

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/172566/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Chen, Chong, Wang, Tao, Mao, Dong, Liu, Ying and Cheng, Lianglun 2025. A multi-scale graph pyramid attention network with knowledge distillation towards edge computing robotic fault diagnosis. Expert Systems with Applications 260 , 125469. 10.1016/j.eswa.2024.125469

Publishers page: <http://dx.doi.org/10.1016/j.eswa.2024.125469>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



A Multi-scale Graph Pyramid Attention Network with Knowledge Distillation towards Edge Computing Robotic Fault Diagnosis

Chong Chen^a, Tao Wang^{a,b*}, Dong Mao^a, Ying Liu^c and Lianglun Cheng^a

^aGuangdong Provincial Key Laboratory of Cyber-Physical System, Guangdong University of Technology, Guangzhou 510006, China

^bSchool of Automation, Guangdong University of Technology, Guangzhou 510006, China.

^cDepartment of Mechanical Engineering, School of Engineering, Cardiff University, Cardiff CF24 3AA, UK

*Corresponding author

E-mail: chenc2021@gdut.edu.cn, wangtao_cps@gdut.edu.cn, 472499617@qq.com, LiuY81@cardiff.ac.uk, LLCheng@gdut.edu.cn

Abstract

The advanced graph neural networks (GNNs) algorithms and the abundant monitoring data have greatly improved the performance of fault diagnosis of industrial robots. However, collecting massive monitoring data from industrial robots and performing diagnosis in the cloud server is hard to achieve due to the heavy transmission load and computational costs. Edge computing provides a potential solution to this challenge, while it requires the neural network deployed in the edge device to be effective and lightweight. This paper proposes a multi-scale graph pyramid attention network (MsGPAT) enhanced with knowledge distillation (KD) to balance accuracy and efficiency for edge-based fault diagnosis. The approach first integrates multi-scale graph convolutional networks (GCN) to extract spatial features at different granularities from monitoring data. Subsequently, pyramidal attention is incorporated to hierarchically summarize inter-scale dependencies and intra-scale neighbors for multi-resolution temporal modeling. Finally, KD is adopted to transfer salient knowledge from the complex MsGPAT to a tailored compressed model suited for edge resources. An experimental study based on the faulty dataset collected from the real world is conducted. The experimental results show that the MsGPAT can achieve the diagnosis accuracy of 85.3% after KD, which model size is only 1.04MB.

Keywords—Fault diagnosis; Graph neural networks; Industrial robot; knowledge Distillation; Deep learning.

1 Introduction

The industrial robot has become the key equipment in smart manufacturing. The safe operation of industrial robots has become a key concern in recent years. Fault diagnosis is a critical aspect of ensuring the reliable and efficient operation of industrial robots (Zhou, Zeng, Chen, Xiao, & Xiang, 2022). By accurately diagnosing fault, potential risks and operational disruptions can be minimized, ensuring optimal performance, productivity, and safety in industrial robot operations (Pu, Cabrera, Li, & Valente de Oliveira, 2023). With the massive deployment of industrial robots in various manufacturing scenarios, the landscape of fault diagnosis has undergone a significant shift. Initially, the focus was primarily on cloud-based solutions for fault diagnosis. However, as the number of deployed robots has grown rapidly, along with the increasing complexity of their operations, it is challenging for the cloud server to achieve real-time and accurate diagnosis. The sheer volume of data generated by multiple robots, coupled with the need for real-time decision-making, has necessitated a shift towards edge computing solutions (Lu, Lu, An, Wang, & He, 2023). Industrial robots often operate in complex and dynamic environments, where real-time decision-making is crucial (Y. Liu, Chen, Wang, & Cheng, 2022). By deploying the fault diagnosis capability directly at the edge devices, closer to the robots, several advantages can be achieved. It allows for quick and localized fault detection and diagnosis without relying heavily on cloud-based processing, reducing latency, and ensuring timely response. By bringing fault diagnosis to the edge devices near the robots, the diagnostic process can be performed more efficiently and effectively. Therefore, it is necessary to explore the advanced fault diagnosis algorithms which can be deployed in edge devices.

Deep learning has shown promising performance in learning the hidden patterns from the monitoring data (Lin, Shao, Zhou, Cai, & Liu, 2023; Wu, et al., 2023). Recent studies have explored the application of advanced techniques such as Recurrent Neural Networks (RNNs) (Keshun, Puzhou, & Yingkui, 2024), attention mechanisms (C. Chen, Liu, et al., 2023; C. Chen, Wang, Liu, Liu, & Cheng, 2023; Zhuohang Chen, et al., 2022; Fang, et al., 2023; Yasong Li, Zhou, Sun, Chen, & Yan, 2022; Zhang, et al., 2023) and GNNs (Ghorvei, Kavianpour, Beheshti, & Ramezani, 2023; Q. Li, 2023; Long, et al., 2022; Ma, Liu, Guo, Wang, & Ma, 2023; Sun, et al., 2021; Tang, et al., 2021; Wang, et al., 2022; Xu, et al., 2022; Yu, Tang, & Deng, 2023) for establishing accurate fault diagnosis models. While RNNs have been traditionally employed for time-series analysis, their limitations in managing long-term dependencies and spatial interconnections necessitate a more sophisticated approach. In contrast, GNNs expertise in processing structured data, capturing the intricate topologies of interconnected systems, and aggregating information from distant nodes, makes them ideally suited for fault diagnosis scenarios. Furthermore, although Transformers have revolutionized natural language processing by efficiently handling long-range dependencies, they may not be able to inherently account for the spatial structure that is important in fault propagation analysis. Traditional GNN architectures often struggle with the uniform aggregation of neighboring node features, failing to adequately differentiate nodes with similar local structures. To address this issue, Graph Attention Networks (GATs), a subclass of GNNs, were introduced, leveraging attention mechanisms to weigh the importance of different neighbors differently. Such a modification allows GATs to dynamically focus on the most relevant parts of the graph when computing node embeddings, thus enhancing the model's ability to capture complex relational patterns.

However, despite their advancements, GATs and other attention-based GNNs still face limitations, such as scalability and computational cost, especially when dealing with large graphs or multi-scale features.

Existing GNN-based have made significant efforts in analyzing static graphs and achieved decent performance (Zhiwen Chen, Xu, Peng, & Yang, 2021; X. Zhao, et al., 2022). However, existing approaches fall short when applied to learn the spatial patterns in different granularities from the graph data, which may impede the model from learning the key patterns relevant to the fault diagnosis. Moreover, the lack of lightweight consideration poses a challenge for the edge deployment of the model. In this study, we propose an integrated approach to address these gaps. Multi-scale GCN provides spatial graph modeling to learn localized patterns. Pyramid attention summarizes inter-scale and intra-scale dependencies, enabling temporal modeling. Finally, KD transfers inert knowledge from the complex model to a compact student model for edge compatibility. The fusion of techniques captures the multi-scale features of robotic monitoring data. GCN extracts spatial features, pyramid attention models long-range sequential evolution, and KD optimizes model size - together forming an engineered solution catered for diagnosis on resource-constrained edge devices. The approach balances modeling accuracy and complexity through domain-driven design choices. The main contributions of this study are four-fold: (1) A multi-scales graph features extraction scheme that adopts GCN with different receptive fields is designed to extract the spatial and temporal features at different granularities; (2) Subsequently, pyramid attention is introduced to learn the dependency between features extracted from different granularities; (3) A KD strategy is adopted to lower

the model size of the MSGPAT, which can then be deployed in the edge device; (4) An experimental study based on the real-world robotic fault dataset is conducted to reveal the effectiveness of the proposed MsGPAT with KD strategy. The rest of this paper is organized as follows. Section 2 reviews the latest deep learning algorithms on fault diagnosis and the research of KD. Section 3 details the proposed method. The experimental setup is introduced in Section 4 and the results are demonstrated and discussed in Section 5. Finally, Section 6 concludes.

2 Literature Review

2.1 The recent advance of deep learning in fault diagnosis

Recent studies have focused on the development of effective fault diagnosis methods for rotating machinery, as it is essential for ensuring safe, stable and long-term operation. In recent years, GNNs and the Transformers network have gained increasing attention in fault diagnosis. GNNs can effectively extract the topological features from the data. The receptive scale of GCN is hard to be determined in the application of GCN. (T. Li, Zhao, Sun, Yan, & Chen, 2020) proposed a multi-receptive field graph convolutional network (MRF-GCN) by considering the importance of neighbors in graph construction and adaptive receptive field. In this approach, data samples are converted into weighted graphs to better capture the relationships within data. Furthermore, the proposed approach can learn features from different receptive fields, and fuses them together to form a better feature representation. (Sun, et al., 2021) proposed a multi-scale cluster-graph convolution neural network with multi-channel residual network. In this approach, multi-channel residual network (MCRN) was first introduced to extract the feeble features in

the signal. Secondly, a multi-scale cluster-graph convolution neural network was proposed for fault diagnosis, utilizing the finite graph data of signal and different scales generated by an autoencoder graph generation layer. (Ghorvei, et al., 2023) proposed a deep subdomain adaptation graph convolution neural network for cross-domain bearing fault diagnosis. In this study, a graph convolution neural network (GCNN) was utilized to model the data structure. (Xu, et al., 2022) proposed an oversmoothing relief GCN (OsR-GCN) method to address the oversmoothing problem in deploying GCN. Euclidean distance-based method and the structure analysis-based method were first deployed to construct the graph data, which was then fed into OsR-GCN. The OsR-GCN adopts a weight coefficient, which was optimized by an improved particle swarm optimization algorithm, to address the oversmoothing issue. (Tang, et al., 2021) proposed a fault diagnosis method based on the Graph Cardinality Preserved Attention Network (GCPAT) which can accurately diagnose motors under varying working conditions and be generalized to the transient state. The diagnosis results were obtained by analyzing signal-converting graphs.

