Information Fusion 134 (2026) 104339

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full Length Article

A multi-channel data fusion-enabled multilevel graph-guided framework for fault diagnosis in rotating machinery under extreme biased data ★★

Yue Yu [a,*], Hamid Reza Karimi [a,*], Pradeep Kundu [b], Enrico Zio [c,d], Ke Feng [e]

[a] Department of Mechanical Engineering, Politecnico di Milano, via La Masa 1, Milan, 20156, Italy
[b] Department of Mechanical Engineering, KU Leuven, Bruges Campus, Bruges, Belgium
[c] Energy Department, Politechnico di Milano, Milano, Italy
[d] CRC, MINES Paris-PSL University, Sophia Antipolis, France
[e] Key Laboratory of Education Ministry for Modern Design & Rotor-Bearing System, Xi'an Jiaotong University, Xi, 710049, China

ARTICLE INFO

Keywords:
Rotating machinery
Fault diagnosis
Multi-channel data
Extreme biased data
Global and local feature fusion learning
Feature inductive learning

ABSTRACT

Fault diagnosis based on multi-channel data plays a crucial role in rotating machinery monitoring. By leveraging signals acquired from multiple sensors, more comprehensive fault-related information can be extracted, thereby improving diagnostic accuracy. This paper proposes a novel Multi-channel data fusion-enabled Multilevel Graph-guided Framework for Diagnosis (MSGFD) to address fault diagnosis under extreme data imbalance. First, an efficient preprocessing strategy is developed to transform multi-channel signals into structured representations suitable for graph-based learning. Subsequently, a MultiGraph construction mechanism is introduced to capture discriminative and complementary fault information through four distinct graph topologies. To address the challenge of limited supervision in extremely imbalanced scenarios, a multilevel learning architecture integrating a Graph Multilayer Perceptron (MLP) and a Graph Transformer is designed to jointly model local and global feature dependencies. Furthermore, a deep divergence-based clustering (DDC) loss is incorporated to enhance inter-class separability and intra-class compactness. Extensive experiments conducted under various imbalance settings demonstrate the robustness and effectiveness of the proposed method across multiple fault categories. The source code is publicly available at: https://github.com/Polimi-YuYue.

1. Introduction

Rotating machinery is a key element of industrial systems [1]. Failures of rotating components cause productivity and safety problems, and must thus be prevented [2,3]. For this, Machine Learning (ML) can be used to establish a link between monitoring data and the health states of rotating machinery [4]. For example, Cai et al. proposed a digital twin-driven fault diagnosis method integrating virtual-real data and Bayesian networks to accurately identify composite faults in subsea production control systems [5]. Qin et al. developed a novel interpretable waveform segmentation model by integrating the class-weighted Lovasz-softmax loss and physics-informed denoising loss to enable bearing fault diagnosis [6]. However, manual feature extraction can be time-consuming, labor-intensive, and biased.

To address this issue, Deep Learning (DL)-based fault diagnosis of rotating machinery has emerged as a story [7–9]. For example, Yan et al. developed a novel partial domain adaptation method to enhance domain-invariant feature learning for fault diagnosis [10]. Chen et al. built a multi-scale and multi-structure information-embedded unsupervised graph transfer framework to improve cross-domain fault diagnosis performance under variable working conditions [11]. Qin et al. proposed an adaptive intermediate class-wise distribution alignment paradigm to enhance fault diagnosis by transfer learning [12]. Ding et al. built a digital twin-assisted dual transfer framework for rolling bearing fault diagnosis [13]. Qin et al. developed an intelligent squirrel cage with integrated piezoelectric and triboelectric components to achieve rotating speed monitoring and bearing fault diagnosis, which provides a new option for online condition monitoring of aero-engines [14]. Li

★ Source code can be available at: https://github.com/Polimi-YuYue

★★ This research is supported by the scholarship from the China Scholarship Council (CSC), China under grant CSC N202308130067 and in part by the Horizon Marie Skłodowska-Curie Actions program under grant 101073037 .

* Corresponding authors.
E-mail addresses: yue.yu@polimi.it (Y. Yu), hamidreza.karimi@polimi.it (H.R. Karimi), pradeep.kundu@kuleuven.be (P. Kundu), enrico.zio@polimi.it (E. Zio), kefeng@xjtu.edu.cn (K. Feng).

https://doi.org/10.1016/j.inffus.2026.104339
Received 7 May 2025; Received in revised form 11 March 2026; Accepted 30 March 2026
Available online 4 April 2026

Nomenclature

Abbreviations

CNN	Convolutional Neural Network
CWT	Continuous Wavelet Transform
DDC	Divergence-based Clustering
DL	Deep Learning
EBD	Extreme Biased Data
FPR	False Positive Rate
GCN	Graph Neural Network
GLFF	Global and Local Feature Fusion
HHT	Hilbert–Huang Transform
IoT	Internet of Things
ML	Machine Learning
MLP	Multilayer Perceptron
MST	Minimum Spanning Tree
PDF	Probability Density Function
ROC	Receiver Operating Characteristic Curve
STFT	Short-Time Fourier Transform
TPR	True Positive Rate
WT	Wavelet Transform

et al. presented a significant advancement by being one of the first attempts to introduce neuro-symbolic AI into the field of Intelligent Fault Diagnosis. By proposing the Deep Expert Network, the authors successfully bridge the gap between data-driven deep learning and symbolic expert knowledge, offering a unified method that addresses the critical "black-box" limitations of traditional AI in safety-critical industrial assets [15]. Wang et al. present a significant and well-motivated hybrid approach that successfully integrates physics-informed constraints with sequential attention-based deep learning, enabling accurate, interpretable, and uncertainty-aware fatigue crack growth prediction under variable-amplitude loading [16]. Wang et al. proposed a novel and important two-stage RUL prediction method that leverages degradation angle-based change-point identification to more accurately model stage-dependent Wiener process drift, significantly improving long-term degradation tracking and prediction accuracy [17]. Qin et al. developed a unified modeling method for complex rotor systems, capable of revealing the coupling mechanisms among rotating component, which facilitates accurate prediction and fault diagnosis of rotating machinery [48].

However, despite the good performance of DL in fault diagnosis of rotating machinery, two main limitations remain [18–20]. First, DL-based studies usually process single-channel signals to generate embeddings, neglecting the importance of multi-channel signals for fault diagnosis. Indeed, single-channel sensor data cannot fully capture the dynamics of rotating machinery, resulting in unsatisfactory diagnostic results, whereas multi-channel data can capture either consistent or complementary diagnostic information for the same fault category [21–23]. Ma et al. et al. proposed a novel supply-demand deviation model based on Taguchi theory. This approach not only strengthens the mathematical foundation for analyzing communication networks but also fills a gap in understanding the coupling mechanisms between energy supply-demand deviation and communication bandwidth and power [47]. Yu et al. introduced multi-channel signals within a DL framework by calculating channel importance; however, the method completely relies on intermediate fusion and may not accurately measure feature alignment [24]. Different from single-fusion-based DL approaches for fault diagnosis with multi-channel data, we revisit the principles of data fusion and propose a multilevel framework of feature extraction and fusion for fault diagnosis.

Also, rotating machinery typically operates under normal conditions, making it difficult to obtain a large number of samples of various fault types. For limited or imbalanced fault samples in the collected data, Wang et al. employed a style-based generative adversarial network to perform rebalancing [25]. From the perspective of DL-based learning, the lack of valid information in these data-level methods limits model performance and alters data distribution [26,27]. Motivated by the ability of time-frequency methods to transform RGB images from multi-channel data and aggregate more representative characteristics [28,29], we propose a multi-level strategy for more effective and robust data preprocessing to facilitate the simultaneous representation of temporal and spectral characteristics under extreme biased data. Moreover, a MultiGraph-based augmentation strategy is designed to enhance fault diagnosis accuracy.

In this work, we propose a novel multi-channel data fusion-enabled multilevel graph-guided framework (MSGFD) for fault diagnosis under extreme biased data. The framework includes data preprocessing, MultiGraph generation, feature inductive learning, local and global feature fusion, and optimal fault diagnosis. By leveraging a novel data preprocessing technique, we first convert multi-channel signals into RGB images to overcome the limitation of insufficient information under extreme biased data, which are then used as input data for subsequent fault diagnosis. Then, to improve the capability of fault diagnosis under extreme biased data, we propose MultiGraph generation to construct reasonable and powerful graph topologies for discriminative and complementary fault-based information. After that, we develop a feature inductive learning (FIL) module to eliminate the heterogeneity of RGB images and exploit the common feature spaces. With the assistance of the FIL module, the proposed Graph MLP and Transformer can effectively capture both local and global features. Moreover, to address the challenge of deficient supervision under extreme biased data, we propose an optimal fault diagnosis approach that effectively regularizes the training of MSGFD. This is achieved by incorporating deep divergence-based clustering (DDC) loss, ensuring both inter-category separability and intra-category compactness.

The contributions of this paper are summarized as follows:

- We present a novel multi-channel data fusion-enabled multilevel graph-guided method (MSGFD) for fault diagnosis under extreme biased data. Additionally, we design a novel data preprocessing technique to capture the simultaneous representation of temporal and spectral characteristics without expert knowledge and experience.
- We propose the FIL module to explore common feature spaces using MultiGraph-based topologies. Furthermore, the advantages of the graph MLP and transformer are leveraged to extract local and global features compatible with extreme biased data.
- We present optimal fault diagnosis under the DDC loss constraint, which benefits category clustering under extreme biased data.
- We perform extensive experiments in four case studies to show the feasibility and effectiveness of the proposed MSGFD framework.

The remainder of this article proceeds as follows. In Section 2, convolutional neural network (CNN), graph neural network (GCN), and Graph Transformer are briefly introduced. In Section 3, the proposed MSGFD is illustrated. The superiority and effectiveness of the MSGFD in fault diagnosis under extreme biased data are validated through four case studies in Section 4. In Section 5, discussion on MSGFD is conducted. The conclusions of this work are presented in Section 6.

All abbreviations and notations used in this paper are summarized in the Appendix at the end of the manuscript, as presented in Table 12.

2. Related works

2.1. Convolutional neural network

CNN, as a classical type of artificial neural network, is usually consisted of a convolutional layer, an activation layer, a pooling layer, a fully-connected layer, and an output layer. It is widely applied in various fields (i.e., object detection, image classification, natural language processing, etc) due to its powerful feature extraction ability [30–32], as shown in Fig. 1.

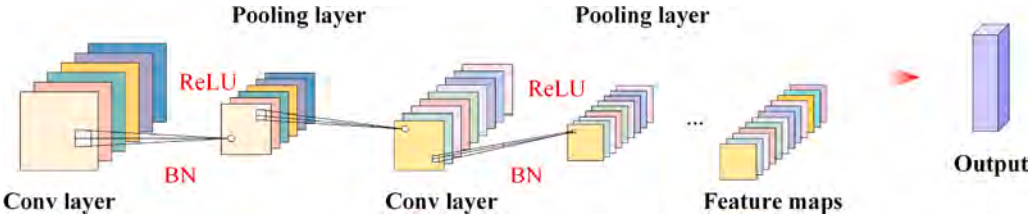

Fig. 1. Architecture of the CNN.

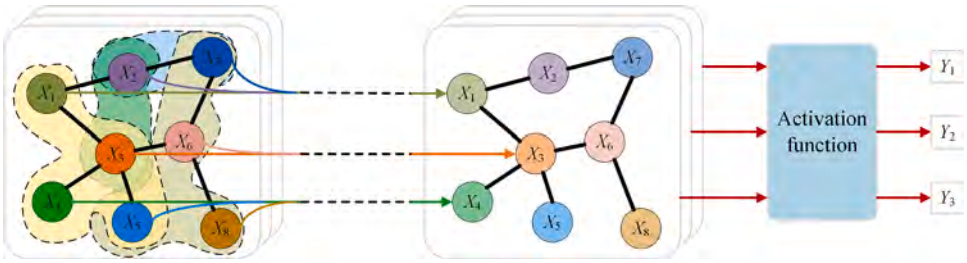

Fig. 2. Framework of graph convolutional network.

The convolutional layer, activation layer, and pooling layer are introduced to extract specific features, improve feature non-linearity, and reduce the size of extracted features, separately. The typical formula for CNN can be written as follows:

$$y_{\text{conv}}^{m+1} = \mathrm{F}\left(\sum_{i=1}^{C}\sum_{j=1}^{R} w_{i,j}^{m} * x_i^m + b_{i,j}^m\right) \tag{1}$$

$$y_{\text{pool}}^{m+1} = \text{pooling}_S\left(y_{\text{conv}}^{m+1}\right) \tag{2}$$

where y_{conv}^{m+1} and y_{pool}^{m+1} refer to the $m+1$-th output and pooling features, respectively. $\mathrm{F}(\cdot)$ and $\text{pooling}_S(\cdot)$ stand for the activation function (i.e., Sigmoid or Tanh) and the pooling operation (i.e., max-pooling or average-pooling), separately. $w_{i,j}^m$ and $b_{i,j}^m$ are the weight and bias of the convolutional layer, respectively. x_i^m is the input feature map. C and R represent the spatial dimensions of the convolutional kernel. m denotes the layer index. $*$ represents the convolution operator. S is the pooling window size.

2.2. Graph neural network

The quality of the graph data $G = \{V, A, E, F\}$ is essential to the Graph Neural Network (GNN), which consists of the node set V, adjacency matrix A, edge set E, and feature matrix F, as shown in Fig. 2 [33,34]. Given that graph data $x \in \mathcal{R}^{N\times S}$, the output through spectral graph convolution can be expressed as

$$Y = g_{\theta_l} *_G x = U g_{\theta_l} U^T x \tag{3}$$

where g_{θ_l} and $*_G$ refer to the graph spectral filters and spectral graph convolution, respectively. θ_l is the learnable parameters. $(\cdot)^T$ is the transpose operation. U means eigenvectors of symmetric normalized graph Laplacian matrix L, defined as:

$$L = I_N - D^{-1/2} A D^{-1/2} \tag{4}$$

where D and I_N stand for the diagonal degree matrix and identity matrix, respectively.

To further improve effectiveness and achieve globalization, Defferrard et al. and Kipf and Welling introduced Chebyshev polynomial expansion into the convolution kernel g_θ, written as

$$g_{\theta_l} \approx \sum_{k=0}^{K-1} \theta_{k_l} T_k(\tilde{\Lambda}) \tag{5}$$

with $\tilde{\Lambda} = 2\Lambda/\lambda_{\max} - I_N$, where $\lambda_{\max}$ means the maximum value of Λ. K and $T_k(\tilde{\Lambda})$ refer to the order of Chebyshev polynomial expansion and diagonal element function, separately.

Therefore, the updated output Y of the spectral graph convolution can be expressed as follows:

$$Y = \sum_{k=0}^{K-1} \theta_{k_l} U T_k(\tilde{\Lambda}) U^T x \tag{6}$$

Finally, the output data $x' \in \mathcal{R}^{N\times M}$ of one GCN layer is expressed as

$$x' = \text{Cheb}(x, W_l) = Y W_l \tag{7}$$

where $\text{Cheb}(\cdot)$ and $W_l \in \mathcal{R}^{S\times M}$ refer to the Chebyshev graph convolution and learnable parameter, respectively.

2.3. Graph transformer

The Graph Transformer is a DL-based framework that adapts the Transformer architecture to process non-Euclidean data. It integrates the advantages of GNNs in processing non-Euclidean data with the strength of the Transformer framework in capturing long-range dependencies, as shown in Fig. 3.

For a graph $G = \{V, A, E, F\}$ where V, A, E, F are the set of nodes, the adjacency matrix, the set of edges and the features, respectively, the self-attention mechanism in the Graph Transformer is expressed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V \tag{8}$$

where $Q \in \mathbb{R}^{n\times d}$, $K \in \mathbb{R}^{n\times d}$, and $V \in \mathbb{R}^{n\times d}$ refer to the query, key and value matrices, respectively. d_k indicates the dimension of the key matrix and M denotes the structural matrix, written as follows:

$$M_{ij} = f\left(d_{ij}\right) \tag{9}$$

where $f(\cdot)$ is the mapping function and $d_i j$ indicates the shortest connection between nodes i and j.

