Causal-Trivial Attention Graph Neural Network for Fault Diagnosis of Complex Industrial Processes

Hao Wang, Ruonan Liu, *Member, IEEE*, Steven X. Ding, Qinghua Hu, *Senior Member, IEEE*, Zengxiang Li, and Hongkuan Zhou

***Abstract*—In modern industrial systems, components have complex interactions with each other, which makes it become a challenging task to identify the operational conditions of industrial systems. Considering that an industrial system, the embedded components and their interactions can be expressed as nodes and edges in a graph, respectively. Therefore, graph representation algorithms are powerful tools for fault diagnosis of industrial systems. As one of the most commonly used graph representation algorithms, graph neural networks (GNN) mainly follow the law of "learning to attend." GNN extract training data features learn the statistical correlations between features and labels, resulting in the attended graph favoring for accessing noncausal features as a shortcut for prediction. This shortcut feature is unstable and depends on the data distribution characteristics in the training dataset, which reduces the generalization ability of the classifier. By performing the causal analysis of GNN modeling for graph representation, the results show that shortcut features act as confounding factors between causal features and predictions, causing classifiers to learn wrong correlations. Therefore, to discover patterns of causality and weaken the confounding effects of shortcut features, a causal-trivial attention graph neural network strategy is proposed. First, node and edge representations are given by estimating soft masks. Second, through disentanglement, both causal features and shortcut features are obtained from the graph. Third, the backdoor adjustment of the causal theory is parameterized to combine each causal feature with a variety of shortcut features. Finally, comparative experiments on the three-phase flow facility dataset illustrate the effectiveness of the proposed method.**

***Index Terms*—Causal intervention, complex industrial processes, fault diagnosis, graph neural networks (GNN).**

Manuscript received 12 February 2023; revised 25 April 2023; accepted 24 May 2023. Date of publication 8 June 2023; date of current version 19 January 2024. This work was supported in part by the National Science and Technology Major Project under Grant 2017-I-0007-0008, in part by the National Natural Science Foundation of China under Grant 62206199, in part by the Tianjin Applied Basic Research Project under Grant S22QNA927, in part by the Alexander von Humboldt Foundation under Grant 1226831, in part by the CCF-Baidu Pinecone Foundation under Grant CCF-BAIDU OF2022020, and in part by the Open Research Fund of State Key Laboratory of High Performance Complex Manufacturing, Central South University under Grant Kfkt2022-10. Paper no. TII-23-0472. *(Corresponding author: Ruonan Liu.)*

Hao Wang, Ruonan Liu, and Qinghua Hu are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China, and also with the Key Laboratory for Machine Learning of Tianjin, Tianjin University, Tianjin 300350, China (e-mail: 3020244104@tju.edu.cn; ruonan.liu@tju.edu.cn; huqinghua@tju.edu.cn).

Steven X. Ding is with the School of Automation, University of Duisburg-Essen, 47057 Duisburg, Germany (e-mail: steven.ding@uni-due.de).

Zengxiang Li is with the Digital Research Institute, ENN Group, Langfang Economic and Technological Development Zone Garbage Health Landfill, Hebei 301739, China (e-mail: lizengxiang@enn.cn).

Hongkuan Zhou is with the Science and Technology on Thermal Energy and Power Laboratory, Wuhan 2nd Ship Design and Research Institute, Wuhan 430205, China (e-mail: hongkuanzhou@foxmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2023.3282979.

Digital Object Identifier 10.1109/TII.2023.3282979

ACRONYMS AND ABBREVIATIONS

AUC	Area under curve.
CTA-GNN	Causal-trivial attention graph neural network.
GAT	Graph attention networks.
GNN	Graph neural networks.
IAGNN	Interaction-aware graph neural networks.
KNN	*K*-nearest neighbor.
MLPs	Multilayer perceptrons.
PKT-MCNN	Progressive knowledge transfer-based multi-task convolutional neural network.
ROC	Receiver operating characteristic.
SCM	Structural causal model.
t-SNE	*t*-distribution stochastic neighbor embedding.
TFF	Three-phase flow facility.

I. Introduction

As TECHNOLOGY continues to evolve in the industrial field, the cost of running industrial systems and processes in a factory increases exponentially. Therefore, an effective diagnosis system is needed to monitor the industrial process, replace manual detection methods, reduce maintenance costs to secure industrial systems. The diagnosis system includes multiple measurement devices, but because of the scale and complexity of today's industrial processes, these device readouts have high dimensions and complex interactions [1]. Therefore, manual fault identification in the past is not advisable, and it was necessary to rely on the deep architecture of the multilayer nonlinear data processing unit in the deep learning algorithm for feature learning to identify faults. After research, the fault diagnosis of complex industrial processes is a classification

problem using multivariate time-series signals, which has received extensive attention [2], [3], [4]. Due to the complex interaction and close connection between various components, once a fault occurs in a certain position, multiple components may produce abnormal readings, affecting the operation of the entire process [5]. In addition, different faults can lead to abnormal readings of different components, and the relationship between components is difficult to find. For fault diagnosis, it is necessary to extract the interactions between multiple components and learn hidden information in time-series signals.

Most of the current mainstream industrial process fault diagnosis methods are model-based and data-driven techniques [6], but both of them have their limitations. As the systems become more and more complex and changeable, model-based methods [7] may not be as suitable for modern industrial systems, so scholars have proposed data-driven methods. The data-driven methods detect anomalous variables or identify fault types by extracting data features and performing statistical analysis [8] or learning discriminative feature [9], [10] based on these features.

The existing data-driven fault diagnosis methods mostly emphasize the signals of several independent components in industrial systems, while ignoring the interactions between different components. Often, there are complex interactions among components. For example, if a certain part of the system fails, multiple components related to it will generate abnormal signals. By fusing data from multiple components, fault diagnosis can be better achieved. To mine complex interactions among components, graph data with topological structure [11] can accomplish this task. Therefore, structural property graphs are used to describe industrial process data, where each component corresponds to a node, and the edges between nodes can be learned based on the similarity of component signals. The component signals in different faulty modes are different, and the learned edges are also different. Then, different topological graph structures are obtained. By learning such a graph structure, the hidden fault information can be mined, which allows us to identify the fault type [9], enables to transform the task into a graph recognition problem.

In order to distinguish the topological graph structures of different faulty modes, it is very important to get a graph construction method that can represent the signal information of components and the interaction between components. The interaction among components can be explored by building a graph based on the similarity relations among the components, and the K-nearest neighbor (KNN) [12] algorithms can be used to obtain the edges connecting the components. There is also a need for a graph classification algorithm that can learn fault information-oriented graph representations while maintaining graph specificity, thereby distinguishing topological graphs of different faults.

