizer employed was stochastic gradient descent (SGD), the

weight decay coefficient was set to 0.02, the dropout rate was set to 0.3, and the total number of training epochs was 100.

Table 2. Experimental environment parameters

Environment Parameter	Value
Operating system	Linux system
Number of training epochs	100
Optimizer	SGD
Batch size	32
Learning rate	0.008
Weight decay	0.02
Dropout rate	0.3

The selection of the above parameters was based on the optimization results of preliminary experiments, which enabled a favorable balance between convergence speed and generalization capability. The length of each one-dimensional raw acoustic emission data sample used in this chapter was 2,048 data points, and the two-dimensional image data were color three-channel images with a size of 64×64×3.

The conversion from one-dimensional signals to MTF images was implemented using the MarkovTransitionField module from the Python pyts library. The number of states was set to 4, with a quantile-based binning strategy. The MTF matrix size was 100×100, and the saved images were resized to 224×224 pixels, with the Jet colormap applied. The first 2,048 sampling points of each sample were used for transformation. The dataset was divided into training, validation, and test sets in a ratio of 7:2:1, after which MTF transformation was performed independently on each subset. The resulting three-channel color images were saved in a consistent format and maintained a one-to-one correspondence with their original one-dimensional signal samples via indexing. Fig. 10 presents the two-dimensional Markov Transition Field images, from which the textural features corresponding to different fault types can be visually distinguished.

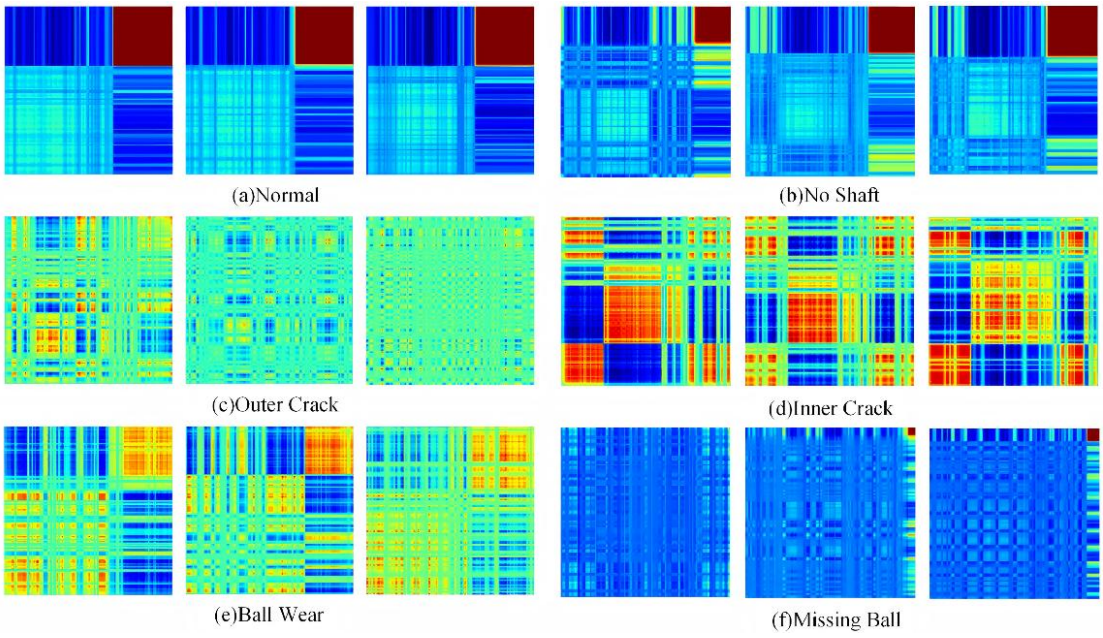

Figure. 10. Markov Transition Field images, (a)–(f) correspond to different fault types

4.2 Experimental Analysis and Evaluation

In this experimental validation, multiple performance evaluation metrics were adopted, including precision, recall, F1-score, and loss value, to comprehensively assess the effectiveness of the multimodal fusion model. Precision is used to evaluate the proportion of samples predicted as positive that are actually positive, recall is used to measure the proportion of actual positive samples correctly identified, and the F1-score is the harmonic mean of precision and recall [30], which can comprehensively reflect the classification performance of the model. The mathematical expressions of the above evaluation metrics are as follows:

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{14}$$

$$F_1 - Score = \frac{2TP}{2TP + FP + FN} = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{15}$$

Where TP denotes the number of fault instances correctly classified, TN denotes the number of non-fault instances correctly classified, FP denotes the number of non-fault instances incorrectly classified as faulty, and FN denotes the number of fault instances incorrectly classified as normal.