Data limitation is one of the challenges in fault diagnosis modelling. In order to address the data imbalance and noise issues, (Ma, et al., 2023) proposed an approach for rolling bearing faults diagnosis based on limited samples. The approach consists of a multi-order graph embedding stacked denoising autoencoder optimized by an improved sine-cosine algorithm (MGE-ISCA-SDAE) and a collaborative central domain adaptation (CCDA). The MGE-ISCA-SDAE was designed to extract the time-frequency faulty features from the raw data, while the CCDA model is used to transfer the fault samples in a new domain. (Yu, et al., 2023) proposed

a graph-weighted reinforcement network (GWRNet). An adjacency matrix was first created by calculating the Euclidean distance of time- and frequency-domain characteristics of the limited data samples, which can pre-classify the nodes. (Long, et al., 2022) proposed a self-adaptation graph attention network via meta-learning for few-shot fault diagnosis. (Wang, et al., 2022) reported a few-shot learning method named dual graph neural network (DGNNNet) with residual blocks for few-shot fault diagnosis modelling.

In recent years, Transformers have been studied in fault diagnosis modelling. The computational cost is a major concern in the Transformers network. Due to the mechanism of the MSA and the stacked usage of the Transformers blocks, the computational burden of the Transformers network is much heavier than other deep learning algorithms. In order to lower the computational cost of the Transformers network, (Fang, et al., 2023) proposed a modified convolution unit based on the transformer architecture, utilizing a self-activation function to enhance the accuracy of bearing fault diagnosis. This integration seamlessly combines the strengths of transformers and CNNs, ensuring a lightweight and high-performing model. Chen et al. (C. Chen, Wang, et al., 2023) proposed lightweight convolutional Transformers (LCT). LCT adopts convolutional blocks to lower the feature map size. The linear spatial reduction mechanism is introduced into the Transformers block to reduce the computational cost of MSA. Finally, the output of the Transformers block is further squeezed via a sequence pooling layer. Besides the modification of the tokenizer and the MSA, researchers have modified Transformers to achieve better fault diagnosis performance. In order to address the domain shifted issue in fault diagnosis, (Zhuohang Chen, et al., 2022) proposed a multi-channel

calibrated transformer with shifted windows, which utilizes non-overlapping windows to compute self-attention, capturing the relationships between sequences of split patches. (Y. Li, et al., 2022) proposed a variational attention-based transformer network, which can mine the association relationships within signals to improve the performance of fault diagnosis. In this network, Variational inference is utilized to constrain attention weights by sampling from Dirichlet distributions, while Laplace approximation is employed to enable reparameterization. (Zhang, et al., 2023) proposed a digital twin-driven partial domain adaptation network for bearing fault diagnosis. A Transformer network is adopted for the cross-domain fault diagnosis in the digital twin fault diagnosis framework.

2.2 The studies of knowledge distillation in manufacturing

In order to deploy deep learning into the edge device such as field programmable gate array (FPGA), it is necessary to shrink the size of the neural network and keep its knowledge level. In the research on fault diagnosis, (Ji, et al., 2022) adopted CNN as a teacher and student network for KD. The student only has one convolutional layer and a fully-connected layer, which is low in parameters and computational costs. (F. Li, Chen, He, & Zhou, 2021) investigated KD for unknown fault category diagnosis. The method utilizes KD to extract knowledge between classes allowing data of various categories to contain information from other categories while maintaining information of their own categories. (Deng, et al., 2021) proposed a hierarchical-split block and KD approach to address the data imbalance issue in fault diagnosis modelling. Soft labels produced through KD can reflect the similarities between categories. By learning the soft labels, information about all categories is considered during

each epoch of model training, rather than solely focusing on the information related to the current samples. (Guan, Qiao, Zhai, & Wang, 2022) presented an incremental fault diagnosis framework, which incorporates adaptive KD and representative exemplar selection (MEMAR) to enhance the fault diagnosis capabilities. The proposed framework addresses the challenge of handling new incoming samples by leveraging KD to identify potential correlations between old and new samples. By applying adaptive constraints on parameter optimization, the model is updated effectively. Additionally, a genetic algorithm is employed to selectively retain representative exemplar subsets, enabling the model to better adapt to dynamic data. These advancements contribute to the improvement of fault diagnosis accuracy and the ability to handle incremental data in evolving systems. To address the issue of class imbalance by utilizing only new task data to update the model, (F. Li, et al., 2021) proposed an intelligent fault diagnosis method based on the layer regeneration network under class-imbalanced samples. The approach recognizes that the data contain information from other categories that may be masked by information from their own categories. To leverage this hidden knowledge, KD is employed to extract and transfer knowledge between classes, thereby enhancing the learning of other categories.

Besides its application in fault diagnosis, KD has also been widely studied in other scenarios of manufacturing. (Yixin Li, Hu, Liu, Ryan, & Wang, 2023) proposed a KD approach to lower the model complexity and computational load of the energy consumption model in additive manufacturing. A teacher assistant component was introduced within a teacher-student architecture. Initially, a teacher ensemble comprising three baseline models was constructed to

enhance accuracy. Subsequently, a teacher ensemble is incorporated to bridge the capacity gap between the ensemble and the simplified student model, resulting in reduced complexity. (J. Zhao, et al., 2023) considered the practical scenario of cross-domain few-shot learning, where the base dataset used for training the model and the target defect dataset applied to the model are obtained from domains with significant differences and proposed an enhanced relation-prototypical network with KD. This improved model effectively addresses the challenges of cross-domain few-shot learning in defect classification. (Z. Liu, et al., 2022) proposed an anchor-free defect detection method based on KD. KD is introduced into defect detection modelling to lower the detection model complexity.

2.3 A Brief Summary

From the existing studies, it can be seen that advanced GNNs have been investigated in fault diagnosis. Advanced GNNs such as GCN and GAT are particularly well-suited for fault diagnosis as they can capture the spatial dependencies and structural patterns in the data. By incorporating graph convolutional and MSA layers, GAT can aggregate information from neighboring nodes in the graph, enabling them to capture local relationships among different fault-related variables or components. This allows GAT to capture the inherent dependencies and interactions between faults. While existing techniques have made progress, a gap remains in developing methods that can holistically capture the spatial and temporal dependencies inherent to robotic monitoring data, while maintaining feasible model complexity for edge deployment. To address this need, we propose an integrated approach that takes advantage of multi-scale graph modeling, pyramidal attention, and KD tailored specifically for the intricacies

and constraints of robotic fault diagnosis on edge devices.

3 Methodology

In this section, the flowchart and details of the proposed approach are elaborated. Figure 1 demonstrated the overall flowchart of the proposed approach. Firstly, data collection and pre-processing were conducted. The feedback current signals from the servo drivers of the transmission system of the industrial robot are collected. The feedback current data can be obtained directly from the existing device, without requiring the installation of additional sensors for data collection. Then, the frequency domain representation of feedback current data is obtained via fast Fourier transformation. Subsequently, the graph data is constructed via fast Fourier transform and k -nearest neighbor (k -NN) algorithm, which can effectively find the representative features of each time window (T. Li, et al., 2022). In the next stage, the constructed graph data is then sent into both the teacher network and student network for KD. A large-size MsGPAT is adopted as the teacher network, while a tiny MsGPAT is adopted as the student network. After the distillation training of both networks, a tiny MsGPAT with satisfactory performance can be obtained. Finally, the performance of the tiny MsGPAT is evaluated and it is integrated into an edge device for deployment.