Graph neural networks (GNN) [13] algorithms applied in various fields have shown excellent graph recognition performance and can be used to learn error-oriented graph representations, which is an efficient way to fuse information from multiple components through a message passing mechanism. However, due to the complex structure of the graph and the large amount of signals, it is very important to find the key parts of the input graph and filter out the irrelevant parts with the help of the powerful representation learning ability of GNN [14]. For example, finding normal signals and abnormal signals in multiple components is of vital importance. The abnormal signals belong to the key parts, and different subgraphs are classified into corresponding fault types based on the key parts. Attention [15] and pooling [16] learning methods that are currently widely used adopt the law of "learning to attend." These methods mine the hidden mutual information between the attended graph and the real label, leading to the utilization of shortcut features for decision making. These features arise from factors such as sample selection or environmental noise, and that they are discriminative but noncausal. However, this ability is limited to the same distribution as the training dataset. The distribution characteristics of the test data are usually different from the training data, resulting in learning shortcut features that can only show good performance on the training set and have poor generalization ability, which hinders them in key deployment in the application. Then, a method should be used to avoid learning shortcut features, but to find the essence that affects the classification effect, that is, causal features.

To address this issue, a causal-trivial attention graph neural network (CTA-GNN) strategy [17] is proposed. This strategy promotes the attended graph to learn the causal features of the input and alleviates the interference of the shortcut features, thereby maximizing their causal influence on the predicted labels, which solves the above problems. Specifically, an attention module is first added to the input graph to generate estimates of causal and shortcut features. Then, the backdoor adjustment formula is parameterized based on the causality, combining causal estimates with shortcut estimates and making stable predictions. Finally, the CTA-GNN strategy is used for graph recognition, and experimental results on the three-phase flow facility (TFF) dataset demonstrate the effectiveness of CTA-GNN.

The main contributions of our work are summarized as follows:

1) Toward the current generalization problem of attention-based and pooling-based GNN in fault diagnosis, a causal GNN framework is proposed, which attributes the problem to the confounding effect of shortcut features.
2) Aiming at the problem that shortcut features and causal features are difficult to deal with, a CTA-GNN strategy is proposed to filter out shortcut features while mining causal features. The strategy is divided into three stages: estimate the soft mask, disentanglement, and causal intervention.
3) Experiments on TFF datasets demonstrate the effectiveness of CTA-GNN, and more visualizations and in-depth analysis demonstrate the interpretability and rationality of CTA-GNN.
4) By comparing with existing fault diagnosis algorithms, it is proved that the CTA-GNN model can filter out shortcut features while learning causal features, perform more stable classification, and have better generalization ability.

II. Problem Formulation and Preliminaries

A. Problem Formulation

1) Component Signal Fragment: The extraction of signals is usually obtained by components. Different components are distributed in various positions in the industrial system, so the generated signals form n original measurement variables. During time t, the signal fragment generated by the ith component is $s_i = (s_i^1, s_i^2, s_i^3 \ldots, s_i^t)$. However, since the industrial system runs for a long time, the signal fragment obtained has a large span and is usually difficult to handle, so obtaining several signal fragments through window sliding is needed, which can be expressed as $w_j = (s_i^{t-m+1}, s_i^{t-m+2}, \ldots, s_i^t) \in \Omega$. Since signal fragments are stable over a short period of time and do not change greatly, they can be used as input for graph-structured modeling.

2) Input Graph: The input graph is denoted by $G = \{V, E\}$ with vertices $v_i \in V$ and edges $e_{i,j} \in E$, where vertices represent components in industrial systems and the edges represent the correlations between them. The adjacency matrix $A \in \mathbb{R}^{n\times n}$ is used to record the details of the entire graph, where $A[i, j] = 1$ if edge $(v_i, v_j \in E)$, otherwise $A[i, j] = 0$. The node features can describe the component signal fragment, which is expressed symbolically as $X \in \mathbb{R}^{n\times m}$, m is the size of signal fragments. $GConv(\cdot)$ represents the GNN module, where $H \in \mathbb{R}^{n\times d}$ represents the node representation matrix.

B. Attention Mechanism in GNN

The attention mechanism can focus on key information and filter out unimportant information. In GNN, the attention mechanism can be used on nodes or edges, which can help us find the key parts of the whole graph, and these key subgraphs can better help us accomplish the task goal.

For the edge-level attention mechanism, the attention matrix $M_{edge} \in \mathbb{R}^{n\times n}$ is constructed using parameters and node representations. Some studies pass weighted messages to diffuse node information and aggregate information from other nodes to represent node information.Then get the updated node representations H′:

$$H' = GConv(A \odot M_{\text{edge}}, H). \tag{1}$$

For the node-level attention mechanism, $M_{\text{node}} \in \mathbb{R}^{n\times 1}$ denotes the attention matrix, which can be obtained using a neural network. To get the most attentive node representations, some studies give the self-attention masks

$$H' = GConv(A, H \odot M_{\text{node}}). \tag{2}$$

In the above two equations, $\odot$ represents the Hadamard product, that is, the product of corresponding elements. Then, perform further pooling operations on the output node representation H^{out} and give the graph representation h_G by the readout function $f_{\text{readout}}(\cdot)$

$$h_G = f_{\text{readout}}(h_i^{\text{out}} | i \in V). \tag{3}$$

Finally, the graph representation is transformed into a probability distribution z_G. The classifier Φ can be used

$$z_G = \Phi(h_G). \tag{4}$$

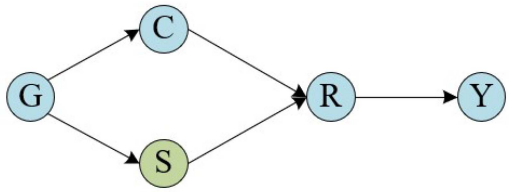

Fig. 1. Structural causal model.

They minimize the following experiential risks, following the law of "learning to attend":

$$\mathcal{L}_{CE} = -\frac{1}{|\mathcal{D}|}\sum_{G\in\mathcal{D}} \mathbf{y}_G^{\top}\log(z_G) \tag{5}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss [19] computed on the training data $\mathcal{D}$. y_G is the ground-truth label. Since this empirical loss will depend on the distribution characteristics and statistical correlations of the training data. Thus, this learning strategy obtains predictive shortcut features without finding key causal features.

C. Causal-Trivial Attention

To solve the above problems, both causal and trivial attention mechanisms are needed on the input graph to find both causal and shortcut features. Since the causal feature is the essential feature to distinguish different fault topology graphs, the corresponding label of the graph representation learned by the causal attended graph should be considered as ground-truth. While the trivial attended graph is complementary to the causal attended graph, the labels corresponding to the graph representations it learns cannot fully distinguish fault representations, so its predictions are averaged over all classes. Our goal is to learn these two attended graph representations to obtain causal features and shortcut features, which can be applied to fault diagnosis.

III. Proposed Fault Diagnosis Method

In the following, a CTA-GNN framework based on information of multivariate signal fragments is proposed. First, the existing problems in GNN learning are analyzed from the perspective of causality, and shortcut features are identified as confounding factors between causal features and predictions. Then propose a CTA-GNN framework to weaken confounding effects and improve model generalization. The framework consists of three essential parts: 1) estimate the soft masks, giving node and edge representations; 2) disentanglement, get causal graph and trivial graph through two loss functions; and 3) causal intervention, the causal intervention diagram is obtained through the backdoor adjustment formula, and the learning goal of CTA-GNN is given.