To ensure optimal classification performance of the image branch, this paper compares the performance of ResNet variants with different depths, including ResNet18, ResNet34, and ResNet50. The experimental results are presented in Table 3.

Table 3. Classification performance of ResNet with different depths

ResNet Depth	Accuracy	Loss Value	F1-score
Resnet18	89.4%	0.36	89.3%
Resnet34	94.1%	0.25	92.5%
Resnet50	92%	0.266	90.1%

As shown in the table, ResNet34 achieved the highest accuracy and F1-score, with the lowest loss value of 0.25. Although ResNet50 has a deeper architecture, its accuracy was slightly lower than that of ResNet34, indicating that excessively deep networks are prone to overfitting given the current image size and dataset scale. Therefore, ResNet34 was ultimately selected as the backbone feature extraction network for the two-dimensional image data in this study.

The proposed model was evaluated on six types of bearing conditions, namely Normal, No Shaft, Ball Wear, Missing Ball, Outer Crack, and Inner Crack, which correspond to loaded healthy condition, unloaded healthy condition, ball wear, missing ball, outer-race crack, and inner-race crack, respectively. The classification performance of the model is presented in Table 4.

Table 4. Validation of the model performance for each fault type

Fault Type	Precision	Recall	F1-Score
Normal	96.9%	96.5%	96.7%
No_Shaft	98%	97.8%	97.9%
Ball_Wear	92.4%	89.5%	90.9%
Missing_Ball	93.9%	95.5%	94.7%

Outer_Crack	94.2%	99.1%	96.6%
Inner_Crack	97.8%	94.58%	96.2%
Macro Avg	95.5%	95.5%	95.5%
Weighted Avg	95.4%	95.5%	95.4%

As shown in the table, the model achieved precision, recall, and F1-score exceeding 0.93 across all fault categories. Among them, the diagnostic performance for the No Shaft condition was the best, with a precision of 98% and an F1-score of 97%. The recall for Ball Wear was 89.5%, which was relatively lower. Overall, the average F1-score of the model reached 95.5%, indicating that the model maintains good classification balance across different fault types.

The experimental results are further evaluated using a confusion matrix, as shown in Fig. 11. The diagonal entries represent the correctly classified proportions for each fault category, where darker colors indicate better classification performance.

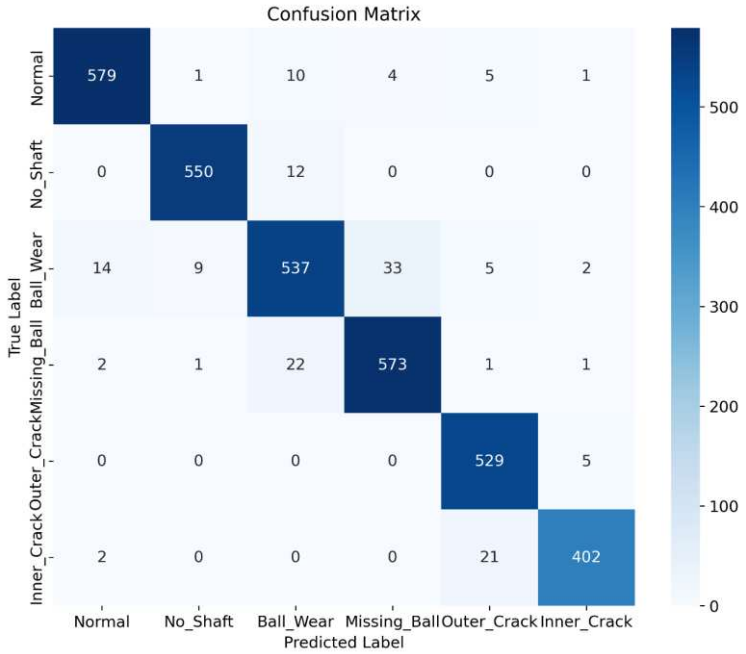

Figure. 11. Confusion matrix of the proposed model

The PR curve is used to evaluate the model's recognition capability on imbalanced datasets. As shown in Fig. 12(a), a curve that is closer to the upper-right corner and a higher AP value indicate a better trade-off between precision and recall. The AUC-ROC curve is employed to measure the model's ability to distinguish between positive and negative classes. As shown in Fig. 12(b), an AUC value closer to 1 indicates superior classification performance.

