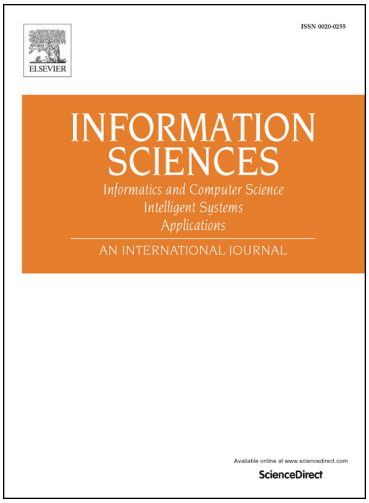
Journal Pre-proof



Casual Inference-Enabled Graph Neural Networks for Generalized Fault Diagnosis in Industrial IoT System

Zhao Zhang, Qi Li, Shenbo Liu, Zhigang Zhang, Wei Chen et al.

PII: S0020-0255(24)01633-5
DOI: https://doi.org/10.1016/j.ins.2024.121719
Reference: INS 121719

To appear in: *Information Sciences*

Received date: 13 May 2024
Revised date: 7 November 2024
Accepted date: 24 November 2024

Please cite this article as: Z. Zhang, Q. Li, S. Liu et al., Casual Inference-Enabled Graph Neural Networks for Generalized Fault Diagnosis in Industrial IoT System, *Information Sciences*, 121719, doi: https://doi.org/10.1016/j.ins.2024.121719.

Casual Inference-Enabled Graph Neural Networks for Generalized Fault Diagnosis in Industrial IoT System

Zhao Zhang[a,*], Qi Li[b], Shenbo Liu[a], Zhigang Zhang[a], Wei Chen[a] and Lijun Tang[a]

[a]School of Physics and Electronic Science, Changsha University of Science and Technology, Changsha, 410114, China.
[b]School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China

ARTICLE INFO

Keywords:
Fault Diagnosis
Graph Neural Networks
Causal Inference
Structural Causal Model
Industrial Internet of Things

ABSTRACT

Data-driven fault diagnosis plays a crucial role in diagnosing the operational status within the Industrial Internet of Things (IIoT) systems. Although Graph Neural Networks (GNNs) have recently gained traction in fault diagnosis by adeptly modeling complex dependencies present in high-dimensional sensor measurements, they still grapple with the challenges presented by varying working conditions and pervasive environmental noise, which can significantly hinder their generalization capabilities. Hence, we propose Causal Inference-Enabled Graph Neural Networks (CIE-GNN) for generalized fault diagnosis in large-scale IIoT systems. Specifically, we establish a structural causal model for the GNN-based fault diagnosis model, revealing that the non-causal factors lead to spurious correlations and act as the confounders, thus impairing the generalization performance of fault diagnosis models. To rectify this issue, we design the disentangled transformation module and the causal disentanglement regularization based on mutual information minimization strategy, facilitating the effective decoupling between the fault-causal factors and the non-causal factors in the representation level. Additionally, we propose the random pairing-based backdoor adjustment regularization to mitigate the negative effects of non-causal factors. Extensive experiments and rigorous theoretical analysis validate the generalization capabilities of CIE-GNN across diverse working environments.

1. Introduction

With the rapid proliferation of Industrial Internet of Things (IIoT) technologies, the deployment of a multitude of sensors enables the full potential of Industry 4.0 to be harvested in manufacturing processes, energy system, and etc. Prognostics and Health Management (PHM) systems have become increasingly prevalent across these IIoT-enabled systems [1], allowing for proactive maintenance to take place, leading to a reduction in the unplanned downtime. As a pivotal data analysis component within the PHM systems, Fault diagnosis [2]-[5] has been extensively employed to assess the health status of equipment and predict the occurrence of failures with the monitoring sensor data. In recent years, the amalgamation of data-driven intelligent fault diagnosis combined with deep neural networks has garnered significant attention from both industry and academia [4, 6, 7]. These intelligent methods autonomously capture and distill intricate fault patterns, ultimately culminating in a heightened level of accuracy in fault identification.

Despite the remarkable advancements in data-driven intelligent fault diagnosis, the growing complexities arising from high-dimensional sensor measurements within large-scale IIoT systems pose formidable challenges. Recently, Graph Neural Networks (GNN)[8]-based fault diagnosis methods [9]-[22] have gained prominence within the realm of IIoT, achieving state-of-the-art performance on the existing fault diagnosis benchmarks. These fault diagnosis models founded on GNN seamlessly integrate information gleaned from diverse sensor readings and model intricate interactions form graph structures, ultimately capturing fine-grained fault information for fault diagnosis. For example, motivated by the successful application of GNN in modeling complex interactions within graph data, an interaction-aware GNN was developed for fault diagnosis in complex industrial systems [9], which showcased reliable and superior diagnostic performance in the context of the three-phase flow facility and power system. Besides, Li et al. [10] proposed a spatial-temporal aware Graph Convolutional Network (GCN) model capable of learning temporal evolutions and complex interaction relationships among sensor measurements, thereby leading to accurate fault identification in industrial processes.

Nevertheless, when deployed in real-world, large-scale IIoT scenarios, GNN-based fault diagnosis methods still encounter pronounced research gaps. The real-time changes in working conditions, stemming from factors such as equipment wear and diverse service environments [23], give rise to distribution discrepancies of collected data. Such inconsistencies undermine the exiting GNN-based fault diagnosis models' ability to generalize from training on previous data to accurately performing in real-world scenarios. Besides, the existing GNN-based fault diagnosis models are predominantly developed and trained in controlled laboratory environments [24]. It is worth noting that the exiting GNN-based fault diagnosis methods are based on statistical learning and tend to overfitting to spurious correlations in the training data, thereby lacking the generalization for the diverse working conditions or the prevalent environmental noise that may occur in practice.

*Corresponding author
zhangzhao@csust.edu.cn (Z. Zhang)
ORCID(s): 0000-0001-9165-1439 (Z. Zhang)

To bridge the aforementioned research gap, it is imperative to develop robust and reliable GNN-based fault diagnosis models that can remain insensitive to spurious features induced by varying working conditions or environmental noise. Recent studies [25, 26] suggest that causal inference could pave the way for learning the causal relationship (invariant under distribution shifts) while eliminating the effect of spurious correlations. Herein, to explore the reasons behind the generalization failure of GNN-based models, we first construct the structural causal model (SCM) [27] that elucidates the causal relationship. The SCM analysis allows the learned graph representations to be disentangled into fault-causal factors and non-causal factors. The fault-causal factors directly contribute to the fault diagnosis and remain invariant across diverse working environments, while the non-causal factors are influenced by the environmental noise, or working conditions. The non-causal factors can be viewed as confounders causing the spurious correlation and deteriorating the generalization ability.

Motived by these insights, we propose the Causal Inference-Enabled Graph Neural Network (CIE-GNN) for generalized fault diagnosis in large-scale IIoT systems. Notably, we design the disentangled transformation module and the causal disentanglement regularization for the effective disentanglement of fault-causal and non-causal subgraphs. To mitigate the negative impact of non-causal factors and enhance generalization, we further design the backdoor adjustment regularization. Extensive experiments conducted on two realistic industrial fault diagnosis datasets validate the superiority of the proposed CIE-GNN over the existing GNN-based FD methods, showcasing robustness to environmental noise and the potential to generalize across diverse working conditions. Moreover, unlike the exiting methods lacking in solid theoretical support, we also provide the rigorous theoretical analysis proving that CIE-GNN guarantees the FD model to achieve great generalization. In summary, the contributions of our work can be summarized as follows:

- We propose the casual inference-enabled graph neural network for generalized fault diagnosis in large-scale IIoT systems, which consists of the disentangled transformation module, the causal disentanglement module, and the backdoor adjustment module.
- We propose the Mutual Information (MI) minimization-based causal disentanglement regularization to effectively decouple the fault-causal and non-causal subgraphs, which designs the MI-estimator neural network for approximating the variational version of MI upper bound between them. Besides, we propose the random pairing-based backdoor adjustment regularization to eliminate the confounding effect, which can boost the diversity of the non-causal features and implement the causal intervention on the graph representation level.
- We provide a thorough theoretical analysis for the proposed casual inference-enabled graph neural network to prove its generalization characteristics.

The remainder of this paper is organized as follows. Section 2 describes the literature review. Section 3 formulates the fault diagnosis problem and introduces the preliminaries of GNN-based fault diagnosis. Section 4 elaborates the proposed methodology. The experimental results on two publicly fault detection datasets are discussed in Section 5. Section 6 concludes this paper.

2. Related work

In this section, we first discuss the existing approaches for GNN-based fault diagnosis. Then, we introduce some representative works that incorporate causal inference into fault diagnosis tasks.

2.1. GNN-based fault diagnosis

Recent advancements in the data-driven fault detection algorithms have been primarily witnessed in the Transformer-based methods [28]-[32], and the GNN-based methods [9]-[22]. The Transformer-based fault diagnosis techniques prioritize the temporal modeling of sensor data, designing sophisticated Transformer modules such as CLFormer [28] and Convformer-NSE [29]. Conversely, the GNN-based approaches excel in capturing complex interactions within the data. This paper focuses on the GNN-based fault detection methods, and presents a comparative analysis with Transformer-based methods in the experimental section.

Several studies based on GNN have demonstrated their potential to enhance fault diagnostic performance in industrial field, including GCN [11, 12], Graph Attention Network (GAT) [13, 14, 15], Graph Isomorphism Network (GIN) [16, 17], and so on. The detailed descriptions for the exiting GNN-based fault diagnosis methods are provided in Table. 1. Liu et al. [14] constructed a fault diagnosis model based on GAT, facilitating robust feature extraction via multiple attention heads. In [17], a GIN-based fault diagnosis model was proposed to mine intricate relationships of large-scale topological graph structures within the context of mechanical fault diagnosis. The proposed STAGED model in [10] combined GCNs, long short-term memory networks, and attention mechanisms to learn comprehensive representations for multi-series data, thereby enabling accurate fault diagnosis. Yin et al. [12] presented a fault diagnosis method based on GCN, effectively capturing the spatial information to enhance diagnosis performance. Extending the GCN framework, several works [18, 19] utilized the ChebyNet to approximate the graph convolutional kernel, thereby improving representations of multiple receptive field features. To further boost the processing capability of large-scale graph data, the intelligent fault diagnosis models based on the GraphSAGE network were proposed in [20, 21]. Furthermore, the authors in [22] applied High-order Graph Convolutional Network (HoGCN) for fault diagnosis, capturing

Table 1
Overview of the existing data-driven fault diagnosis methods

Method	Considering variable working condition	Considering noise addition	Incorporating causal inference	Involving theoretical analysis
STAGED [10]	✗	✗	✗	✗
GAT-based [13, 14, 15]	✗	✗	✗	✗
GCN-based [11, 12]	✗	✗	✗	✗
ChebyNet-based [18, 19]	✗	✗	✗	✗
GraphSage-based [20, 21]	✗	✗	✗	✗
GIN-based [16, 17]	✗	✗	✗	✗
HoGCN [22]	✗	✗	✗	✗
CAL [38]	✗	✗	✓	✗
CIE-GNN	✓	✓	✓	✓

contextual information and mutual influences between sensor nodes.

In summary, these studies pave the way for more accurate and efficient fault diagnosis by harnessing the power of GNNs to analyze graph-structured data and model intricate relationships within industrial control systems. Nonetheless, when applied to real-world, large-scale IIoT scenarios, GNN-based fault diagnosis methods still grapple with challenges primarily stemming from environmental noise and fluctuating operational conditions, which can compromise the methods' effectiveness. To bridge the research gap outlined in Table 1, our method incorporates causal inference into the GNN-based fault diagnosis framework, aiming to boost resilience against environmental noise and enhance generalization amidst varying operational conditions.

2.2. Causal inference for fault diagnosis

To enhance generalization and trustworthiness, researchers have incorporated causality into fault diagnosis methodologies [33]-[39]. Hanif et al. [33] introduced a framework for fault diagnosis using continuous Bayesian networks and causal inference, highlighting its potential advantages. By integrating causal inference and analysis into fault diagnosis, the Sparse Causal Residual Neural Network (SCRNN) was proposed to simultaneously extract multi-lag linear and nonlinear causal relations [34]. Additionally, the Deep Causal Factorization Network (DCFN) was proposed for cross-machine bearing fault diagnosis, incorporating causal task factorization and feature factorization modules to reconstruct causal mechanisms [35]. Uchida et al. [36] introduced a novel fault diagnosis method based on the causality between process variables and a monitored index for fault detection. Likewise, the causal consistency network was proposed to build a generalized causal bearing fault diagnosis module in cross-machine scenarios [37], by learning the invariance of faulty causality. However, these methods don't directly apply to GNN-based fault diagnosis models and fail to address the distribution shift in graph structures caused by the diverse working environments.