A. Insights Into GNN From a Causal Perspective

According to the process of the GNN model, the relationship between variables into a structural causal model (SCM) can be built, as shown in Fig. 1. In this figure, the arrows represent the causal relationship, and the model can clearly show the causal

relationship between the five variables. SCM is explained as follows:

1) $C \leftarrow G \rightarrow S$: The variable S represents the shortcut feature. It is often caused by the chance of sample selection or learned environmental noise features. The variable C represents the causal feature. It can reflect the essential properties of the graph G. The causal effect is established due to the coexistence of shortcut features S and causal features C.
2) $C \rightarrow R \leftarrow S$: The variable R represents the graph representation obtained from both the shortcut feature S and the causal feature C and then get the input of GNN learning strategy.
3) $R \rightarrow Y$: After obtaining the graph representation R, it is used as the basis for learning a classifier that we can use to classify the input graph, denoted by the variable Y.

By learning the structural causal model of GNN, it has been found that there is a backdoor path between C and Y, i.e., $C \leftarrow G \rightarrow S \rightarrow R \rightarrow Y$. There is no doubt that the shortcut feature S is a confounding factor between C and Y. Because of the existence of this backdoor path, C and Y form a false relationship, and the shortcut feature S needs to be avoided. In order for the model to classify graphs according to the causal feature C and obtain accurate classification results, blocking the backdoor path is of vital importance.

B. Backdoor Adjustment to Block the Backdoor Path

In order to remove the influence of the confounding factor S, and then make the model use causal features for classification, it is necessary to eliminate the backdoor path. The solution provided by causality theory can be used: perform do calculation on the causal feature C, and then get $P_m(Y|C) = P(Y|do(C))$ to block the backdoor path. The characteristics of the shortcut will not change due to blocking the backdoor path, so the marginal probability $P(S = s)$ is stable under the intervention, that is, $P(s) = P_m(s)$. In addition, the causal effect between C and S has nothing to do with Y's response to C and S, they are independent of each other, and then the conditional probability $P(Y|C, s)$ is also constant, that is, $P_m(Y|C, s) = P(Y|C, s)$. Finally, after the causal intervention, the causal feature C and the shortcut feature S are independent, that is, $P_m(s|C) = P_m(s)$.

Based on the above statement, the following equation can be obtained:

$$\begin{aligned} P(Y|\text{do}(C)) &= P_m(Y|C) \\ &= \sum_{s\in\tau} P_m(Y|C,s)P_m(s|C) \quad \text{(Bayes Rule)} \\ &= \sum_{s\in\tau} P_m(Y|C,s)P_m(s) \quad \text{(Independency)} \\ &= \sum_{s\in\tau} P(Y|C,s)P(s) \end{aligned} \tag{6}$$

where τ represents the confounding set; $P(Y|C, s)$ represents the conditional probability of the causal feature C and confounding factor s, and $P(s)$ is the prior probability of confounding factor s. Equation (6) is often referred to as backdoor adjustment, which can help us block the backdoor path and eliminate confounding effects. However, the confounder set τ is difficult to obtain and cannot directly interfere with graph data. In the following, a solution will be given.

C. Causal and Trivial Attended Graph

When the input graph $G = \{V, E\}$ is obtained, the soft mask is expressed as $M_{\text{edge}} \in \mathbb{R}^{n\times n}$ and node features as $M_{\text{node}} \in \mathbb{R}^{n\times 1}$. Given a soft mask M, its complementary mask can be expressed as $\overline{M} = 1 - M$. If a graph is represented in another form $G = \{A, X\}$, where A records the detailed structure information of the graph and X is a matrix used to describe node characteristics. Then, a graph can be divided into two graphs: $G_1 = \{A \odot M_{\text{edge}}, X \odot M_{\text{node}}\}$ and $G_2 = \{A \odot \overline{M}_{\text{edge}}, X \odot \overline{M}_{\text{node}}\}$. After research [20], it can be considered that the classes of graphs can often be derived from more essential causal features. Thus, the attended graph that aggregates the causal features is defined as the causal attended graph G_c, and the corresponding graph is the trivial attended graph G_t. However, in practical applications, the attended graph with ground-truth is usually not directly usable. Therefore, the two types of attended graph need to be obtained through learning masks: $G_c = \{A \odot M_{\text{edge}}, X \odot M_{\text{node}}\}$ and $G_t = \{A \odot \overline{M}_{\text{edge}}, X \odot \overline{M}_{\text{node}}\}$.

D. Causal-Trivial Attention Graph Neural Network (CTA-GNN)

To achieve the aforementioned backdoor adjustment, the CTA-GNN framework is proposed. The overview of CTA-GNN is given in Fig. 2.

1) Calculating Soft Masks: First, an attention module is required to filter out causal and shortcut features. Then, according to the obtained features, causal proposals and trivial proposals are generated. Denote the GNN-based encoder by $f(\cdot)$ and the input graph by $G = \{A, X\}$, the nodes are represented as follows:

$$H = f(A, X). \tag{7}$$

In order to obtain the attention score, it can be done separately from the node-level and edge-level perspectives. Then two multi-layer perceptrons (MLPs) are used: $MLP_{\text{node}}(\cdot)$ and $MLP_{\text{edge}}(\cdot)$. For node v_i and edge (v_i, v_j) can get

$$\alpha_{c_i}, \alpha_{t_i} = \sigma(MLP_{\text{node}}(h_i)) \tag{8}$$

$$\beta_{c_{ij}}, \beta_{t_{ij}} = \sigma(MLP_{\text{edge}}(h_i \| h_j)) \tag{9}$$

where $\alpha_{c_i}, \beta_{c_{ij}}$ is the node-level attention score of node v_i in the causal attended graph and the edge-level attention score of edge (v_i, v_j) and $\alpha_{t_i}, \beta_{t_{ij}}$ are used for trivial attended graphs. $\sigma(\cdot)$ is the softmax function, $\|$ represents the stitching operation. Obviously, $\alpha_{c_i} + \alpha_{t_i} = 1$, $\beta_{c_{ij}} + \beta_{t_{ij}} = 1$.

The attention scores $\alpha_{c_i}, \alpha_{t_i}, \beta_{c_{ij}}, \beta_{t_{ij}}$ are used to construct the soft masks $M_{\text{node}}, \overline{M}_{\text{node}}, M_{\text{edge}}, \overline{M}_{\text{edge}}$. Finally, a preliminary representation of the causal and trivial attended graphs is obtained using the graph G: $G_c = \{A \odot M_{\text{edge}}, X \odot M_{\text{node}}\}$ and $G_t = \{A \odot \overline{M}_{\text{edge}}, X \odot \overline{M}_{\text{node}}\}$.