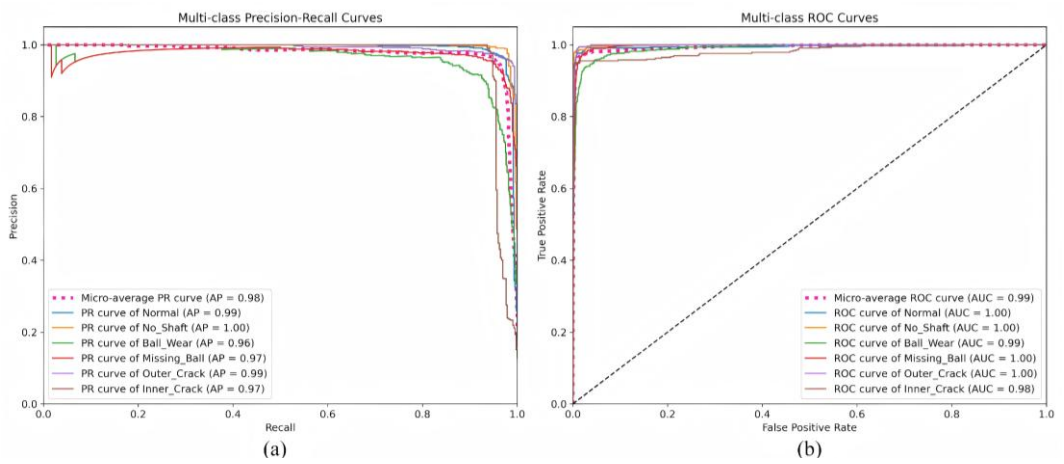

Figure. 12. (a) PR curve, (b) AUC curve

The AUC values for all six fault types in the figure all exceeded 0.99, indicating that the model possesses strong discriminative capability across different fault categories. The AP values for each class in the PR curve also remained at a high level, further validating the superiority of the proposed model in fault diagnosis tasks.

The error heatmap provides an intuitive visualization of classification errors, as shown in Fig. 13. The values in Fig. 13(a) represent the number of misclassifications, while those in Fig. 13(b) represent the percentage. Darker colors indicate higher numbers or proportions of misclassifications at the corresponding positions. The error heatmap facilitates the identification of which fault categories are prone to confusion with each other.



Figure. 13. (a) Error heatmap (count), (b) Error heatmap (percentage)

It can be intuitively observed from the figure that the vast majority of samples are correctly classified, while the misclassified samples are mainly concentrated among a few easily confusable categories.

The experimental analysis also provides threshold analysis curves for evaluating the classification performance of the model under different decision thresholds, as shown in Fig. 14, where the trends of precision, recall, and F1-score for each fault type are plotted against varying thresholds. The analysis results indicate that the model maintains a high F1-score across a wide range of thresholds, demonstrating strong robustness and stability.