Similar to the classic Transformer, Graph Transformer can employ multi-head attention:

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(\text{head}_1, \text{head}_2, \ldots, \text{head}_h\right) W^O \tag{10}$$

with

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{11}$$

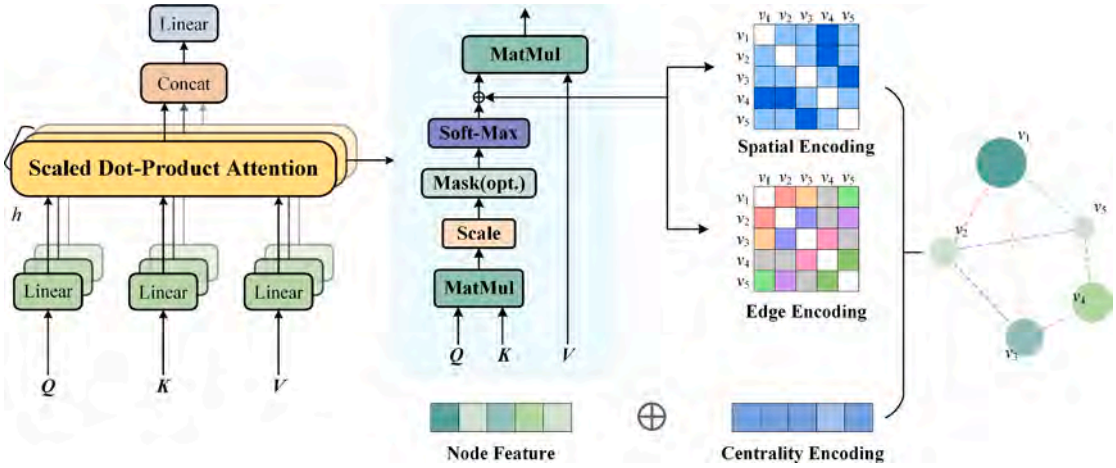

Fig. 3. Structure of graph transformer.

where h is the total number of attention heads. W^O is the output linear projection matrix. W_i^Q, W_i^K, and W_i^V are the learnable projection weight matrices for the i-th head, respectively.

Therefore, the node feature update process in the Graph Transformer can be formulated as follows:

$$h_i^{(l+1)} = \mathrm{LayerNorm}\left(h_i^{(l)} + \mathrm{MLP}\left(\mathrm{LayerNorm}\left(h_i^{(l)} + \sum_{j\in\mathcal{N}(i)\cup\{i\}} \alpha_{ij}^{(l)} W^V h_j^{(l)}\right)\right)\right) \tag{12}$$

where $h_i^{(l)}$ is the representation of node i in layer l, $\alpha_{ij}^{(l)}$ is the weight matrix, $\mathcal{N}(i)$ denotes the set of neighbors of node i, $\mathrm{LayerNorm}(\cdot)$ means layer normalization. MLP represents a Multilayer Perceptron.

3. The proposed method

As illustrated in Fig. 4, we propose a fusion-enabled multilevel graph-guided network (MSGFD) for optimal fault diagnosis using multi-channel data under extreme biased data. The MSGFD framework comprises five main components: data preprocessing, MultiGraph generation, feature inductive learning, global and local feature fusion (GLFF), and optimal fault diagnosis. We first consider the limitations of conventional time-frequency techniques and introduce an efficient data preprocessing method that generates RGB time-frequency images while enhancing computational efficiency. To effectively capture inter-sample relationships, a feature inductive learning module is designed to extract representative and informative features from a MultiGraph topology composed of four different graph types. Subsequently, the GLFF module is employed to enhance robustness and fault representativeness. Finally, a DDC loss is integrated to regularize the model training, further improving diagnostic performance under extreme biased data.

3.1. Data preprocessing

Transforming one-dimensional non-stationary signals into time-frequency images enables simultaneous by capturing variations in both time and frequency [35]. Common time-frequency transformation methods include the Short-Time Fourier Transform (STFT), Continuous Wavelet Transform (CWT), Wavelet Transform (WT) and Hilbert-Huang Transform (HHT). However, these methods have certain limitations. For example, selecting wavelet bases in the CWT relies too much on human expertise and experience, making it a labor-intensive and time-consuming process. The drawback of the STFT is that its time and frequency resolutions cannot be simultaneously optimized, leading to a trade-off between the two. To address these limitations, we propose a novel data preprocessing technique to convert multi-channel signals into time-frequency images, as depicted in Fig. 5.

Multi-channel signals are collected from the mechanical system using a data acquisition system at the same sampling frequency. The multi-channel signals are represented as:

$$x_{mn} = \left\{x_{ij}, x_{ij} \in R^{3\times m}\right\}, i = 1, 2, 3, j = 1, 2, \ldots, m \tag{13}$$

where x_{mn} denotes the generated samples; 3 and m stand for the number of channels and sampling points, respectively. Herein, triaxial accelerometers collect vibration data along the x, y, and z axes; therefore, the number of channels is inherently set to three.

Three-channel vibration signals under different health conditions are randomly segmented without overlap to generate samples with a size of $S_t = k \times k \times 3$. The samples $\left\{x_{1n}, \ldots, x_{mn}, \ldots, x_{3n}\right\}$ are separately normalized and converted into a pixel matrix. The detailed process of the pixel matrix is expressed as follows:

$$PM^i(a,b) = \frac{L^i((a-1)\times k + b) - \min\left(L^i\right)}{\max\left(L^i\right) - \min\left(L^i\right)} \times 255 \tag{14}$$

where $PM^i(a,b)$ denotes the pixel matrix of channel-axis signals, N represents the number of samples, $L(\cdot)$ represents the values of the samples.

After normalization, the three-channel vibration signals are converted into three RGB pixel matrices (ranging from 0 to 255) through the aforementioned process, and saved as RGB images through the $k \times k$ matrices. Hence, the RGB images are expressed as

$$RGB_{img}(a,b,c) = \left(PM^i(a,b), c\right) \tag{15}$$

where c denotes the channel index, corresponding to 1, 2, and 3.

3.2. Multigraph generation

Most existing GCN-based fault diagnosis methods consider only a single graph construction strategy, overlooking the fusion and analysis of multiple graph types. However, different graph construction methods may yield varying diagnostic performance depending on the dataset and model characteristics. Therefore, in this section, we propose a multigraph generation strategy that integrates information from multiple graph construction learning methods and leverages their respective strengths to enhance fault diagnosis performance under extreme biased data.

3.2.1. Multigraph

The process of multigraph generation is proposed to acquire discriminative and complementary fault-based information through four graph topologies from different graph construction methods (i.e., KNNGraph,

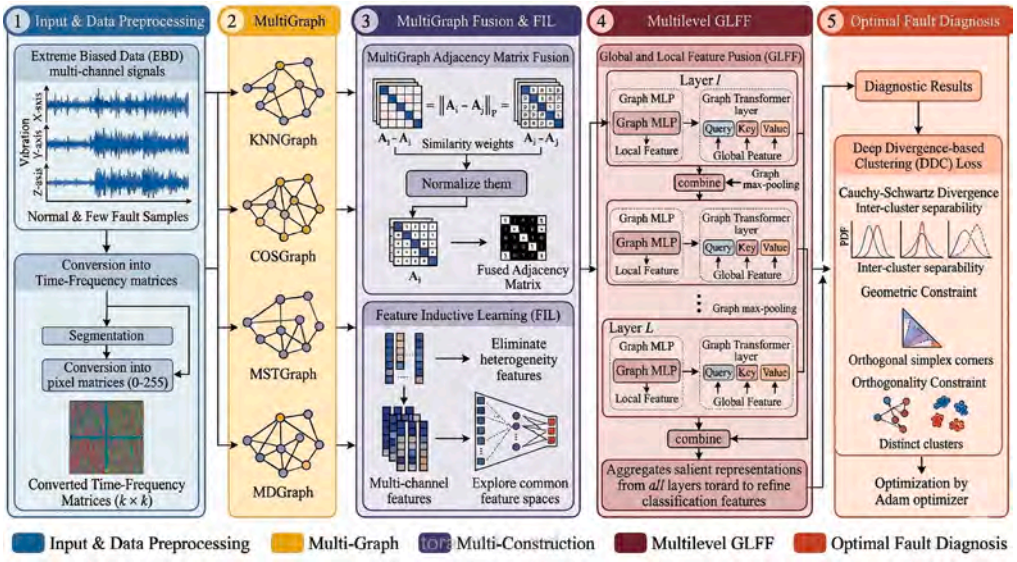

Fig. 4. Overall framework of the proposed method.

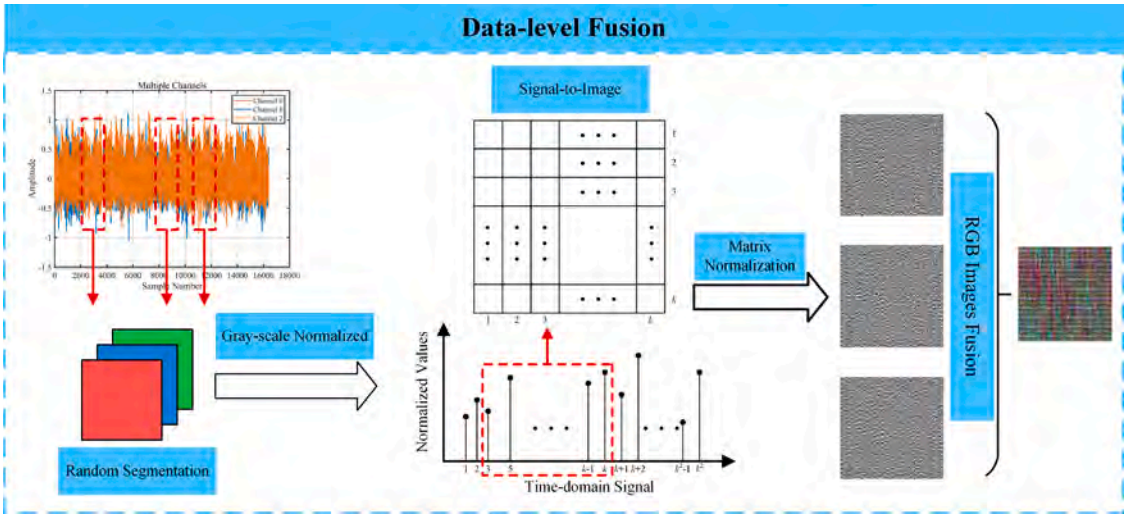

Fig. 5. Illustration of data preprocessing technique.

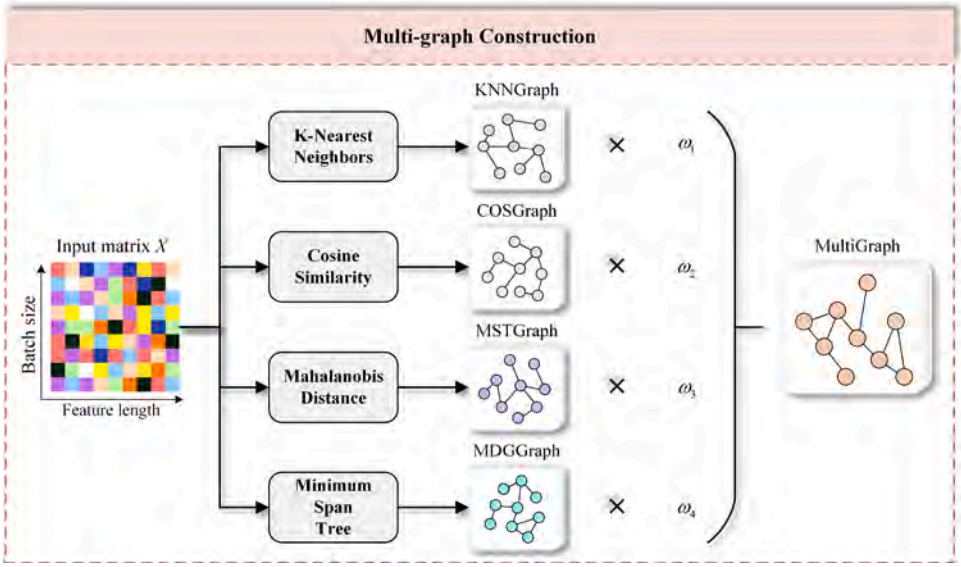

Fig. 6. Schematic diagram of multi-graph construction.

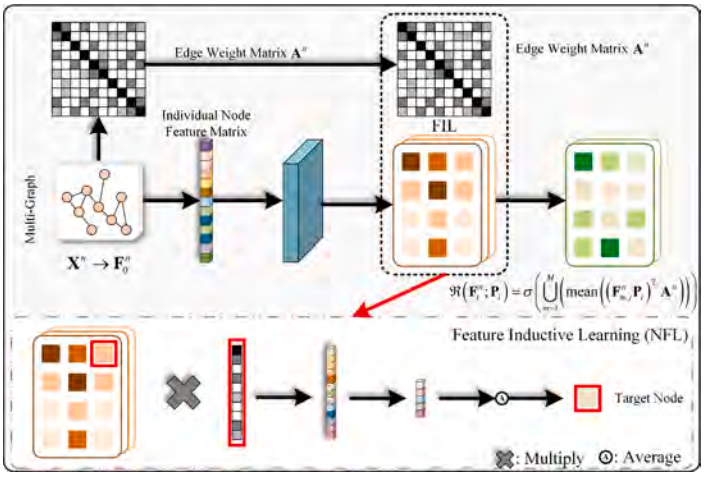

Fig. 7. Detailed process of FIL.

COSGraph, MSTGraph, and MDGraph), as shown in Fig. 6. The detailed construction process of these four graphs is described in Appendix.

First, the similarity weights of the four adjacency matrices of the different graphs can be denoted as

$$w_{ij} = \exp\left(-\left\|A_i - A_j\right\|_F\right) \tag{16}$$

where $\|\cdot\|_F$ represents the Frobenius norm. w_{ij} measures the similarity between adjacency matrices i and j.

After normalization, the weight can be further expressed as follows:

$$\hat{w}_i = \frac{\sum_j w_{ij}}{\sum_i \sum_j w_{ij}} \tag{17}$$

where i and j refer to the adjacency matrix i and adjacency matrix j, respectively.

Then, the fused adjacency matrix of multi-graph can be denoted as

$$A_{\text{fused}} = \sum_{i=1}^{4} \hat{w}_i A_i \tag{18}$$

where A_i means the four adjacency matrices.

Subsequently, the fused adjacency matrix is binarized to make it sparse and reduce the computational burden, as written below:

$$A_{\text{binary}} = \mathcal{I}\left(A_{\text{fused}} > \text{median}\left(A_{\text{fused}}\right)\right) \tag{19}$$

where $\mathcal{I}(\cdot)$ means an indicator function, median$(\cdot)$ denotes the median.

3.3. Feature inductive learning

In Huang et al. [36], Wang et al. [37] and other studies, a 'competition' phenomenon among different modal features during the gradient descent training process has been observed: the gradient updates within the shared network may suppress or dominate one another, potentially weakening or eliminating certain modal information during fusion. To better exploit the potential feature spaces, we propose the feature inductive learning (FIL) strategy, which is designed to eliminate the heterogeneity of fused RGB images from multi-channel signals, as illustrated in Fig. 7. The detailed process of feature inductive learning is illustrated in Fig. 8.

To enhance the interaction between nodes and mitigate channel competition during the global and local feature fusion learning, we introduce an adjacency matrix to explicitly model inter-node relationships, written as follows:

$$F_{i+1}^n = \mathcal{H}\left(F_i^n; \mathbf{P}_i\right) = \sigma\left(\bigcup_{m=1}^{M}\left(\text{mean}((F_{m,i}^n \mathbf{P}_i)^{\mathrm{T}} \mathbf{A}^n)\right)\right) \tag{20}$$

where $\mathbf{A}$ denotes the edge weight feature matrix. σ is the non-linear activation function. $\mathcal{H}(\cdot)$ denotes the feature induction operation, which projects the feature representation into a shared latent space using the learnable projection matrix $\mathbf{P}_i$. $\bigcup$ denotes the concatenation aggregation operation across M channels.

To explicitly address the feature competition, we introduce a shared latent space driven by the projection matrix $\mathbf{P}$. First, we initialize a random matrix $\mathbf{P}$ to represent this shared feature space. The primary role of $\mathbf{P}$ is to act as an alignment mechanism: it maps the heterogeneous, unaligned representations from different RGB channels into a unified dimensional space. By forcing the multi-channel features to project into this common subspace, $\mathbf{P}$ directly mitigates feature competition. Instead of individual channels updating their weights independently during backpropagation-which often leads to dominant channels suppressing weaker ones-the shared projection ensures that gradient updates are applied cooperatively.

Then, the matrix $\mathbf{P}$ is optimized over I iterations using the graph structure and multi-channel data. This shared optimization pathway enables the network to encode complementary information across different channels smoothly, preventing conflicting gradient signals. Next, the edge weight matrix and pooling operations are used to further align and improve the feature inductive learning. Finally, these aligned multi-channel features are fused to obtain a more robust and homogeneous representation. Therefore, the loss function of feature inductive learning is expressed as follows:

$$\mathcal{L}_{\text{FIL}} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}\left(y^n, \mathcal{F}\left(\sum_{i=1}^{I}\left(F_i^n \mathbf{P}_i\right)^{\mathrm{T}} \mathbf{A}^n\right)\right) \tag{21}$$

Finally, the optimization objective function of the FIL is summarized as follows:

$$\min_{\mathbf{P}} \frac{1}{n}\sum_{n=1}^{N}\mathcal{L}\left(y^n, \mathcal{F}\left(\sum_{i=1}^{I}\left(F_i^n \mathbf{P}_i\right)^{\mathrm{T}} \mathbf{A}^n\right)\right) \tag{22}$$

where n is the number of samples.