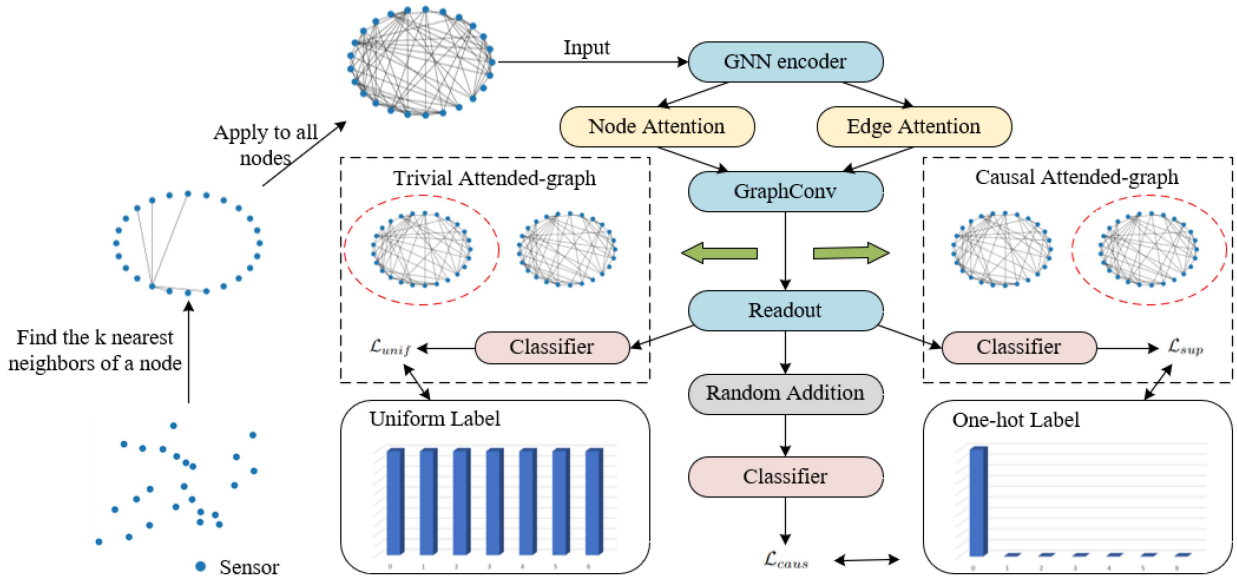

Fig. 2. Overview of CTA-GNN.

2) Disentanglement: Initial attended graphs are created by calculating soft masks. To obtain causal and trivial attended graphs, representations of the attended graphs can be obtained using GNN modules, respectively. Finally, the category of the input graph is predicted by the readout function and the classifier:

$$h_{G_c} = f_{\text{readout}}(GConv_c(A \odot M_{\text{edge}}, X \odot M_{\text{node}}))$$
$$z_{G_c} = \Phi_c(h_{G_c}) \tag{10}$$

$$h_{G_t} = f_{\text{readout}}(GConv_t(A \odot \overline{M}_{\text{edge}}, X \odot \overline{M}_{\text{node}}))$$
$$z_{G_t} = \Phi_t(h_{G_t}). \tag{11}$$

The purpose of causal attended graphs is to estimate causal features whose representations can be classified as ground-truth labels. Correspondingly, the supervised loss on the graph classification problem is defined as

$$\mathcal{L}_{\text{sup}} = -\frac{1}{|\mathcal{D}|}\sum_{G\in\mathcal{D}} \mathbf{y}_G^{\top}\log(z_{G_c}). \tag{12}$$

In contrast, trivial attended graphs are designed to approximate noncausal trivial patterns. Therefore, the prediction of the trivial attended graph can be motivated for all fault classes known, and the unified loss on the graph classification problem is defined as

$$\mathcal{L}_{\text{unif}} = \frac{1}{|\mathcal{D}|}\sum_{G\in\mathcal{D}} KL(\mathbf{y}_{\text{unif}}, z_{G_t}) \tag{13}$$

where KL is the KL-Divergence and $\mathbf{y}_{\text{unif}}$ is the uniform distribution. Causal features are distinguished from trivial features by optimizing the above two objectives. However, a previous study [21] showed that real-world graph data are noisy, which undoubtedly leads to a larger correlation between the causal part and the label than between the full graph and the label. Furthermore, due to the existence of trivial patterns, the causal attended graphs obtained by the above disentanglement methods are unlikely to eventually converge to the full graph.

3) Causal Intervention: Backdoor adjustment can effectively weaken the confounding effect by stratifying the confounding factors and pairing each layer of the target causal attended graph with the trivial attended graph to form the intervened graph. Due to the irregularity of the graph data, it prevents us from intervening at the data level, only implicitly at the representation level, and proposes a loss guided by backdoor adjustments

$$z_{G'} = \Phi(h_{G_c} + h_{G_{t'}}) \tag{14}$$

$$\mathcal{L}_{\text{caus}} = -\frac{1}{|\mathcal{D}|\cdot|\hat{\tau}|}\sum_{G\in\mathcal{D}}\sum_{t'\in\hat{\tau}} \mathbf{y}_G^{\top}\log(z_{G'}) \tag{15}$$

where $z_{G'}$ is the classification result of implicit intervened graph G' in classifier Φ. h_{G_c} is the representation of the causal attended graph G_c. $h_{G_{t'}}$ is the representation of layered $G_{t'}$. $\hat{\tau}$ is an ensemble of estimates of layered trivial attended graphs that yield trivial features present in the training data.

The random addition method is used to intervene (14). In addition, (15) is called the causal intervention loss [22] at the representation level. Due to the shared nature of causal features, it enables the intervention graph to make stable predictions across different strata. At last, the learning objective of CTA-GNN is given, which is the total loss

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_1\mathcal{L}_{\text{unif}} + \lambda_2\mathcal{L}_{\text{caus}} \tag{16}$$

where λ_1 and λ_2 are constants that control the degree of disentanglement and causal intervention, they are hyperparameters that can be tuned.

IV. Experiment Results and Comparisons

A. Datasets

The TFF [23] designed by Cranfield University is one of the typical industrial systems. This facility can be used to control a pressurized system, with sensors distributed at different places in the system to measure water flow, oil flow, and air flow. The

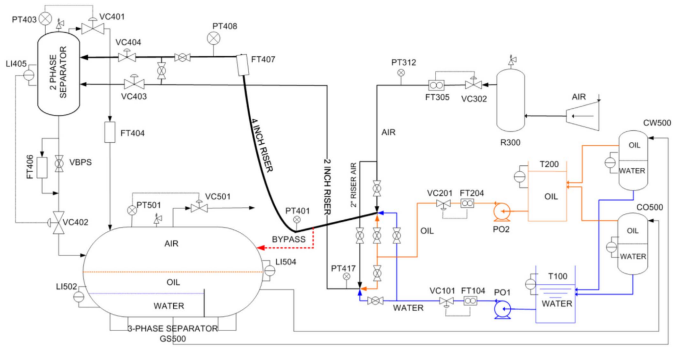

Fig. 3. Overall structure and sensor distribution of the three-phase flow facility.

description of the facility is shown in Fig. 3. The system is an industrial-scale workbench of authenticity and sophistication that works under different operating conditions and provides experimental data. Pipes and gas–liquid two-phase separators with different pore sizes and geometries constitute the test setup. During the experiment, a total of two process inputs, including air flow and water flow setpoints, were continuously modified to simulate variable operating conditions and finally obtain corresponding data. It has a total of 24 components distributed in different places in the system. Their locations are predesigned to measure density, temperature, pressure, and flow. There are five types of water flow and four types of air flow, and we can choose one of them as the process input. After calculation, 20 different input combination types can be created. TFF dataset can be downloaded from https://www.mathworks.com/matlabcentral/fileexchange/50938-a-benchmark-case-for-statistic-process-monitoringcranfield-multiphase-flow-facility.