Recent works have delved into how to use the casual theory in the realm of graph learning. Sui et al. [38] proposed the Causal Attention Learning (CAL) to decouple the causal graph patterns and trivial graph patterns, which designed the uniform classification loss based on KL-Divergence to implement the causal disentanglement. By designing an instrumental variable for causal intervention on graphs, a causal intervention graph neural network (CIGNN) [39] was proposed to achieve stable and reliable fault diagnosis in complex industrial processes. Unlike the causal inference-enabled paradigm described above, this paper designs unique causal disentanglement module and backdoor adjustment module for GNN-based fault diagnosis, rendering it suitable to learn environment-invariant graph representations under distribution shifts and mitigate the cofounding effects. Moreover, this is the first study with a theoretical derivation that proves the generalization of our proposed method on diverse working environments.

3. Preliminaries

This section provides a clear problem definition for fault diagnosis in large-scale IIoT systems, along with an overview of a typical GNN-based fault diagnosis model.

3.1. Problem statement

To monitor the operational efficiency of critical infrastructures in IIoT, the Supervisory Control and Data Acquisition (SCADA) systems are deployed for the collections of massive sensor data. Multiple sensors generate raw measurements, denoted as $\mathcal{X} = [X_1^{(M)}, ..., X_T^{(M)}] \in \mathbb{R}^{M\times T}$, which represent as a set of M sensor measurements with the length of T. A sliding window technique is involved to partition the raw time series into segments with a fixed window length d, which facilitating the analysis of sensor measurements over time. Concretely, for the m-th sensor measurement, the t-th sub-sequence is denoted as $S_t^m = [X_t^m, ..., X_{t+d-1}^m] \in \mathbb{R}^d$.

In the realm of industrial processes, the dynamic interactions among sensors that contribute to fault propagation manifest in intricate and interconnected ways. A graph $\mathcal{G} = (\mathcal{E}, \mathcal{X})$ is introduced to elucidate these interactions, where $\mathcal{X}$ denotes the node attributes of M sensors with their raw measurements, and $\mathcal{E}$ represents the adjacency matrix for M sensors. Herein, each sensor measurement is treated as a node, and the interactions between two sensor measurements

are considered as edges. Within the t-th window, the m-th node attribute is denoted as $S_t^m \in \mathbb{R}^d$, the edge between the i-th node and the j-th node is formulated if and only if the distance between them is less than a pre-defined threshold,

$$\mathcal{E} = \left\{ \left\langle \mathcal{E}_i, \mathcal{E}_j \right\rangle : i, j \in [M], i \neq j, d\left(S^i, S^j\right) \leq r \right\} \quad (1)$$

where d denotes as the distance function, r represents the predefined distance threshold.

3.2. Fault diagnosis based on GNN

Recently, GNNs have demonstrated tremendous potential in fault diagnosis due to their ability to effectively model and analyze complex dependencies among industrial processes. In detail, the complex relationships of multiple sensor measurements construct the graph data $\mathcal{G} = (\mathcal{X}, \mathcal{E})$, and then GNNs leverage the constructed graph $\mathcal{G}$ to identify the patterns associated with the fault labels $\mathcal{Y}$. Here, the graph encoder is defined as Φ to update the graph representation $\mathcal{Z}_\mathcal{G} \in \mathbb{R}^{M \times h}$, where h is the dimension of the graph embedding. The graph representation captures the learned embeddings of the nodes and edges via the weighted message passing and aggregation mechanism. Subsequently, the readout operations are employed to generate a concise graph representation and the classifier are utilized to project the graph representation onto a probability distribution. This projection Ψ enables the mapping of the high-level graph representation to a meaningful probability distribution, facilitating the downstream task of fault diagnosis. The entire process can be summarized as follows:

$$\mathcal{Z}_\mathcal{G} = \Phi\left(\mathcal{G}; \theta_\Phi\right), \hat{y}_\mathcal{G} = \Psi\left(\mathcal{Z}_\mathcal{G}; \theta_\Psi\right) \quad (2)$$

where the graph encoder Φ can be implemented by the GNN layers (e.g., GCN, GAT, and so on) and Ψ denotes as the readout function and the classifier.

Lastly, the training loss between the predicted fault labels $\hat{y}_\mathcal{G}$ and the ground truth labels $y_\mathcal{G}$ is computed using a cross-entropy loss function,

$$\begin{aligned} \mathcal{L}_{CE} &= -\frac{1}{D_\mathcal{G}} \sum_i^{D_\mathcal{G}} \log p\left(y_\mathcal{G}^i \circ \Psi\left(\mathcal{Z}_\mathcal{G}^i\right)\right) \\ &= -\frac{1}{D_\mathcal{G}} \sum_i^{D_\mathcal{G}} y_\mathcal{G}^i \log \hat{y}_\mathcal{G}^i \end{aligned} \quad (3)$$

where $D_\mathcal{G}$ denotes as the training dataset of the graph data $\mathcal{G}$, $\mathcal{Z}_\mathcal{G}^i$ is the graph representation for the i-th sample. By minimizing this loss function, the GNN model learns to accurately predict fault labels based on the learned graph representations.

4. Methodology

Based on the above-mentioned analysis, the traditional GNN-based fault diagnosis approach is directly learning the statistical correlations between the constructed input graphs and the corresponding fault labels. There exists a risk of unintentionally capturing non-causal features (induced by working conditions or environmental noises). Consequently, these might introduce undesired dependencies or spurious correlations, leading to biased predictions in fault diagnosis. To this end, it is imperative to mitigate the potential pitfalls associated with non-causal features. Firstly, we first construct Structural Causal Model (SCM) and identify environmental confounders as the key reason for the failure of GNN-based Fault Diagnosis (FD) models to generalize on diverse working environments. Then we design the Causal Inference-Enabled Graph Neural Network (CIE-GNN) for FD model to foster its robustness and reliability. Finally, we provide rigorous theoretical proof of CIE-GNN that can achieve great generalization.

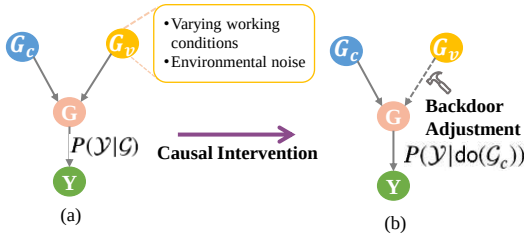


Figure 1: The structural causal model for the GNN-based fault diagnosis model.

4.1. Structural Causal Model For GNN-based fault diagnosis

To explore the reasons behind the failure of GNN-based models to generalize on diverse working environments, we construct a SCM to scrutinize the modeling of GNN-based FD from a causal perspective, as illustrated in Fig.1. It describes the causal relationships among the key variables in the GNN-based FD model. Here, $\mathcal{G}$ represents the observable variables of the constructed graph data, $\mathcal{Y}$ represents the observable variables of fault labels, $\mathcal{G}_c$ denotes the latent variables of fault-causal factors, and $\mathcal{G}_v$ is the latent variables of non-causal factors. The links between variables indicate cause-effect relationships: cause → effect. We elaborate the main causal relationships as follows.

- $\mathcal{G}_c \rightarrow \mathcal{G}$, $\mathcal{G}_v \rightarrow \mathcal{G}$. The fault-causal variables $\mathcal{G}_c$ and the non-causal variables $\mathcal{G}_v$ naturally coexist in the graph data $\mathcal{G}$. In the context of fault diagnosis, the fault-causal factors are primarily determined by fault-specific information, while the non-causal factors are induced by working conditions or environmental noise.

- $\mathcal{G} \rightarrow \mathcal{Y}$. In the existing GNN-based FD model, the constructed graph data $\mathcal{G}$ directly cause the fault labels $\mathcal{Y}$, implying that the classifier leverages the graph representations of $\mathcal{G}$ to generate predictions $\mathcal{Y}$. The conventional learning strategy of GNN-based FD model takes both the fault-causal feature and the non-causal feature as input to distill discriminative information.

As depicted in Fig. 1(a), the constructed SCM reveals the causal relationships in model training for the GNN-based FD model. $\mathcal{G}_v$ acts as the environmental confounder and directly optimizing $P(\mathcal{Y}|\mathcal{G})$ leads the GNN-based FD model to learn the shortcut relationship between $\mathcal{G}_v$ and $\mathcal{Y}$, which is highly correlated with the working environments. During the model training process, there is a tendency to use this easily captured shortcut relationships for model prediction. However, this shortcut relationships are highly sensitive to the working environments. When the environment of the test dataset is different from that of the training dataset, this relationship becomes unstable and invalid. The FD model that excessively learns environment-sensitive (e.g., noise-sensitive) relationships in the training data will struggle to accurately identify novel data during the testing phase, resulting in a fault diagnosis accuracy decrease. Hence, we find that the presence of the backdoor path can induce a spurious correlation between $\mathcal{G}_c$ and $\mathcal{Y}$, the confounding factor $\mathcal{G}_v$ is the key reason for the generalization failure of GNN-based FD model.

Through the above analysis, we can improve the generalization ability of GNN-based FD model by guiding the model to eliminate the influence of confounder $\mathcal{G}_v$ and uncover stable relationships behind the training data, specifically those that are less sensitive to working environment changes. Specifically, we learn stable correlations $\mathcal{G}_c$ and $\mathcal{Y}$ by optimizing $P(\mathcal{Y}|\mathsf{do}(\mathcal{G}_c))$ instead of $P(\mathcal{Y}|\mathcal{G})$. In causal theory, the "do-operation" [41] signifies removing the dependencies between the target variable and other variables. As shown in Fig. 1(b), by cutting off the correlations between $\mathcal{G}_v$ and $\mathcal{G}$, the model no longer learns the unstable correlations between $\mathcal{G}_v$ and $\mathcal{Y}$. This do-operation blocks the unstable backdoor path, enabling the GNN-based FD model to capture the desired causal relationship that remains invariant under diverse working environments.

Definition 1. *(do-operation). The intervention posterior resulting from the action of removing the backdoor path is given by,*

$$\begin{aligned} P\left(\mathcal{Y}\left|\mathsf{do}\left(\mathcal{G}_c\right)\right.\right) &= P_m\left(\mathcal{Y}\left|\mathcal{G}_c\right.\right) \\ &= \sum_{v\in\mathcal{V}} P_m\left(\mathcal{Y}\left|\mathcal{G}_c,\mathcal{G}_v\right.\right) P_m\left(\mathcal{G}_v\left|\mathcal{G}_c\right.\right) \end{aligned} \tag{4}$$

where do denotes as the do-operation, P_m represents the modified probability distribution, and $\mathcal{V}$ represents the confounder set of the non-causal factors. For the intervention distribution P_m, we have the following two rules:

Rule 1. *(Independency): The fault-causal variable $\mathcal{G}_c$ and the non-causal variable $\mathcal{G}_v$ are independent under the do-operation, such that $P_m\left(\mathcal{G}_v\left|\mathcal{G}_c\right.\right) = P_m\left(\mathcal{G}_v\right)$.*

Rule 2. *(Invariance): The marginal probability remains invariant under the intervention since removing the backdoor path does not affect the non-causal variable $\mathcal{G}_v$. Likewise, the conditional probability remains invariant because the relationship between $\mathcal{Y}$, $\mathcal{G}_c$, and $\mathcal{G}_v$ is unrelated to the causal effect between $\mathcal{G}_c$ and $\mathcal{G}_v$. Thus, $P_m\left(\mathcal{G}_v\right) = P\left(\mathcal{G}_v\right)$, $P_m\left(\mathcal{Y}\left|\mathcal{G}_c,\mathcal{G}_v\right.\right) = P\left(\mathcal{Y}\left|\mathcal{G}_c,\mathcal{G}_v\right.\right)$*

Based on the above rules, Eq. (4) can be rewritten as,

$$P\left(\mathcal{Y}\left|\mathsf{do}\left(\mathcal{G}_c\right)\right.\right) = \sum_{v\in\mathcal{V}} P\left(\mathcal{Y}\left|\mathcal{G}_c,\mathcal{G}_v\right.\right) P\left(\mathcal{G}_v\right) \tag{5}$$

where $P\left(\mathcal{Y}\left|\mathcal{G}_c,\mathcal{G}_v\right.\right)$ represents the conditional probability given the fault-causal variable $\mathcal{G}_c$ and the non-causal variable $\mathcal{G}_v$, $P\left(\mathcal{G}_v\right)$ is the prior probability of $\mathcal{G}_v$. Eq. (5) is usually called backdoor adjustment [40], which is a powerful tool to eliminate the confounding effect.