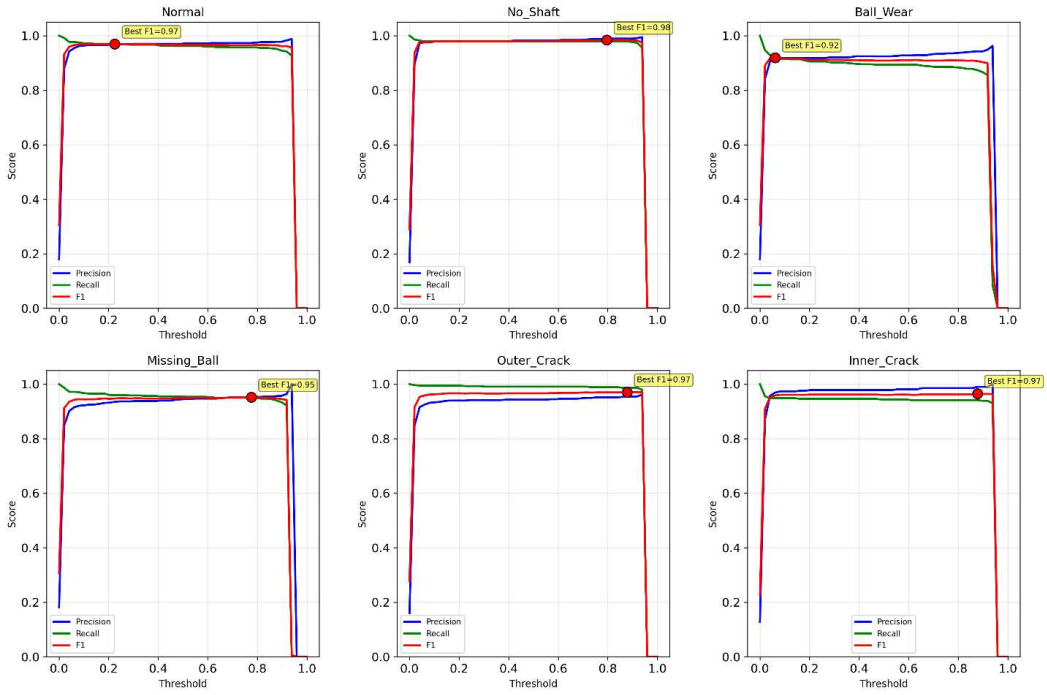

Figure. 14. Threshold analysis curves

For fault categories with distinct features, such as Normal and No Shaft, the model consistently achieves F1-scores above 0.95 within the threshold range of 0.4 to 0.6. For fault types such as Ball Wear and Inner Crack, the optimal threshold exhibits a slight deviation.

To further evaluate the performance of the multimodal model, comparative experiments were conducted with multiple mainstream algorithms, including both one-dimensional signal processing methods and two-dimensional image processing methods. All comparative experiments adopted the same dataset division and evaluation metrics to ensure fairness of the results. The experimental results are presented in Table 5.

Table 5. Performance comparison of different models

Model	Accuracy	Recall	Average Precision	Training Time	F1-score
1D-CNN	91.45%	91.33%	92.21%	40.43 minutes	91.52%
1D-KNN	39.45%	40%	46%	11.48 minutes	33.32%
1D-Transformer	85.25%	85.25%	85.51%	40.68 minutes	85.12%
1D-Naïve Bayes	43.27%	44%	46%	0.95 minutes	41%
1D-Support Vector Machine	33.97%	35%	37%	2.27 minutes	33.85%
2D-CNN	70.67%	70.24%	69.78%	1063.10 minutes	70.2%
2D-Mobilenet	73.35%	72.79%	72.11%	361.38 minutes	72.31%
2D-Vit	90.14%	90.41%	91.14%	960.24 minutes	90.22%
2D-VGG 16	87.46%	88.32%	87.65%	934.64 minutes	87.54%
Ours	95.74%	95.5%	95.6%	572.67 minutes	95.83%

Among the one-dimensional signal processing models, 1D-CNN achieved the best performance, with an accuracy of 91.45% and an F1-score of 91.52%. In contrast, traditional machine learning methods,

such as KNN, Naïve Bayes, and SVM, exhibited inferior performance, indirectly indicating that deep learning models possess certain advantages in processing high-dimensional data. Among the two-dimensional image processing models, 2D-ViT achieved an accuracy of 90.14%, 2D-VGG16 achieved 87.46%, 2D-MobileNet achieved 73.35%, and the standard 2D-CNN achieved only 70.67%, suggesting that deeper network architectures have a positive impact on improving diagnostic performance.

In terms of training time, the multimodal fusion model required 572.67 minutes, which falls between those of 2D-MobileNet and 2D-VGG16, indicating that the model achieves a favorable balance between accuracy and training efficiency.

The experimental analysis also introduces three-dimensional line charts and performance radar charts to demonstrate the comprehensive performance of each model across multiple dimensions, as shown in Fig. 15(a). In the three-dimensional line chart, the X-axis represents the model names, the Y-axis represents different metrics, and the Z-axis represents the actual performance values. The radar chart encompasses five evaluation dimensions, including accuracy, recall, average precision, and F1-score, where each axis corresponds to an evaluation metric and higher values on the axes indicate better performance. A larger area enclosed by the radar chart in Fig. 15(b) signifies superior comprehensive performance. The comparison among the models in the figure clearly demonstrates that the proposed multimodal model (in red) achieves satisfactory performance across all metrics.