Due to the complex operating environment of the system, twenty sets of inputs were obtained by modifying the air flow and water flow, and they were simulated to obtain three sets of data. For the fault dataset, the system simulates a total of six faults, which are used to represent some typical faults that may occur in practice. It should be noted that the faults will not occur until after a period of time under normal conditions, and are not immediately generated faults. After a fault occurs, it will endanger the system, and when it is severe enough, it will return to a normal state. At this point, the fault state will be suspended. Therefore, the data generated by each fault type have transition information from the initial state to the fault state. In order to improve the generalization ability of the model, considering both the steady-state condition and the changing condition, there are multiple datasets representing the same fault type. Each component output is preprocessed by max–min normalization. In order to better extract fault feature information, normal data are deleted from fault data, and fragments with 50 s information are taken as samples.

B. Experimental Settings

1) Current Baseline Methods: To demonstrate the performance of our CTA-GNN method in real industrial systems, the CTA-GNN method is compared with existing baseline methods: Graph attention networks (GAT) [24], progressive knowledge transfer-based multitask convolutional neural network (PKT-MCNN) [25], interaction-aware graph neural networks (IAGNN) [26].

1) *GAT:* In order to obtain the edge weight and measure the importance of the edge, GAT will learn according to the node features. Structured data can be directly applied to GAT for industrial fault classification.
2) *PKT-MCNN:* In order to perform knowledge transfer for multitask CNN models, PKT-MCNN designs a coarse-to-fine framework to obtain a hierarchical structure of fault types, which is finally used for large-scale fault diagnosis.
3) *IAGNN:* IAGNN classifies the fault using an interactive perception and fused data approach, learning multiple interactions between components and extracting fault features from each subgraph. Among them, when extracting subgraph fault features, a message passing mechanism is used.

2) Evaluation Indicators and Parameter Settings: In order to provide stronger evidence, the accuracy, Micro F1, Macro F1, and confusion matrix are used to compare with other methods. Extensive experiments were performed using various baseline methods and the CTA-GNN model, selecting the hyperparameters that provided the best results for optimal performance and training efficiency for comparison. For the TFF dataset, the input graph contains a total of 24 nodes, corresponding to 24 sensors. The feature size of each node is 50, which represents the intercepted 50 s signal segment. In addition, the maximum epoch is 100 and the learning rate is 0.001. In the loss function, λ_1 is set to 1.0 and λ_2 is set to 0.5.

C. Fault Classification Performance

When comparing the fault diagnosis performance of different methods, with the help of the model confusion matrix during the test, after visualization, it can be used to intuitively understand the performance of each model in samples of different types of faults. The visualization results of the confusion matrix of each model are shown in Fig. 4. In addition, in order to avoid class imbalance in the test data, we use the receiver operating characteristic (ROC) curve to evaluate different classification models, and then use the area under curve (AUC) to compare the models more clearly. The microaverage ROC curve and the macroaverage ROC curve are shown in Fig. 5. Finally, as the training process progresses, the performance during the test will also change. To examine this process, record their F1 scores on the test set and observe their changes. The F1 score is one of the important indicators to measure the classification accuracy of the model. It comprehensively utilizes the accuracy and recall, which is more effective and reliable. The Micro F1 score and Macro F1 score results are shown in Fig. 6. Table I shows the comparison of classification performance of different models on the TFF dataset. Table II show the computational time of different models.

Compared with the baseline methods, the CTA-GNN model performs best on the TFF dataset. Recent results demonstrate that graph embeddings learned by CTA-GNN can effectively reveal fault characteristics in process industries. By comparing with the performance of GAT, it is found that the CTA-GNN

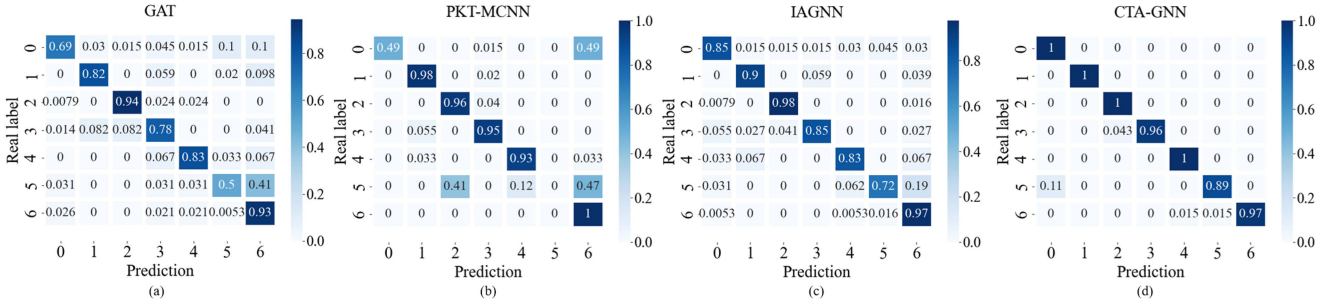

Fig. 4. Confusion matrix comparison. (a) GAT. (b) PKT-MCNN. (c) IAGNN. (d) CTA-GNN.

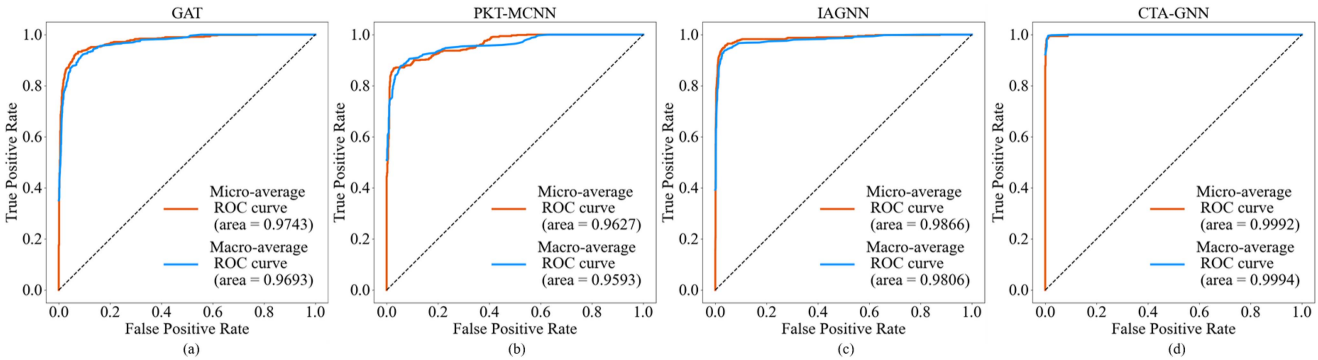

Fig. 5. ROC curve comparison. (a) GAT. (b) PKT-MCNN. (c) IAGNN. (d) CTA-GNN.