4.2. Causal inference-enabled GNN for fault diagnosis

As discussed in the abovementioned SCM analysis, there exist two significant obstacles that hinders the implementations of Eq. (5): i) The fault-causal variable $\mathcal{G}_c$ and the non-causal variable $\mathcal{G}_v$ are typically unobservable and challenging to obtain. ii) It is an arduous task to estimate the effects of causal intervention on these graph data. Here, we propose a novel causal inference-enabled GNN for fault diagnosis named CIE-GNN to tackle the above challenges. The proposed CIE-GNN comprises three key components, including the disentangled transformation module, the causal disentanglement module, and the backdoor adjustment module, as shown in Fig.2 .

4.2.1. Disentangled transformation

Inspired by the causality analysis, our primary focus is to decouple the input whole graph $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ into the fault-casual subgraph $\mathcal{G}_c = \left(\mathcal{X}_c, \mathcal{E}_c\right)$ and the non-causal subgraph $\mathcal{G}_v = \left(\mathcal{X}_v, \mathcal{E}_v\right)$. To achieve this, we design the disentangled transformation module that facilitates the projection of the full graph $\mathcal{G}$ into distinct subspaces at the graph representation level. Concretely, our proposed disentangled transformation module employs the learnable representational masks to encode fault-casual and non-causal information. These representational masks act as filters, selectively distilling the relevant features associated with the fault-casual and non-causal subgraphs, thereby enabling a more fine-grained and disentangled representation of the input graph. For simplicity, the graph representation of the input whole graph retains the notation $\mathcal{G}$. The disentangled transformation can be expressed as,

$$\mathcal{G}_c = \left(\mathcal{X}_c, \mathcal{E}_c\right) = \left(\mathcal{X} \circ \mathcal{T}_c, \mathcal{E} \circ \mathcal{Q}_c\right) \tag{6}$$

$$\mathcal{G}_v = \left(\mathcal{X}_v, \mathcal{E}_v\right) = \left(\mathcal{X} \circ \mathcal{T}_v, \mathcal{E} \circ \mathcal{Q}_v\right) \tag{7}$$

where $\mathcal{T} \in \mathbb{R}^M$ and $\mathcal{Q} \in \mathbb{R}^{M\times M}$ are the corresponding learnable masks implemented on the node representations and the edge representations, and $\circ$ denotes as the element-wise dot operation. $\mathcal{T}_c$ and $\mathcal{T}_v$ are used to extract specific semantics in the fault-casual subspace and the non-causal

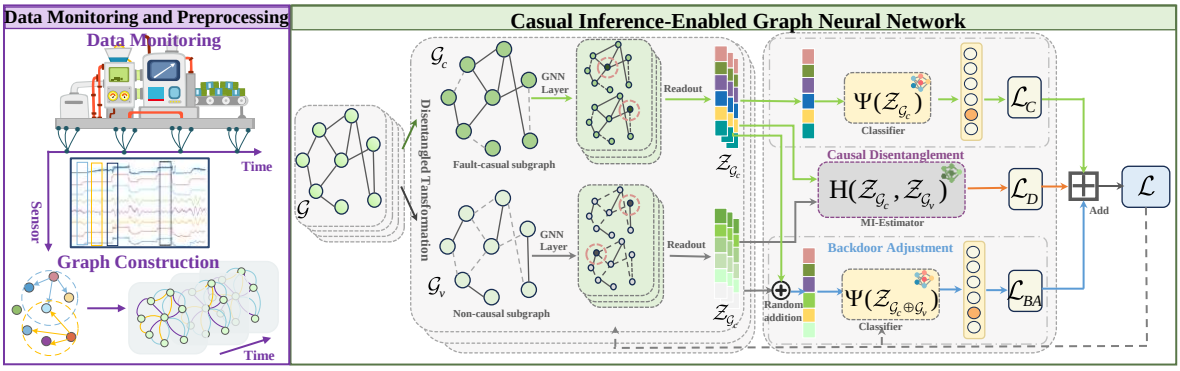


Figure 2: The framework of the proposed causal inference-enabled graph neural network for fault diagnosis.

subspace, respectively, as are $\mathcal{Q}_c$ and $\mathcal{Q}_v$. Each element of the learnable masks can be obtained by,

$$\mathcal{T}_c^i = \sigma(W_{\mathcal{T}_c}\mathcal{X}^i + b_{\mathcal{T}_c}), \mathcal{Q}_c^{i,j} = \sigma(W_{\mathcal{Q}_c}(\mathcal{X}^i||\mathcal{X}^j) + b_{\mathcal{Q}_c}) \quad (8)$$

$$\mathcal{T}_v^i = \sigma(W_{\mathcal{T}_v}\mathcal{X}^i + b_{\mathcal{T}_v}), \mathcal{Q}_v^{i,j} = \sigma(W_{\mathcal{Q}_v}(\mathcal{X}^i||\mathcal{X}^j) + b_{\mathcal{Q}_v}) \quad (9)$$

where $W_{\mathcal{T}_c}$, $b_{\mathcal{T}_c}$, $W_{\mathcal{T}_v}$, and $b_{\mathcal{T}_v}$ are the learnable parameters. $\sigma(x)$ serves as the activation function, mapping x to a value within the range of 0.0 to 1.0. $||$ denotes as the vector concatenation. Hence, by learning these disentangled transformations, we are capable of capturing the fault-casual subgraphs $\mathcal{G}_c$ and the non-causal subgraphs $\mathcal{G}_v$ from the full graph. Subsequently, we adopt the GNN layers to obtain the corresponding representations of the fault-casual subgraphs and the non-causal subgraphs:

$$\mathcal{Z}_{\mathcal{G}_c} = \Phi\left(\mathcal{G}_c; \theta_{\Phi_c}\right), \mathcal{Z}_{\mathcal{G}_v} = \Phi\left(\mathcal{G}_v; \theta_{\Phi_v}\right) \quad (10)$$

where $\mathcal{Z}_{\mathcal{G}_c}$ and $\mathcal{Z}_{\mathcal{G}_v}$ denote as the graph representations for the fault-casual subgraphs $\mathcal{G}_c$ and the non-causal subgraphs $\mathcal{G}_v$, respectively. Considering that the fault-casual subgraphs $\mathcal{G}_c$ aim to estimate the causal features, we make predictions with $\mathcal{Z}_{\mathcal{G}_c}$. The corresponding supervised classification loss is expressed as:

$$\mathcal{L}_C = -\frac{1}{D_{\mathcal{G}}}\sum_{i}^{D_{\mathcal{G}}} P\left(y_{\mathcal{G}_c}^i \circ \Psi\left(\mathcal{Z}_{\mathcal{G}_c}^i\right)\right) \quad (11)$$

where $\mathcal{L}_C$ is the cross-entropy loss over the training dataset $D_{\mathcal{G}}$, $y_{\mathcal{G}_c}^i$ is the ground-truth label, and Ψ denotes as the readout function and classifiers.

4.2.2. *MI minimization-based causal disentanglement regularization*

To further optimize the learnable mask parameters in the disentangled transformation module and ensure effective decoupling of the fault-casual subgraphs and the non-causal subgraphs, we design the causal disentanglement regularization based on mutual information (MI) minimization. Due to the fact that the causal features dominate the decision of fault identification while the widespread trivial patterns or noise have little impact on fault diagnosis, the mutual dependence between $\mathcal{G}_c$ and $\mathcal{G}_v$ should be small enough. Considering that MI can be used to measures the mutual dependence between two variables, we can minimize the MI values between $\mathcal{G}_c$ and $\mathcal{G}_v$ to diversify the representations from different subspaces. Thereby, by viewing their graph representations as random variables, the causal disentanglement regularization loss can be implemented by minimizing their MI values,

$$\mathrm{H}(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v}) = \mathbb{E}_{P\left(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v}\right)}\left[\frac{P\left(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v}\right)}{P\left(\mathcal{Z}_{\mathcal{G}_c}\right)P\left(\mathcal{Z}_{\mathcal{G}_v}\right)}\right] \quad (12)$$

However, it is intractable to exactly calculate or estimate the value of mutual information of high-dimensional variables. Referring to [42], we estimate the upper bound of MI and minimize MI by reducing the upper bound. Specifically, we define MI contrastive log-ratio upper bound as,

$$\begin{aligned}\hat{H}(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v}) =& \mathbb{E}_{P(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v})}[\log P(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c})] \\ &- \mathbb{E}_{P(\mathcal{Z}_{\mathcal{G}_c})}\mathbb{E}_{P(\mathcal{Z}_{\mathcal{G}_v})}[\log P(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c})]\end{aligned} \quad (13)$$

Considering that Eq. (13) is the expectation over N samples $\{(\mathcal{Z}_{\mathcal{G}_c}^i, \mathcal{Z}_{\mathcal{G}_c}^i)\}_{i=1}^N$ drawn from the joint distribution, can be written as,

$$\hat{H}(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v}) = \frac{1}{N}\sum_{i=1}^{N}\log p(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c}) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\log P(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c}) \quad (14)$$

where $\mathcal{Z}_{\mathcal{G}_v}^i$ and $\mathcal{Z}_{\mathcal{G}_c}^i$ are the samples of $\mathcal{Z}_{\mathcal{G}_c}$ and $\mathcal{Z}_{\mathcal{G}_v}$, respectively. Unfortunately, the conditional relation between variables is unavailable. we use a variational distribution $q_\pi(\mathcal{Z}_{\mathcal{G}_v}^i|\mathcal{Z}_{\mathcal{G}_c}^i)$ with parameter π to approximate $P(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c})$. Consequently, a variational version is defined by:

$$\hat{H}_v(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v}) = \frac{1}{N}\sum_{i=1}^{N}\left[\log(q_\pi(\mathcal{Z}_{\mathcal{G}_v}^i|\mathcal{Z}_{\mathcal{G}_c}^i) - \frac{1}{N}\sum_{j=1}^{N}\log q_\pi(\mathcal{Z}_{\mathcal{G}_v}^j|\mathcal{Z}_{\mathcal{G}_c}^i))\right] \quad (15)$$

where $q_\pi(\cdot|\cdot)$ is the variational approximation. To implement the variational approximation, we design the MI-estimator model with a Multilayer Perceptron (MLP) neural

network. The MI-estimator neural network is optimized with maximizing the loglikelihood of $q_\pi(\cdot|\cdot)$, and it is alternate-trained with our GNN encoders. The detailed training process is given in Alg. 1. After the adequately training, we can approximate H through $\hat{H}_v$. To this end, the proposed MI minimization-based causal disentanglement loss can be expressed as,

$$\mathcal{L}_D = \hat{H}_v\left(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v}\right) \tag{16}$$

4.2.3. Random pairing-based backdoor adjustment regularization

As analyzed in Section 4.1, implementing backdoor adjustment is one promising solution to alleviate the confounding effect. Herein, we design the backdoor adjustment module to approximate $P(y|\text{do}(\mathcal{G}_c))$ in Eq. (5). To achieve it, we follow two assumptions in [43]: i) $P(G_v^i) = 1/|\mathcal{V}|$, where we assume a uniform prior for the non-causal factors. In this study, the non-causal factors can be modeled by the noise originating from the working environment. The noise being uniformly distributed throughout the whole dataset is widely accepted and commonly considered in the existing study [44]. ii) $P(y|\mathcal{G}_c = \mathcal{G}_c^i, \mathcal{G}_v = \mathcal{G}_v^j) = P(y|\mathcal{G}_c^i \oplus \mathcal{G}_v^j)$, where $\mathcal{G}_c^i \oplus \mathcal{G}_v^j$ is the intervened graph and $\oplus$ denotes the vector combination. Overall, the two assumptions are reasonable and practical for handling noisy environments.

Based on the two assumptions, the backdoor adjustment given in Eq. (5) can be written as,

$$P\left(y\left|\text{do}\left(\mathcal{G}_c = \mathcal{G}_c^i\right)\right.\right) = \frac{1}{|\mathcal{V}|}\sum_{j=1}^{|\mathcal{V}|} P\left(y\,|\mathcal{G}_c^i \oplus \mathcal{G}_v^j\right) \tag{17}$$

where $\mathcal{V}$ is the confounder set that collects all the potential non-causal patterns from training data. Theoretically, backdoor adjustment requires traversing the non-causal graph data from any possible confounder set and combining with the causal graph data to compose the intervened graph data. Nevertheless, due to the specific attributes of the graph structures, it is impractical to generate the intervened graph data in the original graph data level. Through the aforementioned casual disentanglement, we can stratify the confounder apart from the causal factor and acquire every stratification of non-causal graph representations. Thus, we implement the backdoor adjustment in the graph representation level.