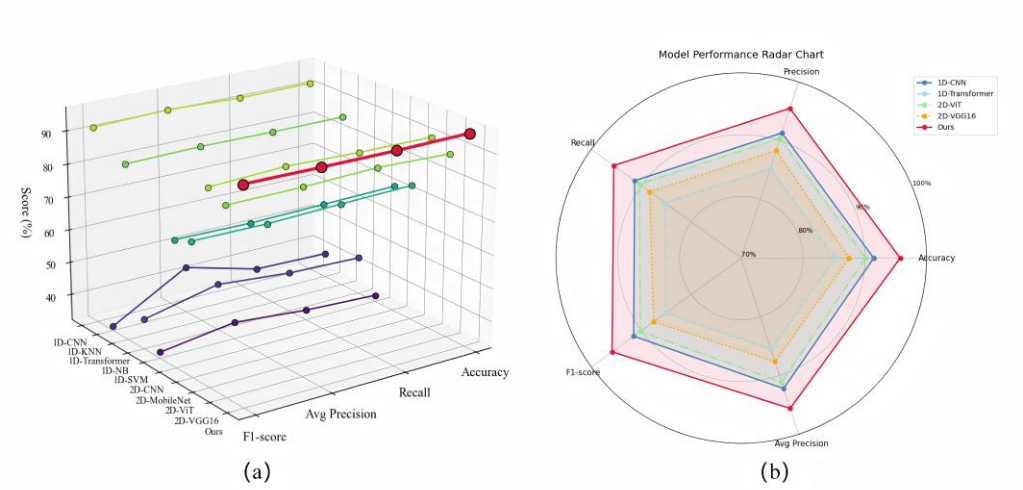

Figure. 15. (a) Three-dimensional line chart of different models, (b) Radar chart comparing the performance of different models

This paper also designed an analysis of the effects of different attention mechanisms to validate the effectiveness of SK-Attention in ResNet34, as shown in Table 6, which evaluates the performance of various attention mechanisms. All experiments were conducted with identical hyperparameter settings to ensure fair comparison. The experimental results demonstrate that SK-Attention achieved the highest performance, with a validation accuracy of 95.74% and an F1-score of 95.83%, exhibiting the best performance among the evaluated mechanisms in ResNet34.

Table 6. Performance comparison of different attention mechanisms

Attention Mechanism	Precision	Macro F1-Score	Weighted F1-Score	Total Training Time

SK-Net	95.74%	95.83%	95.74%	572.67 minutes
Wavelet-Kernel Attention	95.48%	95.66%	95.5%	5361.53 minutes
Cross-Attention	95.06%	95.08%	95.06%	434.64 minutes
ECA-Net	94.89%	94.94%	94.89%	373.52 minutes
PSANet	94.88%	94.92%	94.87%	432.41 minutes
DSW	94.47%	94.48%	94.46%	499.27 minutes
Co-Attention	65.05%	64.27%	64.75%	361.86 minutes

Meanwhile, Wavelet-Kernel Attention ranked second with an accuracy of 95.48% and a macro F1-score of 95.66%, exhibiting a performance gap of only 0.26 percentage points compared with SK-Net. Wavelet-Kernel Attention integrates wavelet transform with the attention mechanism, enabling feature enhancement in both the time and frequency domains, which offers certain advantages for analyzing acoustic emission signals.

In terms of training efficiency, ECA-Net demonstrated the most favorable performance, completing training in 373.52 minutes. In contrast, Co-Attention yielded abnormal experimental results, with an accuracy of only 65.05%, substantially lower than all other methods. It is inferred that Co-Attention is designed for collaborative attention computation between two modalities; however, in the present framework, the substantial disparity in feature dimensions between the two branches led to an imbalance in attention weight learning, thereby preventing effective convergence of the Co-Attention mechanism.

To validate the effectiveness of each component of the proposed model, ablation experiments were designed in this study, as presented in Table 7. The results demonstrate that the complete model achieves the best performance, with an accuracy of 95.74% and an F1-score of 95.83%. Compared with the single-modality 2D-ResNet34 and 1D-dual GRU network, multimodal fusion improved the accuracy by 0.48 and 2.33 percentage points, respectively, verifying the effectiveness of the multimodal fusion strategy. Meanwhile, the accuracy of 2D-ResNet34+SK-Attention was only 93.07%, which is lower than the 95.26% accuracy of 2D-ResNet34, indicating that simply adding SK-Attention does not guarantee overall performance improvement. The complete model also achieved an accuracy improvement of 1.03 percentage points over ResNet34+dual GRU network, further validating the beneficial contribution of SK-Net.