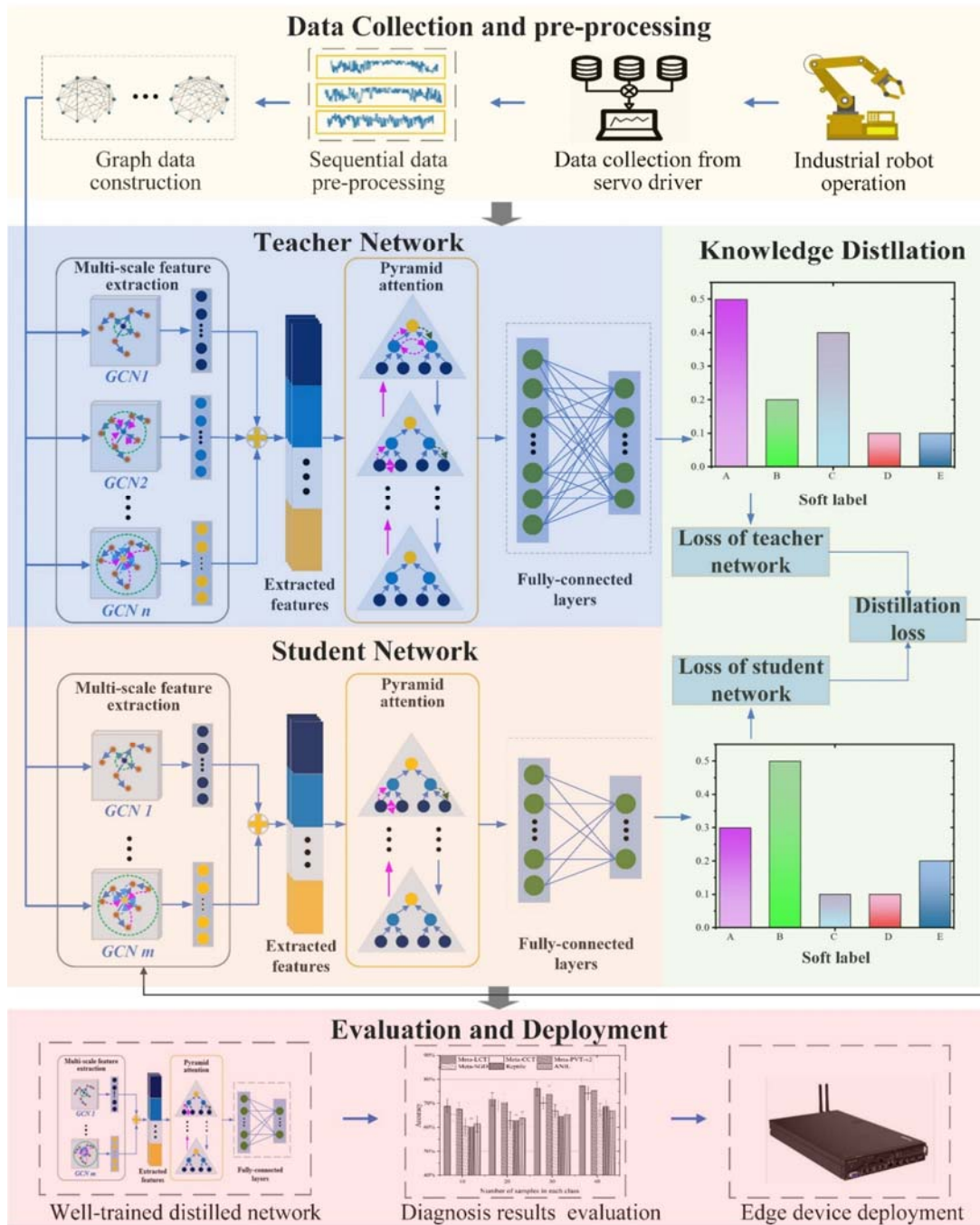


Figure 1. The overall flow-chart of the proposed approach.

3.1 Multi-scale graph pyramid attention network

MsGPAT contains two main parts which are the multi-scale feature extraction module and the pyramid attention module. The feature extracted from the multi-scale feature extraction module is then sent into the pyramid attention module. Finally, the features extracted from the pyramid

attention layer are then sent into fully connected layers for fault diagnosis. The multi-scale GCN provides localized spatial graph modeling to uncover inter-component relationships. Pyramid attention summarizes features hierarchically to capture long-range temporal fault evolution.

In the teacher and student networks, GCNs are employed to perform convolution operations directly on the graph data. This allows for the extraction of spatial features at different granularities, which is essential for capturing inter-component relationships in the industrial robot. The number of GCN layers in the teacher network is denoted by $N_{teacher}$, and in the student network by $N_{student}$. The mathematical formulation for a single GCN layer can be represented as follows:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}) \quad (1)$$

, where $H^{(l)}$ is the input matrix to the l^{th} GCN layer, $\tilde{A}$ is the adjacency matrix of the graph with added self-connections, $\tilde{D}$ is the degree matrix corresponding to $\tilde{A}$, $W^{(l)}$ is the weight matrix for the l^{th} GCN layer, σ is the activation function.

The number of GCN layers is a significant parameter that influences the model's capacity to learn complex patterns. A larger number of layers in the teacher network allows for the extraction of deeper hierarchical features, while the student network, with potentially fewer layers, benefits from the distilled knowledge, achieving a balance between model size and performance. KD is implemented to transfer knowledge from the teacher network to the student network. The student network, with fewer GCN layers ($N_{student} < N_{teacher}$), learns to mimic the behavior of the teacher network while maintaining a compact architecture suitable for

deployment on edge devices.

In the multi-scale feature extraction block, to obtain different granularities of temporal features, multiple GCN modules are integrated and perform feature extraction with different scale C . Multi-scale GCNs operate by aggregating information from multiple graph convolutional layers with varying receptive fields or neighborhood sizes. Each layer captures information at a specific scale or level of detail, allowing the model to learn features at different granularities. This enables the GCN to capture local patterns and fine-grained details in lower layers while incorporating global context and higher-level relationships in deeper layers. GCN is a type of neural network designed to process graph-structured data, where the relationships between the data points are represented by edges in a graph. GCN is able to learn the key features from the graph embedding data.

The different granularities of temporal features are concatenated and transformed into the input of the next level of pyramid attention through a linear layer. In order to extract multi-granularity feature information, the data is output as G_{in} after passing through GCN at different scales. Here, B represents the batch size, and D_K represents the feature dimension, which is used as the input of the pyramid attention, which can be expressed as:

$$G_{in} = B \times (GCN_L + GCN_{L/C} + GCN_{L/C^2} + \dots + GCN_{L/C^n}) \times D_K \quad (2)$$

In GCN, the filter $f_{\theta'}$ is approximated by a polynomial. The truncated expansion of the Chebyshev polynomials can be used to approximate f_{θ} . The K^{th} order approximation can be

expressed as:

$$f_{\theta'} = \sum_{k=0}^{K-1} \theta_k \Lambda^k = \sum_{k=0}^{K-1} \theta'_k T_k(\tilde{\Lambda}) \quad (3)$$

where K is the maximum order of Chebyshev polynomials, $\Lambda = \frac{2\Lambda}{\lambda_{max}} - I_N$ is the rescaled eigenvalues, λ_{max} denotes the largest eigenvalues of L . θ is a vector of polynomial coefficients. $T_k(\tilde{\Lambda})$ is the Chebyshev polynomial of order k , and it can be determined by the following recurrence relation, that is:

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x) \quad (4)$$

,where $T_0(x) = 1$ and $T_1(x) = x$.

After approximating the filter by Chebyshev polynomials, the GCN of a node feature x with a filter f_{θ} can be expressed as follows:

$$h = U \begin{bmatrix} \theta'_1 T_1(\tilde{\Lambda}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \theta'_K T_K(\tilde{\Lambda}) \end{bmatrix} U^T x = \sum_{k=0}^{K-1} \theta'_k T_k(\tilde{L})x \quad (5)$$

where $\tilde{L}$ is the rescaled Laplacian matrix.

The pyramid attention uses a tree-like structure to propagate global information, which can effectively capture the sequential patterns within the data. Compared to other neural networks that can also obtain global temporal features, such as recurrent networks, long short-term memory networks, and self-attention networks, it has lower time and space complexity (S. Liu, et al., 2021). The comparison between pyramid attention and other prevailing networks is shown in Figure 2. The pyramid attention uses a tree-like efficient feature propagation structure, where each node only focuses on its parent, neighbor, and child node features, and through this relationship. It ultimately globally propagates local feature information. Since each node only

focuses on limited features, its complexity is controllable and very low, achieving the goal of propagating global features while reducing complexity and improving computational efficiency.

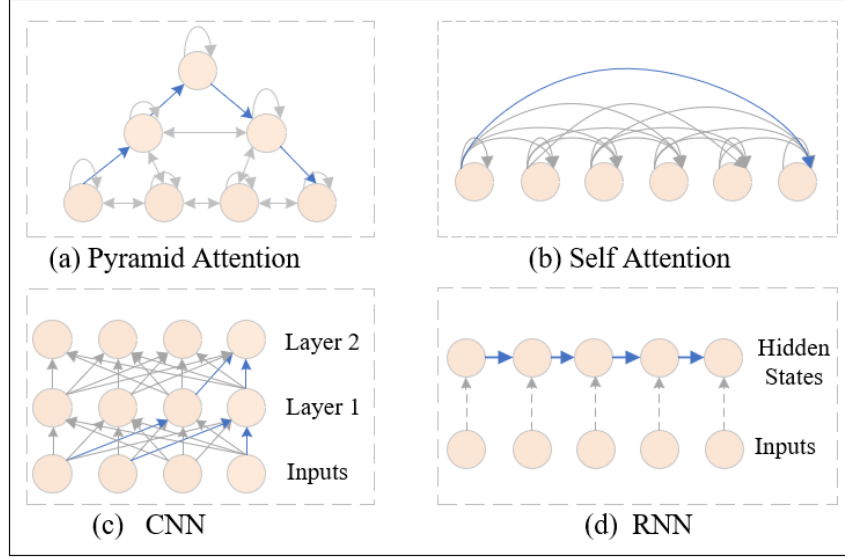


Figure 2. The comparison between pyramid attention and other prevailing networks.

X and Y are denoted as the input and output of a single self-attention head. In a MSA layer, multiple self-attention heads are used to extract the temporal pattern from different aspects. The input data X is first converted into three matrices, which are denoted as Q , K and V , which can be expressed as:

$$Q = XW^q \quad (5)$$

$$K = XW^k \quad (6)$$

$$V = XW^v \quad (7)$$

,where W^q , W^k , and $W^v \in \mathbb{R}^{L \times D_K}$. The senior features y_i obtained from MSA can be expressed as:

$$y_i = \sum_{l=1}^L \frac{\exp(q_i K_l^T / \sqrt{d_k}) v_l}{\sum_{l=1}^L \exp(q_i K_l^T / \sqrt{d_k})} \quad (8)$$

In the MSA mechanism, the number of query-key dot products that need to be calculated and stored dictates the time and space complexity of the self-attention mechanism. Since all Q-K pairs are computed and stored in the self-attention mechanism, the resulting time and space complexity is $O(L^2)$.

After obtaining the multi-scale features G_{in} through the encoding layers, a directed graph is formed with C node features. Then, the $g_l^{(s)}$ node only needs to focus on the adjacent nodes $N_l^{(s)}$ of the same scale (including itself), C child nodes $K_l^{(s)}$ with it as the parent node, and its parent node $P_l^{(s)}$. Finally, a set of features $G_l^{(s)}$ is formed. The process can be expressed as follows:

$$\begin{cases} G_l^{(s)} = N_l^{(s)} \cup K_l^{(s)} \cup P_l^{(s)} \\ N_l^{(s)} = \{g_l^{(s)}: |j - l| \leq \frac{A-1}{2}, 1 \leq j \leq \frac{L}{C^{s-1}}\} \\ K_l^{(s)} = \{g_l^{(s-1)}: (j-1)C < j < lC\} \text{ if } s \geq 2 \text{ else } 0 \\ P_l^{(s)} = \{g_l^{(s+1)}: j = \lfloor \frac{l}{C} \rfloor\} \text{ if } s \leq S-1 \text{ else } 0 \end{cases} \quad (9)$$

, where $s = 1, \dots, S$ represents the scale (from bottom to top layer), and l represents the l^{th} node under this scale.