Further, given that the GNN model parameters are usually updated through minibatch, the types of non-causal features are constrained to the batch size, substantially limiting their diversity and making the backdoor adjustment less effective. To address this issue, we adopt the random pairing the causal features with the non-causal features from the enhanced confounder set. Concretely, if the batch size cannot be enlarged owning to the computation resource limit, we can apply a mixed way combined two data augmentation methods to generate more non-causal feature proxies to boost the diversity of the confounder set. The augmented non-causal samples can be interpolated by the randomly selected non-causal features (a pair $(\mathcal{Z}_{\mathcal{G}_v}^j, \mathcal{Z}_{\mathcal{G}_v}^k)$), denoted by

$$\mathcal{Z}_{\mathcal{G}_v}^{j'} = \begin{cases} \tau\mathcal{Z}_{\mathcal{G}_v}^j + (1-\tau)\mathcal{Z}_{\mathcal{G}_v}^k, & \text{all-feature mix,} \\ \mathbf{M}\circ\mathcal{Z}_{\mathcal{G}_v}^j + (1-\mathbf{M})\circ\mathcal{Z}_{\mathcal{G}_v}^k, & \text{partial-feature mix,} \end{cases} \tag{18}$$

where $\tau \in [0,1]$ controls the interpolation degree, $\mathbf{M}$ is the binary mask implemented on the feature channel dimension. Then, for each $\mathcal{Z}_{\mathcal{G}_c}^i$, we randomly select different $\mathcal{Z}_{\mathcal{G}_v}^{j'}$ and apply the random addition on them to estimate the effects of causal intervention in the graph representation level. It pushes the predictions of such intervened graph $\mathcal{Z}_{\mathcal{G}_c}^i \oplus \mathcal{Z}_{\mathcal{G}_v}^{j'}$ to be invariant and stable across different non-causal patterns, due to the shared fault-causal patterns.

Thereby, we design the backdoor adjustment loss to implement the causal intervention on the GNN-based FD model, which is defined as,

$$\mathcal{L}_{BA} = -\frac{1}{D_\mathcal{G}}\frac{1}{|\mathcal{V}|}\sum_{i}^{D_\mathcal{G}}\sum_{j}^{|\mathcal{V}|} P\left(y_\mathcal{G}^i \circ \Psi\left(\mathcal{Z}_{\mathcal{G}_c}^i \oplus \mathcal{Z}_{\mathcal{G}_v}^{j'}\right)\right) \tag{19}$$

where the fault-causal representations $\mathcal{Z}_{\mathcal{G}_c}$ and the non-causal representations $\mathcal{Z}_{\mathcal{G}_v}$ are derived from Eq. (10), and $\mathcal{V}$ is the confounder set that collects the potential non-causal patterns from training data.

As such, the whole training loss of CIE-GNN can be defined as the sum of the supervised classification loss $\mathcal{L}_C$, the causal disentanglement loss $\mathcal{L}_D$, and the backdoor adjustment loss $\mathcal{L}_{BA}$:

$$\mathcal{L} = \mathcal{L}_C + \alpha\mathcal{L}_D + \beta\mathcal{L}_{BA} \tag{20}$$

where α and β are hyper-parameters that determine the strength of causal disentanglement and backdoor adjustment, respectively. The detailed procedures of the proposed casual inference-enabled GNN fault diagnosis method is provided in Alg.1.

4.2.4. Theoretical analysis

Ultimately, we provide the theoretical analysis for our fault diagnosis model based on the proposed CIE-GNN. Based on the implementation process of our CIE-GNN model, it is potential for the proposed CIE-GNN to generalize in diverse working environments.

Lemma 1. *Let* $P\left(\mathcal{Z}_{\mathcal{G}_c}^i \left| \mathcal{Z}_{\mathcal{G}_v} = \mathcal{Z}_{\mathcal{G}_v}^j \right.\right)$ *be the class-conditional density function of the i-th fault-causal variable given the j-th non-causal variable. It can be proved that, when the optimal optimization of CIE-GNN is achieved, it will lead to,*

$$P\left(\mathcal{Z}_{\mathcal{G}_c}^i \left| \mathcal{Z}_{\mathcal{G}_v} = \mathcal{Z}_{\mathcal{G}_v}^j \right.\right) = P\left(\mathcal{Z}_{\mathcal{G}_c}^i\right), \forall \mathcal{Z}_{\mathcal{G}_v}^j \tag{21}$$

It indicates that in the latent space of fault-causal graph representations, the probability density will be invariant to different non-causal graph representations. Considering that non-causal factors typically stem from a range of working conditions and environmental noises, the fault-causal

Algorithm 1: Casual inference-enabled GNN

Input: The raw measurements of multiple sensors $\mathcal{X}$ and the corresponding fault labels $\mathcal{Y}$
Output: The CIE-GNN model

Construct graph data $\mathcal{G}$ based on Eq. (1) ;
Initialize the network parameters of the CIE-GNN model;
for *r = 1 to* R_g **do**
  Implement the causal disentanglement transformation to generate the fault-causal subgraph $\mathcal{G}_c$ and the non-causal subgraph $\mathcal{G}_v$ based on Eq.(6)-(7) ;
  Obtain their corresponding graph representations $\mathcal{Z}_{\mathcal{G}_c}$ and$\mathcal{Z}_{\mathcal{G}_v}$ based on Eq.(10);
  for $t = 1$ *to* R_{mi} **do**
    Sample $\left\{\left(\mathcal{Z}^i_{\mathcal{G}_c}, \mathcal{Z}^i_{\mathcal{G}_v}\right)\right\}_{i=1}^{N}$ from the batch;
    Calculate the log-likelihood $\mathcal{L}\left(q_\pi\right) = \frac{1}{N}\sum_{j=1}^{N} \log q_\pi\left(\mathcal{Z}^j_{\mathcal{G}_v} \middle| \mathcal{Z}^i_{\mathcal{G}_c}\right)$;
    Update q_π by maximization $\mathcal{L}\left(q_\pi\right)$;
  Calculate the MI upper bound $\hat{H}(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v})$ based on the updated q_π;
  Obtain the causal distanglement loss by Eq. (16), the supervised classification loss by Eq.(11), and the backdoor adjustment loss by Eq. (19);
  Update the network parameters of the CIE-GNN model based on the whole loss $\mathcal{L}$;
Return the CIE-GNN model;

factors are independent of the working environments. The independence guarantees that the proposed CIE-GNN can maintain invariant across diverse environments and ensures the generalization performance.

Proof. We assume that $q_\pi(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c})$ can be achieved as the accurate approximation for $p(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c})$ due to the high expressiveness of neural networks. Thus, after training the MI-estimator neural network, $\hat{H}_v$ can be tight bound to the MI contrastive log-ratio upper bound $\hat{H}$. By definition, the gap Δ between $\hat{H}$ and H is given by,

$$\begin{aligned}\Delta &\triangleq \hat{H}(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v}) - H(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v}) \\ &= \mathbb{E}_{p(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v})}[\log p(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c})] - \mathbb{E}_{p(\mathcal{Z}_{\mathcal{G}_c})}\mathbb{E}_{p(\mathcal{Z}_{\mathcal{G}_v})}[\log p(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c})] \\ &- \mathbb{E}_{p(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v})}[\log p(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c}) - \log p(\mathcal{Z}_{\mathcal{G}_v})] \\ &= \mathbb{E}_{p(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v})}[\log p(\mathcal{Z}_{\mathcal{G}_v})] - \mathbb{E}_{p(\mathcal{Z}_{\mathcal{G}_c})}\mathbb{E}_{p(\mathcal{Z}_{\mathcal{G}_v})}[\log p(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c})]\end{aligned} \tag{22}$$

When the MI contrastive log-ratio upper bound is minimized to the optimum, it implies that $\Delta = 0$. Based on Eq. (22), this equality can be rewritten using the definitions of entropy and expected values:

$$\begin{aligned}&\int\int p(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v}) \log p(\mathcal{Z}_{\mathcal{G}_v}) d\mathcal{Z}_{\mathcal{G}_c} d\mathcal{Z}_{\mathcal{G}_v} \\ &- \int\int p(\mathcal{Z}_{\mathcal{G}_c}) \log p(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c}) \log p(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c}) d\mathcal{Z}_{\mathcal{G}_v} d\mathcal{Z}_{\mathcal{G}_c} = 0\end{aligned} \tag{23}$$

Which can be simplified as,

$$\begin{aligned}&\int\int p(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v})(\log p(\mathcal{Z}_{\mathcal{G}_v}) - \log p(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c})) d\mathcal{Z}_{\mathcal{G}_c} d\mathcal{Z}_{\mathcal{G}_v} \\ &= \int\int p(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v}) \log \frac{p(\mathcal{Z}_{\mathcal{G}_v})}{p(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c})} d\mathcal{Z}_{\mathcal{G}_c} d\mathcal{Z}_{\mathcal{G}_v} = 0\end{aligned} \tag{24}$$

The above equality holds universally only if $\frac{p(\mathcal{Z}_{\mathcal{G}_v})}{p(\mathcal{Z}_{\mathcal{G}_v}|\mathcal{Z}_{\mathcal{G}_c})} = 1$. It implies that,

$$P\left(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v}\right) = P\left(\mathcal{Z}_{\mathcal{G}_c}\right) P\left(\mathcal{Z}_{\mathcal{G}_v}\right) \tag{25}$$

The above condition signifies that $\mathcal{Z}_{\mathcal{G}_c}$ and $\mathcal{Z}_{\mathcal{G}_v}$ are statistically independent. Then, let's consider the conditional probability $P(\mathcal{Z}_{\mathcal{G}_c}|\mathcal{Z}_{\mathcal{G}_v}) = P(\mathcal{Z}_{\mathcal{G}_c}, \mathcal{Z}_{\mathcal{G}_v})/P(\mathcal{Z}_{\mathcal{G}_v})$. By substituting Eq. (25), it can be simplified to $P(\mathcal{Z}^i_{\mathcal{G}_c}|\mathcal{Z}_{\mathcal{G}_v} = \mathcal{Z}^j_{\mathcal{G}_v}) = P(\mathcal{Z}^i_{\mathcal{G}_c}), \forall \mathcal{Z}^j_{\mathcal{G}_v}$, thus completing the proof. □

5. Experiment

5.1. Experimental settings

5.1.1. Dataset

In this work, we evaluate the performance of the proposed casual inference-enabled graph neural networks on two industrial fault diagnosis datasets, including **Tennessee Eastman (TE) process simulation dataset** [45] and **Three-phase Flow Facility (TFF) simulation dataset** [46]. The TE dataset and the TFF dataset were widely chosen for assessing the fault diagnosis method in IIoT-enabled systems [10, 47], because they involve a dispersed network of sensors throughout a manufacturing facility. It is an alternative to the standard Industry 4.0 setup, in which a plethora of decentralized sensor networks report their recordings to a central processing unit.

The TE dataset[1], developed by the Eastman Chemical Company, serves as a widely recognized benchmark for evaluating the efficacy of various fault diagnosis approaches. The TE process encompasses five critical operation units: input feed, reactor, separator, stripper, and compressor, involving a total of 33 variables. This dataset has gained significant traction in both research and industrial settings due to its inclusion of the normal type and 21 distinct fault categories. To facilitate comprehensive evaluation, the dataset

[1] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi 10.7910/DVN/6C3JR1#

Table 2
Descriptions of TE dataset and TFF dataset

Data set	Fault case	Description	Type	Fault case	Description	Type
TE	1	A/C feed ratio, B composition constant	Step	12	Condenser cooling water inlet temperature	Random variation
	2	B composition, A/C ratio constant	Step	13	Reactor kinetics	Slow drift
	3	D feed temperature	Step	14	Reactor cooling water valve	Sticking
	4	Reactor cooling water inlet temperature	Step	15	Condenser cooling water valve	Sticking
	5	Condenser cooling water inlet temperature	Step	16	Variation coefficient of the steam supply of heat exchange	Random variation
	6	A feed loss	Step	17	Variation coefficient of heat transfer in reactor	Random variation
	7	C header pressure loss-reduced availability	Step	18	Variation coefficient of heat transfer in condenser	Random variation
	8	A, B, C feed composition	Random variation	19	Unknown	Unknown
	9	D feed temperature	Random variation	20	Unknown	Unknown
	10	C feed temperature	Random variation	21	A feed temperature	Random variation
	11	CReactor cooling water inlet temperature	Random variation			
TFF	1	Air line blockage	Incipient	4	Open direct bypass	Incipient
	2	Water line blockage	Incipient	5	Slugging conditions	Intermittent
	3	Top separator input blockage	Incipient	6	Pressurization of the 2 th line	Abrupt

incorporates a diverse range of fault scenarios and varying working conditions, rendering it a robust and challenging resource for assessing the performance of fault diagnosis methodologies. Detailed information regarding the dataset can be found in Table 3. Besides, the TFF dataset [2], conducted at Cranfield University, offers a rigorously controlled setting for measuring the flow of water, oil, and air into a pressurized system. Within the SCADA system, 24 sensors are strategically placed to monitor critical process variables, encompassing pressure, flow rate, temperature, and density at various points within the system. This dataset involves 6 fault categories, thereby providing valuable resources for studying fault diagnosis in complex systems, particularly those involving multi-phase flows.