3.4. Global and local feature fusion learning

In this section, we propose a novel GLFF framework for feature extraction and fusion, leveraging the advantages of both Graph MLP and Graph Transformer. By establishing and integrating global and local features, the GLFF framework effectively captures comprehensive and representative information, as illustrated in Fig. 9.

After the representative feature learning, the feature representation is denoted as F_I^n. The detailed process of the GLFF framework is, then,

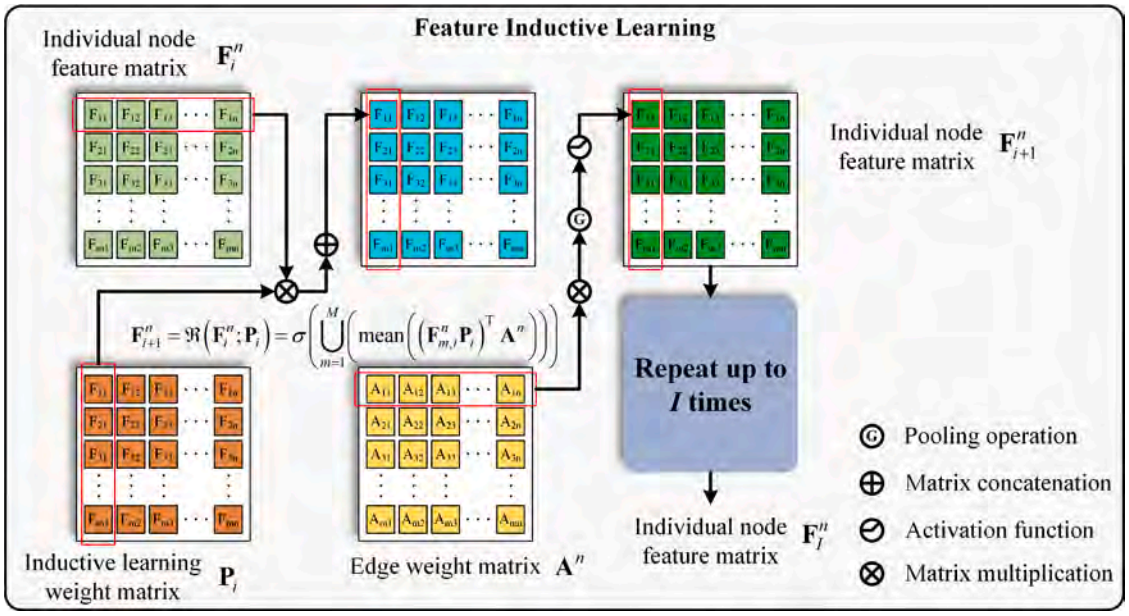

Fig. 8. Flowchart of feature inductive learning.

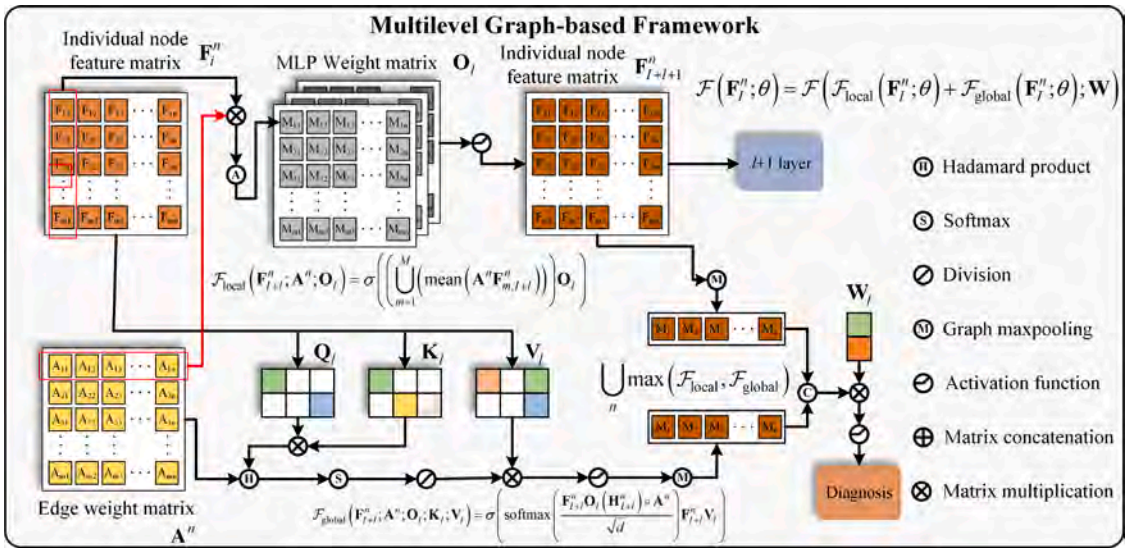

Fig. 9. Illustration of global and local feature fusion learning.

described as follows:

$$\mathcal{F}\left(\mathbf{F}_I^n;\theta\right) = \mathcal{F}\left(\mathcal{F}_{local}\left(\mathbf{F}_I^n;\boldsymbol{\theta}\right) + \mathcal{F}_{global}\left(\mathbf{F}_I^n;\boldsymbol{\theta}\right);\mathbf{W}\right) \tag{23}$$

where $F_{local}(\cdot)$, $F_{global}(\cdot)$ and $F(\cdot)$ refer to the Graph MLP, Graph Transformer and feature extraction, respectively.

As illustrated in Fig. 9, the Graph MLP and Graph Transformer focus on the local and global regions of the representative features in the first layer. The corresponding formulas are as follows:

$$\mathcal{F}\left(\mathbf{F}_{I+l}^n;\mathbf{A}^n;\mathbf{O}_l;\mathbf{Q}_l;\mathbf{K}_l;\mathbf{V}_l\right) = \mathcal{F}_{local}\left(\mathbf{F}_{I+l}^n;\mathbf{A}^n;\mathbf{O}_l\right) + \mathcal{F}_{global}\left(\mathbf{F}_{I+l}^n;\mathbf{A}^n;\mathbf{Q}_l;\mathbf{K}_l;\mathbf{V}_l\right) \tag{24}$$

$$\mathcal{F}_{local}\left(\mathbf{F}_{I+l}^n;\mathbf{A}^n;\mathbf{O}_l\right) = \sigma\left(\left(\bigcup_{m=1}^{M}\left(\text{mean}\left(\mathbf{A}^n\mathbf{F}_{m,I+l}^n\right)\right)\right)\mathbf{O}_l\right) \tag{25}$$

$$\mathcal{F}_{global}\left(\mathbf{F}_{I+l}^n;\mathbf{A}^n;\mathbf{Q}_l;\mathbf{K}_l;\mathbf{V}_l\right) = \sigma\left(\text{softmax}\left(\frac{\left(\mathbf{F}_{I+l}^n\mathbf{Q}_l\left(\mathbf{F}_{I+l}^n\mathbf{K}_l\right)^{\mathrm{T}}\right)\circ\mathbf{A}^n}{\sqrt{d}}\right)\mathbf{F}_{I+l}^n\mathbf{V}_l\right) \tag{26}$$

where **Q**, **K** and **V** refer to the Query, Key and Value vectors of the Graph Transformer, respectively, **O** denotes the learnable weight parameters of the Graph MLP, l represents the number of feature fusion learning layers. The detailed process of the proposed GLFF is illustrated in Fig. 10.

In each feature fusion learning layer, local features extracted by the Graph MLP are passed to the next layer, whereas global features extracted by the Graph Transformer are not. This design allows each layer to recompute global information independently, enabling different layers to capture global features at varying scales, thereby enhancing the representational capacity and adaptability to multi-scale information. Additionally, global features are obtained through the attention mechanism, and directly propagating them would increase computational cost. In contrast, local features are extracted via fully connected operations, which are computationally more efficient.

In each feature fusion learning layer, a graph max-pooling operation is applied after extracting the local and global features to obtain the salient representations of each layer. Then, $\mathcal{F}(\cdot)$ aggregates these salient representations from all feature fusion layers and optimizes the

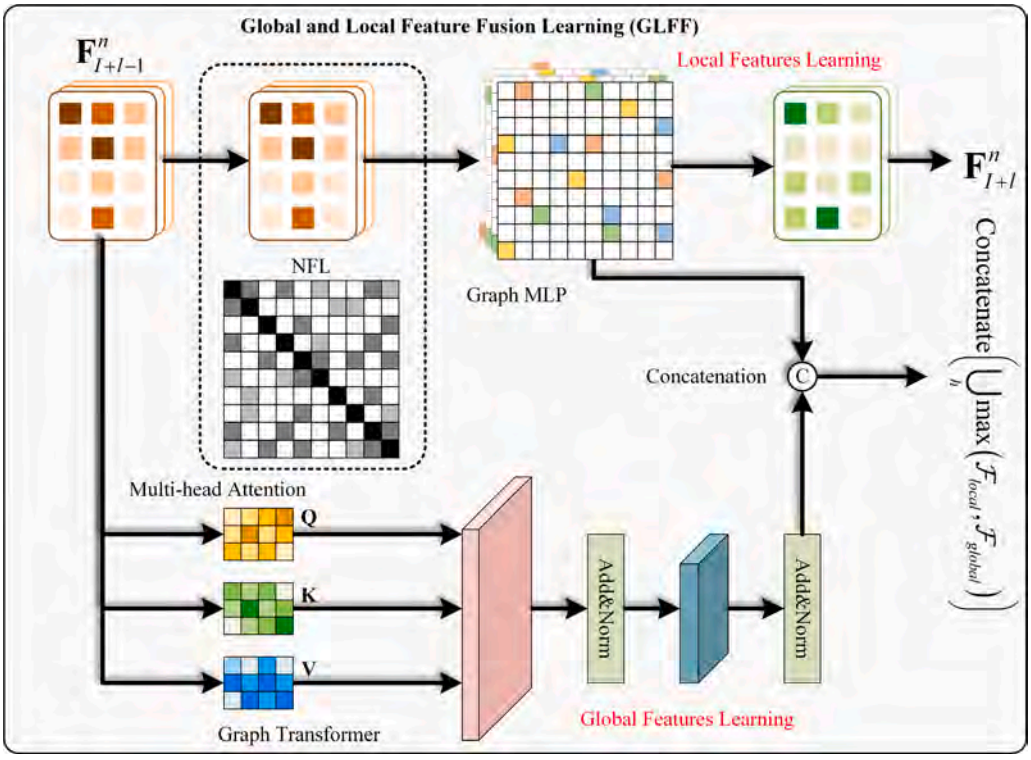

Fig. 10. Detailed process of the proposed GLFF.

weight matrix to refine the classification features, ultimately improving the fault diagnosis performance. Therefore, the classification function is defined as follows:

$$\mathcal{F}(\cdot)=\sum_{l=1}^{L}\sigma\left(\mathbf{W}_l\left(\text{concatenate}\left(\bigcup_h \max\left(F_{local},F_{global}\right)\right)\right)\right) \tag{27}$$

where L represents the number of feature fusion learning layers, $\bigcup_h \max(\cdot)$ denotes the graph maxpooling operation, $\mathbf{W}_l$ indicates the learnable weight parameters of l-th feature fusion learning layer.

The loss function of global and local feature fusion learning is written as follows:

$$\mathcal{L}_{\text{GLFF}}=\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(\hat{y},f(\mathbf{X}^n;\mathbf{P};\mathbf{O};\mathbf{Q};\mathbf{K};\mathbf{V};\mathbf{W})) \tag{28}$$

3.5. Optimal fault diagnosis

To achieve optimal fault diagnosis, a deep divergence-based clustering (DDC) loss is introduced to emphasize both inter-category separability and intra-category compactness, while also leveraging the geometric structure of the space. The detailed process of optimal fault diagnosis is illustrated in Fig. 12.

The DDC loss mainly consists of three parts, with the first part being the Cauchy-Schwartz divergence, which measures the similarity between probability distributions to enhance inter-cluster separability. It is written as follows:

$$D_{CS}=-\log\left(\frac{1}{k}\sum_{i=1}^{k-1}\sum_{j>i}\frac{\mathbb{E}_{\mathbf{h}\sim p_i}(p_j(\mathbf{h}))}{\sqrt{\mathbb{E}_{\mathbf{h}\sim p_i}(p_i(\mathbf{h}))\mathbb{E}_{\mathbf{h}\sim p_j}(p_j(\mathbf{h}))}}\right) \tag{29}$$

where c means the number of categories and p_i denotes the probability density function (PDF) of the i-th category, representing the distribution of data points assigned to that specific category in the feature space. Therefore, Eq. (29) is updated and integrated into a unified deep neural network, represented as:

$$D_{CS}=\frac{1}{k}\sum_{i=1}^{k-1}\sum_{j>i}\frac{\mu_i^T\mathbf{K}\mu_j}{\sqrt{\mu_i^T\mathbf{K}\mu_i\mu_j^T\mathbf{K}\mu_j}} \tag{30}$$

where $\mathbf{K}$ expresses the Gaussian kernel matrix; μ_i and μ_j are the column and row of the diagnostic results, respectively.

The second term of DDC loss is a geometric constraint, which is used to encourage the clustering assignments to align with the simplex corners, effectively spreading the clusters while maintaining their geometric structure, defined as:

$$\begin{cases} D_{GS}=\frac{1}{k}\sum_{i=1}^{k-1}\sum_{j>i}\frac{\gamma_i^T\mathbf{K}\gamma_j}{\sqrt{\gamma_i^T\mathbf{K}\gamma_i\gamma_j^T\mathbf{K}\gamma_j}} \\ B=\left[B_{ab}\right]=\exp\left(-\left\|\alpha_a-e_b\right\|^2\right) \end{cases} \tag{31}$$

where $[\cdot]$ is the column values of B. α_a is the soft cluster assignment and e_b represents the corners of the orthogonal simplex The visualization of exponential decay function on the simplex. is shown in Fig. 11.

The third term of DDC loss is to enforce orthogonality among categories, ensuring that different clusters remain distinct. A common way to achieve this is by minimizing the correlation between cluster assignments, which can be mathematically expressed as:

$$D_{EO}=triu(C^TC) \tag{32}$$

where $triu(\cdot)$ denotes the sum of elements of the triangles on C^TC. C is the cluster assignment probability matrix.

Therefore, based on the three terms mentioned above, the DDC loss function is written as follows:

$$\mathcal{L}_{DDC}=D_{CS}+D_{GS}+D_{EO} \tag{33}$$

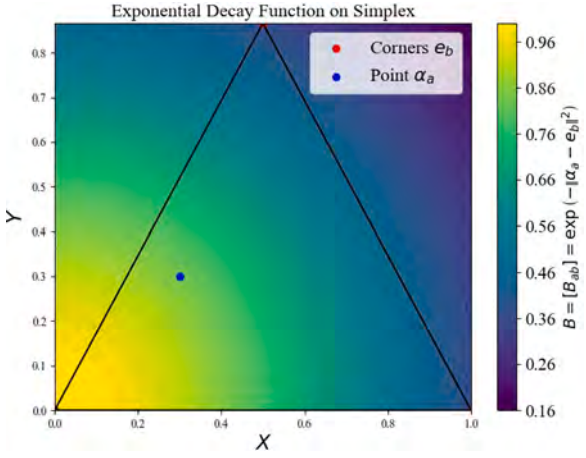

Fig. 11. Visualization of exponential decay function on simplex.

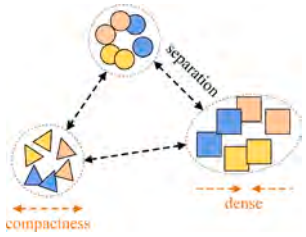

Fig. 12. Visualization of optimal fault diagnosis.

3.6. Overall framework

The overall loss function of our proposed MSGFD is written as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FIL}} + \mathcal{L}_{\text{GLFF}} + \alpha\mathcal{L}_{DDC} \tag{34}$$

where α means the trade-off parameter.

Therefore, the optimization of the overall loss function is performed using the Adam optimizer with back-propagation, and the detailed algorithm of the proposed framework is given in Algorithm 1.

Algorithm 1 The Proposed Algorithm.