In pyramid attention, each node only needs to pay attention to its adjacent three-scale neighboring node features, so the maximum path length is controlled at $O(L)$. In contrast, traditional methods are based on sequences, and in order to obtain the features of the first L nodes, their receptive field length is at least $O(L)$. Fine-grained features are aggregated into coarse-grained features and propagated in a tree path, so the time and space complexity of the

model are linearly related to the number of nodes, controlled at $O(N)$, where N is the number of nodes. The key features related to fault can be learned by all nodes in the model after being propagated at N times. The output y_i for the pyramid attention layer can be expressed as:

$$y_i = \sum_{l \in g_l^{(s)}} \frac{\exp(q_i K_l^T / \sqrt{d_k}) v_l}{\sum_{l \in g_l^{(s)}} \exp(q_i K_l^T / \sqrt{d_k})} \quad (10)$$

The specific process of the multi-scale feature extraction is shown in the second part of Figure 3. The features in different granularities are concatenated and fed into a fully connected layer to adjust the appropriate size. After the multi-scale features are obtained, the fused features are then sent into the pyramid attention layer to learn the hidden patterns related to the fault. After the data is processed by the pyramid attention module, the sequential features that are strongly related to the robotic fault can be extracted. Finally, the extracted features are fed into connected layers for fault category classification.

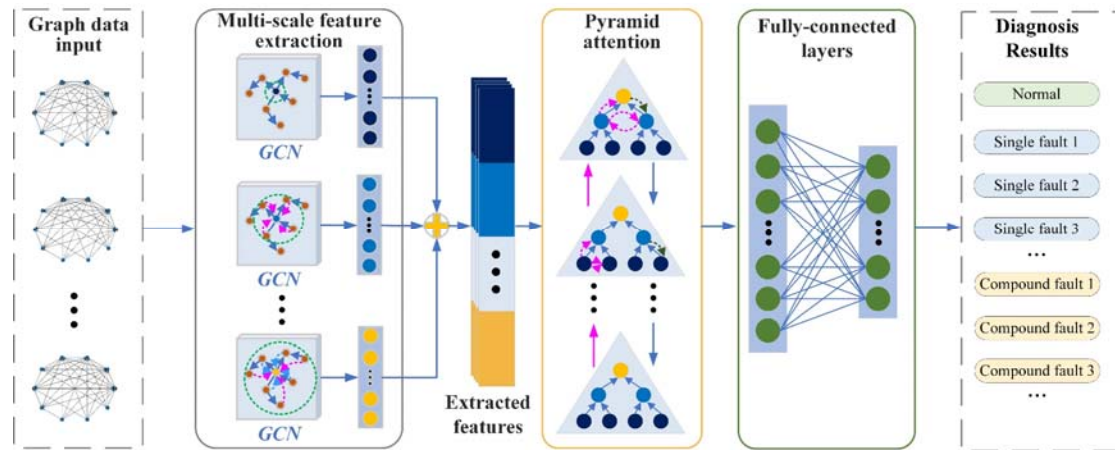


Figure 3. The structure of the proposed MsGPAT.

3.2 Knowledge distillation

The goal of KD is to transfer knowledge from a large, complex model (teacher model) to a smaller, more efficient model (student model). The student model aims to mimic the behavior and performance of the teacher model while maintaining a smaller model size and computational requirements. In KD, the teacher-student framework is established, where the teacher model, which is usually a deep and complex network, acts as the source of knowledge, and the student model, which is a smaller and more lightweight network, learns from the teacher model.

The distillation loss function quantifies the discrepancy between the outputs of the teacher and student models. It consists of two components: the knowledge transfer component and the regularization component. The knowledge transfer component measures the similarity between the predictions of the teacher and student models. It typically utilizes a soft target, which is a smoothed version of the teacher's output, to provide more informative supervision to the student model. The most common form of the knowledge transfer component is the cross-entropy loss:

$$Loss_{KT} = CrossEntropy(Softmax_{Teacher}, Softmax_{Student}) \quad (11)$$

Let $p_T(y | X)$ denote the probability distribution over the classes output by the teacher model for a given input X , and $p_S(y | X)$ denote the corresponding distribution predicted by the student model. Then, the distillation loss, which quantifies the dissimilarity between the teacher and student models' predictions, can be formulated using the Kullback-Leibler (KL) divergence

as follows:

$$L_{\text{distill}} = D_{\text{KL}}(p_T(y | X) || p_S(y | X)) = \sum_y p_T(y | X) \log \frac{p_T(y|X)}{p_S(y|X)} \quad (11)$$

In this formulation, y represents the class labels, and the summation is over all possible classes. The KL divergence measures the difference between the probability distribution predicted by the teacher model and that predicted by the student model, penalizing the student for diverging from the teacher's predictions.

Temperature scaling is a technique used to soften the logits (log-SoftMax outputs) of both the teacher and student models. It involves dividing the logits by a temperature parameter, which controls the softness of the output distributions. The temperature scaling allows the student model to explore a wider range of solutions and improves the knowledge transfer from the teacher model. The softened logits q_i can be expressed as:

$$q_i = \frac{\exp(z_i/T)}{\sum \exp(z_i/T)} \quad (13)$$

, where T is the temperature parameter. z_i is the output of the i^{th} node.

The total loss used for training the student model is a combination of the knowledge transfer component and the regularization component, weighted by hyperparameters:

$$Loss_{Total} = \alpha * Loss_{KT} + \beta * Loss_{Reg} \quad (14)$$

, where α and β are hyperparameters that control the relative importance of the two loss components.

The teacher model is used to generate soft targets by applying temperature scaling to the logits of its predictions on the training data. The student model is trained using the soft targets generated by the teacher model. The total loss is computed based on the distillation loss function, and the student model parameters are updated using gradient-based optimization methods.

The performance of the student model is evaluated on a validation set or test set to assess its accuracy and generalization capability. The student model is considered successful if it achieves comparable performance to the teacher model while being more computationally efficient and compact.

4 Experimental Setup

4.1 Data Collection

To evaluate the effectiveness of the proposed approach, a fault injection experiment was conducted on a six-axis industrial robot. In this experiment, faulty components including reducers and motors were obtained from real-life manufacturing scenarios to replace the normal ones. The drivers in the transmission system were utilized to collect the feedback current of the third-axis motor. The faulty motors and reducers were collected from actual industrial robots from manufacturing works. The faulty parts have minor issues such as abnormal sounds or oil leakage. The faulty parts can continue to be used in the industrial robot operation. Although these faults were subtle, they have the potential to escalate into more severe equipment failures in the future, making their diagnosis crucial. In the experiment, the faulty robot was then deployed for the operation, and the three-phase current signals were captured using a Hall

element in the motor driver. The current signals were collected from the same current signal. Specifically, the currents of the q-axis, with a frequency of 1Hz, were extracted from the collected signals and utilized for subsequent modeling and analysis.

Our experimental study was designed to eliminate the influence of external variables, ensuring that any observed variations were directly attributable to the specific fault conditions introduced. This standardization across tests enables a direct comparison between healthy and faulty states, indicating the effects of different fault types on the current waveform. The dataset used in this study consisted of seven distinct categories, which were further classified into three types of single faults, three types of faults, and a normal category. The faulty parts are the reducers and motors, commonly found in industrial robots. To facilitate data labeling, the one-hot encoding technique was employed. It involves creating a binary vector where each category is represented by a unique index within the vector. By applying one-hot encoding to the dataset, each of the seven categories was labeled, enabling subsequent analysis and modeling of the fault data. After the dataset was labelled, it was then segmented by time window and transformed into a graph structure using the k -NN graph construction approach (T. Li, et al., 2022). The constructed graph data

In Figure 4, from the time domain, it can be seen that the faulty patterns between normal, single-fault and compound faults are obvious. However, the observed differences for different faulty categories in the time-domain signals do not necessarily correlate linearly with the severity of the fault or its impact on operational performance. The time-frequency images for different

fault categories based on Continuous Wavelet Transform (CWT) are presented. The analysis utilizing CWT offered a more interpretable insight into the dynamics of faults. The visualizations of CWT reveal the distribution of signal energy across both temporal and spectral domains, each fault type demonstrating distinct characteristics. Under the 'Normal' operational state, the CWT visualization shows a tightly confined and uniform energy profile at lower frequencies, which is reflective of the robot's stable operation. Conversely, the 'Reducer 3 fault' condition is characterized by the CWT displaying sporadic energy peaks, implying intermittent disruptions at particular intervals that align with the mechanical origin of the fault. The 'Motor 2 fault' scenario, on the other hand, is illustrated by discrete energy across elevated frequencies, signifying a wider range of electrical anomalies. The compound faults such as 'Reducer 1 and Motor 2 fault', exhibited a sophisticated blend of these distinctive patterns, with various energy peaks occurring at different frequencies and times, mirroring the compounded complexity of the faults. Since the faulty data is hard to collect and it causes the data imbalance issue, we have adjusted the samples in each category to alleviate the imbalanced issue. After the adjustment, there are 1400, 1640, 1150, 990, 2005, 1405, and 1900 graph samples in the categories of Normal, Reducer 3, Motor 2, Reducer 4, Reducer1& Motor2, Reducer1& Reducer3 and Reducer3& Reducer4.