5.1.2. Benchmarks

To rigorously validate the effectiveness of the proposed method, we conduct a comprehensive comparative analysis against the state-of-the-art Transformer-based and GNN-based fault diagnosis methods. The details of these benchmarks are listed as follows:

- **CLFormer**: CLFormer [28] leverages a lightweight Transformer based on convolutional embedding and linear self-attention for fault diagnosis. The CLFormer block contains three modules, including embedding, projector, and forward module.
- **Convformer-NSE**: Convformer-NSE [29] integrates CNN and Transformer for fault diagnosis, and a novel channel attention mechanism based on multi-head self-attention is modified by adapting convolutional projection.
- **STAGED**: STAGED [10] combines GCN, long short-term memory networks, and attention mechanisms to implement fault diagnosis.
- **ChebyNet**: The ChebyNet-based fault diagnosis method in [19] utilizes Chebyshev polynomial approximations for graph convolutional operations.
- **GraphSage**: The GraphSage-based method [21] samples and aggregates information from the neighborhood of each node on large graphs.
- **HoGCN**: HoGCN [22] utilizes a higher-order graph neural network to learn features of different orders and capture graph structures of various granularity.
- **GCN**: The GCN-based method [12] leverages the foundational GNN architectures and aggregates information from neighboring nodes to update node representations.
- **GIN**: The GIN-based method [17] is designed to be invariant to the ordering of nodes in a graph, making it robust to fault diagnosis.
- **GAT**: The GAT-based method [14] employs multiple head attention mechanisms to dynamically assign weights to edges based on the features of the nodes connected by those edges.

[2] http://depts.washington.edu/control/LARRY/TE/download.html

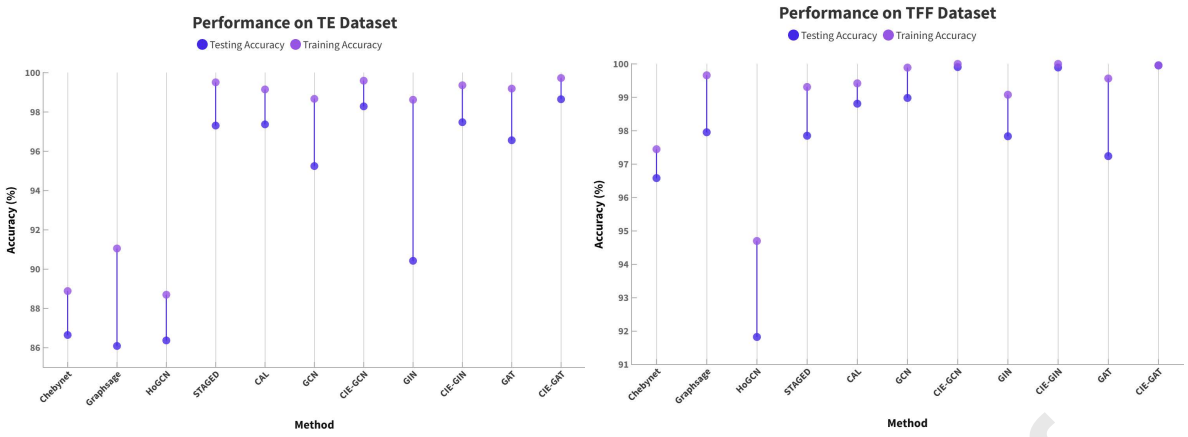


Figure 3: The fault diagnosis performances on the training and test sets of TE dataset and TFF dataset.

- **CAL**: CAL [38] leverages the graph attention learning on the basis of GCN model to identify causal patterns and mitigate the confounding effect. Specifically, it designs the KL-Divergence-based uniform classification loss for the trivial attended-graph and the backdoor adjustment loss.

5.1.3. Implementation Details

In this paper, we implement three casual inference-enabled GNN methods, including CIE-GCN, CIE-GIN, and CIE-GAT, to investigate the effectiveness in fault diagnosis tasks. To be specific, the graph encoders of CIE-GCN, CIE-GIN, and CIE-GAT are based on GCN, GIN, and GAT, respectively. To ensure a fair comparison, the corresponding compared GNN-based fault diagnosis methods share the same network architecture of the graph encoder as our proposed methods. For both datasets, the distance function adopts the cosine distance function, the pre-defined threshold is fixed to 1. It is worth noting that the proposed methods and the compared methods are implemented using PyTorch on a server equipped with NVIDIA RTX 2080 Ti and Intel(R) Xeon(R) W-2255 CPU@3.70GHz. This computational environment ensures the efficient training and evaluation of the models. To foster reproducibility and facilitate further research in this domain, we have made the source codes for the proposed method implementations publicly available online[3].

Besides, considering that the Gaussian noise is widely used in the existing studies [29, 44], we employ the random Gaussian noise injection to simulate the environmental noise, which is given by,

$$\tilde{\mathcal{X}}_{\mathrm{n}} = \gamma_n \cdot \mathcal{X} + \varepsilon_n; \gamma_n \sim \mathcal{N}(1, \sigma_{\mathrm{n}}), \varepsilon_n \sim \mathcal{N}(0, \sigma_{\mathrm{n}}) \quad (26)$$

where $\tilde{\mathcal{X}}_{\mathrm{n}}$ denotes as the noisy data and $\mathcal{X}$ corresponding the original data without noise. ε_n and γ_n refers to the added random noise term and the scale factor respectively, both following the Gaussian distribution. The noisy data is uniformly distributed across the entire dataset. The parameter σ_{n} is leveraged to regulate the noise intensify, such that the varied values of σ_{n} can be used to analyze the noise effect on the fault diagnosis.

5.1.4. Evaluation metrics

In order to evaluate the performance of the proposed fault diagnosis models, we provide the fault diagnosis accuracy on test dataset subjected to varying levels of noise, which can be expressed as,

$$\mathcal{A}_{\sigma_{\mathrm{n}}} = \sum_{\tilde{x}, y \sim \tilde{D}_n} \mathbb{1}(F(\tilde{x}) = y) / |\tilde{D}_n| \quad (27)$$

where $\tilde{D}_n$ is the noisy test dataset with the noise intensity level σ_n, $F(\tilde{x})$ is the predicted result of the fault diagnosis model. Further, to quantify the generalization capability of the proposed fault diagnosis model against noisy data, we introduce two novel metrics: the Gi-Score [48] (Gi) and the Pal-Score [48] (Pal). The two metrics allow us to compare the accuracy of the proposed model with that of an idealized model, whose accuracy remains unaffected by noise perturbations. The Gi-Score is calculated by taking the ratio of the area between the Perturbation Cumulative Density (PCD) curve of the idealized network and that of the actual network, divided by the total area below the idealized network's PCD curve. The Pal-Score is computed by taking the ratio of the area under the PCD curve for the top perturbation magnitudes divided by the area for the bottom magnitudes. The lower Gi-Score and the higher Pal-Score indicates the model can maintain better performance even under noisy environments.

We also assess our method to evaluate the computation complexity in terms of the number of model parameters, the GPU memory costs, and the training and inference time each epoch. Besides, we measure the fault diagnosis performances for the by precision, recall, and F1-Score. These metrics are calculated by $Prec = TP/(TP + FP)$, $Rec = TP/(TP + FN)$, and $F1 = 2 \cdot Prec \cdot Rec/(Prec + Rec)$, where TP corresponds to True Positive, FP represents False Positive, FN denotes as False Negative.

[3] https://github.com/zhangzhao156/CIE-GNN

Table 3
The fault diagnosis performances with different noise intensities on TE dataset

Method	$\sigma_n = 0.03$	$\sigma_n = 0.05$	$\sigma_n = 0.07$	$\sigma_n = 0.09$	$\sigma_n = 0.1$	$Gi \downarrow$	$Pal \uparrow$
CLFormer	91.60 ± 1.89	89.82 ± 1.90	87.92 ± 1.94	85.29 ± 2.74	83.20 ± 1.39	0.4723	0.9547
Convformer-NSE	92.39 ± 3.71	91.13 ± 3.38	89.54 ± 4.92	87.01 ± 2.60	80.64 ± 3.13	0.4662	0.9620
STAGED	95.14 ± 1.23	95.34 ± 1.45	94.62 ± 1.53	94.12 ± 1.70	93.95 ± 1.78	0.4445	0.9908
ChebyNet	87.02 ± 0.64	86.75 ± 0.68	86.52 ± 0.71	86.78 ± 0.48	83.14 ± 0.75	0.4850	0.9942
GraphSage	89.39 ± 1.52	86.78 ± 1.77	86.03 ± 1.35	85.34 ± 1.56	85.04 ± 1.30	0.4832	0.9727
HoGCN	87.45 ± 1.38	86.69 ± 1.32	85.80 ± 1.36	86.73 ± 1.74	86.92 ± 1.59	0.4853	0.9907
GCN	95.02 ± 1.46	94.88 ± 1.74	94.13 ± 1.65	92.69 ± 1.37	92.78 ± 1.79	0.4469	0.9837
GIN	91.57 ± 3.06	89.71 ± 2.70	89.30 ± 2.86	88.17 ± 3.96	87.02 ± 1.74	0.4695	0.9789
GAT	97.11 ± 0.62	97.98 ± 0.51	97.02 ± 0.61	93.39 ± 1.11	94.48 ± 0.59	0.4343	0.9760
CAL	96.96 ± 0.90	96.47 ± 1.45	96.56 ± 1.71	94.18 ± 3.37	96.46 ± 1.72	0.4376	0.9861
CIE-GCN	98.18 ± 0.41	97.99 ± 0.47	$\mathbf{97.98 \pm 0.50}^{\star}$	$\mathbf{97.20 \pm 0.64}^{\star}$	$\mathbf{95.73 \pm 1.05}^{\star}$	**0.4300**	**0.9949**
CIE-GIN	96.67 ± 0.60	96.67 ± 0.45	95.92 ± 0.46	95.99 ± 0.54	94.19 ± 0.98	0.4381	0.9939
CIE-GAT	$\mathbf{98.58 \pm 0.54}^{\star}$	$\mathbf{98.35 \pm 0.47}^{\star}$	97.03 ± 0.71	97.17 ± 0.66	95.14 ± 0.89	0.4301	0.9861

Table 4
The fault diagnosis performances with different noise intensities on TFF dataset

Method	$\sigma_n = 0.03$	$\sigma_n = 0.05$	$\sigma_n = 0.07$	$\sigma_n = 0.09$	$\sigma_n = 0.1$	$Gi \downarrow$	$Pal \uparrow$
CLFormer	99.36 ± 0.08	99.16 ± 0.69	97.9 ± 0.49	95.68 ± 3.70	96.15 ± 2.96	0.4271	0.9751
Convformer-NSE	99.59 ± 0.03	98.38 ± 0.26	97.27 ± 2.66	97.32 ± 1.91	97.68 ± 3.15	0.4283	0.9829
STAGED	99.35 ± 0.71	98.72 ± 0.84	97.91 ± 0.75	98.50 ± 1.30	97.89 ± 1.62	0.4266	0.9916
ChebyNet	96.83 ± 0.95	96.25 ± 0.99	96.18 ± 0.72	95.43 ± 0.88	95.57 ± 1.20	0.4381	0.9923
GraphSage	97.91 ± 0.59	97.55 ± 0.69	97.41 ± 0.72	97.18 ± 0.62	94.87 ± 0.92	0.4320	0.9955
HoGCN	93.71 ± 1.14	92.54 ± 1.95	91.51 ± 0.42	92.87 ± 0.55	92.04 ± 0.75	0.4562	0.9899
GCN	99.00 ± 0.84	98.97 ± 0.48	98.76 ± 0.66	97.50 ± 0.77	94.55 ± 1.30	0.4261	0.9913
GIN	97.99 ± 1.34	97.79 ± 1.00	95.86 ± 1.13	94.47 ± 1.22	94.68 ± 1.32	0.4345	0.9721
GAT	97.56 ± 0.52	94.30 ± 0.57	91.99 ± 0.93	91.36 ± 0.76	90.11 ± 1.23	0.4483	0.9556
CAL	99.87 ± 1.02	99.59 ± 3.07	98.56 ± 2.33	96.58 ± 1.37	99.75 ± 1.89	0.4242	0.9783
CIE-GCN	$\mathbf{99.90 \pm 0.05}^{\star}$	$\mathbf{99.90 \pm 0.06}^{\star}$	99.28 ± 0.22	$\mathbf{99.61 \pm 0.14}^{\star}$	$\mathbf{99.52 \pm 0.14}^{\star}$	**0.4213**	**0.9959**
CIE-GIN	99.85 ± 0.08	99.82 ± 0.08	$\mathbf{99.57 \pm 0.12}^{\star}$	99.29 ± 0.25	98.50 ± 0.35	0.4214	0.9954
CIE-GAT	99.77 ± 0.60	98.72 ± 0.59	98.82 ± 0.34	97.51 ± 0.63	94.45 ± 1.16	0.4256	0.9891

5.2. Experimental results

5.2.1. Performance evaluation

To synthetically evaluate the fault diagnosis performances of the proposed CIE-GNNs, we conduct extensive experiments on two industrial process datasets and compare the results with the state-of-art fault diagnosis methods.