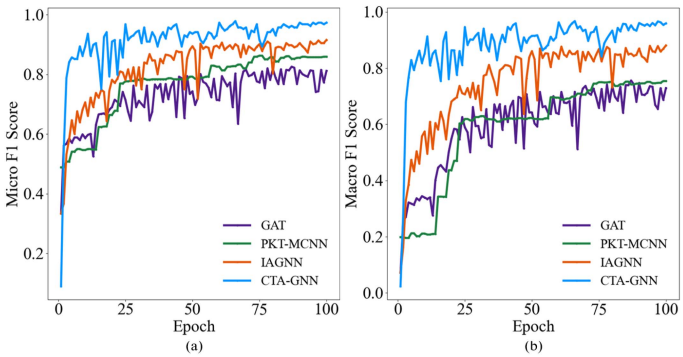

Fig. 6. F1 score comparison. (a) Micro F1 score. (b) Macro F1 score.

TABLE I
CLASSIFICATION PERFORMANCE COMPARISON

	Micro F1	Macro F1	Micro-average AUC	Macro-average AUC
GAT	0.8260	0.7561	0.9743	0.9693
PKT-MCNN	0.8946	0.7903	0.9728	0.9843
IAGNN	0.9209	0.8792	0.9869	0.9823
CTA-GNN	**0.9735**	**0.9533**	**0.9905**	**0.9942**

The bold values indicate that the CTA-GNN model outperforms other comparison methods in terms of Micro F1, Macro F1, Micro-average AUC and Macro-average AUC.

TABLE II
COMPUTATIONAL TIME COMPARISON

	Computational time / seconds
GAT	0.322835
PKT-MCNN	0.119127
IAGNN	4.331401
CTA-GNN	**0.100290**

The bold values indicate that the CTA-GNN model computation time is shorter than other comparative methods.

model can comprehensively utilize the original node information and learn the correlation between node information, so as to perform better. This is because the GAT model uses an attention mechanism to calculate weights for the features of adjacent nodes in the graph and aggregate node feature information. However, the graph structure is not considered when calculating the weight of adjacent node features. So GAT does not focus on the interactions between different components. This suggests that complex interactions between different components are also part of the important fault signature. The graph structure descriptions of different fault types are used as the input of the CTA-GNN model as shown in Fig. 7, while GAT uses fully connected graphs. Such a graph has edge weights without sparse operations, which will bring a lot of noise to the fault features, making it difficult to learn the fault features. This undoubtedly makes it difficult to distinguish between different fault types.

Different from GAT, PKT-MCNN is a structure learning algorithm based on the clustering of different coarse-grained node fault types. To obtain more general fault information, the PKT-MCNN structure can learn both coarse-grained and fine-grained tasks. But compared with CTA-GNN, PKT-MCNN controls the model to accomplish different learning objectives by changing

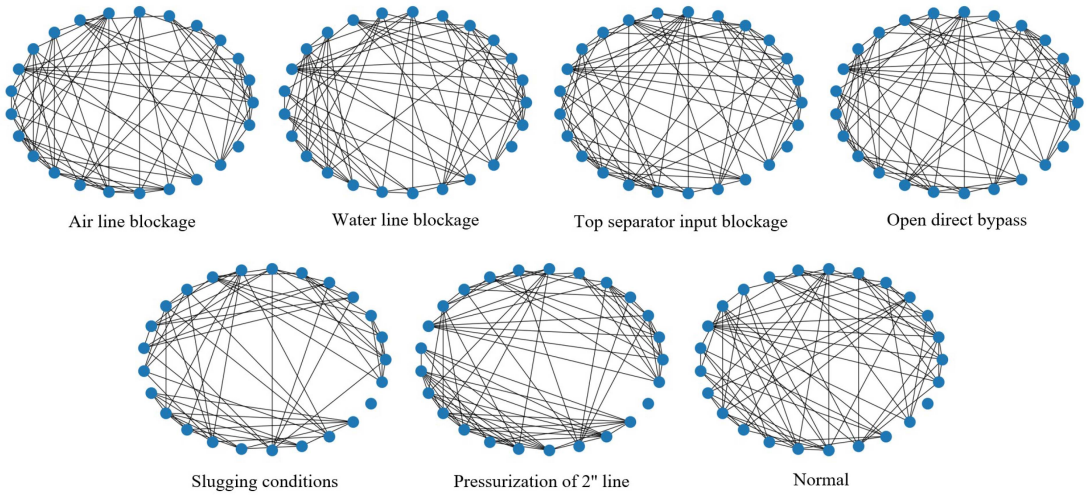

Fig. 7. Graph structure illustration.

the attention weights of coarse-grained and fine-grained tasks. But in reality, the scale of the target classification task is small and the coarse-grained structure is fuzzy, resulting in worse classification performance than CTA-GNN.

IAGNN overcomes the shortcomings of GAT and PKT-MCNN, and adaptively learns the edge weights of heterogeneous graphs composed of component reads through an attention mechanism. Subgraphs of each edge type are then subjected to fault feature extraction using multiple independent GNN blocks. Finally, with the help of a weighted sum function, the subgraph features are aggregated to obtain graph embeddings. Unlike the CTA-GNN model, multiple independent GNN blocks perform fault extraction on subgraphs of each edge type, tending to learn the external correlation between the input image and the real label, and cannot distinguish the more essential causal features, resulting in its classification effect being inferior to CTA-GNN.

CTA-GNN revisits GNN modeling for graph classification from the perspective of causality, alleviating the confounding effect. Different from other baseline models, CTA-GNN uses an attention module to learn causal and shortcut features for a given graph. Each causal feature is then combined with different shortcut features. Fig. 8 shows the simplified 2-D feature maps of original graph data and learned fault features of the CTA-GNN model, via the *t*-distribution stochastic neighbor embedding (t-SNE) method [27]. It can be seen that due to the ubiquity of environmental noise and fault degree, the sample characteristics of the same fault are diverse. After the learning of the CTA-GNN model, the features of the same fault are aggregated together, making it easy to distinguish.

In terms of computational time, by measuring the time of each test epoch, the results shown in Table II are obtained. The CTA-GNN model has the shortest computational time, which means it can get classification results faster. For the other mentioned baseline models, the computational time is longer than that of the CTA-GNN model, especially the IAGNN model is the slowest.