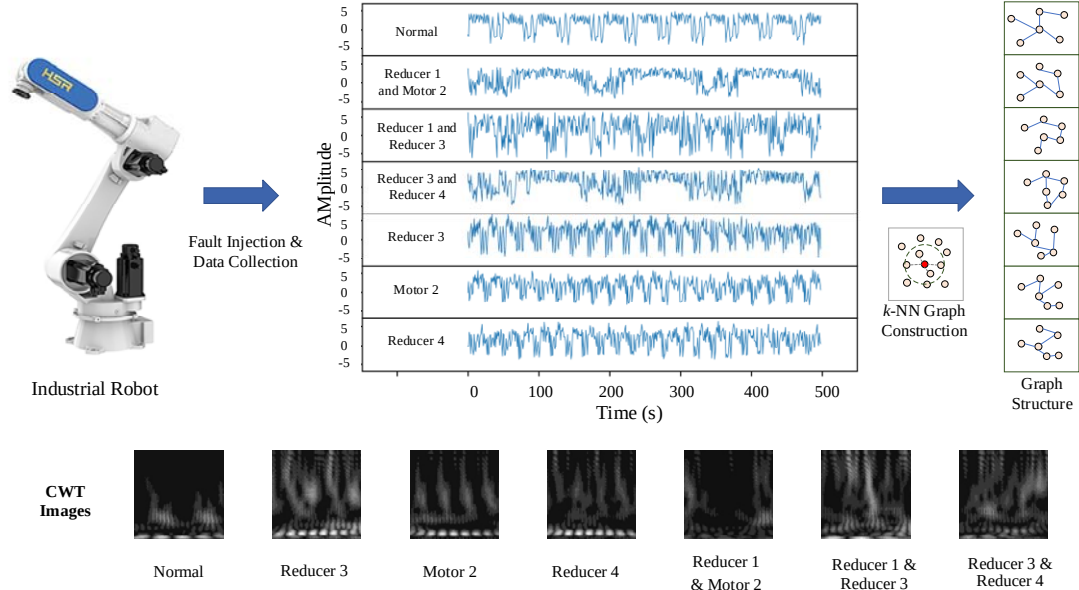


Figure 4. The comparison of signals from different faulty categories.

4.2 Benchmarking Experiments

In this study, two experiments were set up to evaluate the proposed approach. In the first experiment, the algorithm performance of the proposed MSGPAT was revealed. Four graph-based algorithms were adopted for benchmarking experiments. The four graph-based algorithms are GCN (Zhiwen Chen, et al., 2021), GAT (Jiang, Li, Wu, & Li, 2022), SuperGAT (Kim & Oh, 2022), GAT V2 (Brody, Alon, & Yahav, 2021), and multi-scale deep graph convolutional networks (MS-DGCNs) (X. Zhao, et al., 2022). The data pre-processing of the benchmarking algorithms was the same as the proposed algorithm. k -NN was adopted to construct the graph data in this study. The number of nodes in the GCN layer and the data length used for graph construction are essential to the algorithm performance. An experiment was implemented to reveal the impact of both factors on the algorithm performance. The data used in graph construction is controlled by the value of k . Hence, it is worthwhile to identify the relationship between the value of k and the diagnosis performance. The proposed MsGPAT and

the benchmarking algorithms were used for fault diagnosis modelling. In this study, the hyperparameters of the key parameters were determined after several trials. The number of nodes in feedforward layers was set as 512 in the teacher network to enhance the learning capability. The learning rate was set as 0.0001 to ensure stable convergence. The training epoch was set as 30 to allow for sufficient training iterations while avoiding overfitting. The batch size was set as 32 as it strikes a good balance between computational efficiency and effective use of gradient information. Adam was adopted as the optimizer due to its effectiveness in adapting the learning rate. The fault diagnosis accuracy was adopted as the evaluation metric. Meanwhile, 5-fold cross-validation was used to get the average and standard deviation of the results. Besides, the t-SNE technique was adopted in this experiment to further study the insights of the fault diagnosis results.

Table 1. The parameters setting of the proposed MsGPAT.

Parameter	Value
Data length used in graph construction	4/8/16/32/64
Number of nodes in feedforward layer	512
Learning Rate	0.0001
Training Epochs	30
Batch Size	32

In the second experiment, the effectiveness of the KD was evaluated. MsGPAT was adopted in KD. A large-size MsGPAT was designed as the teacher network, while a tiny MsGPAT was designed as the student network. The key parameter that has a significant impact on the model size and model performance is the number of nodes in feedforward layers. The impact of the

number of nodes in the feedforward layer on the model performance and model size was revealed. All tests were performed on an Ubuntu 16.0 server featuring an Intel i9-10920X 3.50Ghz CPU and an Nvidia GeForce RTX 3090 graphics card. The experimental environment is Python 3.6.13 and Pytorch 1.9.1 package was used to develop the algorithms.

5 Experimental Results

5.1 The results of benchmarking experiments of GNNs

In the first experiment, five GNN algorithms were adopted for benchmarking comparison. In the comparison of the GNN algorithms, the MsGPAT stands out in terms of accuracy, which is a critical measure of model performance. Despite its relatively larger model size of 9.02MB, MsGPAT significantly surpasses the other models by achieving an impressive accuracy of $91.3\pm2.76\%$. While other algorithms such as GCN, GAT, SuperGAT, GAT V2, and MS-DGCNs also offer commendable performances within their respective scopes. MS-DGCNs achieve the second highest diagnosis accuracy in this case which is 86.3%. The GCN presents the smallest model size (0.61MB), while its diagnosis accuracy is the worst in this case.

Table 2. The comparison of diagnosis accuracy and model size of different algorithms.

Algorithms	Accuracy	Model size (MB)
GCN (Zhiwen Chen, et al., 2021)	$81.3\pm2.88\%$	0.61
GAT (Jiang, et al., 2022)	$82.5\pm2.89\%$	1.64
GAT V2 (Brody, et al., 2021)	$83.6\pm2.97\%$	3.26
SuperGAT (Kim & Oh, 2022)	$83.4\pm3.47\%$	1.63
MS-DGCNs (X. Zhao, et al., 2022)	$86.3\pm2.48\%$	6.68
MsGPAT (proposed)	$91.3\pm2.76\%$	9.02

Figure 5 demonstrates the results of an experiment investigating the relationship between the

number of nodes and the accuracy of the proposed MsGPAT model, with variations in the sampling length. The x-axis represents the number of nodes, ranging from 4 to 64, while the y-axis shows the accuracy percentage. The bars indicate different sampling lengths: gray for 128, light gray for 256, dark gray for 384, and white for 512. As the number of nodes increases, the overall trend suggests a steady rise in accuracy for all four sampling lengths. However, the magnitude of improvement varies among the sampling lengths. When the number of nodes is low (4 to 16), the accuracy gains are more substantial. As the number of nodes continues to grow (from 32 to 64), the accuracy improvements become minor. Despite this, the accuracy remains consistently high across all sampling lengths, indicating robustness in the model's ability to handle various input sizes.

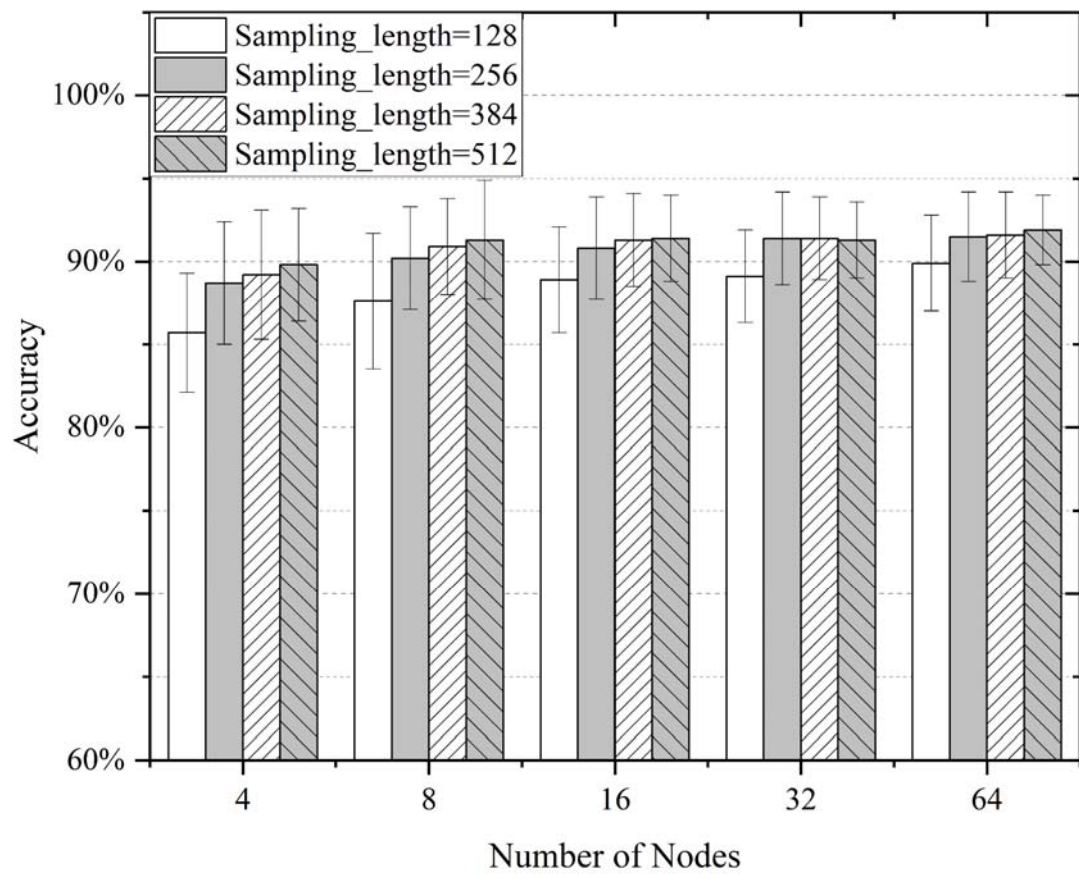


Figure 5. The impact of sampling data length and number of nodes in GCN layer on diagnosis accuracy.