Table 7 Ablation experiment results

Model Configuration	Accuracy	Recall	Average Precision	Training Time	F1-score
Complete model	95.74%	95.5%	95.6%	572.67 minutes	95.83%
2D-ResNet34	95.26%	95%	95%	774.87 minutes	95.41%
1D-dual GRU network	93.41%	93.46%	93.55%	17.77 minutes	93.4%
ResNet34 + dual GRU network	94.71%	93.44%	93.45%	600.67 minutes	93.6%

To more intuitively verify the effectiveness of the multimodal fusion strategy, t-SNE dimensionality reduction visualization was performed on the 288-dimensional features output by the fusion layer, as

shown in Fig.16. In the figure, different colors represent the six fault categories. It can be observed that each category forms a distinct cluster in the feature space, with clear inter-class boundaries and compact intra-class aggregation, demonstrating that the fused features possess strong discriminative capability for different fault types and validating the effectiveness of the multimodal fusion strategy.

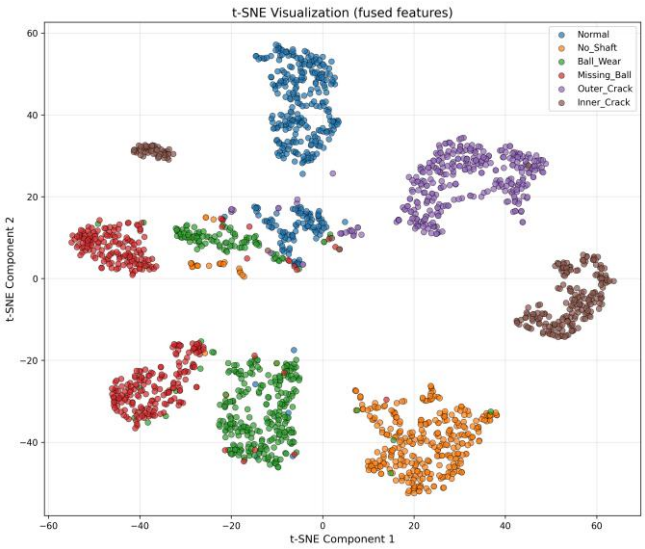

Figure.16. t-SNE Visualization (fused features)

5. CONCLUSION

This paper takes rolling bearings as the research subject and, in response to the problem of insufficient information in single-modality acoustic emission fault diagnosis methods, proposes a deep learning-based multimodal diagnostic model. A systematic investigation is conducted around the characteristic analysis of acoustic emission signals, relevant deep learning theories, multimodal network design, and experimental validation.

Validated on the actual collected dataset, the proposed multimodal model achieved an average correct classification rate of 95.74% across the six fault types, with a macro-average F1-score of 95.83%. The confusion matrix reveals that the correct classification rate for each category exceeds 97%. Comparative experiments demonstrate that the multimodal fusion model outperforms the image branch alone and the signal branch alone by 3.2% and 5.7% in accuracy, respectively, confirming the advantage of fusing information from both modalities. In terms of attention mechanism comparison, SK-attention achieved the highest validation accuracy, verifying the superiority of multi-scale adaptive receptive fields in image feature extraction.

Although the proposed model has achieved satisfactory diagnostic performance on the experimental dataset, certain limitations remain. First, in terms of model efficiency, the current training time is relatively long, which still falls short of the real-time monitoring requirements in practical industrial scenarios. Future work may consider model compression or lightweight network design to reduce computational cost and training time while maintaining accuracy, thereby making the model more suitable for deployment on edge computing devices.

The experimental data used in this study were all acquired from a fault test rig under fixed rotational speed conditions, without involving dynamic operating scenarios. In actual industrial environments, the

rotational speed and load of bearings vary over time, and the data distribution as well as fault characteristics may shift under certain conditions. Future work can introduce transfer learning methods to validate the generalization capability of the model under varying operating conditions. In addition, this study primarily addresses the classification of bearing faults, while the detection and early warning of incipient weak faults have not yet been explored. Given the sensitivity of acoustic emission technology to early-stage faults, subsequent research can investigate early fault warning methods based on the proposed model, for instance, by analyzing the temporal evolution trends of feature vectors to track fault progression.