Firstly, we rigorously appraise the fault diagnosis performance of our proposed CIE-GNNs in the absence of noise. The experimental results on both the training and test datasets are presented in Fig. 3. Results show that CIE-GNNs consistently outperform the existing GNN-based fault diagnosis methods, with a significantly reducing in the performance gap between the training and test datasets. Especially for TFF dataset, the accuracy of our method on the test set closely aligns with that on the training set. Note that the training and test datasets are collected under different working conditions and confront with distribution shifts. Existing GNN-based fault diagnosis methods often

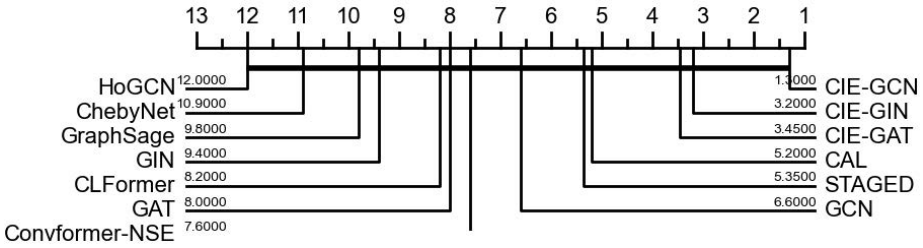


Figure 4: Wilcoxon–Holm post-hoc test on datasets for all baseline methods.

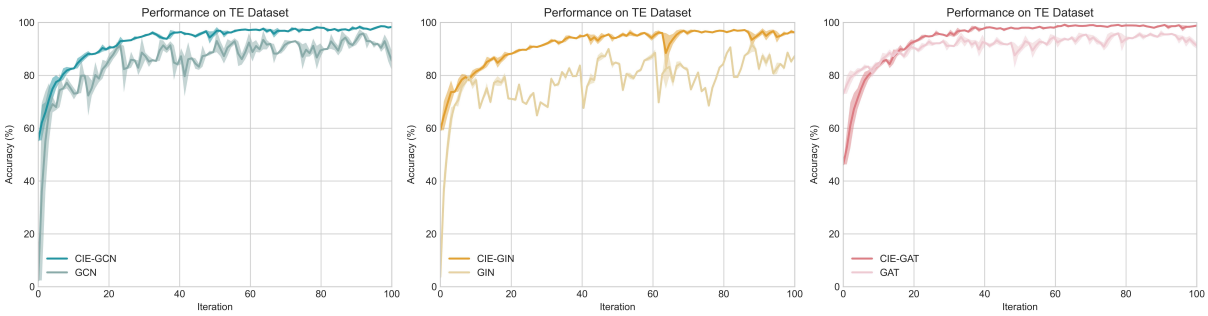


Figure 5: The accuracy curve of the proposed CIE-GNNs and the compared GNNs over the training process on TE dataset.

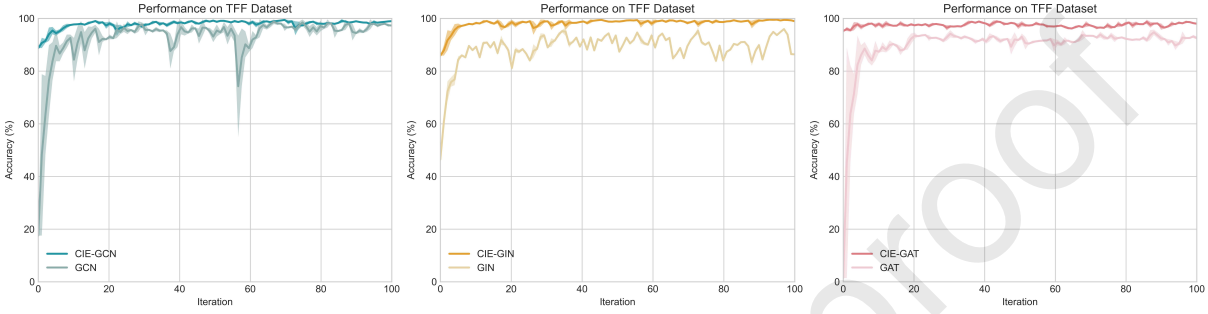


Figure 6: The accuracy curve of the proposed CIE-GNNs and the compared GNNs over the training process on TFF dataset.

suffer from learning spurious correlations between shortcut features and fault labels. These shortcut features may easily change outside the training distribution, thereby leading to substantial performance degradation when applied to the test dataset. On the contrary, CIE-GNNs employ the causal inference theory to eliminating environmental confounding factors and learning invariant graph representations, thus enhances the model's generalization under various working conditions. CAL, also empowered by causal inference, demonstrates superior performance over other GNN-based baselines when dealing with distribution shifts. This reiterates the efficacy of causal inference theory in enhancing the generalization capabilities of FD model.

Secondly, we conduct a thorough evaluation of the fault diagnosis performances of the proposed CIE-GNNs versus the state-of-the-art fault diagnosis methods under noisy conditions. The mean and variance of the fault diagnosis accuracies derived from repeated experiments are delineated in Table 3 and Table 4, with the best results highlighted in bold and marked with ⋆. As evident in Table 3, the accuracies of GNN-based fault diagnosis methods diminish with escalating the noise intensity. The performance degradation can be attributed to the introduction of noise that widens the distribution shifts between the training and testing datasets and deteriorates the generalization ability of GNN-based models under novel data distributions. Compared to the exiting methods, CIE-GNNs achieves the highest performance across both datasets, consistently exhibits higher mean accuracy, lower variance, lower Gi-Score, and higher Pal-Score. For instance, on TE dataset, with a noise level of 0.1, CIE-GNNs outperform the correspond GNN baselines

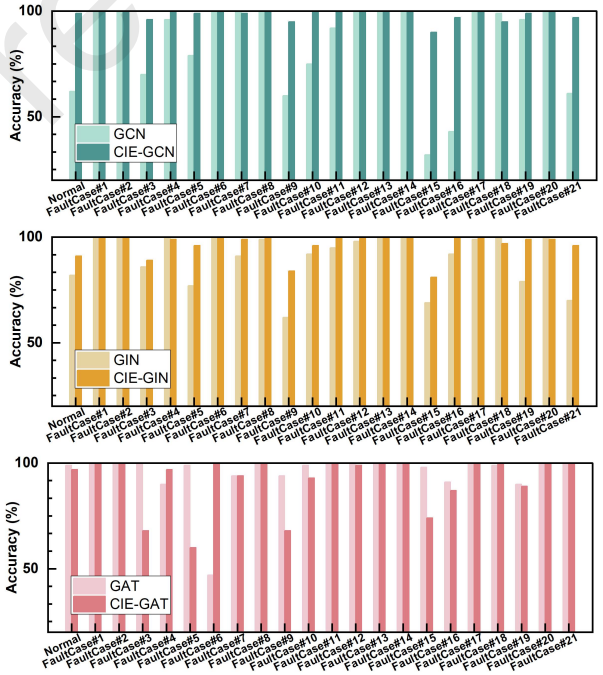


Figure 7: The fault diagnosis accuracy of each category on TE Dataset.

with improvements ranging from 0.66% to 7.17%, indicating its generalization capability in handling distribution shifts. The performance improvements own to its causal relationship modeling, capturing the environment-invariant graph representations under distribution shifts.

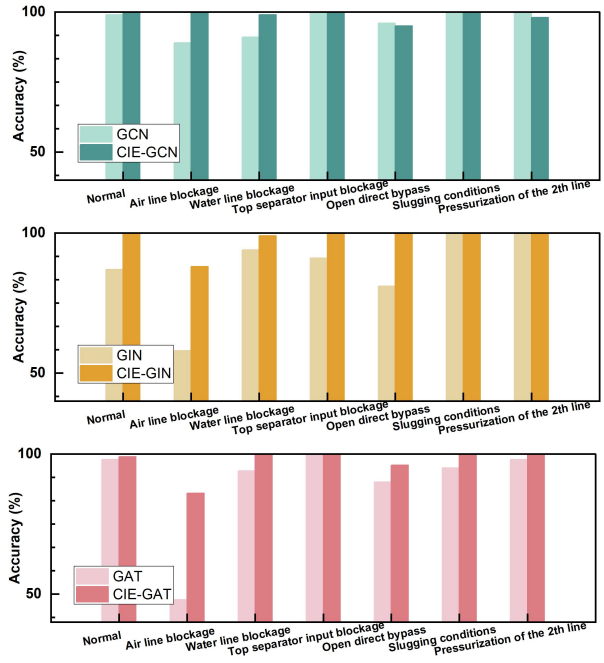


Figure 8: The fault diagnosis accuracy of each category on TFF Dataset.

In an effort to comprehensively assess the performances of all methods on two datasets, we delve into the statistical significance among them. Initially, we incorporate the Friedman rank-sum nonparametric statistical test with a confidence level of 95%. By utilizing the average rank of all results outlined in Table 3 and Table 4, we obtain a p-value of 1.208e-11 from the initial test. This remarkably low p-value provides compelling evidence to reject the null hypothesis about the absence of any significant statistical disparity among compared methods. Subsequently, to more accurately quantify statistically significant differences, we employ the Wilcoxon-Holm post-hoc test and visualize the results in a critical difference diagram, drawing upon average ranks. Fig. 4 lucidly delineates that the proposed CIE-GCN, CIE-GIN, and CIE-GAT emerge as the top three ranked approaches. This outstanding performance serves as a powerful testament to the enhanced generalization ability of our approach in achieving meaningful advancements over existing methods. Additionally, CIE-GNNs surpasses CAL thanks to its unique designs of the MI minimization-based causal disentanglement regularization and the random pairing-based backdoor adjustment regularization, capturing stable and invariant causal rationales from complex dependencies among variables.

Moreover, to further explore the performance disparity between the proposed CIE-GNNs and the corresponding GNN-based fault methods, we conduct extensive experiments under a noise level of 0.07. Fig. 5 and Fig. 6 depict their accuracy curves over the training process on TE dataset and TFF dataset, respectively. Under the noisy environment, the proposed CIE-GCN, CIE-GIN, and CIE-GAT exhibit higher fault diagnosis accuracy, faster convergence speed, and diminished fluctuations compared to the corresponding GNN methods. This is because the proposed causal disentanglement regularization and backdoor adjustment regularization help the model speed up the convergence speed under distribution shifts. Additionally, Fig. 7 and Fig. 8 showcase the fault diagnosis accuracy for each fault category on both datasets. Relative to the corresponding GNN-based methods, the proposed CIE-GNNs significantly enhance the diagnostic accuracies across all fault categories. For instance, CIE-GCN surpasses GCN on TE dataset, with the accuracy improvements of 32% and 43% in detecting Fault case#15 and Fault case#16 is, respectively. This stark improvement further underscores the effectiveness of the proposed CIE-GCN in learning distinguishable representation distributions in the noisy environments.