Input: Dataset with multiple channels x_{mn}, the hyperparameter α.
Output: Diagnostic results y;

1: Initialize the neural network.
2: Compute the feature inductive learning loss by Eq. (21).
3: Compute the global and local feature fusion learning loss by Eq. (28).
4: Compute the DDC loss by Eq. (33).
5: Optimize the overall loss Eq. (34) by Adam optimizer. The specific stochastic gradients update by back-propagation are shown as follows (γ is the learning rate):
Network parameters update of the feature inductive learning loss: $\theta_f \leftarrow \theta_f - \gamma\nabla_{\theta_f}\frac{\partial L_{total}(\theta_f)}{\partial\theta_f}$
Network parameters update of the global and local feature fusion learning loss: $\theta_g \leftarrow \theta_g - \gamma\nabla_{\theta_g}\frac{\partial L_{total}(\theta_g)}{\partial\theta_g}$
Network parameters update of the DDC loss: $\theta_c \leftarrow \theta_c - \gamma\nabla_{\theta_c}\frac{\partial L_{total}(\theta_c)}{\partial\theta_c}$
6: Obtain the final diagnostic results y.

4. Case study

4.1. Dataset description

4.1.1. BJTU-RAO dataset

This section presents the subway train bogie fault diagnosis dataset (BJTU-RAO), provided by the State Key Laboratory of Advanced Rail Autonomous Operation at Beijing Jiaotong University, China [38]. The experimental platform consists of a gearbox, a motor and two axle boxes, as shown in Fig. 13. This experimental platform is capable of collecting multi-sensor vibration signals from the gearbox under various working conditions. The dataset classifies the health states of the gearbox into nine categories: normal (G_0), gear cracked tooth (G_1), gear worn tooth (G_2), gear missing tooth (G_3), gear chipped tooth (G_4), bearing inner race fault (G_5), bearing outer race fault (G_6), bearing rolling element fault (G_7) and bearing cage fault (G_8), as depicted in Fig. 14.

In this paper, including the normal state, a total of nine gearbox health states are used in the experiment, as described in Table 1. The sampling frequency of the multi-source sensor data is 64 kHz. Each sample in the dataset used for this experiment has a length of 1024, without overlap between samples and a total of 250 samples for each health state. In the BJTU-RAO bogie dataset, the different health states of the multi-sensor data correspond to the same operating conditions, with the motor speed set at 60 Hz and the lateral load set at 0 kN.

4.1.2. GearEccDataset

This multi-channel gear eccentricity data (GearEccDataset) is collected from the SMART Group, Faculty of Science and Technology, University of Macau, as shown in Fig. 15 [39]. The test rig mainly consists of a drive motor, a load motor, a torque sensor, two couplings, two load gearboxes and a test gearbox. Using this experimental platform, we tested 11 gears with varying levels of eccentricity, as described in Table 2. During the experiment, each gear operated under four different conditions, with the drive motor speeds set at 600 rpm, 900 rpm, 1200 rpm, and 1500 rpm, respectively. Eleven sensors were installed at different positions on the test bench to collect multi-channel signals, including acoustic, current and vibration signals, as shown in Fig. 16. Herein, sensors 1, 2, 3, 4, 5, 6, 7, and 8 operated at a sampling frequency of 51.2 kHz, whereas the remaining sensors operated at 12.8 kHz. However, only sensors 3, 4, 5, 9 and 10 are used to construct the samples. Multi-channel data from eleven gears with different health states under four working conditions were selected for analysis. For each working condition, each health state has 300 samples, with each sample consisting of 1024 data points ((Fig. 17)).

4.1.3. Motor operating condition (MOC) dataset

The multi-channel signals from the motor experimental platform provided by University of Huddersfield were used to validate the feasibility and superiority of the proposed method, which is shown in Fig. 18. The MOC test rig consists of a load display, a DC motor loader, a supply current, an encoder, an induction motor, a sensor box and vibration sensors. The loads of the gearbox are set to 0%, 40% and 80% hp at the same rotating speed. The detailed data acquisition process of MOC dataset is represented in Fig. 19. Two levels of damage severity (i.e., 0.2 mm and 0.5 mm) of inner race and outer race are used for testing. Together with normal state (N) and bearing ball fault (BF), there are six fault levels in total, as shown in Fig. 20. The detailed settings for the MOC dataset are described in Table 3. A triaxial acceleration sensor is placed on the motor for capturing the multi-channel signals with a sampling frequency of 10,240 Hz. To obtain more comprehensive fault information, each sample consists of 1024 data points. Specially, 300 samples are collected for each health state, resulting in a total of 1500 samples. Therefore, 80% of the samples from each health states are

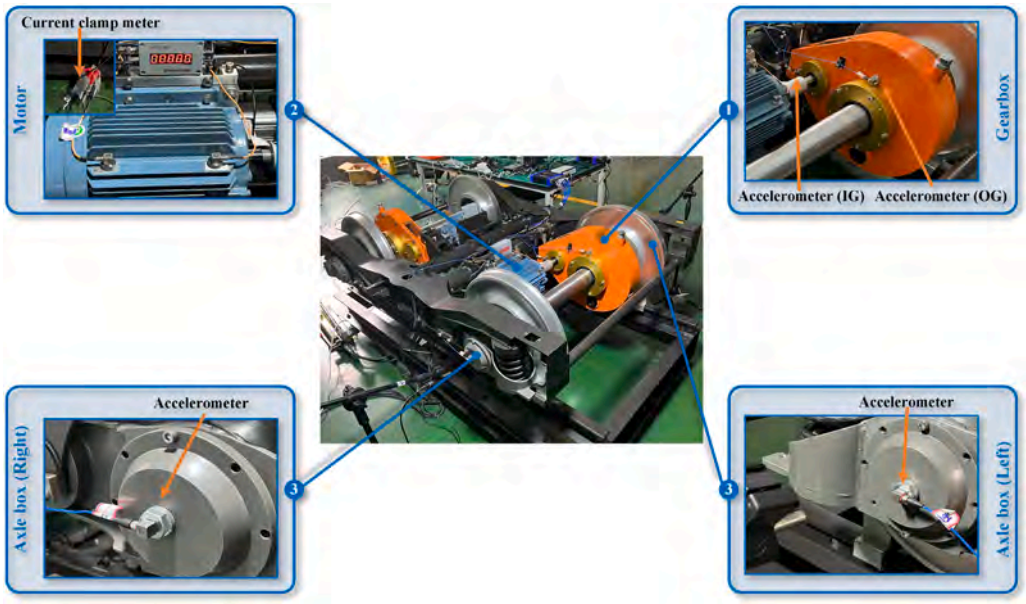

Fig. 13. The experimental platform of BJTU-RAO bogie dataset. 1. Gearbox; 2. Motor; 3. Right axle box; 4. Left axle box.

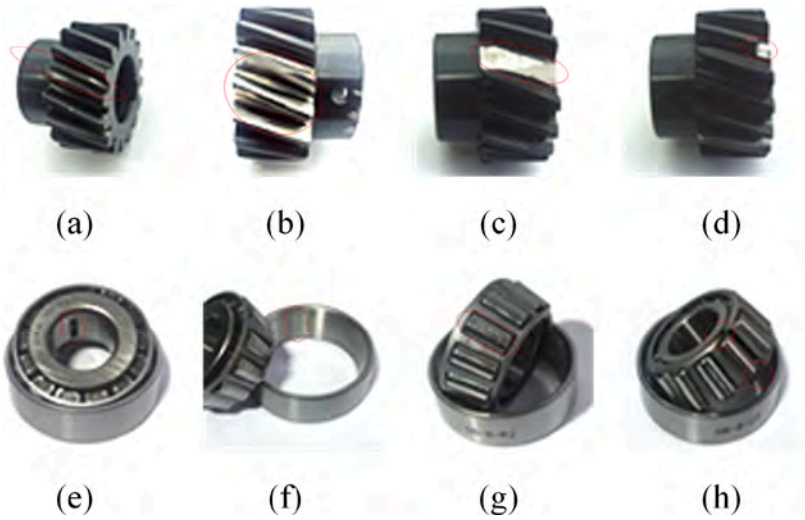

Fig. 14. Eight gearbox fault conditions. (a) gear cracked tooth; (b) gear worn tooth; (c) gear missing tooth; (d) gear chipped tooth; (e) bearing inner race fault; (f) bearing outer race fault; (g) bearing rolling element fault; (h) and bearing cage fault.

Table 1
Detailed settings for the BJTU-RAO dataset.

Condition	N	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8
Type	/	Gear	Gear	Gear	Gear	Bearing	Bearing	Bearing	Bearing
Position	/	/	/	/	/	Inner race	Outer race	rolling element	Cage
Label	0	1	2	3	4	5	6	7	8

Table 2
Detailed settings for the GearEccDataset.

Condition	Normal	Gear Eccentricity									
Degree (mm)	/	0.02	0.04	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2
Label	0	1	2	3	4	5	6	7	8	9	10

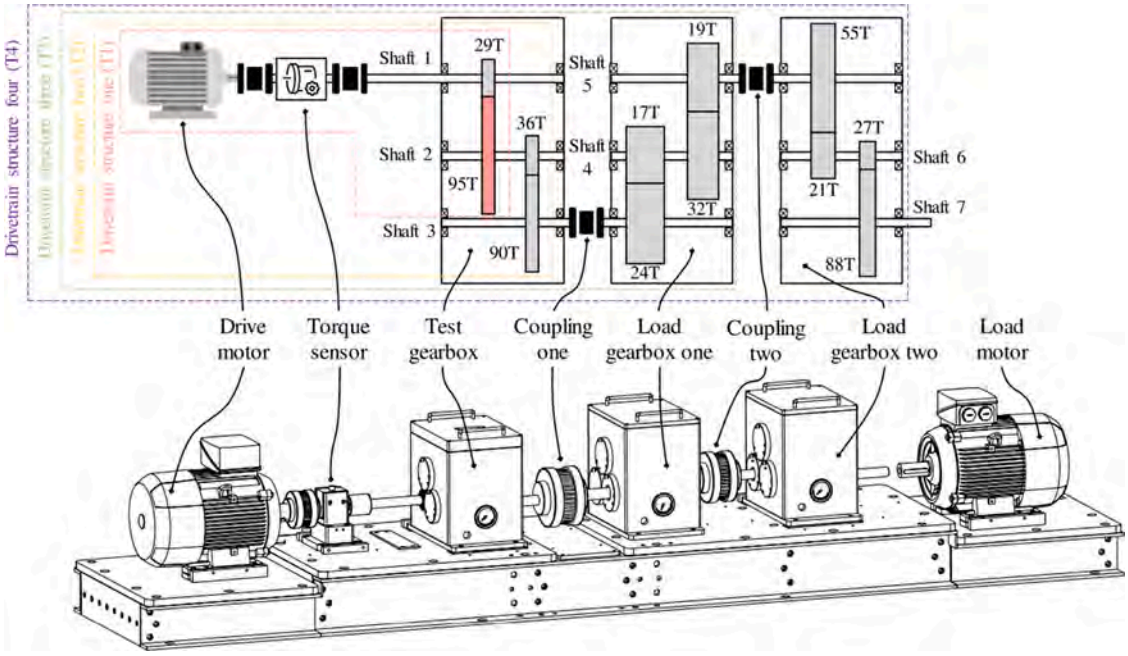

Fig. 15. Schematic diagram of the GearEccDataset platform.

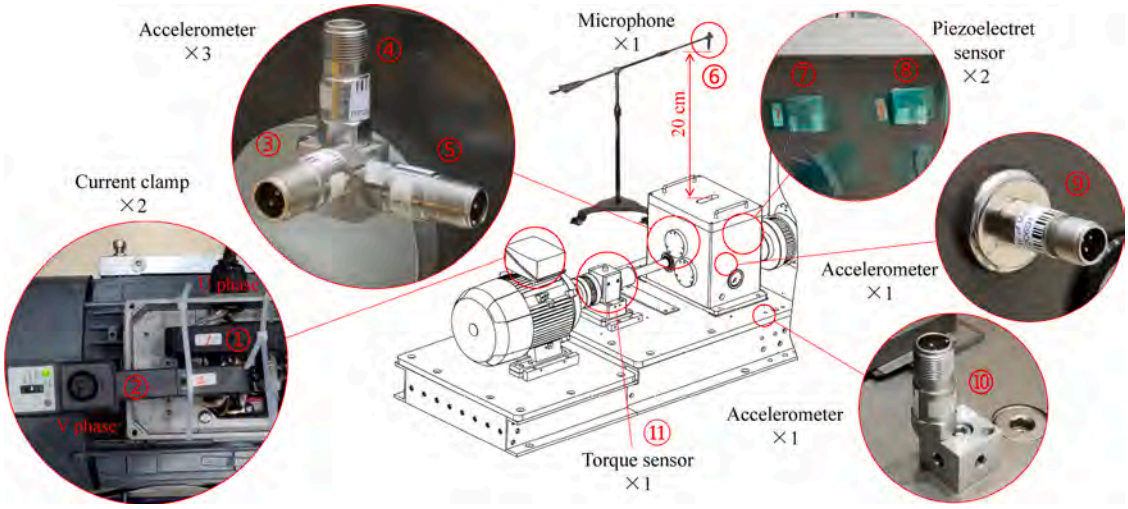

Fig. 16. The detailed sensor locations. 1. Current sensor; 2. Current sensor; 3. Accelerometer; 4. Accelerometer; 5. Accelerometer; 6. Sound sensor; 7. Voltage sensor; 8. Voltage sensor; 9. Accelerometer; 10. Accelerometer; 11. Torque sensor.

Table 3
Descriptions of fault levels on MOC dataset.

Condition	N0	O1	O2	I3	I4	BF5
Severity (mm)	/	0.2	0.5	0.2	0.5	/
Position	/	Outer race	Outer race	Inner race	Inner race	Bearing ball
Label	0	1	2	3	4	5

randomly selected as the training dataset, with the remaining 20% used as the testing dataset.

4.1.4. Rolling mill dataset

To verify the feasibility and superiority of the proposed method, an experimental rolling mill setup was constructed, as shown in Fig. 21. The platform consists of an induction motor, a coupling, a reduction gearbox, a direction-changing gearbox, a cross universal joint, a drive motor, four rolls, a handwheel, a vertical shaker, a horizontal shaker and a data acquisition system. Six accelerometers are mounted on the rolling mill to capture multi-channel signals from six measurement points, with a sampling rate of 10,240 Hz. Four health states of the tested bearing in the lower roller housing are simulated, including normal, inner race fault, outer race fault and element fault, as displayed in Fig. 23. The detailed data acquisition process of rolling mill dataset is illustrated in Fig. 22. The detailed settings for the XJGD dataset is described in

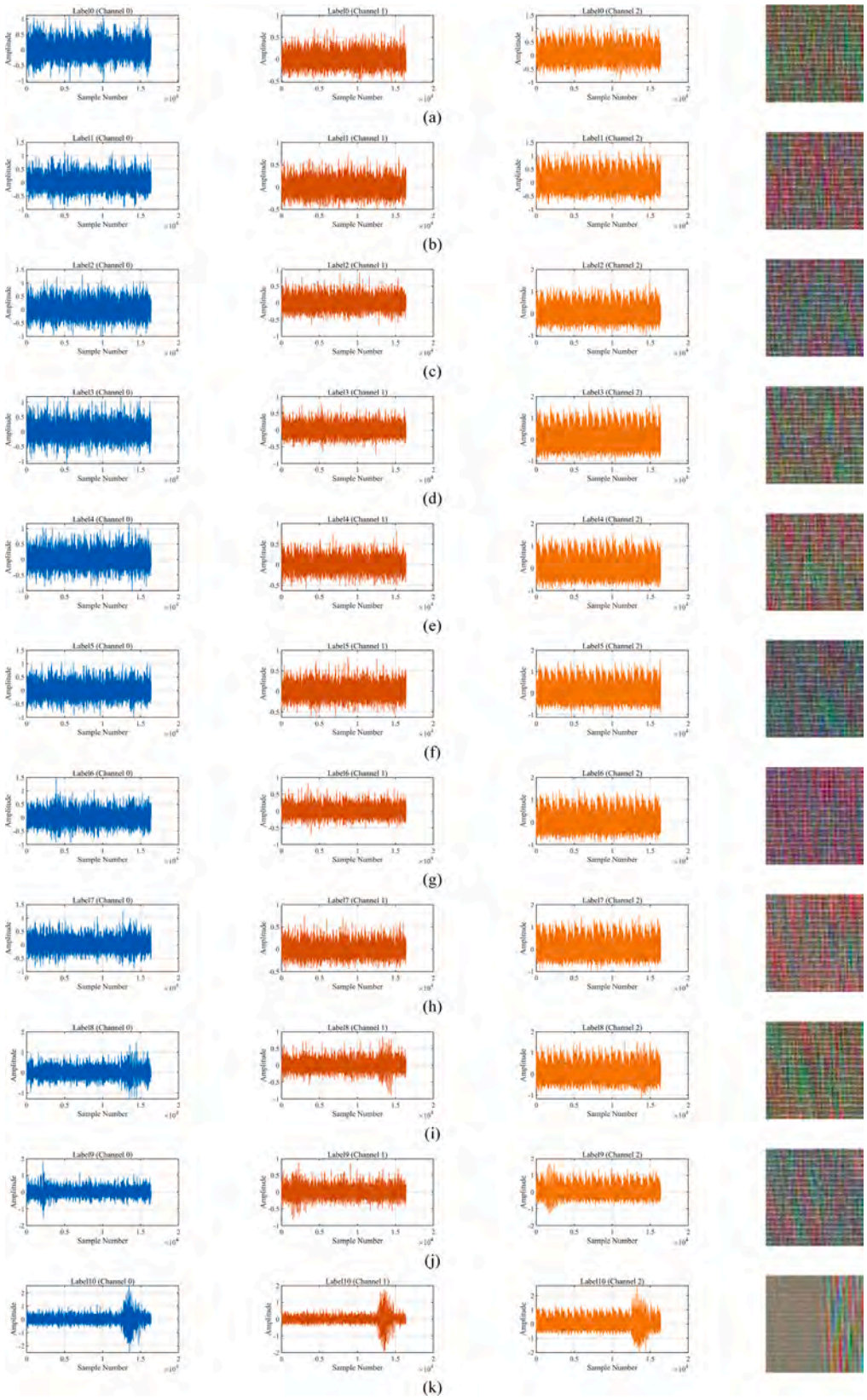

Fig. 17. Time-domain and RGB images on XJGD Dataset.