FUNDING

This work described in this paper was supported by The Research Funding of GDUPT, Research on Lagrangian Coherent Structure Analysis of Heat Sink for Heat Transfer Performance with Piezoelectric Fan, Guangdong Science and Technology Program in 2024, (Project No.: 2024A0505050022), the College Students' Innovation and Entrepreneurship Project of Guangdong University of Petrochemical Technology (Project No.: 25A018).

AUTHOR CONTRIBUTIONS

Yi Chen contributed to Conceptualization, Methodology, Software, and Writing original draft. Wenmin Wen contributed to Data curation, Investigation, and Validation. Wanying Wu contributed to Formal analysis and Visualization. Junqi Liu contributed to Resources and Data curation. Yi Pan contributed to Supervision and Writing review and editing. Yanzuo Chang contributed to Project administration, Funding acquisition, Supervision, and Writing review and editing. All authors have read and agreed to the published version of the manuscript.

REFERENCES

1. Yaqub, M. Z., & Alsabban, A. (2023). Industry-4.0-Enabled Digital Transformation: Prospects, Instruments, Challenges, and Implications for Business Strategies. Sustainability, 15(11), 8553. https://doi.org/10.3390/su15118553
2. Li, H., Sun, Y., Adumene, S. et al. Machinery safety improvement in manufacturing-oriented facilities: a strategic framework. Int J Adv Manuf Technol 138, 4531–4560 (2025). https://doi.org/10.1007/s00170-025-15763-w
3. Heng, A., Zhang, S., Tan, A. C. C., & Mathew, J. (2009). Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing*, *23*(3), 724～739. https://doi.org/10.1016/j.ymssp.2008.06.009
4. Zhu, H., Sui, Z., Xu, J., & Lan, Y. (2024). Fault Diagnosis of Mechanical Rolling Bearings Using a Convolutional Neural Network–Gated Recurrent Unit Method with Envelope Analysis and Adaptive Mean Filtering. Processes, 12(12), 2845. https://doi.org/10.3390/pr12122845
5. Morhain, A., & Mba, D. (2003). Bearing defect diagnosis and acoustic emission. *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, *217*(4), 257-272. 10.1243/135065003768618614

6. Heng, A., Zhang, S., Tan, A. C. C., & Mathew, J. (2009). Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing*, *23*(3), 724～739. https://doi.org/10.1016/j.ymssp.2008.06.009
7. Song, J., Leng, J., Li, J., Wei, H., Li, S., & Wang, F. (2025). Application of Acoustic Emission Technique in Landslide Monitoring and Early Warning: A Review. *Applied Sciences*, *15*(3), 1663. https://doi.org/10.3390/app15031663
8. Shi, H., Liu, J., Sun, Y., Wang, X., & Li, F. (2025). Study on the Temporal–Spatial Evolution Characteristics of Micro-Cracks During Fault Slip Based on Acoustic Emission Monitoring. *Rock Mechanics and Rock Engineering*, *58*(3), 3417～3435. https://doi.org/10.1007/s00603-024-04320-8
9. Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., & Iyengar, S. S. (2018). A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Comput. Surv.*, *51*(5). https://doi.org/10.1145/3234150
10. Ahmed, S. F., Alam, Md. S. B., Hassan, M., Rozbu, M. R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A. B. M., & Gandomi, A. H. (2023). Deep learning modelling techniques: Current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, *56*(11), 13521～13617. https://doi.org/10.1007/s10462-023-10466-8
11. Prosvirin, A., Kim, J. Y., & Kim, J. M. (2018). Bearing fault diagnosis based on convolutional neural networks with kurtogram representation of acoustic emission signals. In *Lecture Notes in Electrical Engineering* (Vol. 474, pp. 21–26). Springer Verlag. https://doi.org/10.1007/978-981-10-7605-3_4
12. Jiang, P., Xia, J., Li, W., Xu, C., and Sun, W. (June 29, 2024). "Innovative Bearing Fault Diagnosis Method: Combining Swin Transformer Deep Learning and Acoustic Emission Technology." ASME. *ASME J. Risk Uncertainty Part B*. March 2025; 11(1): 011102. https://doi.org/10.1115/1.4065754
13. Li, J., Sheng, H., Liu, B., & Liu, X. (2026). Fault Diagnosis and Classification of Rolling Bearings Using ICEEMDAN–CNN–BiLSTM and Acoustic Emission. *Sensors*, *26*(2), 507. https://doi.org/10.3390/s26020507
14. Mao, M., Jiang, Z., Tan, Z., Xiao, W., & Du, G. (2025). Tilting Pad Thrust Bearing Fault Diagnosis Based on Acoustic Emission Signal and Modified Multi-Feature Fusion Convolutional Neural Network. *Sensors (Basel, Switzerland)*, *25*(3), 904. https://doi.org/10.3390/s25030904
15. Wang, R., Dong, E., Cheng, Z., Liu, Z., & Jia, X. (2024). *Transformer-based intelligent fault diagnosis methods of mechanical equipment: A survey*. *Open Physics*, *22*(1). https://doi.org/10.1515/phys-2024-0015
16. Wang, Y., & Vinogradov, A. (2023). Improving the Performance of Convolutional GAN Using History-State Ensemble for Unsupervised Early Fault Detection with Acoustic Emission Signals. *Applied Sciences*, *13*(5), 3136. https://doi.org/10.3390/app13053136
17. Jian, X., Li, W., Guo, X., & Wang, R. (2019). Fault Diagnosis of Motor Bearings Based on a One-Dimensional Fusion Neural Network. *Sensors*, *19*(1), 122. https://doi.org/10.3390/s19010122
18. Shen, F., He, S., Song, D., He, X., Yang, G., & baig, O. (2024). Response characteristics of acoustic emission signal and judgment criteria for different fracture modes. *Engineering Fracture Mechanics*, *296*, 109862. https://doi.org/10.1016/j.engfracmech.2024.109862
19. Kullu O, Cinar E. A Deep-Learning-Based Multi-Modal Sensor Fusion Approach for Detection of Equipment Faults. *Machines*. 2022; 10(11):1105. https://doi.org/10.3390/machines10111105