From the experimental results, the fault diagnosis effect of the CTA-GNN model on the TFF dataset is better than that of the baseline method. The causal characterization method using

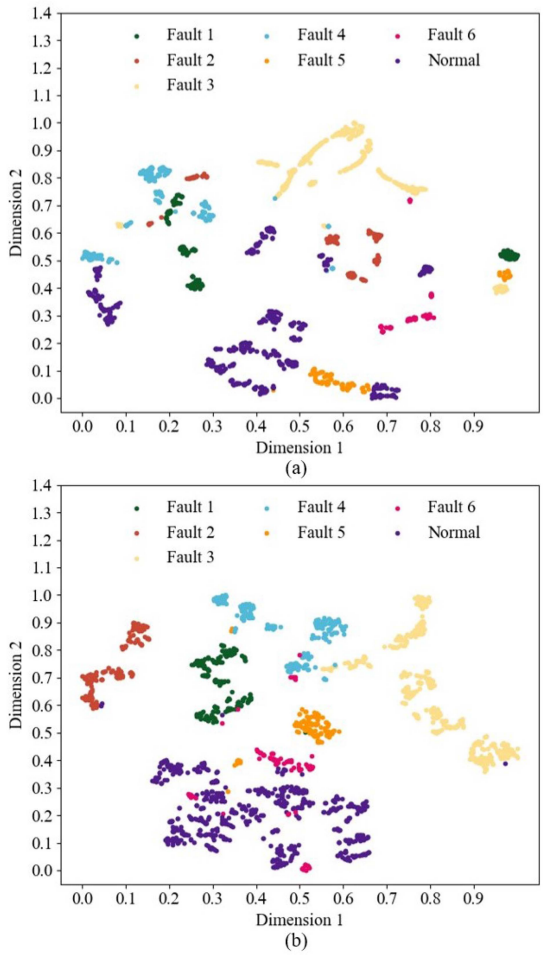

Fig. 8. Use the *t*-SNE method to visualize the results. (a) Original graph data space. (b) CTA-GNN learning space.

the CTA-GNN model can reveal the essential causes of failures in the process industry. The CTA-GNN model is capable of simultaneously learning causal features and shortcut features,

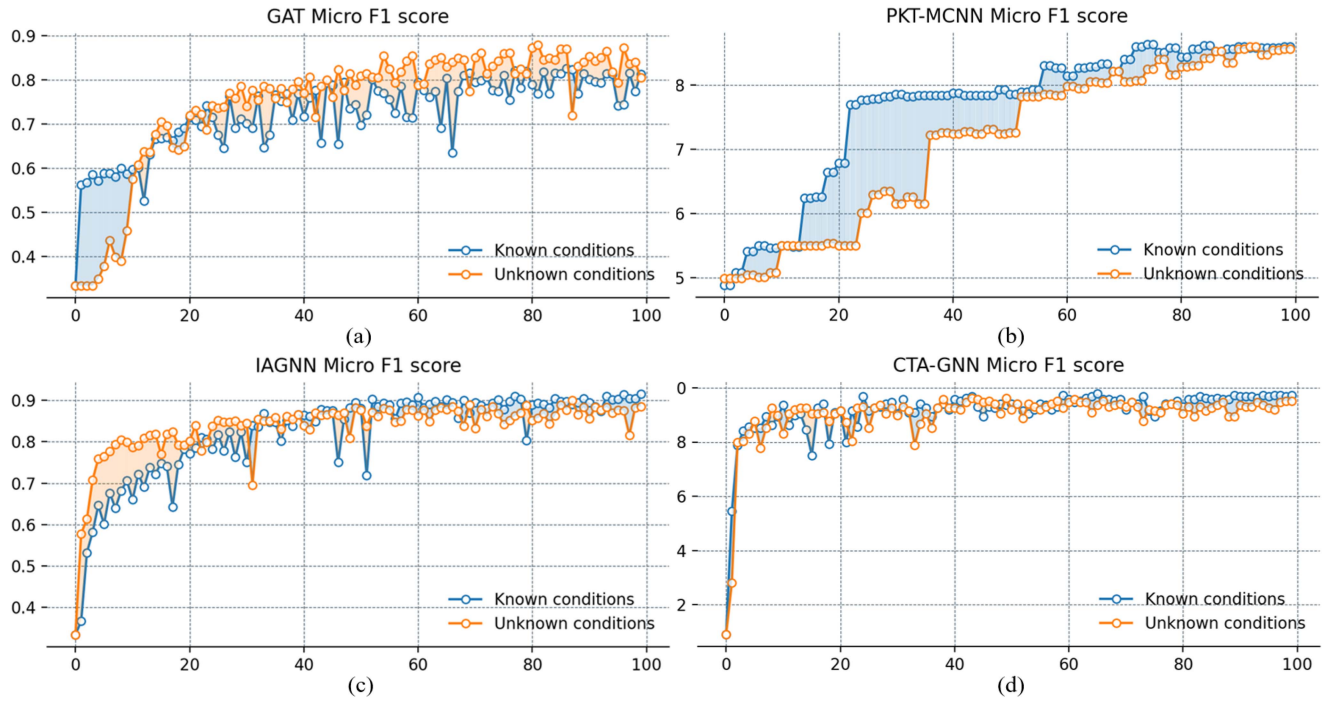

Fig. 9. Micro F1 score change comparison. (a) GAT. (b) PKT-MCNN. (c) IAGNN. (d) CTA-GNN.

combining each causal feature with different shortcut features, and finally obtaining excellent fault diagnosis results.

D. Ablation Experiment

In order to prove the generalization performance of the model, it is assumed that a certain sensor malfunctions, resulting in all signal fragments generated by it being 0. In this case, the KNN algorithm is used for the signal fragments generated by the 24 components, and a new topological map is calculated and input into the original model. A normal condition is defined as a known condition, and a condition where a sensor malfunctions is defined as an unknown condition. Micro F1 score is used as an evaluation indicator to observe changes in model performance, as shown in Fig. 9. The results show that the proposed CTA-GNN model performs best under unknown conditions, which is almost consistent with the fault diagnosis under known conditions. For the PKT-MCNN model, the failure of the sensor has a great influence on the classification effect of the model, indicating that the PKT-MCNN model is not stable enough. While the GAT model and the IAGNN model are also affected by sensor failures, their generalization ability is still inferior to that of the CTA-GNN model. Therefore, in the case of a certain sensor failure, the CTA-GNN model can still maintain a stable classification, indicating the strong generalization ability of the model.

V. Conclusion

In this article, first transform the complex industrial fault diagnosis problem into a graph recognition task, and then re-understand the GNN modeling process of graph recognition with reference to causality theory. Current baseline methods prefer to utilize shortcut features to perform calculations and give prediction results, but in fact, these features are obfuscators between causal features and predictions. The shortcut features build a backdoor path that misleads GNN model to learn spurious correlations. With the purpose of weakening confounding effects, a CTA-GNN strategy is used for industrial fault diagnosis tasks. It forces the GNN model to learn and utilize causal features and ignores the shortcut part. Experiments on the TFF dataset show that the CTA-GNN model achieves satisfactory results in various fault diagnosis tasks, which verifies its effectiveness. Future research includes applying the CTA-GNN strategy to open set recognition, which requires not only accurate diagnosis of known faults using causal features, but also effective identification of unknown faults to prevent new faults from hiding and affecting industrial production.

References

[1] Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3757–3767, Jun. 2015.