5.2.2. Complexity analysis

To further assess the efficacy of the proposed CIE-GNN framework, we delve into its computational complexity analysis. The computational complexity of CIE-GNN is determined by the training process of the CIE-GNN model, parameterized by θ, and the MI-Estimator model, parameterized by q_π. Updating the network parameters of the CIE-GNN model involves $R_g \cdot |D_\mathcal{G}| \cdot |\theta|$ floating-point operations, whereas updating the MI-Estimator model necessitates $R_g \cdot R_{mi} \cdot |D_{\mathcal{Z}_\mathcal{G}}| \cdot |q_\pi|$ floating-point operations. Consequently, the CIE-GNN model's time complexity is formulated as $O(R_g \cdot |D_\mathcal{G}| \cdot |\theta| + R_g \cdot R_{mi} \cdot |D_{\mathcal{Z}_\mathcal{G}}| \cdot |q_\pi|)$. Fig. 9 illustrates the average consuming time for training and inference of the proposed CIE-GNNs per epoch. Notably, the inference time for the proposed CIE-GNNs are all under 1 second on both datasets. Particularly for TE dataset, the inference time of CIE-GCN and CIE-GIN are as low as 0.31 seconds. These results underscore the CIE-GNNs' operational efficiency and their potential for real-world applications. Furthermore, the training time of the proposed CIE-GNNs significantly surpasses the inference time, attributed to the additional training phase required for the MI-Estimator's model parameter updates. In this experiment, R_{mi} is set to 5, resulting in considerable training duration for the MI-estimator. Thus, selecting an appropriate value for R_{mi} is crucial for reducing training time and ensuring convergence. Additionally, the training and inference time on TFF dataset exceed those on TE dataset, which can be ascribed to TFF dataset's larger training data volume.

Furthermore, we examine both the model complexity and the operating efficiency of all these methods on TE dataset, as illustrated in Table 5. Except for STAGED, the proposed CIE-GNNs exhibit a higher number of model parameters in comparison to GNN-based fault diagnosis methods. This disparity arises due to the additional network parameters required by the proposed CIE-GNNs to facilitate disentangled transformations for extracting fault-causal features and non-causal features. Likewise, the training time for CIE-GCN experiences an increase of 9.52 seconds over that of GCN, and the inference time rises by 0.03 seconds. We attribute the time increment to the incorporation of the

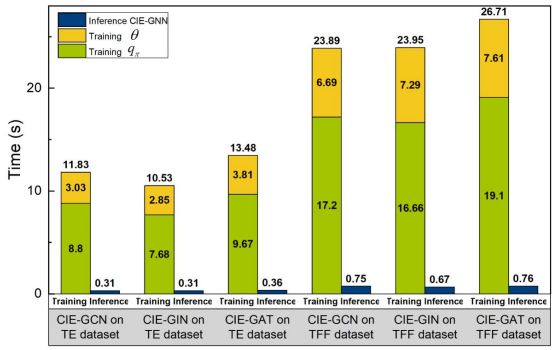


Figure 9: Training time and inference time of CIE-GNNs on TE Dataset and TFF dataset.

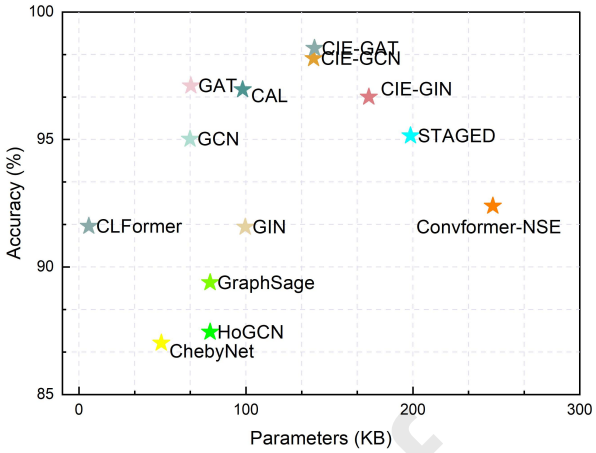


Figure 10: Comparison of average accuracy and model complexity on TE Dataset.

Table 5
Complex analysis of each model on TE dataset.

Method	Praramters (KB)	GPU Memory Costs (MB)	Training Time (s)	Inference Time (s)
CLFormer	5.88	0.66	8.76	0.39
Convformer-NSE	247.88	8.44	15.54	0.82
STAGED	198.62	12.13	2.25	0.30
ChebyNet	49.30	6.92	2.11	0.28
GraphSage	78.49	9.97	1.97	0.24
HoGCN	78.49	9.96	2.05	0.23
GCN	66.53	11.93	2.31	0.28
GIN	99.55	16.80	2.48	0.25
GAT	67.04	135.94	2.84	0.29
CAL	98.04	235.32	3.13	0.39
CIE-GCN	140.56	432.17	11.83	0.31
CIE-GIN	173.58	445.80	10.53	0.31
CIE-GAT	141.07	657.26	13.48	0.36

causal disentanglement loss and the backdoor adjustment loss in our model training. Nonetheless, given the enhanced performance of CIE-GCN relative to GCN, this increment in computation time is deemed acceptable, especially for the inference time. As for the Transformer-based fault diagnosis methods, CLFormer has the smallest model parameter size, but its training and inference times are not the shortest. This discrepancy stems from the lightweight Transformer architecture of CLFormer, which consumes extra time for temporal modeling.

Fig. 10 delineates a comparative analysis regarding the average fault diagnosis accuracy versus the model parameter size across different methods on TE dataset. CIE-GAT and CIE-CCN exhibit significantly superior fault diagnostic performances relative to the other methods, while still maintaining a moderate number of model parameters. Particularly, CIE-GAT boasts a model parameter size of 141.07K, reaching an impressive average fault diagnostic accuracy of 97%. Moreover, in contrast to the optimal Transformer-based method, Convformer-NSE, our CIE-GNNs effectuate a parameter size reduction by over 50% while improving the average accuracy by over 5%. Although CIE-GNNs require slightly more time compared to GNN-based fault diagnosis methods, their substantial performance improvements make the tradeoff between performance and efficiency acceptable.

5.2.3. Further analysis

1) Effect of the proposed causal disentanglement and backdoor adjustment mechanisms: To investigate the effect of the MI minimization-based causal disentanglement and the random pairing-based backdoor adjustment mechanisms, Fig. 11 presents the results of the ablation study that compares CIE-GCN and its two variants: 'w/o LD' (CIE-GCN trained without the causal disentanglement loss) and 'w/o LBA' (CIE-GCN trained without the backdoor adjustment loss). Results reveal that CIE-GCN training without the backdoor adjustment loss leads to a significant accuracy decline. This performance degradation can be attributed to the fact that it relies on the intervened graph for fault diagnosis. In the absence of the backdoor adjustment regularization, the random operations performed on the intervened graph can substantially deteriorate the fault diagnosis performance. This underscores the crucial role of the backdoor adjustment mechanism in eliminating environment confounding factors. In addition, CIE-GCN training without the causal disentanglement loss results in a slight accuracy decrease and a variance increase. These results corroborate the significance of incorporating the causal disentanglement regularization to capture invariant graph representations and enhance the stability of CIE-GCN under noisy environments. Overall, the ablation experiments highlight the importance of all modules in CIE-GCN for improving fault diagnosis performance.

Furthermore, in order to scrutinize the influence of the MI minimization-based causal disentanglement regularization, Fig. 12 provides the MI values of $\mathcal{Z}_{\mathcal{G}_c}$ and $\mathcal{Z}_{\mathcal{G}_v}$ at various training stages. Concretely, the MI values calculated correspond to the average values associated with $\mathcal{Z}_{\mathcal{G}_c}$ and $\mathcal{Z}_{\mathcal{G}_v}$ for each category. $T1$, $T2$, and $T3$ represent the 1-st, 10-th, and 30-th training round, respectively. It can be found that with the increase of training iterations, the MI value of $\mathcal{Z}_{\mathcal{G}_c}$

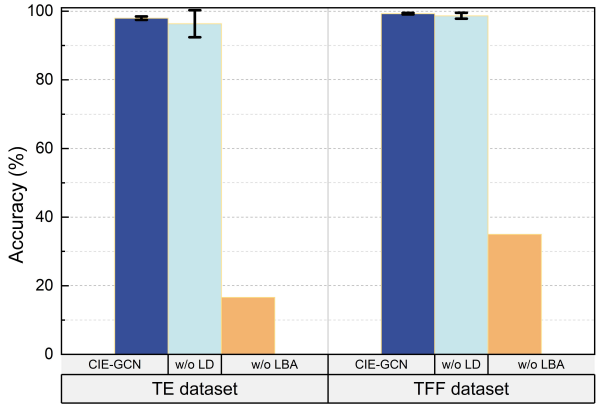


Figure 11: Ablation analysis of the proposed loss on TE dataset and TFF dataset.

and $\mathcal{Z}_{\mathcal{G}_v}$ consistently diminishes. This provides evidences that the proposed causal disentanglement regularization successfully reduces the MI of $\mathcal{Z}_{\mathcal{G}_c}$ and $\mathcal{Z}_{\mathcal{G}_v}$, thereby achieving causal decoupling. In addition, we offer visualizations of the intervened graph representations for CIE-GCN at different training stages via T-SNE (t-distributed Stochastic Neighbor Embedding), as depicted in Fig. 13 and Fig. 14. Results reveal that with the increasing number of training iterations, the representations of the same class progressively converge, the variance within the class shrinks, and the representations of different classes are further separated. This further attest to the effectiveness of the causal disentanglement regularization and its ability to extract causal representations with higher levels of generalization and discrimination.

2) Interpretability analysis: To delve deeper into the contribution of each component within the proposed CIE-GCN model, we present the fault diagnosis performances of the fault-casual subgraph $\mathcal{G}_c$, the non-causal subgraph $\mathcal{G}_v$, and the intervened subgraph $\mathcal{G}_c \oplus \mathcal{G}_v$, as outlined in Table 6, Fig. 15 and Fig. 16. The experimental findings reveal that, across both datasets, the fault-causal subgraph attains remarkably high fault diagnosis accuracies, while the non-causal subgraph exhibits notably lower accuracies. Particularly on TFF dataset, the fault-causal subgraph achieves an accuracy of 99.63%, contrasting sharply with the 5.69% accuracy of the non-causal subgraph. Specifically, the non-causal subgraph leading to considerable misclassifications on TFF dataset, with Fault Case #6 being erroneously classified as various fault types. This misclassification occurs independent of the class distribution, as depicted in Fig. 16. Additionally, it is noteworthy that the intervened subgraph performs slightly better than the fault-causal subgraph, providing further evidences of the effectiveness of the backdoor adjustment regularization.

Additionally, to provide interpretability analysis for the fault-causal subgraph $\mathcal{G}_c$ and the non-causal subgraph $\mathcal{G}_v$ learned by the proposed CIE-GCN, Fig. 17 present their graphs visualizations, where the size of the nodes and the width of the edges are proportional to their importance. The

Table 6
The fault diagnosis performances of $\mathcal{G}_c$, $\mathcal{G}_v$, and $\mathcal{G}_c \oplus \mathcal{G}_v$ on TE dataset and TFF dataset.

Dataset	$\mathcal{G}_c \oplus \mathcal{G}_v$	$\mathcal{G}_c$	$\mathcal{G}_v$
TE	97.22	96.98	4.95
TFF	99.67	99.63	5.69

fault-causal subgraphs $\mathcal{G}_c$ and the non-causal subgraphs $\mathcal{G}_v$ exhibit obviously different graph structures, due to the effectiveness of the causal disentanglement module. Besides, the fault-causal subgraphs of different categories present distinct patterns of graph structures, while the non-causal subgraphs of different categories have no significant differences. These observations also underscore that the fault-causal subgraph is directly pertinent to fault categories, whereas the non-causal subgraph influenced by working conditions or environmental noise, contributes negligibly to fault diagnosis.

3) Parameter sensitivity analysis: We carry out the parameter sensitivity analysis to discuss the strength of causal disentanglement and backdoor adjustment. Considering that the strength of causal disentanglement and backdoor adjustment are determined by the regularization factor α and β in the loss function, respectively, Fig. 18 reports the experimental results of CIE-GCN with varying α and β values. Results reveal that the performances of CIE-GCN improves with the increase in the strength of backdoor adjustment. This further confirms the crucial role of the proposed backdoor adjustment regularization in enhancing fault diagnostic performance. For TE dataset, when β reaches 0.8 and α reaches 0.003, CIE-GCN achieves the highest performance. For TFF dataset, when β is set within the range of 0.8 to 1.0, α in the range of 0.001 to 0.003, CIE-GCN achieve commendable performance. These experimental results provide a basis for us to choose the appropriate regularization factors in the experiment.