After the investigation of data length and number of nodes in the GCN layer, we selected 384 and 16 as the sampling length and number of nodes in the GCN layers for the subsequent experiments. In order to better understand the performance of the GNN algorithms in this experiment, the confusion matrices of the classification results are shown in Figure 6. It can be seen that all the algorithms show similar performance in the classification of normal and single fault categories, while the performance varies in the classification of compound faults categories. t-SNE was employed to visualize the distribution of learned features. The results of t-SNE are illustrated in Figure 7. It can be seen from the figures that the boundary between the normal category and the faulty categories is clear. Among all the faulty categories, the classification of the single faults is better than the compound faults. It can be seen that the samples in Reducer1_and_Reducer3_faulty and Reducer3_and_Reducer4_faulty are hard to be corrected classified by the benchmarking algorithms.

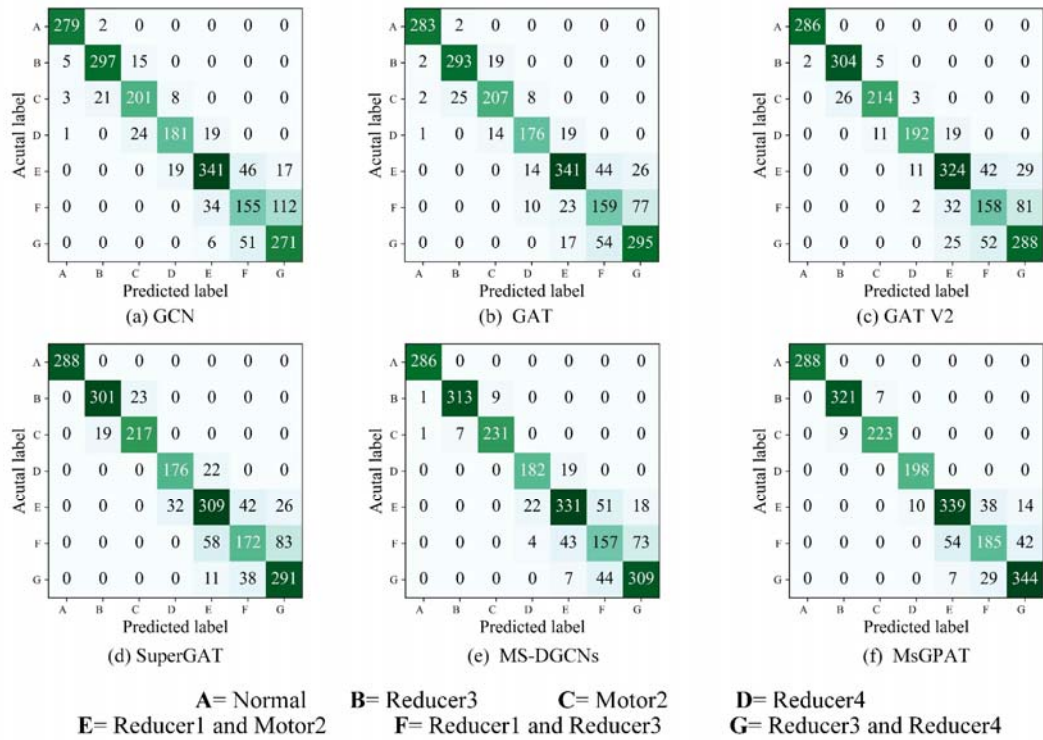


Figure 6. The confusion matrix of fault diagnosis using different GNNs.

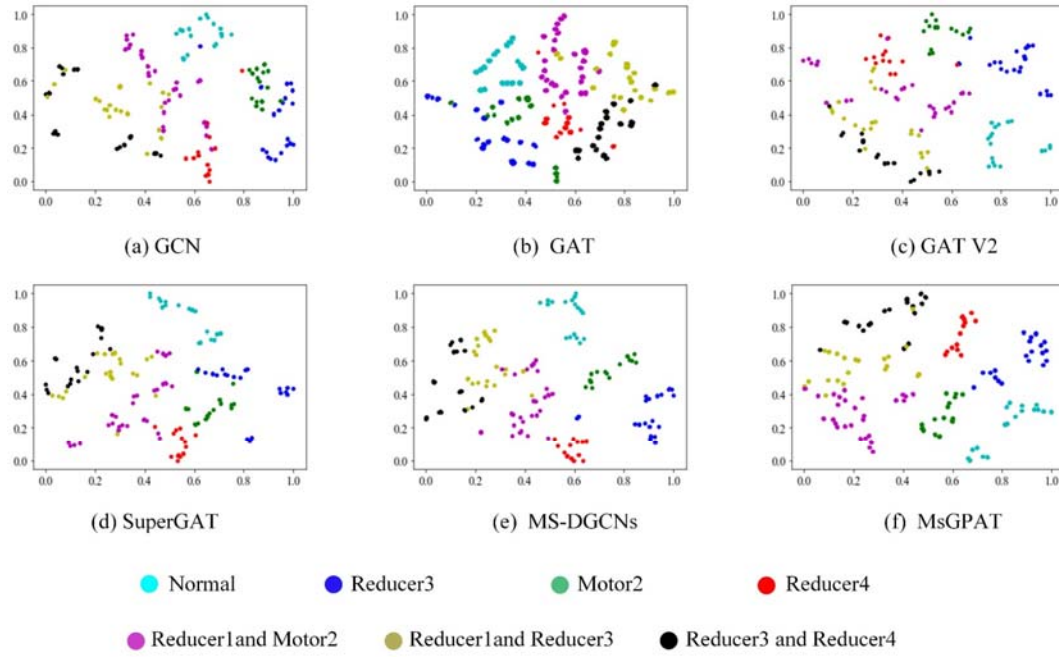


Figure 7. The results of t-SNE for fault diagnosis using different GNNs.

The value of k in k -NN graph construction has great impact on the dimension of the constructed graph. If k is set in a small value, the graph may not contain sufficient faulty patterns. If k is set in a large value, the dimension of the constructed graph becomes large, which greatly increases the training burden of the GNNs. The impact of different k values on the algorithm performance of GNNs is demonstrated in Figure 8. Across all k values, MsGPAT consistently outperforms the other algorithms in diagnosis accuracy, with its performance increasing from 89.4% to 90.9% as k increases from 4 to 32, and maintaining a high accuracy of 90.8% when k arrives 64. Other benchmarking algorithms including GAT, GCN, GAT V2, and MS-DGCNs also show a similar trend of increasing accuracy with increasing k values.

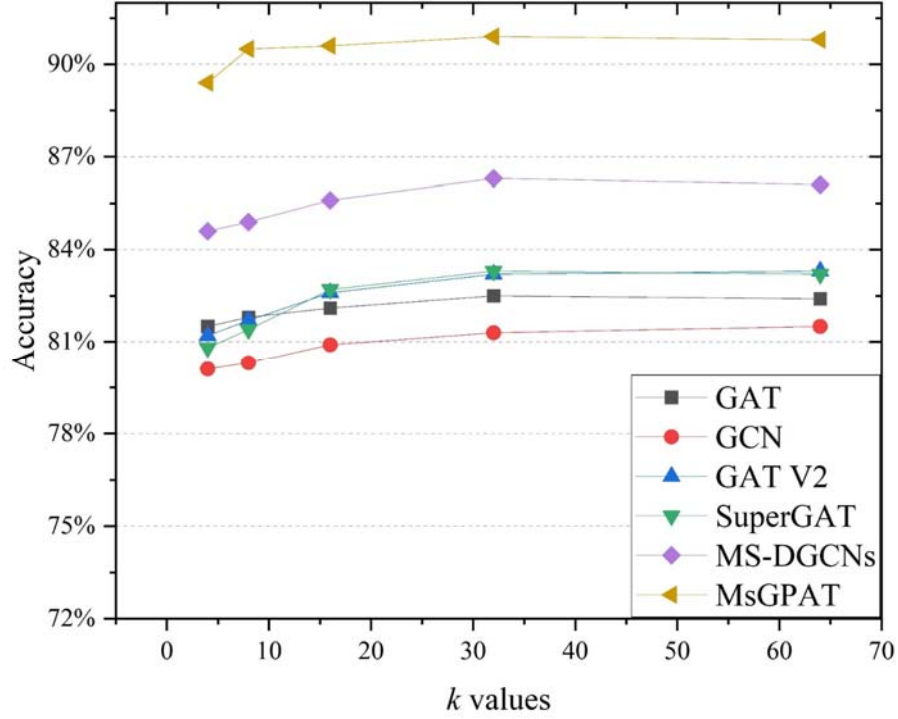


Figure 8. The impact of different k values on the algorithm performance of GNNs.

5.2 The results of knowledge distillation for MsGPAT

In the experiment of KD, a large size MsGPAT was adopted as the teacher network and a small-size MsGPAT was adopted as the student network. The results are shown in Figure 9. It can be seen that there is a promising correlation between the number of nodes in the student network and its accuracy. Starting at a lower accuracy of 85.3% with 64 nodes, the student network exhibited a steady increase in accuracy, reaching a significant 90.4% when the number of nodes was expanded to 512. This trend underscores the effective transfer of knowledge from the teacher network to the student network, demonstrating the successful implementation of the KD process. On the other hand, the teacher network maintained a relatively high and stable accuracy of around 91%, showcasing its robust performance. However, it's worth noting that the model size proportionally increased with the number of nodes, implying a trade-off between

the model performance and its size. Despite this, the KD from the large MsGPAT to the small MsGPAT proved effective, marking a significant advancement in optimizing the balance between model size and accuracy.