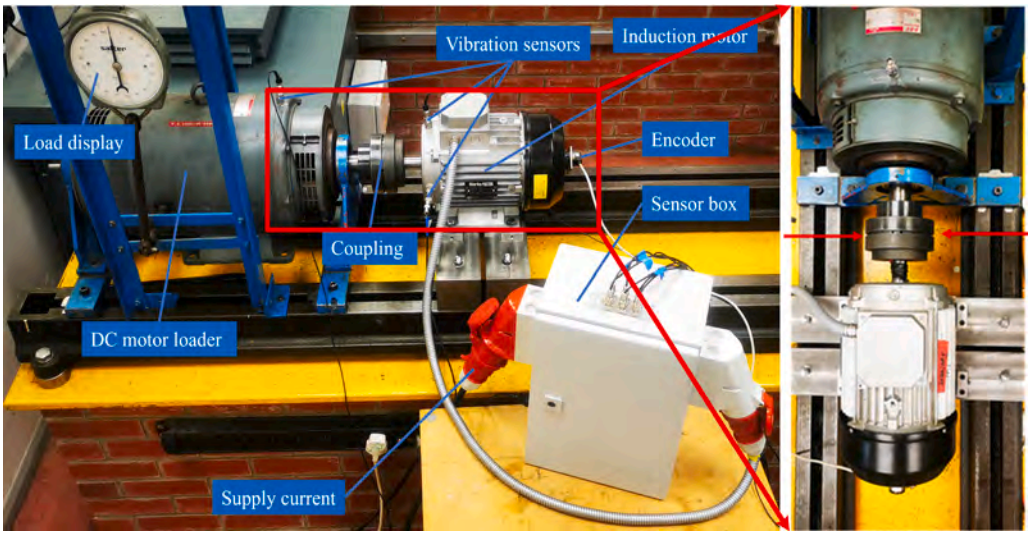

Fig. 18. The test rig of MOC dataset.

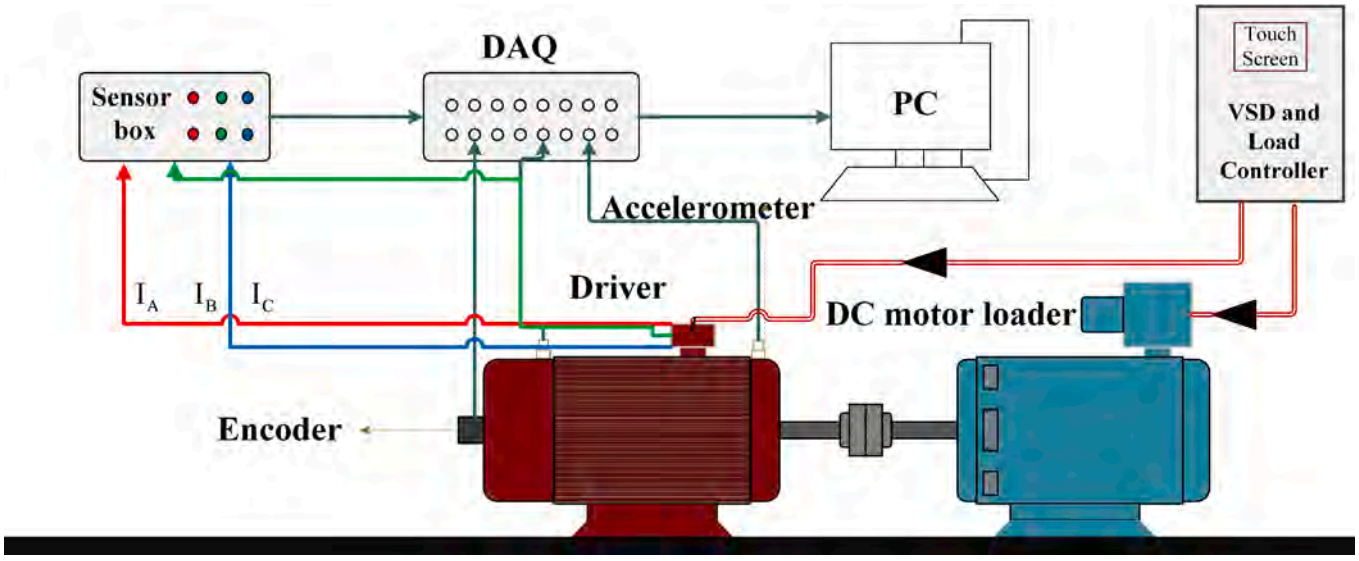

Fig. 19. Data acquisition system.

Table 4
Descriptions of health conditions on rolling mill dataset.

Condition	NC	IRF	ORF	REF
Position	/	Inner race	Outer race	Rolling element
Label	0	1	2	3

Table 4. Additionally, each experiment was conducted at three different rotating speeds (i.e., 10 Hz, 11.67 Hz, and 13.33 Hz) and a constant load (i.e., 0 hp). In each case, 800 samples were randomly selected for model training, whereas the remaining 400 samples were used for testing.

4.2. Methods for comparison

We evaluate the performance of the proposed MSGFD by comparing it with state-of-the-art (SOTA) fault diagnosis methods under extreme biased data. All methods in all experiments were repeated ten times to mitigate randomness and ensure fairness. The comparison methods are listed below:

1) Proposed MSGFD.
2) Improved CNN based on multi-sensor data (MSICNN) [40]. MSICNN integrates improved one-dimensional and two-dimensional CNNs for multi-sensor data. It leverages group normalization, global average pooling and dropout to enhance robustness and generalization.
3) Improved DBN based on multi-sensor data [41]. MSIDBN leverages improved Restricted Boltzmann Machines (RBMs) and Fast Fourier Transform (FFT) to enhance feature extraction under limited datasets.
4) Improved GCN based on multi-channel data (IMCGCN) [42]. IMCGCN builds a parallel graph processing framework to enhance graph quality and facilitate multi-channel feature fusion for fault diagnosis.
5) Multi-sensor information fusion Transformer (MSIFT) [33]. MSIFT integrates data-level preprocessing, feature extraction, and a multi-source transformer with a dual-stream diagnostic predictor to address limited and imbalanced challenges.
6) Two-stage importance-aware GCN based on multi-sensor data (I^2SGCN) [43]. I^2SGCN integrates multi-source sensors and a subgraph learning strategy to enhance cross-domain fault diagnosis by addressing graph quality, domain adaptation, and data limitations.

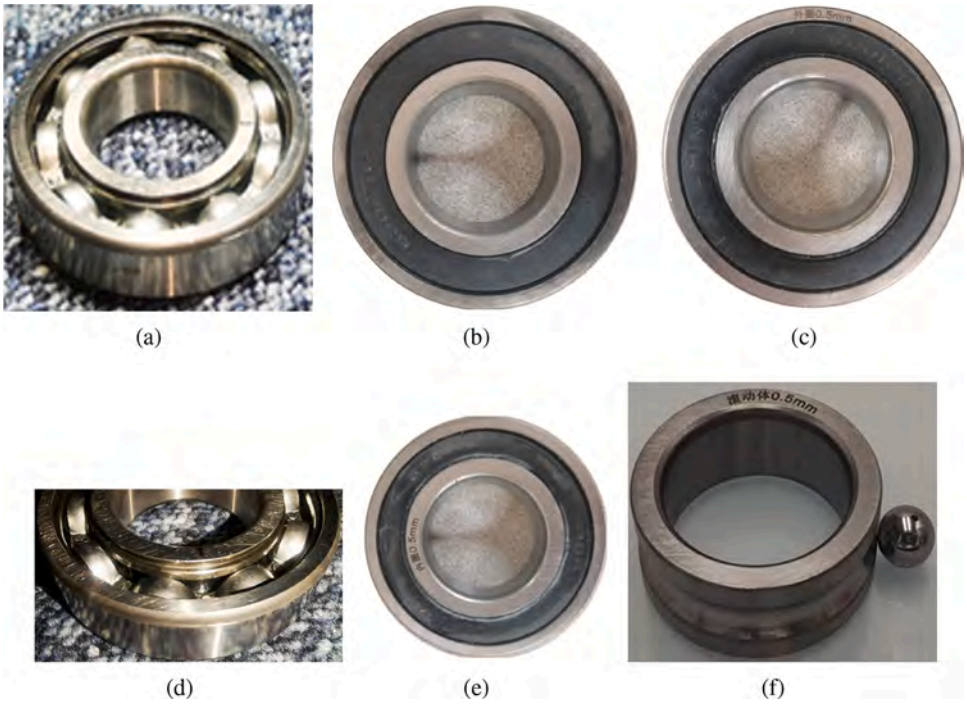

Fig. 20. Descriptions of fault levels on MOC dataset. (a) N0; (b) O1; (c) O2; (d) I3; (e) I4; (f) B5.

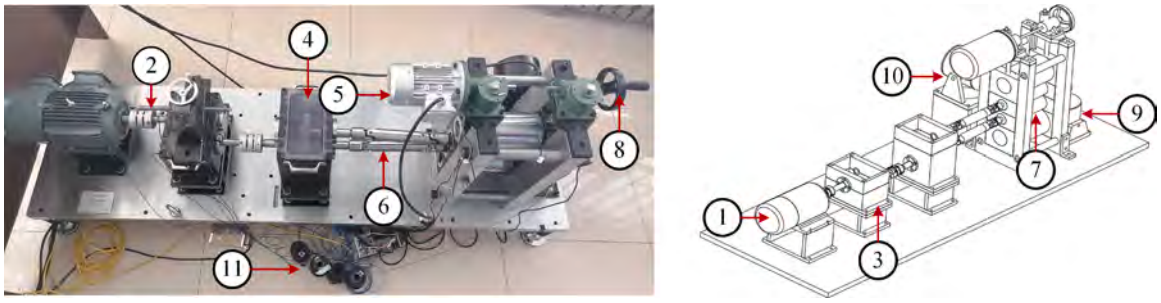

Fig. 21. Rolling mill experimental platform.

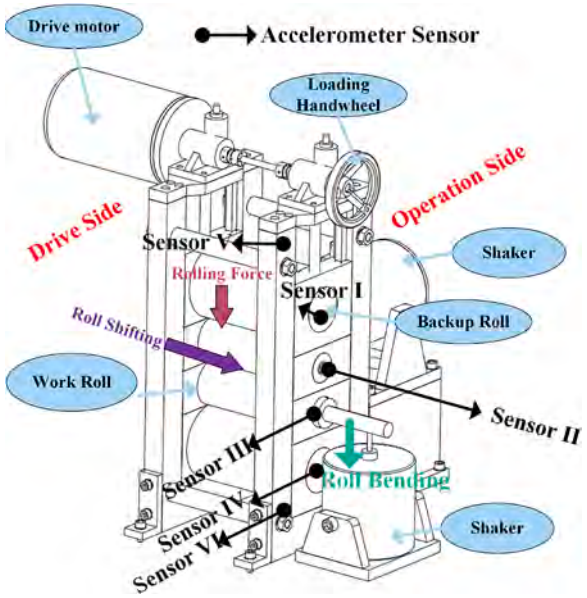

Fig. 22. Detailed data acquisition process.

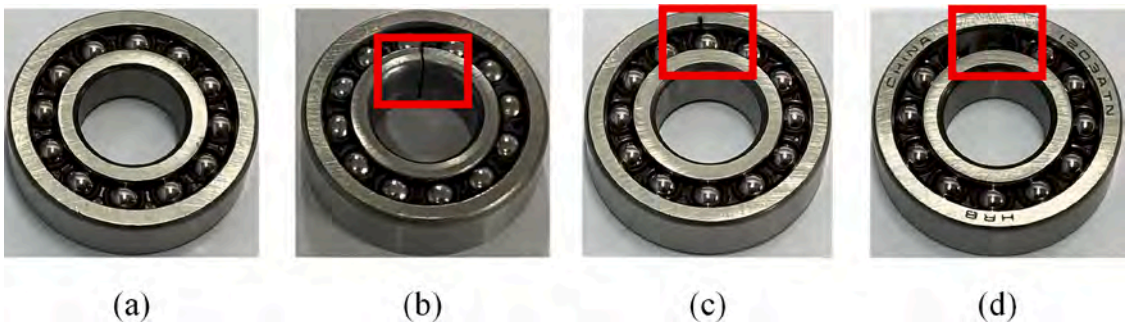

Fig. 23. Four different fault gears in the lower roller housing.

Table 5
Detailed data allocation strategy on the four case studies.

Tasks	EBD ratios	Training Datasets	Testing Samples
A_1	0.1	0 × 300, (1, 2, 3, 4, 5, 6, 7, 8) × 30	(0, 1, 2, 3, 4, 5, 6, 7, 8) × 100
A_2	0.2	0 × 300, (1, 2, 3, 4, 5, 6, 7, 8) × 60	(0, 1, 2, 3, 4, 5, 6, 7, 8) × 100
A_3	0.3	0 × 300, (1, 2, 3, 4, 5, 6, 7, 8) × 90	(0, 1, 2, 3, 4, 5, 6, 7, 8) × 100
B_1	0.1	0 × 300, (1, 2, 3, 4, 5) × 30	(0, 1, 2, 3, 4, 5) × 100
B_2	0.2	0 × 300, (1, 2, 3, 4, 5) × 60	(0, 1, 2, 3, 4, 5) × 100
B_3	0.3	0 × 300, (1, 2, 3, 4, 5) × 90	(0, 1, 2, 3, 4, 5) × 100
C_1	0.1	0 × 300, (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) × 30	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10) × 100
C_2	0.2	0 × 300, (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) × 60	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10) × 100
C_3	0.3	0 × 300, (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) × 90	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10) × 100
D_1	0.1	0 × 300, (1, 2, 3) × 30	(0, 1, 2, 3) × 100
D_2	0.2	0 × 300, (1, 2, 3) × 60	(0, 1, 2, 3) × 100
D_3	0.3	0 × 300, (1, 2, 3) × 90	(0, 1, 2, 3) × 100

* EBD represents the extreme biased data.
* ∗ 0 represents the normal condition, whereas other numbers correspond to fault conditions.
* ∗∗ The values 30, 60, 90, 100, and 300 refer to the number of samples.

Table 6
Experimental results of different methods on the four case studies.

Tasks	Accuracy (%)					
A_1	82.42±1.21	83.85±1.23	82.91±1.19	91.05±0.92	90.21±0.99	**91.45±0.48**
A_2	85.49±1.19	85.31±1.20	85.13±1.16	95.27±0.58	95.26±0.56	**96.61±0.49**
A_3	88.58±0.97	88.35±0.89	90.21±0.74	98.35±0.56	98.29±0.45	**98.68±0.47**
B_1	82.12±1.01	84.42±0.98	83.76±0.75	91.33±0.67	90.64±0.53	**91.78±0.46**
B_2	86.35±0.96	86.12±0.89	85.98±0.84	95.45±0.42	95.38±0.45	**96.85±0.38**
B_3	89.22±0.88	88.95±0.82	90.84±0.69	98.12±0.48	98.05±0.39	**98.32±0.42**
C_1	80.14±1.40	80.07±1.42	80.68±1.44	90.12±0.81	89.25±0.91	**90.83±0.83**
C_2	83.23±1.28	82.31±1.35	84.24±1.21	95.18±0.61	94.62±0.78	**95.34±0.62**
C_3	89.17±0.89	89.21±0.87	90.31±0.91	96.27±0.45	96.32±0.44	**97.18±0.42**
D_1	82.63±1.29	83.52±1.22	81.68±1.37	90.67±0.94	91.38±0.87	**92.85±0.71**
D_2	84.89±1.11	85.21±1.08	84.27±1.18	**96.84±0.53**	96.28±0.54	96.83±0.50
D_3	87.34±1.02	87.95±1.03	89.19±0.95	**97.86±0.48**	96.78±0.46	98.11±0.44
Average	85.13	85.44	85.77	94.71	94.37	**95.40**
Methods	MSICNN	MSIDBN	IMCGCN	MSIFT	I^2SGCN	MSGFD

* The bold fonts denote the best performance.
* ∗ The underlined fonts represent the second best performance.
* ∗∗ ± means the standard deviation.