[2] A. Mahapatro and P.-M. Khilar, "Fault diagnosis in wireless sensor networks: A survey," *IEEE Commun. Surveys Tut.*, vol. 15, no. 4, pp. 2000–2026, Fourth Quarter 2013.

[3] L. Qin, X. He, and D. Zhou, "A survey of fault diagnosis for swarm systems," *Syst. Sci. Control Eng. Open Access J.*, vol. 2, no. 1, pp. 13–23, 2014.

[4] D.-T. Hoang and H.-J. Kang, "A survey on deep learning based bearing fault diagnosis," *Neurocomputing*, vol. 335, pp. 327–335, 2019.

[5] L. Ma, J. Dong, K. Peng, and C. Zhang, "Hierarchical monitoring and root-cause diagnosis framework for key performance indicator-related multiple faults in process industries," *IEEE Trans. Ind. Inform.*, vol. 15, no. 4, pp. 2091–2100, Apr. 2019.

[6] H. Chen, B. Jiang, S. X. Ding, and B. Huang, "Data-driven fault diagnosis for traction systems in high-speed trains: A survey, challenges, and perspectives," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 1700–1716, Mar. 2022.
[7] J. Marzat, H. Piet-Lahanier, F. Damongeot, and E. Walter, "Model-based fault diagnosis for aerospace systems: A survey," *Proc. Inst. Mech. Eng., J. Aerosp. Eng.*, vol. 226, pp. 1329–1360, 2012.
[8] T. Yang, S. Hong, S. Bing, and T. Shuai, "A novel dynamic weight principal component analysis method and hierarchical monitoring strategy for process fault detection and diagnosis," *IEEE Trans. Ind. Electron.*, vol. 67, no. 9, pp. 7994–8004, Sep. 2020.
[9] K. Zhong, M. Han, T. Qiu, and B. Han, "Fault diagnosis of complex processes using sparse kernel local Fisher discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1581–1591, May 2020.
[10] S. Mahadevan and S. L. Shah, "Fault detection and diagnosis in process data using one-class support vector machines," *J. Process. Control*, vol. 19, no. 10, pp. 1627–1639, Dec. 2009.
[11] D. Wu and J. Zhao, "Process topology convolutional network model for chemical process fault diagnosis," *Process. Saf. Environ. Protection*, vol. 150, pp. 93–109, Jun. 2021.
[12] A. Boutet, A.-M. Kermarrec, N. Mittal, and F. Taiani, "Being prepared in a sparse world: The case of kNN graph construction," in *Proc. IEEE 32nd Int. Conf. Data Eng.*, 2016, pp. 241–252.
[13] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.
[14] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
[15] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–26.
[16] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4438–4445.
[17] Y. Sui, X. Wang, J. Wu, M. Lin, X. He, and T.-S. Chua, "Causal attention for interpretable and generalizable graph classification," in *Proc. Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 1696–1705.
[18] S. Hu, H. Wang, J. Zhang, W. Kong, Y. Cao, and R. Kozma, "Comparison analysis: Granger causality and new causality and their applications to motor imagery," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 7, pp. 1429–1444, Jul. 2016.
[19] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2019.
[20] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[21] W. Lin, H. Lan, and B. Li, "Generative causal explanations for graph neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 1–14.
[22] C. Mao, A. Cha, A. Gupta, H. Wang, J. Yang, and C. Vondrick, "Generative interventions for causal learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3947–3956.
[23] C. Ruiz-Carcel, Y. Cao, D. Mba, L. Lao, and R. T. Samuel, "Statistical process monitoring of a multiphase flow facility," *Control Eng. Pract.*, vol. 42, pp. 74–88, 2015.
[24] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representation*, 2018, pp. 1–12.
[25] Y. Wang et al., "Coarse-to-fine: Progressive knowledge transfer-based multitask convolutional neural network for intelligent large-scale fault diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 2, pp. 761–774, Feb. 2023.
[26] D. Chen, R. Liu, Q. Hu, and S. X. Ding, "Interaction-aware graph neural networks for fault diagnosis of complex industrial processes," *IEEE Trans. Neural Netw. Learn. Syst.*, Dec. 2021, early access, doi: 10.1109/TNNLS.2021.3132376.
[27] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

Hao Wang is currently working toward the bachelor's degree in artificial intelligence with the College of Intelligence and Computing, Tianjin University, Tianjin, China.

His research interests include deep learning, computer vision, and fault diagnosis of mechanical systems.

Ruonan Liu (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2013, 2015, and 2019, respectively.

She was a Postdoctoral Researcher with the School of Computer Science, Carnegie Mellon University, in 2019. She currently is an Associate Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. Her research interests include machine learning, intelligent manufacturing, and computer vision.

Steven X. Ding received the Ph.D. degree in electrical engineering from Gerhard-Mercator University, Duisburg, Germany, in 1992.

From 1992 to 1994, he was a Research and Development Engineer with Rheinmetall GmbH, Germany. From 1995 to 2001, he was a Professor of Control Engineering with the University of Applied Science Lausitz, Senftenberg, Germany, and the Vice President from 1998 to 2000. He is currently a Full Professor of Control Engineering and the Head of the Institute for Automatic Control and Complex Systems (AKS) with the University of Duisburg-Essen, Germany. His research interests include model-based and data-driven fault diagnosis, fault tolerant systems, real-time control, and their application in industry with a focus on automotive systems, chemical processes, and renewable energy systems.

Qinghua Hu (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science and engineering and the Ph.D. degree in aerospace from the Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively.

He was a Postdoctoral Fellow with the Department of Computing, Hong Kong Polytechnic University, from 2009 to 2011. He is currently the Dean of the School of Artificial Intelligence, the Vice Chairman of the Tianjin Branch of China Computer Federation, the Vice Director of the SIG Granular Computing and Knowledge Discovery, and the Chinese Association of Artificial Intelligence. He is currently supported by the Key Program, National Natural Science Foundation of China. He has published over 200 peer-reviewed papers. His current research is focused on uncertainty modeling in Big Data, machine learning with multimodality data, intelligent unmanned systems. He is an Associate Editor of the IEEE TRANSACTIONS ON FUZZY SYSTEMS, Acta Automatica Sinica, and Energies.

Zengxiang Li received the B.S. degree from the Shanghai University of Electric Power, Shanghai, China, in 2003, the M.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 2006, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2012, all in computer science and engineering.

He was a Research Scientist with High Performance Computing, A*STAR, Singapore, from 2012 to 2020. He is currently the Executive Vice President of Digital Research Institute of ENN Group. His research interests include, distributed system, graph computing, Industrial IoT, Blockchain, AI, federated learning, incentive mechanism and privacy-preserving technology. He has published more than 50 high-quality papers on ACM/IEEE Transactions, reputable journals and conferences.

Hongkuan Zhou received the B.S. degree in electronic and information engineering and the Ph.D. degree in information and communication engineering from Huazhong University of Science and Technology, Wuhan, China, in 2014 and 2019, respectively.

He is currently with the Science and Technology on Thermal Energy and Power Laboratory, Wuhan, China. His current research interests include fault diagnosis and deep learning.