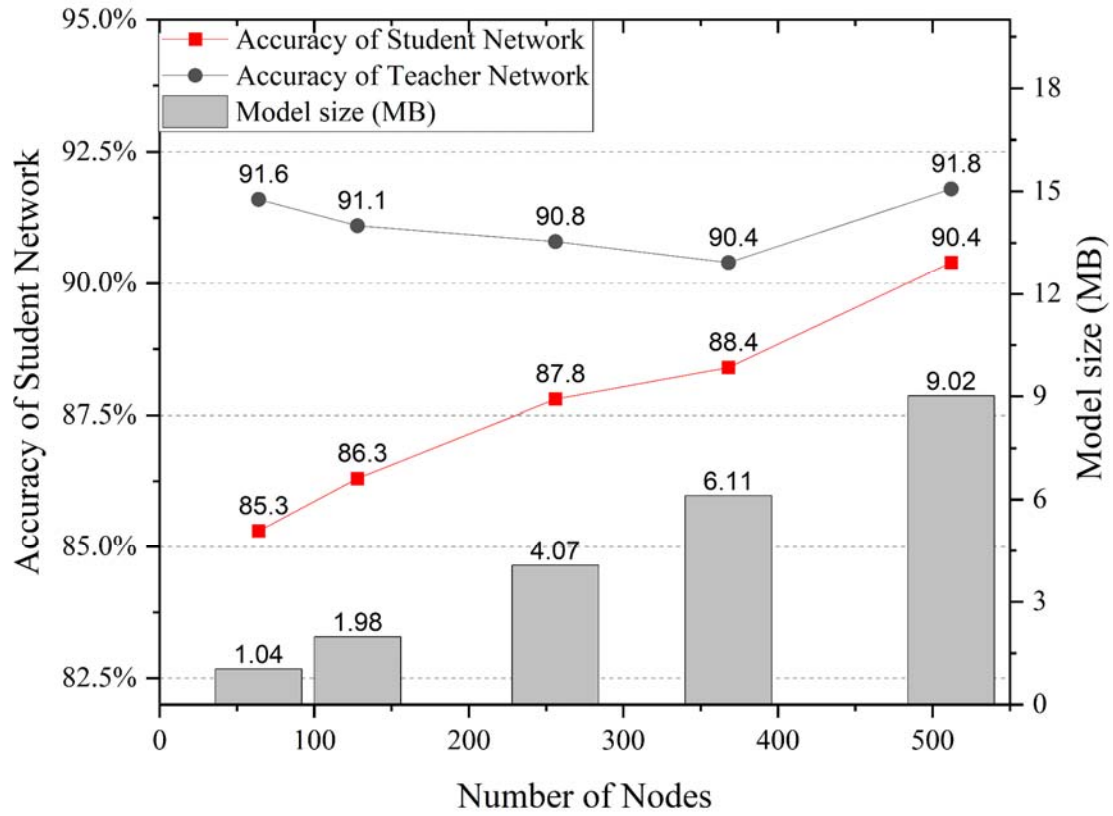


Figure 9. The impact of the number of nodes on the algorithm performance of MsGPAT.

In the comparison experiment between MsGPAT with and without KD, the results in table 3 illustrate that the models utilizing KD consistently outperformed their counterparts that did not apply the technique across all levels of node numbers. Specifically, at the lower end of 64 nodes, the model with KD achieves significantly higher accuracy, leading by 7.8%. This substantial gain in performance indicates the considerable impact of KD on enhancing the model's accuracy, especially when the model size is small. As we increased the number of nodes, the accuracy of both models improved. However, the advantage of applying KD began to decrease, with the

difference in accuracy narrowing from 4.9% at 128 nodes to a mere 0.3% at 512 nodes. Although the benefit of KD diminishes as the model size grows, it's evident that KD offers a notable improvement in accuracy for smaller models, validating its efficacy in the performance enhancement of MsGPAT. An additional experiment that uses GCN (1.73MB) without KD was implemented, which achieved an accuracy of 81.3%. When KD was applied, the accuracy was promoted to 83.7%. Such results indicate that the GCN can show better performance than the proposed MsGPAT without KD when the model parameters are limited. However, when KD is applied, it cannot be helpful for GCN to improve its feature learning capability.

Table 3. The comparison of diagnosis accuracy of MsGPAT with and without KD.

	Number of Nodes	Size (MB)	Accuracy (with KD)	Accuracy (without KD)	Difference
MsGPAT	64	1.04	85.3 $\pm$ 3.13%	77.5 $\pm$ 3.96%	7.8%
	128	1.98	86.3 $\pm$ 2.81%	81.4 $\pm$ 3.54%	4.9%
	256	4.07	87.8 $\pm$ 2.76%	85.3 $\pm$ 3.69%	2.5%
	384	6.11	88.4 $\pm$ 2.82%	87.1 $\pm$ 2.97%	1.3%
	512	9.02	90.4 $\pm$ 2.77%	90.1 $\pm$ 2.66%	0.3%
GCN	256	1.73	83.7 $\pm$ 2.44%	81.3 $\pm$ 2.10%	2.4%

5.3 Discussion

With the advance of GNNs, an accurate fault diagnosis model can be established, which is beneficial to the operation and maintenance of industrial robots. However, the large model size of GNN impedes the deployment in edge devices. While GCN has been widely successful in graph representation learning, it has certain limitations that the proposed MsGPAT addresses. Firstly, GCN Lack of multi-scale sensitivity. Traditional GCNs aggregate neighborhood information at a fixed scale, which may lead to a loss of fine-grained details or fail to capture macro-level structures in the graph. Secondly, GCN suffers from over-smoothing problems. As

the number of layers increases in a GCN, there is a tendency for node representations to become indistinguishable due to the repeated application of smoothing operations. This phenomenon, known as over-smoothing, can limit the depth of effective GCN architectures and hinder their ability to learn complex hierarchical patterns within the graph data. Different from GCN, the proposed MsGPAT can extract features from different granularities by the multi-scale feature extraction module and overcome over-smoothing issues by the pyramid attention mechanism.

This study proposes a MsGPAT with KD to address this challenge. The proposed MsGPAT has demonstrated superior performance in terms of accuracy across different experiments. In the first experiment, compared to other prevailing algorithms including GCN, GAT, SuperGAT, GAT V2, and MS-DGCNs, MsGPAT consistently achieved the highest diagnosis accuracy, making it the top performer among all the benchmarking algorithms. GAT adopts GCN for feature extraction and MSA to learn the node representation. Different from GAT, the proposed MsGPAT designs multi-scale GCNs to capture the temporal-spatial features from different granularities and a pyramid attention to learn the node representation. Compared with the MS-DGCNs, MsGPAT achieved better performance with the introduction of pyramid attention. Pyramid attention can effectively learn the faulty patterns obtained from multi-scale GCNs, which greatly leverages the algorithm performance without significantly increasing the model size. When evaluating the algorithm performance against different k values in k -NN, the algorithm performance of MsGPAT remains remarkably consistent. The size of the constructed graph data is affected by the k values. When the k values arrive 32, the diagnosis accuracy of all the algorithms are not further promoted, which indicates that 32 is a suitable number for

fault diagnosis modelling.

By comparing the CWT images in Figure 4 for different fault conditions, we observed that the proposed method captures the minor differences that traditional methods often overlook. For example, the CWT image for the reducer 4 single fault displays a unique pattern, characterized by higher frequency content and more erratic fluctuations compared to other faults, whose characteristics are complex. The graph structure effectively encodes the temporal and spatial dependencies within the signals, allowing the model to identify the underlying relationships between different fault conditions. Our proposed algorithm, equipped with multi-scale feature extraction can capture these intricate signal features from different levels. Subsequently, the pyramid attention mechanism can identify the key patterns relevant to the faulty category from the multi-scale features. It focuses on the most discriminative features, enabling the model to distinguish between similar-looking faults and improve classification accuracy. The experimental results provide evidence of the benefits of integrating graph-based representations and attention mechanisms in our model, ultimately leading to enhanced performance and accuracy in fault diagnosis for industrial robots.

The high diagnosis accuracy of MsGPAT becomes evident in the implementation of KD. The process of KD, where the learned knowledge from a larger, complex teacher model is transferred to a smaller student model, has proven to be effective in enhancing the accuracy of MsGPAT. When the model size of the student network was reduced to 1.04MB, the student network achieved 85.3% diagnosis accuracy, which is acceptable for the actual deployment.

KD allows a smaller MsGPAT model to achieve comparable performance to much larger MsGPAT models. This is particularly valuable in edge computing scenarios where computational resources are limited. The training of the teacher and student models can be accomplished in the cloud side and the student models can be deployed on the edge side.

In this study, the spatial-temporal features were collected via multi-scale GCNs, which utilize three different brunches for feature extraction. Such a design increases the computational in feature extraction. GCN provides localized spatial modeling to uncover inter-component relationships crucial for diagnosis. Pyramid attention builds on this by enabling long-range temporal modeling across multiple timescales to capture fault evolution. This integrated design allows holistic modeling of the spatial dependencies and temporal dynamics of robotic monitoring data for enhanced diagnosis. KD bridges the accuracy-complexity gap by transferring knowledge from the accurate but complex model to a compressed one suited for edge resources. Meanwhile, fault diagnosis is always restricted by the insufficiency of faulty samples. The performance of the proposed MsGPAT can be jeopardized when the faulty samples are limited. Meta-learning is a useful technique for few-shot learning, which will be further investigated in conjunction with the proposed MsGPAT in future works. Finally, the distilled MsGPAT can achieve 85.3% diagnosis accuracy for the testing set. However, the experiment was conducted in a high-performance server, which has not been tested in the edge computing device. Therefore, future work will focus on the deployment and fine-tuning of the MsGPAT in the edge device such as FPGA.