4.3. Implementation details

All methods in these experiments were implemented in PyTorch and conducted on an NVIDIA RTX 3070 GPU with an Intel i5-10400F CPU. The networks are optimized using the minibatch stochastic gradient descent (SGD) algorithm for 300 epochs, with a training batch size of 8 and a testing batch size of 8. The weight decay, momentum, and learning rate are set to 0.0005, 0.9 and 0.001, respectively. The detailed dataset allocation strategy based on the four case studies is given in Table 5.

4.4. Diagnosis accuracy

As shown in Table 6, our proposed MSGFD achieves the best performance across twelve tasks under extreme biased data, with an average accuracy of 95.4%. Compared with multi-channel data-based methods (MSICNN and MSIDBN), those based on simple feature concatenation perform relatively worse, highlighting the limitations of traditional feature fusion approaches. By appropriately incorporating graph topology into multi-channel data-based methods, the improved models (IMCGCN and I^2SGCN) demonstrate a clear advantage in capturing inter-sample relationships, achieving average accuracies of 85.77% and 94.37%, respectively. Meanwhile, effective feature exchange and fusion are introduced into MSIFT to enhance information flow, resulting in an average accuracy of 94.71%. Following these comparisons, the proposed MSGFD extracts representative and valuable features using MultiGraph topologies through the FIL module within a global and local feature fusion framework to achieve optimal fault diagnosis, which results in the superior performance across four case studies under extreme biased data.

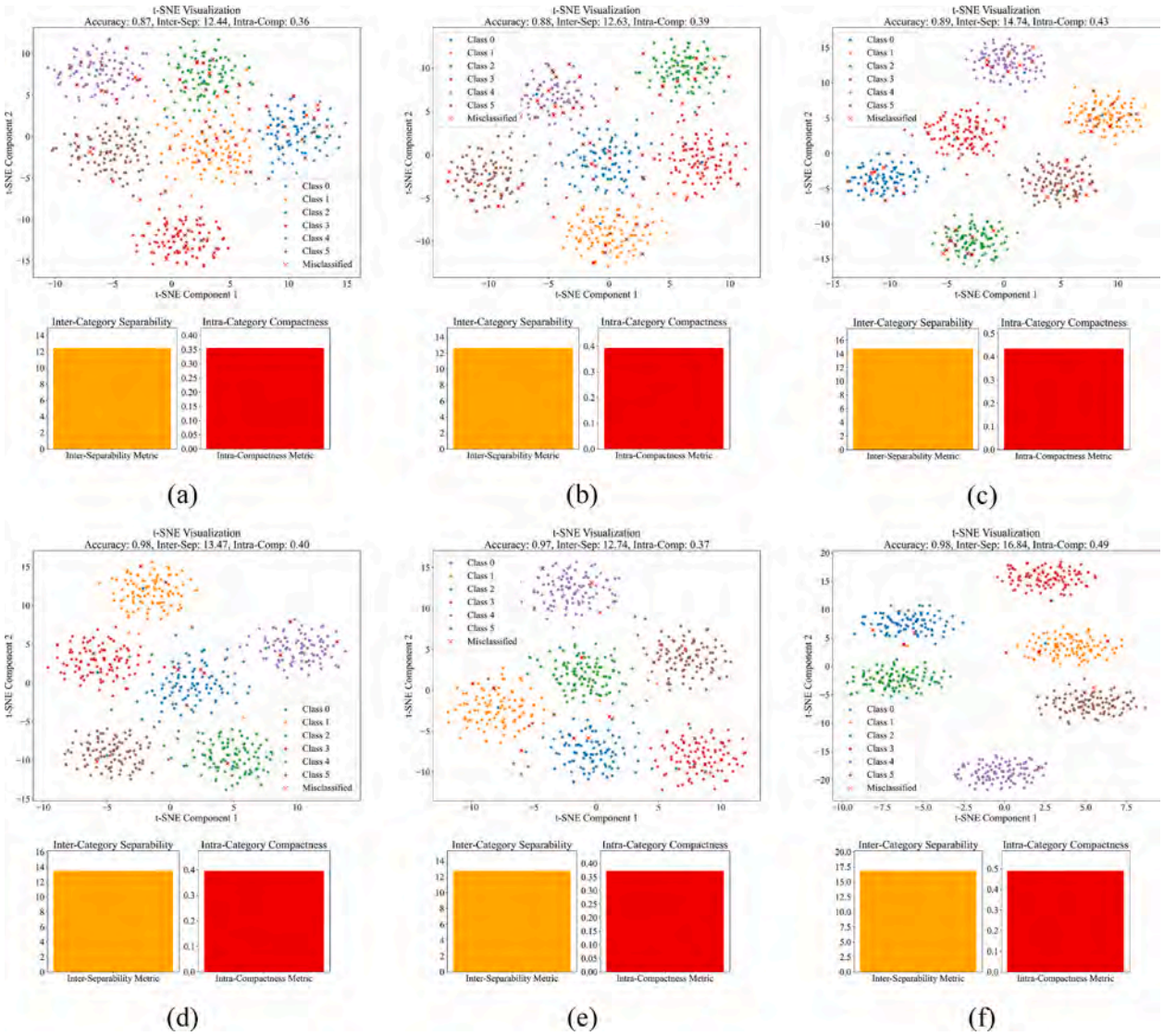

Fig. 24. t-SNE visualization comparing feature distributions across methods on D_3 dataset. (a) MSICNN; (b) MSIDBN; (c) IMCGCN; (d) MSIFT; (e) I^{2}SGCN ; (f) MSGFD.

4.5. Feature visualization

To demonstrate the effectiveness of optimal fault diagnosis, specifically the impact of DDC loss, the t-distributed stochastic neighbor embedding (t-SNE) technique on D_3 is employed, as illustrated in Fig. 24. Meanwhile, the inter-separability and intra-compactness metrics are introduced to evaluate how well different categories are separated from one another and how tightly the data points within each category are clustered, respectively. Therefore, the inter-separability and intra-compactness metrics can be obtained by the following equations:

$$\text{Inter-Separability} = \frac{\sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K} d\left(c_i, c_j\right)}{K(K-1)} \tag{35}$$

$$\text{Intra-Compactness} = \frac{1}{\frac{1}{N}\sum_{i=1}^{K}\sum_{x\in C_i} d\left(x, c_i\right) + \varepsilon} \tag{36}$$

where $d\left(c_i, c_j\right)$ represents the Euclidean distance between centers of category i and category j, K means the number of categories, $d\left(x, c_i\right)$ expresses the Euclidean distance from point x to its category center, C_i is the set of all points in class i, N denotes the total number of points across all categories, ε is a small constant to prevent division by zero.

It can be seen that features from different categories in the proposed MSGFD are well divided into six parts. The visualization results demonstrate that the proposed method has learned the discriminative and representative features through the DDC loss, which are important for realizing fault diagnosis under extreme biased data.

4.6. Classification performance

To evaluate the classification performance of all methods, we visualize the results for different fault categories using the confusion matrix on D_3, as shown in Fig. 25.

As can be seen from the Table, the proposed MSGFD accurately diagnoses most samples across various health conditions. The accuracy for Label 0 reached 99%, with only one sample misclassified as Label 1. Most importantly, the accuracies of all health conditions on D_3 reach more than 90%. In contrast, the worst-performing method (MSICNN) achieved only 83% accuracy for Label 0, which falls short of meeting real-world engineering requirements.

Fig. 25. Confusion matrices on D_3. (a) MSICNN; (b) MSIDBN; (c) IMCGCN; (d) MSIFT; (e) I^{2}SGCN ; (f) MSGFD.

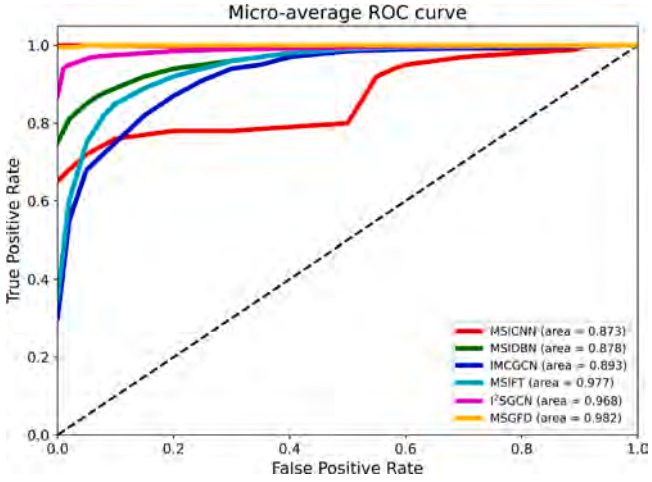

Fig. 26. ROC analysis of the proposed MSGFD.

Furthermore, we further investigate the diagnosis performance on D_3 through the Receiver Operating Characteristic Curve (ROC), as illustrated in Fig. 26. The horizontal axis and the vertical axis represent the False Positive Rate (FPR) and the True Positive Rate (TPR), respectively.

"The micro-average ROC areas for MSICNN, MSIDBN, IMCGCN, MSIFT, I^{2}SGCN, and MSGFD are 0.873, 0.878, 0.893, 0.977, 0.968, and 0.982, respectively, indicating that the proposed MSGFD method achieves the best diagnostic classification performance.

5. Model discussions

5.1. Analysis of data preprocessing

To evaluate the computational efficiency and diagnostic performance of the data preprocessing used to generate RGB images, several widely-used time-frequency methods (i.e., CWT, S-transformer, and STFT) were considered at A_3. In this comparison, only single-channel signals were transformed into time-frequency images. From Fig. 27, we observe that the processing time of our proposed method, CWT, S-transformer and STFT are 18.5s, 151.7s, 167.3s, and 162.8s, respectively; our proposed method has the shortest runtime, whereas the S-transformer exhibits the longest. Meanwhile, the diagnostic results demonstrate that the proposed data preprocessing technique improves diagnostic accuracy by approximately 1% compared with conventional time-frequency representations under the identical network configurations, thereby confirming its superior feature expressiveness and robustness.

5.2. Loss and accuracy curves

We state the loss and accuracy curves of different methods on A_3 in Fig. 28.

It is seeing that the loss and accuracy curves generated by the proposed MSGFD demonstrate superior performance, as evidenced by

Table 7
Evaluation of each proposed component on rolling mill dataset.

Method	Components						ACC (%)	F1 (%)	AUC (%)
lr	DP	FIL	MG	LF	GF	OFD			
M1		✓	✓	✓	✓	✓	94.27	93.65	95.08
M2	✓		✓	✓	✓	✓	97.52	96.83	98.13
M3	✓	✓		✓	✓	✓	95.79	94.94	96.26
M4	✓	✓	✓		✓	✓	95.13	94.27	96.04
M5	✓	✓	✓	✓		✓	96.38	95.52	96.85
M6	✓	✓	✓	✓	✓		97.02	96.45	97.71
M7		✓	✓				85.12	84.38	86.65
M8			✓	✓	✓		89.45	88.71	90.52
M9		✓		✓	✓		88.67	87.94	89.81
M10				✓	✓	✓	87.92	87.15	88.94
Ours	✓	✓	✓	✓	✓	✓	98.11	97.85	98.54

* DP, MG, LF, GF, and OFD represent the data preprocessing, MultiGraph, local features, global features, and optimal fault diagnosis, respectively.
* $*\pm$ means the standard deviation.

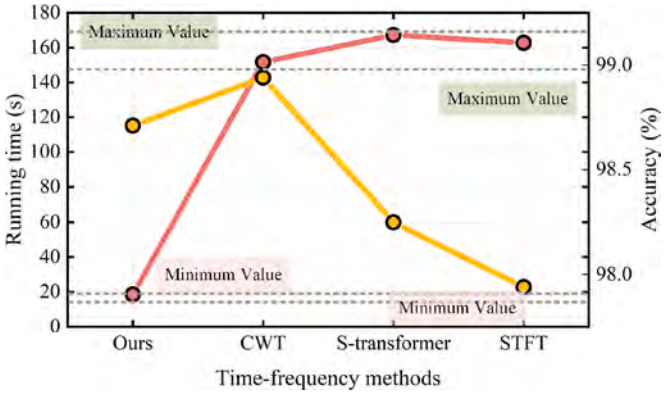

Fig. 27. Running time of different time-frequency methods.

the lowest minimum loss value and highest maximum accuracy value. Meanwhile, the proposed method effectively suppresses oscillating loss during the training process, which can result in a smoother loss curve and faster convergence speed. This indicates that the proposed method not only effectively alleviates loss fluctuations but also improves the convergence speed of network training.

5.3. Ablation study

An ablation study has been performed on six key components of the proposed method on rolling mill dataset. The quantitative results of the ablation study on D_3 is presented in Table 7. Six variants were derived from our proposed MSGFD, named as follows: 1) **M1**: MSGFD without the proposed data preprocessing; 2) **M2**: MSGFD without the FIL module; 3) **M3**: MSGFD without MultiGraph; 4) **M4**: MSGFD without local features; 5) **M5**: MSGFD without global features; 6) **M6**: MSGFD without optimal feature diagnosis; 7) **M7:** MSGFD only with FIL + MG; 8) **M8:** MSGFD only with MG + LF + GF; **M9:** MSGFD only with FIL + LF + GF; **M10:** MSGFD only with LF + GF + OFD.

Compared with **M1**, the proposed method incorporating data preprocessing achieves improvements of 3.84%, 4.2%, and 3.46% in ACC, F1 and AUC, respectively. These improvements can be attributed to the RGB images constructed from multi-channel data, as defined in Eq. 15. This performance advantage demonstrates that the proposed data preprocessing effectively addresses the limitations of extreme biased data by enriching time-frequency characteristics to enhance diagnostic accuracy. The proposed method leverages feature inductive learning to extract representative and valuable features, resulting in improved diagnostic performance with increases of 0.59%, 1.02% and 0.41% in ACC, F1 and AUC over **M2**, respectively. By exploiting complementary instance embeddings in MultiGraph topologies, the proposed method achieves an accuracy of 98.11%, an F1 score of 97.85%, and an AUC of 98.54%, demonstrating improvements of 2.32%, 2.91% and 2.28% in ACC, F1 and AUC over **M3**, respectively. Compared with **M4** and **M5**, the integration of local and global features enhances diagnostic performance, resulting in improvements of 2.98% and 1.73% in ACC, respectively. This confirms that the global and local feature fusion strategy enhances robustness and captures more representative fault features, effectively mitigating the challenges under extreme biased data. In particular, the incorporation of DDC loss enables more optimal fault diagnosis during the training of the MSGFD framework, resulting in improvements of 1.09% in accuracy, 1.40% in F1-score and 0.83% in AUC compared to **M6**. These ablation experiments show that the tailor-made modules of the proposed MSGFD can comprehensively facilitate the multilevel framework for fault diagnosis under extreme biased data. To further investigate the collaborative and redundant relationships among the components, we introduced variants **M7**-**M10**, each retaining only specific combinations of modules in the proposed MSGFD. The comparison results presented in Table 7 show a significant diagnostic performance degradation when relying on these partial configurations, with accuracy dropping to 85.12%-89.45%, compared with the single-component variants (i.e., **M1**-**M6**). Specifically, **M7**, which retains only FIL and MG, achieves the lowest diagnostic accuracy of 85.12%, indicating that structural learning without comprehensive feature extraction (i.e., LF and GF) is insufficient. In contrast, variants **M8**, **M9**, and **M10** rely heavily on feature extraction modules (i.e., LF and GF) but lack essential supporting components, such as DP or MG, and therefore fail to exceed 90% diagnostic performance. These findings demonstrate that the proposed modules exhibit minimal redundancy and instead maintain strong collaborative relationships. Each component contributes sequentially, forming a cohesive and indispensable pipeline for effective fault diagnosis under extremely biased data conditions.

5.4. Additional statistical metrics

In this section, four kinds of statistical metrics, including recall, precision, F1-score and area under receiver operating characteristic curve (AUC) are introduced to evaluate all algorithms. The detailed quantitative results on task C_3 are summarized in Table 8.

It can be seen that the mean accuracy, precision, recall and F1-score of the proposed NIFD-Net are 98.35%, 98.72%, 98.53% and 98.62%, respectively, which are much higher than SOTA methods. This is because the proposed MSGFD effectively extracts representative and informative features using MultiGraph topologies through the FIL module within a global and local feature fusion framework, enabling optimal fault diagnosis and achieving superior performance.