20. Liu, Q., Tong, W., Bai, H., & Dong, S. (2025). SCBM-Net: A multimodal feature fusion-based dual-channel method for bearing fault diagnosis. Scientific Reports, 15(1), 37904. https://doi.org/10.1038/s41598-025-21665-4
21. Li, D., Pang, Z., Chen, Y., Yang, K., Shao, J., Luo, Y., Zeng, Y., He, C., & Gao, Y. (2025). FD-MVLLM: Fault diagnosis based on multimodal vibration data and large language model for bearing system. *Mechanical Systems and Signal Processing*, *239*, 113226. https://doi.org/10.1016/j.ymssp.2025.113226
22. Zhang, S., Song, B., Zhao, F., Sun, G., & Meng, F. (2025). A decoupling method for compound faults of rolling bearings based on multi-modal bidirectional cross-attention fusion and contrastive learning. *Measurement Science and Technology*, *36*(10), 106131. https://doi.org/10.1088/1361-6501/ae080f
23. Hu, Y., Xu, W., Du, X. (2026). LWCNet: A Physics-Guided Multimodal Few-Shot Learning Framework for Intelligent Fault Diagnosis. *Computers, Materials & Continua*, *87(2)*, 67. https://doi.org/10.32604/cmc.2026.074437
24. K. He, X. Zhang, S. Ren, & J. Sun. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770 ～ 778. https://doi.org/10.1109/CVPR.2016.90
25. Lan, R., Zou, H., Pang, C., Zhong, Y., Liu, Z., & Luo, X. (2021). Image denoising via deep residual convolutional neural networks. Signal, Image and Video Processing, 15(1), 1 ～ 8. https://doi.org/10.1007/s11760-019-01537-x
26. Cho, K., Van Merriënboer, B., Gulçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, October). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1724-1734). https://doi.org/10.3115/v1/D14-1179
27.Zhang, X., Zhang, Y., & Yang, Q. (2026). GRU-Dual Transformer–BiLSTM Fusion for Transformer Fault Diagnosis. *International Journal of Circuit Theory and Applications*, *54*(5), 2582～2600. https://doi.org/10.1002/cta.70143
28. X. Li, W. Wang, X. Hu, & J. Yang. (2019). Selective Kernel Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 510 ～ 519. https://doi.org/10.1109/CVPR.2019.00060
29. Wang, M., Wang, W., Zhang, X., & Iu, H. H.-C. (2022). A New Fault Diagnosis of Rolling Bearing Based on Markov Transition Field and CNN. Entropy, 24(6), 751. https://doi.org/10.3390/e24060751
30. Sujon, K. M., Hassan, R., Choi, K., & Samad, M. A. (2025). Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *Journal of Big Data*, *12*(1), 268. https://doi.org/10.1186/s40537-025-01313-4