6 Conclusions

Fault diagnosis is essential to the operation and maintenance of industrial robots. In order to deploy a deep learning model with satisfactory performance and smaller model sizes, this paper presents a fault diagnosis approach for industrial robots based on MsGPAT and KD. In the MsGPAT, the multi-granularity features are extracted by multi-scale GCN and then the salient features are further learned by the pyramid attention module. Subsequently, KD is introduced to compress the model size of MsGPAT, which enables it to be deployed in the edge device. The experimental results indicate that the proposed MsGPAT shows merits in comparison with the state-of-the-art algorithms. The proposed approach balances modeling performance and complexity for edge computing constraints. It provides an integrated solution catered to the spatial and temporal feature extraction and hardware intricacies of accurate robotic fault diagnosis without excessive data transmission or computing needs. The MsGPAT with KD holds great promise for practical implementation in the operation and maintenance of industrial robots, paving the way for accurate fault diagnosis in edge deployment.

Acknowledgement

Our work is supported by multiple funds in China, including the Key Program of NSFC-Guangdong Joint Funds (U2001201), the National key R & D project (2021YFB3301802), the Natural Science Foundation of Guangdong Province (2020B1515120010). Our work is also supported by Guangdong Provincial Key Laboratory of Cyber-Physical System (2020B1212060069).

References

- Brody, S., Alon, U., & Yahav, E. (2021). How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*.
- Chen, C., Liu, C., Wang, T., Zhang, A., Wu, W., & Cheng, L. (2023). Compound fault diagnosis for industrial robots based on dual-transformer networks. *Journal of Manufacturing Systems*, 66, 163-178.
- Chen, C., Wang, T., Liu, C., Liu, Y., & Cheng, L. (2023). Lightweight Convolutional Transformers Enhanced Meta Learning for Compound Fault Diagnosis of Industrial Robot. *IEEE Transactions on Instrumentation and Measurement*.
- Chen, Z., Chen, J., Liu, S., Feng, Y., He, S., & Xu, E. (2022). Multi-channel Calibrated Transformer with Shifted Windows for few-shot fault diagnosis under sharp speed variation. *ISA transactions*, 131, 501-515.
- Chen, Z., Xu, J., Peng, T., & Yang, C. (2021). Graph convolutional network-based method for fault diagnosis using a hybrid of measurement and prior knowledge. *IEEE Transactions on Cybernetics*, 52, 9157-9169.
- Deng, J., Jiang, W., Zhang, Y., Wang, G., Li, S., & Fang, H. (2021). HS-KDNet: A Lightweight Network Based on Hierarchical-Split Block and Knowledge Distillation for Fault Diagnosis With Extremely Imbalanced Data. *IEEE Transactions on Instrumentation and Measurement*, 70, 1-9.
- Fang, H., Deng, J., Chen, D., Jiang, W., Shao, S., Tang, M., & Liu, J. (2023). You can get smaller: A lightweight self-activation convolution unit modified by transformer for fault diagnosis. *Advanced Engineering Informatics*, 55, 101890.
- Ghorvei, M., Kavianpour, M., Beheshti, M. T., & Ramezani, A. (2023). Spatial graph convolutional neural network via structured subdomain adaptation and domain adversarial learning for bearing fault diagnosis. *Neurocomputing*, 517, 44-61.
- Guan, L., Qiao, F., Zhai, X., & Wang, D. (2022). Model Evolution Mechanism for Incremental Fault Diagnosis. *IEEE Transactions on Instrumentation and Measurement*, 71, 1-11.
- Ji, M., Peng, G., Li, S., Cheng, F., Chen, Z., Li, Z., & Du, H. (2022). A neural network compression method based on knowledge-distillation and parameter quantization for the bearing fault diagnosis. *Applied Soft Computing*, 127, 109331.
- Jiang, L., Li, X., Wu, L., & Li, Y. (2022). Bearing fault diagnosis method based on a multi-head graph attention network. *Measurement Science and Technology*, 33, 075012.
- Keshun, Y., Puzhou, W., & Yingkui, G. (2024). Towards efficient and interpretative rolling bearing fault diagnosis via quadratic neural network with BI-LSTM. *IEEE Internet of Things Journal*.
- Kim, D., & Oh, A. (2022). How to find your friendly neighborhood: Graph attention design with self-supervision. *arXiv preprint arXiv:2204.04879*.
- Li, F., Chen, J., He, S., & Zhou, Z. (2021). Layer regeneration network with parameter transfer and knowledge distillation for intelligent fault diagnosis of bearing using class unbalanced sample. *IEEE Transactions on Instrumentation and Measurement*, 70, 1-10.
- Li, Q. (2023). A comprehensive survey of sparse regularization: Fundamental, state-of-the-art methodologies and applications on fault diagnosis. *Expert Systems with Applications*,

- Li, T., Zhao, Z., Sun, C., Yan, R., & Chen, X. (2020). Multireceptive field graph convolutional networks for machine fault diagnosis. *IEEE Transactions on Industrial Electronics*, 68, 12739-12749.
- Li, T., Zhou, Z., Li, S., Sun, C., Yan, R., & Chen, X. (2022). The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study. *Mechanical Systems and Signal Processing*, 168, 108653.
- Li, Y., Hu, F., Liu, Y., Ryan, M., & Wang, R. (2023). A hybrid model compression approach via knowledge distillation for predicting energy consumption in additive manufacturing. *International Journal of Production Research*, 1-23.
- Li, Y., Zhou, Z., Sun, C., Chen, X., & Yan, R. (2022). Variational Attention-Based Interpretable Transformer Network for Rotary Machine Fault Diagnosis. *IEEE transactions on neural networks and learning systems*.
- Lin, J., Shao, H., Zhou, X., Cai, B., & Liu, B. (2023). Generalized MAML for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals. *Expert Systems with Applications*, 230, 120696.
- Liu, S., Yu, H., Liao, C., Li, J., Lin, W., Liu, A. X., & Dustdar, S. (2021). Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations* (pp. 1-20).
- Liu, Y., Chen, C., Wang, T., & Cheng, L. (2022). An attention enhanced dilated CNN approach for cross-axis industrial robotics fault diagnosis. *Autonomous Intelligent Systems*, 2, 1-11.
- Liu, Z., Lyu, W., Wang, C., Guo, Q., Zhou, D., & Xu, W. (2022). D-CenterNet: An Anchor-Free Detector With Knowledge Distillation for Industrial Defect Detection. *IEEE Transactions on Instrumentation and Measurement*, 71, 1-12.
- Long, J., Zhang, R., Yang, Z., Huang, Y., Liu, Y., & Li, C. (2022). Self-Adaptation Graph Attention Network via Meta-Learning for Machinery Fault Diagnosis With Few Labeled Data. *IEEE Transactions on Instrumentation and Measurement*, 71, 1-11.
- Lu, S., Lu, J., An, K., Wang, X., & He, Q. (2023). Edge computing on IoT for machine signal processing and fault diagnosis: A review. *IEEE Internet of Things Journal*.
- Ma, W., Liu, R., Guo, J., Wang, Z., & Ma, L. (2023). A collaborative central domain adaptation approach with multi-order graph embedding for bearing fault diagnosis under few-shot samples. *Applied Soft Computing*, 110243.
- Pu, Z., Cabrera, D., Li, C., & Valente de Oliveira, J. (2023). Sliced Wasserstein cycle consistency generative adversarial networks for fault data augmentation of an industrial robot. *Expert Systems with Applications*, 222, 119754.
- Sun, K., Huang, Z., Mao, H., Qin, A., Li, X., Tang, W., & Xiong, J. (2021). Multi-scale cluster-graph convolution network with multi-channel residual network for intelligent fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, 71, 1-12.
- Tang, Y., Zhang, X., Qin, G., Long, Z., Huang, S., Song, D., & Shao, H. (2021). Graph cardinality preserved attention network for fault diagnosis of induction motor under varying speed and load condition. *IEEE Transactions on Industrial Informatics*, 18, 3702-3712.
- Wang, H., Wang, J., Zhao, Y., Liu, Q., Liu, M., & Shen, W. (2022). Few-Shot Learning for Fault

- Diagnosis With a Dual Graph Neural Network. *IEEE Transactions on Industrial Informatics*, 19, 1559-1568.
- Wu, H., Xu, X. a., Xin, C., Liu, Y., Rao, R., Li, Z., Zhang, D., Wu, Y., & Han, S. (2023). Intelligent fault diagnosis for triboelectric nanogenerators via a novel deep learning framework. *Expert Systems with Applications*, 226, 120244.
- Xu, J., Ke, H., Chen, Z., Fan, X., Peng, T., & Yang, C. (2022). Oversmoothing Relief Graph Convolutional Network-Based Fault Diagnosis Method With Application to the Rectifier of High-Speed Trains. *IEEE Transactions on Industrial Informatics*, 19, 771-779.
- Yu, X., Tang, B., & Deng, L. (2023). Fault diagnosis of rotating machinery based on graph weighted reinforcement networks under small samples and strong noise. *Mechanical Systems and Signal Processing*, 186, 109848.
- Zhang, Y., Ji, J., Ren, Z., Ni, Q., Gu, F., Feng, K., Yu, K., Ge, J., Lei, Z., & Liu, Z. (2023). Digital twin-driven partial domain adaptation network for intelligent fault diagnosis of rolling bearing. *Reliability Engineering & System Safety*, 234, 109186.
- Zhao, J., Qian, X., Zhang, Y., Shan, D., Liu, X., Coleman, S., & Kerr, D. (2023). A knowledge distillation-based multi-scale relation-prototypical network for cross-domain few-shot defect classification. *Journal of Intelligent Manufacturing*, 1-17.
- Zhao, X., Yao, J., Deng, W., Ding, P., Zhuang, J., & Liu, Z. (2022). Multiscale deep graph convolutional networks for intelligent fault diagnosis of rotor-bearing system under fluctuating working conditions. *IEEE Transactions on Industrial Informatics*, 19, 166-176.
- Zhou, X., Zeng, H., Chen, C., Xiao, H., & Xiang, Z. (2022). An attention-enhanced multi-modal deep learning algorithm for robotic compound fault diagnosis. *Measurement Science and Technology*, 34, 014007.