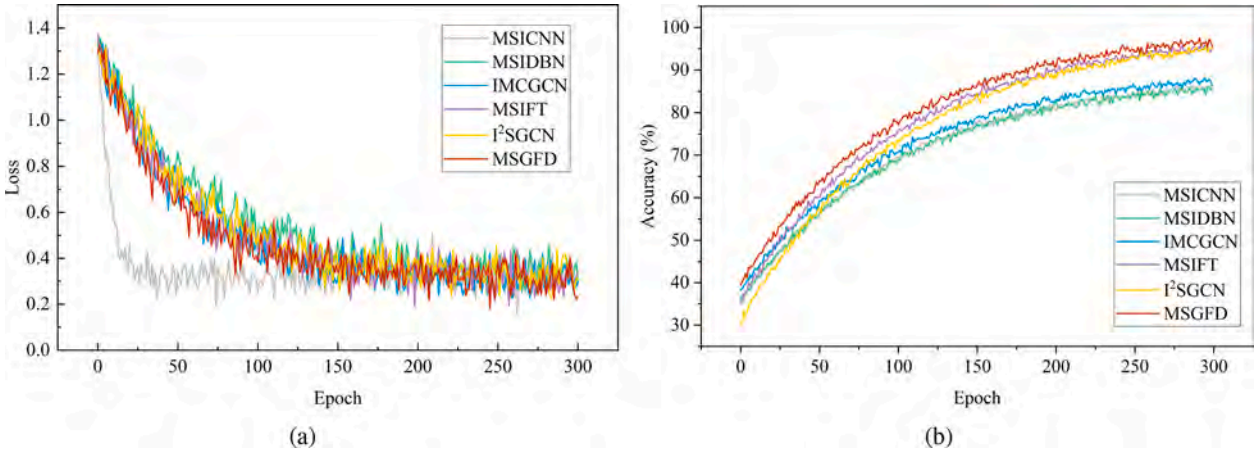

Fig. 28. Visualization of loss and accuracy for different methods. (a) Loss curve; (b) Accuracy curve.

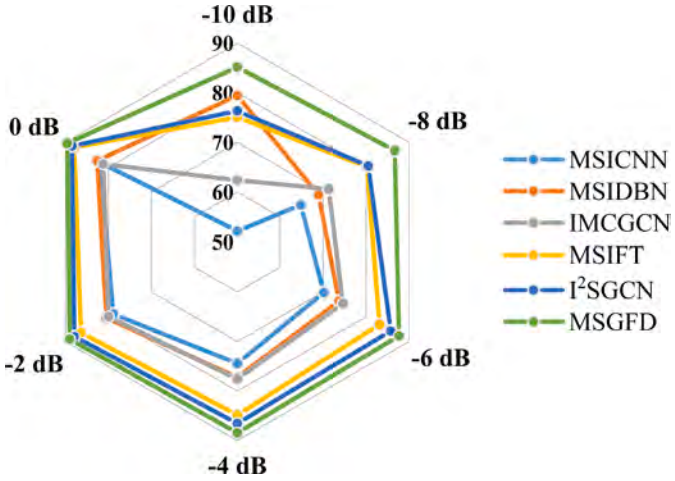

Fig. 29. Diagnostic results on D_1 under different SNRs.

Table 8
Experimental results of C_3 using various metrics.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MSICNN	88.58%	89.42%	88.95%	89.18%
MSIDBN	88.35%	89.10%	88.72%	88.91%
IMCGCN	90.21%	90.85%	90.47%	90.66%
MSIFT	98.35%	98.72%	98.53%	98.62%
I²SGCN	98.29%	98.65%	98.42%	98.53%

Table 9
Analysis of DDC loss.

Methods	D_{CS}	D_{GS}	D_{EO}	ACC(%)
Full DDC	✓	✓	✓	98.11
w/o D_{CS}		✓	✓	97.49
w/o D_{GS}	✓		✓	97.82
w/o D_{EO}	✓	✓		97.36

5.5. Analysis of noise resistance

To investigate the impact of noise, we employ varying levels of Gaussian white noise into multi-channel data to evaluate the robustness of the proposed MSGFD. The signal-to-noise ratio (SNR) is defined as follows:

$$\mathrm{SNR} = 10\log_{10}(P_o/P_n) \tag{37}$$

Table 10
Average ranks for each diagnostic task across different methods.

	MSICNN	MSIDBN	IMCGCN	MSIFT	I²SGCN	MSGFD
Average Rank	5.17	5	4.83	2.08	2.83	1.08

where P_o and P_n represent the powers of the multi-channel signal and noise, respectively. Therefore, we added varying levels of noise, ranging from -10 dB to 2 dB, to the multi-channel data, with the experimental results presented in Fig. 29.

As observed, the diagnostic accuracies of all methods decrease consistently as the SNR decreases. The proposed MSGFD demonstrates the strongest noise resistance across different SNR levels through effective data preprocessing and maintains satisfactory diagnostic performance for all tasks.

5.6. Analysis of DDC loss

To further investigate the effectiveness of each component (i.e., D_{CS}, D_{GS}, and D_{EO}) of the proposed DDC loss on the rolling mill dataset, we conduct an ablation analysis by selectively removing individual terms from the loss function, as summarized in Table 9.

The experimental results presented in Table 9 demonstrate that removing any individual component leads to a degradation in diagnostic performance, indicating that each loss term contributes positively to the overall diagnostic capability under extremely biased data conditions. Specifically, the D_{CS} loss term primarily promotes inter-class separability, while the D_{GS} loss term enhances feature representation by encouraging group-level sparsity. Meanwhile, the D_{EO} loss term plays an important role in improving the confidence of diagnostic clustering. Furthermore, we visualize the category distributions using t-SNE projections. As illustrated in Fig. 30, the full DDC loss produces more compact intra-category structures and clearer inter-category boundaries compared with its variants. These observations indicate that the three loss components work to improve fault diagnostic performance under extreme biased data.

5.7. Statistical tests

In this section, Friedman and Holm post-hoc tests are conducted to assess the superiority and effectiveness of the proposed MSGFD from a statistical perspective. The statistical tests based on twelve diagnostic tasks (from A_1 to D_3) are conducted to quantitatively analyze the diagnostic differences among all methods, and the corresponding ranking results are listed in Table 10.

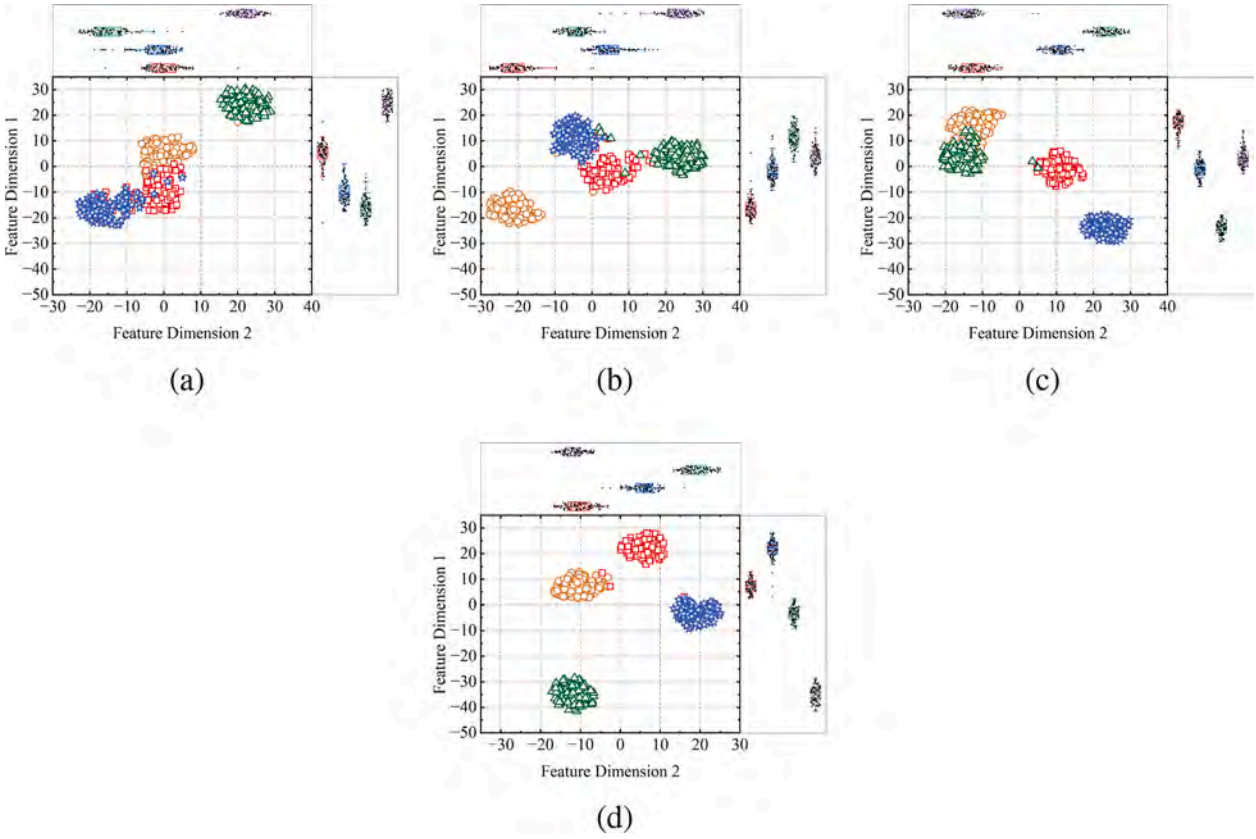

Fig. 30. Category distributions using t-SNE projections. (a) w/o D_{CS}; (b) w/o D_{GS}; (c) w/o D_{EO}; (d) Full DDC.

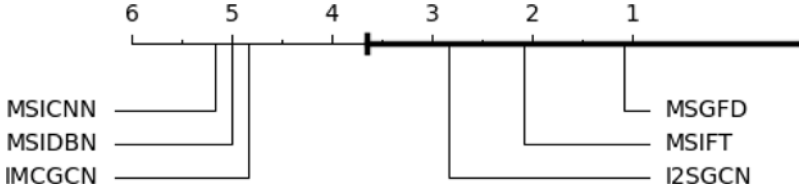

Fig. 31. CD diagram of average rank on the ranking axis.

Table 11
Computational cost comparison of different methods.

Methods	Params (M)	FLOPs (G)	Model Size (MB, FP32)	Relative Latency
MSICNN	~1.8	~0.06	~7.2	Low
MSIDBN	~2.5	~0.04	~10.0	Low
IMCGCN	~4.2	~0.09	~16.8	Medium
MSIFT	~9.6	~0.18	~38.4	Medium–High
I^2SGCN	~8.1	~0.15	~32.4	Medium–High
MSGFD	**17.13**	**0.274**	**68.5**	**High**

Then, we calculated the Friedman statistic, χ_F^2, based on these ranks.

$$\chi_F^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} \left(R_j - \frac{n(k+1)}{2} \right)^2 \tag{38}$$

where k means the number of methods, n represents the number of all tasks in this work, R_j indicates the average ranks for each method.

Finally, we compared the calculated $\chi_F^2 = 60.7836$ value to the chi-square distribution table with degrees of freedom $k - 1 = 6 - 1 = 5$ to determine that there was a statistically significant difference among all methods. A Critical Difference (CD) diagram, often used with the F_t, is employed to represent multiple pairwise comparisons among the various models visually. After the Friedman test showed significance, the Nemenyi post-hoc test was conducted to calculate the Critical Difference (CD), as depicted in Fig. 31.

It can be observed that the methods MSIFT, I^2SGCN, and MSGFD are connected because their rank difference is less than the CD of 2.56, which can indicate no significant performance difference. In contrast, the methods MSICNN, MSIDBN, and IMCGCN are not connected, which can show a statistically significant difference in performance.

5.8. Analysis of computational cost

To evaluate the diagnostic performance of the proposed MSGFD in industrial fault-diagnosis scenarios, particularly for edge devices and online monitoring systems [44], a computational complexity analysis is conducted in this section.

Table 12
Summary of key mathematical notations.

Symbol	Description
x_i^m, y_{conv}^{m+1}	Input feature map and output convolutional feature map at layer m
$w_{i,j}^m$, $b_{i,j}^m$	Weight and bias of the convolutional layer
$G = \{V, A, E, F\}$	Graph structure with nodes V, adjacency matrix A, edges E, and features F
L, I_N, D	Symmetric normalized Laplacian, Identity matrix, and Degree matrix
U, Λ	Eigenvectors and eigenvalues of the Laplacian matrix
$T_k(\tilde{\Lambda})$	Chebyshev polynomial expansion of order k
Q, K, V	Query, Key, and Value matrices in the Graph Transformer
M_{ij}, d_{ij}	Structural matrix and the shortest path connection between nodes i and j
$h_i^{(l)}$	Hidden representation of node i in layer l
$PM^i(a, b)$	Normalized pixel matrix for the i-th channel
A_{fused}, A_{binary}	Fused multi-graph adjacency matrix and its binarized version
$\mathbf{P}_i$	Learnable projection matrix for shared latent space mapping
$\mathbf{O}_l, \mathbf{W}_l$	Learnable weight parameters in Graph MLP and GLFF
$\mathcal{L}_{\text{FIL}}, \mathcal{L}_{\text{GLFF}}, \mathcal{L}_{\text{DDC}}$	Feature Inductive Learning loss, Global-Local loss, and DDC loss
D_{CS}, D_{GS}, D_{EO}	Cauchy-Schwartz divergence, Geometric constraint, and Orthogonality constraint

The results in Table 11 show that the proposed MSGFD architecture contains approximately 17.13 million trainable parameters, corresponding to a model size of approximately 68.5 MB (FP32). Although the proposed MSGFD introduces higher computational complexity compared with other comparison methods, the rapid advancement of industrial computing hardware can significantly improve the deployment efficiency. Therefore, designing more lightweight architectures while maintaining high diagnostic performance will be an important direction for future work.

6. Conclusions

In this work, we propose MSGFD, a multilevel graph-guided framework for fault diagnosis using multi-channel data under extreme data imbalance. The framework unifies data preprocessing, MultiGraph construction, feature inductive learning, global-local feature fusion, and divergence-based optimization within a single architecture. An efficient preprocessing strategy is developed to transform multi-channel signals into structured representations while maintaining computational efficiency. The proposed MultiGraph mechanism captures complementary inter-instance relationships through four distinct graph topologies, thereby enhancing structural robustness. Feature inductive learning further explores shared and discriminative feature spaces, whereas the global-local feature fusion module jointly models short-range and long-range dependencies to alleviate the challenges posed by limited supervision. Finally, a deep divergence-based clustering (DDC) loss is incorporated to regularize the learning process by promoting inter-category separability and intra-category compactness. Extensive experiments conducted on four case studies under various imbalance settings demonstrate the effectiveness, robustness, and superiority of MSGFD, consistently outperforming SOTA methods. In future work, we aim to extend MSGFD to imbalanced open-set domain generalization scenarios by leveraging meta-learning paradigm [45]. Building upon this direction, we will investigate meta-learning-based multi-graph construction strategies to improve cross-machine generalization under few-shot settings [44,46]. Additionally, we plan to incorporate noise-robust and uncertainty-aware learning techniques to better address measurement noise and unknown disturbances. We will also explore online and streaming extensions to facilitate real-time deployment in industrial monitoring environments.

CRediT authorship contribution statement

Yue Yu: Conceptualization, Data curation, Software, Formal analysis, Investigation, Methodology, Resources, Validation, Writing - original draft; **Hamid Reza Karimi:** Conceptualization, Writing - review & editing, Visualization, Supervision, Project administration; **Pradeep Kundu:** Investigation, Methodology, Validation, Writing - review & editing; **Enrico Zio:** Investigation, Methodology, Validation, Writing - review & editing; **Ke Feng:** Investigation, Methodology, Validation, Writing - review & editing.

Data Availability

The authors do not have permission to share private data. The code and model are available at: https://github.com/Polimi-YuYue.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

KNNGraph

Given the feature matrix $X \in \mathbb{R}^{n\times d}$, where n is the number of samples and d is the feature dimension, the KNNGraph is constructed as follows:

1) For each sample X_i(the i-th row of X), find its k nearest neighbors based on a distance metric (e.g., Euclidean distance).

2) Construct a binary adjacency matrix $A \in \mathbb{R}^{n\times n}$ where:

$$A_{ij} = \begin{cases} 1, & \text{if } X_j \text{ is one of the } k \text{ nearest neighbors of } X_i \\ 0, & \text{otherwise} \end{cases} \tag{39}$$

3) The diagonal elements A_{ii} are set to 0 ($A_{ii} = 0$) to exclude self-connections.

Then, to ensure the adjacency matrix represents an undirected graph, it is symmetrized as follows:

$$A = \max\left(A, A^T\right) \tag{40}$$

where A^T means the transpose of A. It is noteworthy that the $\max(\cdot)$ operation ensures that if either A_{ij} or A_{ji} is 1, the resulting matrix A has $A_{ij} = A_{ji} = 1$.

Finally, the final adjacency matrix A satisfies: $A_{ij} \in \{0, 1\}$ (binary connections); $A_{ii} = 0$ (no self-connections); $A = A^T$ (symmetric, representing an undirected graph).

COSGraph

Given the feature matrix $X \in \mathbb{R}^{n\times d}$, where n denotes the number of samples and d is the feature dimension, the cosine similarity matrix $S \in \mathbb{R}^{n\times n}$ is defined as:

$$S_{i,j} = \frac{X_i \cdot X_j}{\|X_i\|\|X_j\|} \tag{41}$$

where X_i and X_j refer to the i-th and j-th rows of the feature matrix X, respectively, $\cdot$ denotes the dot product, $\|X_i\|$ and $\|X_j\|$ are the L2 norms of X_i and X_j, respectively.

Then, given a quantile q (default $q = 75$), the threshold τ is defined as the q-th percentile of the similarity matrix S:

$$\tau = \text{percentile}(S, q) \tag{42}$$

Based on the threshold τ, the adjacency matrix $A \in \mathbb{R}^{n\times n}$ is generated as follows:

$$A_{ij} = \begin{cases} 1, & \text{if } S_{ij} \geq \tau \text{ and } i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{43}$$

To ensure the adjacency matrix is symmetric, the following operation is performed:

$$A = A + A^T \tag{44}$$

where A^T means the transpose of A.

Then, A is normalized:

$$A_{ij} = \begin{cases} 1, & \text{if } A_{ij} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{45}$$

Therefore, the final adjacency matrix A satisfies: diagonal elements are 0 ($A_{ii} = 0$); symmetry ($A = A^T$); connection weights are 1 ($A_{ij} \in \{0, 1\}$).

MSTGraph

Given the feature matrix $X \in \mathbb{R}^{n\times d}$, where n and d refer to the number of samples and the feature dimension, respectively, the Euclidean distance matrix $\mathbb{D} \in R^{n\times n}$ is computed as:

$$D_{ij} = \left\| X_i - X_j \right\|_2 \tag{46}$$

where X_i and X_j are the i-th and j-th rows of the feature matrix, respectively, $\|\cdot\|_2$ denotes the Euclidean (L2) norm.

The minimum spanning tree (MST) is generated by the distance matrix D.The MSTGraph is the subset of edges that join all vertices (samples) together without any loops and with the minimum total weight of the edges.The MSTGraph can be represented as the sparse adjacency matrix $M \in \mathbb{R}^{n\times n}$, where:

$$M_{ij} = \begin{cases} D_{ij}, & \text{if the edge between } i \text{ and } j \text{ is included in the MSTGraph} \\ 0, & \text{otherwise} \end{cases} \tag{47}$$

The adjacency matrix M of MSTGraph is converted to a dense binary adjacency matrix $A \in \mathbb{R}^{n\times n}$

$$A_{ij} = \begin{cases} 1, & \text{if } M_{ij} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{48}$$

To ensure the adjacency matrix represents an undirected graph, it is symmetrized as follows:

$$A = A + A^T \tag{49}$$

where A^T is the transpose of A.

After symmetrization, the matrix is normalized to ensure binary weights:

$$A_{ij} = \begin{cases} 1, & \text{if } A_{ij} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{50}$$

MDGraph

Given a feature matrix $X \in \mathbb{R}^{n\times d}$, where n and d stand for the number of samples and the feature dimension, respectively, the Mahalanobis distance matrix is written as follows:

1) **Estimate the Covariance Matrix.** Compute the empirical covariance matrix $\sum \in R^{d\times d}$ from X,

$$\sum = \text{EmpiricalCovariance}(X) \tag{51}$$

2) **Compute Mahalanobis Distances.** For each pair of samples X_i and X_j, the Mahalanobis distance can be expressed as:

$$Dij = \sqrt{(X_i - X_j)^T \sum^{-1} (X_i - X_j)} \tag{52}$$

where $\sum^{-1}$ means the inverse of the covariance matrix $\sum$. The pairwise distances are stored in a condensed distance vector and, then, converted to a square distance matrix D.

3) **Normalize the Distance Matrix.** Normalize D using the normalize($\cdot$) function:

$$D^{\text{normalized}} = \frac{D_{ij} - \min(D)}{\max(D) - \min(D)} \tag{53}$$

After that, given a quantile q (default $q = 25$), the threshold τ is defined as the q-th percentile of the normalized distance matrix $D^{\text{normalized}}$:

$$\tau = \text{percentile}\left(D^{\text{normalized}}, q\right) \tag{54}$$

Based on the threshold τ, the adjacency matrix $A \in \mathbb{R}^{n\times n}$ is generated as:

$$A_{ij} = \begin{cases} 1, & \text{if } D_{ij}^{normalized} \leq \tau \text{ and } i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{55}$$

The diagonal elements are set to 0 to exclude self-connections:

$$A_{ii} = 0, \ \forall i \tag{56}$$

To ensure the adjacency matrix represents an undirected graph, it is symmetrized as follows:

$$A = A + A^T \tag{57}$$

where A^T is the transpose of A.

Finally, the matrix is normalized to ensure binary weights:

$$A_{ij} = \begin{cases} 1, & \text{if } A_{ij} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{58}$$

References

[1] S. Li, K. Feng, Y. Xu, Y. Li, Q. Ni, K. Zhang, Y. Wang, W. Ding, Cross-modal zero-sample diagnosis framework utilizing non-contact sensing data fusion, Inform. Fusion 110 (2024) 102453.

[2] X. Jiang, Q. Song, H. Wang, G. Du, J. Guo, C. Shen, Z. Zhu, Central frequency mode decomposition and its applications to the fault diagnosis of rotating machines, Mech. Mach. Theory 174 (2022) 104919.

[3] H. Shao, J. Lin, L. Zhang, D. Galar, U. Kumar, A novel approach of multisensory fusion to collaborative fault diagnosis in maintenance, Inform. Fusion 74 (2021) 65–76.

[4] Q. Song, X. Jiang, J. Liu, J. Shi, Z. Zhu, Contrast-assisted domain-specificity-removal network for semi-supervised generalization fault diagnosis, IEEE Trans. Neural Netw. Learn. Syst. 36 (3) (2024) 5403–5416.

[5] C. Yang, B. Cai, Q. Wu, C. Wang, W. Ge, Z. Hu, W. Zhu, L. Zhang, L. Wang, Digital twin-driven fault diagnosis method for composite faults by combining virtual and real data, J. Ind. Inform. Integrat. 33 (2023) 100469.

[6] H. Li, J. Lin, Z. Liu, J. Jiao, B. Zhang, An interpretable waveform segmentation model for bearing fault diagnosis, Adv. Eng. Inf. 61 (2024) 102480.

[7] Y. Xu, K. Feng, X. Yan, R. Yan, Q. Ni, B. Sun, Z. Lei, Y. Zhang, Z. Liu, CFCNN: A novel convolutional fusion framework for collaborative fault identification of rotating machinery, Inform. Fusion 95 (2023) 1–16.

[8] Y. Yu, H.R. Karimi, L. Gelman, J. Tian, P. Mei, A novel multi-source sensor correlation adaptive fusion framework with uncertainty quantification for intelligent fault diagnosis, Reliab. Eng. Syst. Safe. (2025) 111812.

[9] M. Li, J. Huang, F. Zhang, Y. Yu, F. Gu, F. Chu, MSIF-Convformer: a novel end-to-end fault diagnosis framework with multi-source sensors under strong noise, Inform. Fusion (2025) 104000.

[10] Y. Chen, D. Zhang, R. Yan, F. Guo, Q. Xuan, Class-consistent matching attention wavelet networks for partial transfer intelligent diagnosis, IEEE Trans. Neural Netw. Learn. Syst. (2025).

[11] Z. Lei, F. Tian, Y. Su, G. Wen, K. Feng, X. Chen, M. Beer, C. Yang, Unsupervised graph transfer network with hybrid attention mechanism for fault diagnosis under variable operating conditions, Reliab. Eng. Syst. Safe. 255 (2025) 110684.

[12] Q. Qian, J. Luo, Y. Qin, Adaptive intermediate class-wise distribution alignment: a universal domain adaptation and generalization method for machine fault diagnosis, IEEE Trans. Neural Netw. Learn. Syst. (2024).

[13] Z. Li, X. Ding, Z. Song, L. Wang, B. Qin, W. Huang, Digital twin-assisted dual transfer: a novel information-model adaptation method for rolling bearing fault diagnosis, Inform. Fusion 106 (2024) 102271.

[14] L. Qin, L. Zhang, J. Feng, F. Zhang, Q. Han, Z. Qin, F. Chu, A hybrid triboelectric-piezoelectric smart squirrel cage with self-sensing and self-powering capabilities, Nano Energy 124 (2024) 109506.
[15] Q. Li, Y. Liu, S. Sun, Z. Qin, F. Chu, Deep expert network: a unified method toward knowledge-informed fault diagnosis via fully interpretable neuro-symbolic AI, J. Manuf. Syst. 77 (2024) 652–661.
[16] Z. Wang, Y. Ta, W. Cai, Y. Li, Research on a remaining useful life prediction method for degradation angle identification two-stage degradation process, Mech. Syst. Signal Process. 184 (2023) 109747.
[17] Y. Zhang, P. Liu, D. Hu, W. Wang, Advancing fatigue crack growth prognosis in metallic structures: a physics-informed sequential attention approach with uncertainty quantification, Reliab. Eng. Syst. Safe. (2026) 112241.
[18] C. Lin, Y. Kong, Q. Han, T. Wang, M. Dong, H. Liu, F. Chu, An information fusion-based meta transfer learning method for few-shot fault diagnosis under varying operating conditions, Mech. Syst. Signal Process. 220 (2024) 111652.
[19] Z. Ming, B. Tang, L. Deng, Q. Li, Simulation data-driven adaptive frequency filtering focal network for rolling bearing fault diagnosis, Eng. Appl. Artif. Intell. 138 (2024) 109371.
[20] X. Li, S. Yu, Y. Lei, N. Li, B. Yang, Intelligent machinery fault diagnosis with event-based camera, IEEE Trans. Ind. Inf. 20 (1) (2023) 380–389.
[21] T. Yan, X. Xing, D. Wang, K.-L. Tsui, M. Xia, Interpretable degradation tensor modeling through multi-scale and multi-level time-frequency feature fusion for machine health monitoring, Inform. Fusion 117 (2025) 102935.
[22] Z. He, C. Shen, B. Chen, J. Shi, W. Huang, Z. Zhu, D. Wang, A new feature boosting based continual learning method for bearing fault diagnosis with incremental fault types, Adv. Eng. Inf. 61 (2024) 102469.
[23] B. Hou, D. Wang, Z. Peng, K.-L. Tsui, Adaptive fault components extraction by using an optimized weights spectrum based index for machinery fault diagnosis, IEEE Trans. Ind. Electron. 71 (1) (2023) 985–995.
[24] Y. Yu, Y. He, H.R. Karimi, L. Gelman, A.E. Cetin, A two-stage importance-aware subgraph convolutional network based on multi-source sensors for cross-domain fault diagnosis, Neural Netw. 179 (2024) 106518.
[25] Y. Wang, L. Zeng, L. Wang, Y. Shao, Y. Zhang, X. Ding, An efficient incremental learning of bearing fault imbalanced data set via filter styleGAN, IEEE Trans. Instrum. Meas. 70 (2021) 1–10.
[26] X. Jiang, J. Zheng, Z. Chen, Z. Ge, Z. Song, X. Ma, Leveraging transfer learning for data augmentation in fault diagnosis of imbalanced time-frequency images, IEEE Trans. Autom. Sci. Eng. (2024).
[27] K. Wu, W. Tong, J. Xie, F. Wang, B. Huang, D. Wu, Optimal weighted envelope spectrum: an enhanced demodulation method for extracting specific characteristic frequency of rotating machinery, Mech. Syst. Signal Process. 211 (2024) 111165.
[28] J. Huang, F. Zhang, B. Safaei, Z. Qin, F. Chu, The flexible tensor singular value decomposition and its applications in multisensor signal fusion processing, Mech. Syst. Signal Process. 220 (2024) 111662.
[29] Z. Li, H. Jiang, X. Wang, A novel reinforcement learning agent for rotating machinery fault diagnosis with data augmentation, Reliab. Eng. Syst. Safe. 253 (2025) 110570.
[30] D. Wang, C. Chen, Multilevel feature encoder for transfer learning-based fault detection on acoustic signal, Inform. Fusion 121 (2025) 103128.
[31] J. Tian, Y. Yu, H.R. Karimi, F. Gao, J. Lin, A continual test-time domain adaptation method for online machinery fault diagnosis under dynamic operating conditions, Neural Netw. (2025) 108192.
[32] Y. Yu, H.R. Karimi, P. Shi, R. Peng, S. Zhao, A new multi-source information domain adaption network based on domain attributes and features transfer for cross-domain fault diagnosis, Mech. Syst. Signal Process. 211 (2024) 111194.
[33] Y. Yu, H.R. Karimi, L. Gelman, A.E. Cetin, MSIFT: a novel end-to-end mechanical fault diagnosis framework under limited & imbalanced data using multi-source information fusion, Expert Syst. Appl. 274 (2025) 126947.
[34] S. Zhao, L. Bao, C. Hou, Y. Bai, Y. Yu, Multi-source domain adversarial graph convolutional networks for rolling mill health states diagnosis under variable working conditions, Struct. Health Monitor. 23 (6) (2024) 3505–3524.
[35] S. Liu, J. Chen, Y. Feng, Z. Xie, T. Pan, J. Xie, Generative artificial intelligence and data augmentation for prognostic and health management: taxonomy, progress, and prospects, Expert Syst. Appl. 255 (2024) 124511.
[36] Y. Huang, J. Lin, C. Zhou, H. Yang, L. Huang, Modality competition: what makes joint training of multi-modal network fail in deep learning? (Provably), in: International Conference on Machine Learning, PMLR, 2022, pp. 9226–9259.
[37] W. Wang, D. Tran, M. Feiszli, What makes training multi-modal classification networks hard?, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12695–12705.
[38] A. Ding, Y. Qin, B. Wang, L. Guo, L. Jia, X. Cheng, Evolvable graph neural network for system-level incremental fault diagnosis of train transmission systems, Mech. Syst. Signal Process. 210 (2024) 111175.
[39] J. Li, H. Chen, X.-B. Wang, Z.-X. Yang, A comprehensive gear eccentricity dataset with multiple fault severity levels: description, characteristics analysis, and fault diagnosis applications, Mech. Syst. Signal Process. 224 (2025) 112068.
[40] P. Shi, Y. Yu, H. Gao, C. Hua, A novel multi-source sensing data fusion driven method for detecting rolling mill health states under imbalanced and limited datasets, Mech. Syst. Signal Process. 171 (2022) 108903.
[41] Y. Yu, P. Shi, J. Tian, X. Xu, C. Hua, Rolling mill health states diagnosing method based on multi-sensor information fusion and improved DBNs under limited datasets, ISA Trans. 134 (2023) 529–547.
[42] C. Yang, J. Liu, K. Zhou, X. Jiang, X. Zeng, An improved multi-channel graph convolutional network and its applications for rotating machinery diagnosis, Measurement 190 (2022) 110720.
[43] Y. Yu, H.R. Karimi, L. Gelman, X. Liu, A novel digital twin-enabled three-stage feature imputation framework for non-contact intelligent fault diagnosis, Adv. Eng. Inf. 66 (2025) 103434.
[44] C. Wang, J. Yang, H. Jie, Z. Tao, Z. Zhao, A lightweight progressive joint transfer ensemble network inspired by the Markov process for imbalanced mechanical fault diagnosis, Mech. Syst. Signal Process. 224 (2025) 111994.
[45] C. Wang, Z. Shu, J. Yang, Z. Zhao, H. Jie, Y. Chang, S. Jiang, K.Y. See, Learning to imbalanced open set generalize: a meta-learning framework for enhanced mechanical diagnosis, IEEE Trans. Cybern. (2025).
[46] C. Wang, X. Liu, J. Yang, H. Jie, T. Gao, Z. Zhao, Addressing unknown faults diagnosis of transport ship propellers system based on adaptive evolutionary reconstruction metric network, Adv. Eng. Inf. 65 (2025) 103287.
[47] K. Ma, J. Yang, P. Liu, Relaying-assisted communications for demand response in smart grid: Cost modeling, game strategies, and algorithms, IEEE J. Sel. Areas Commun. 38 (1) (2019) 48–60.
[48] S. Zhao, L. Zhang, R. Zhu, Q. Han, Z. Qin, F. Chu, Modeling approach for flexible shaft-disk-drum rotor systems with elastic connections and supports, Appl. Math. Model. 106 (2022) 402–425.