5.2.4. Application for industrial rotating machinery fault diagnosis

In a bid to further assess the performance of the proposed CIE-GNNs in real-world industrial rotating machinery fault diagnosis, we conduct the experimental validation on the XJTUSpurgear Dataset. This dataset is generated from an industrial platform that comprises a driving motor, belt, shaft, gearbox, and other components, as illustrated in Fig. 19. By using the transmission technology of IIoT, twelve accelerometers (PCB333B32) placed on the gearbox can collect and transmit vibration signal data to monitor various positions. Additionally, this dataset encompasses a total of five states, including the normal state and four distinct levels of root cracks, as depicted in Fig. 19.

Table 7 presents the average recall, precision, and F1-Score of all methods from repeated experiments conducted in a noisy environment. CIE-GAT achieves the highest values of Precision, Recall, and F1-Score, with an impressive F1-Score of 97.41%. CIE-GNNs surpass the correspond GNN baselines with improvements ranging from 1.49% to

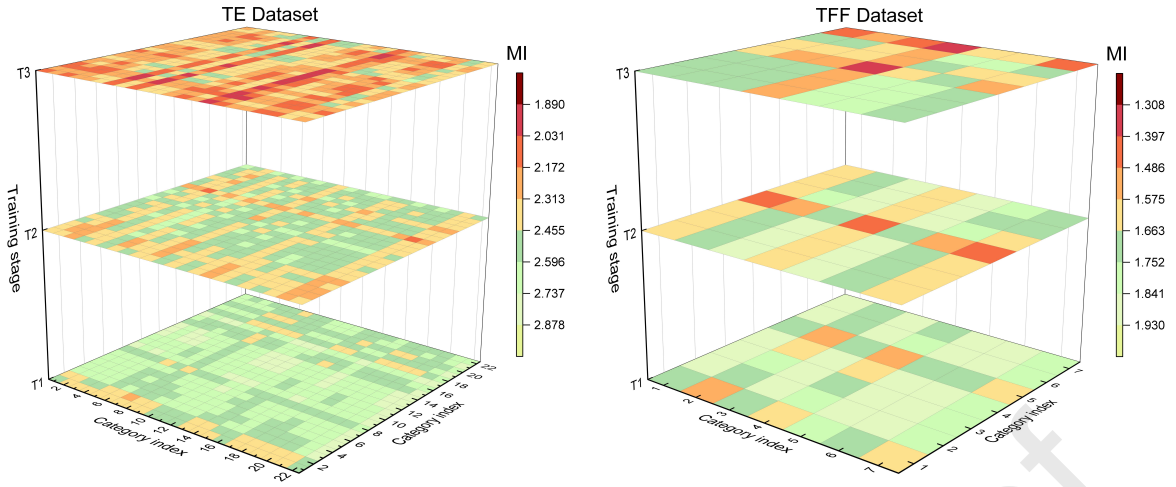


Figure 12: The MI values between the embeddings of $\mathcal{G}_c$ and those of $\mathcal{G}_c$ at differet stage on TE dataset and TFF dataset.

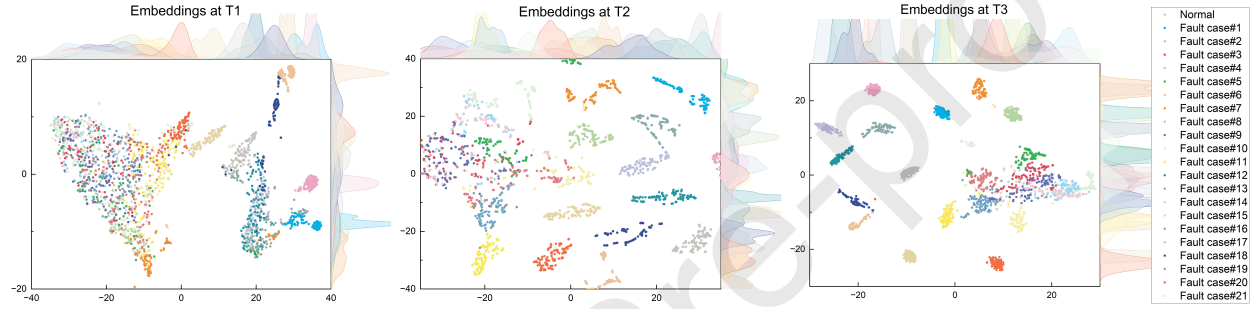


Figure 13: The embedding of CIE-GCN at differet stage on TE dataset.

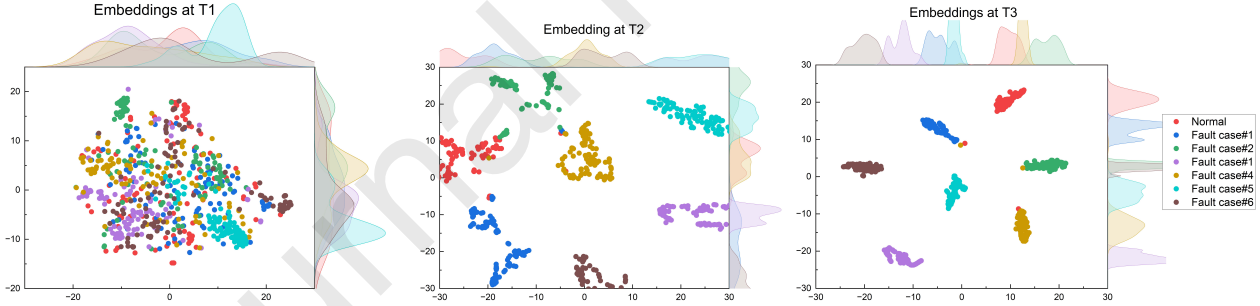


Figure 14: The embedding of CIE-GCN at differet stage on TFF dataset.

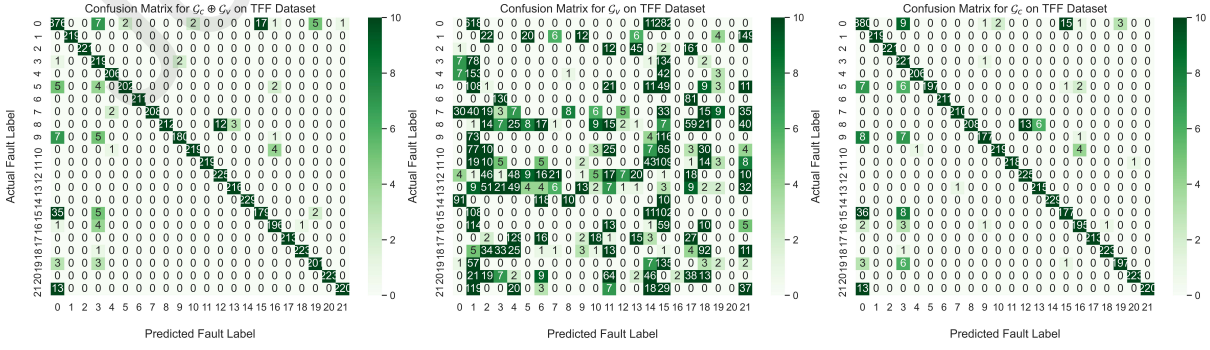


Figure 15: The confusion matrixes of the fault-casual subgraph $\mathcal{G}_c$, the non-causal subgraph $\mathcal{G}_v$, and the intervened subgraph $\mathcal{G}_c \oplus \mathcal{G}_v$ on TE dataset.

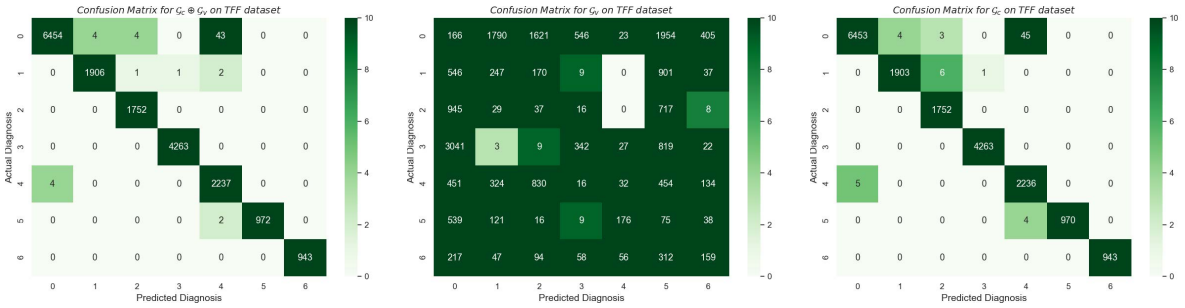
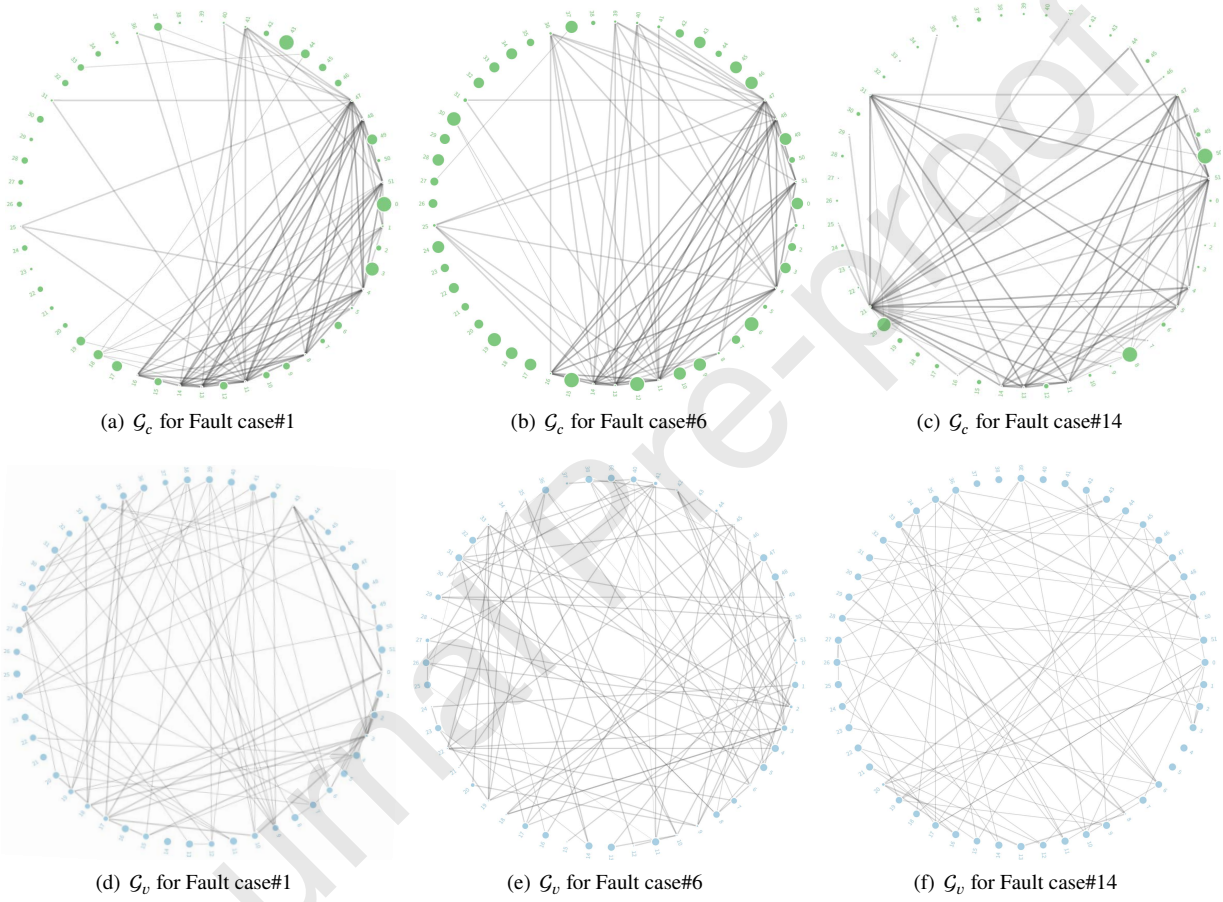


Figure 16: The confusion matrixes of the fault-casual subgraph $\mathcal{G}_c$, the non-causal subgraph $\mathcal{G}_v$, and the intervened subgraph $\mathcal{G}_c \oplus \mathcal{G}_v$ on TFF dataset.



(a) $\mathcal{G}_c$ for Fault case#1 (b) $\mathcal{G}_c$ for Fault case#6 (c) $\mathcal{G}_c$ for Fault case#14

(d) $\mathcal{G}_v$ for Fault case#1 (e) $\mathcal{G}_v$ for Fault case#6 (f) $\mathcal{G}_v$ for Fault case#14

Figure 17: The visualizations of the fault-casual subgraph $\mathcal{G}_c$ and the non-causal subgraph $\mathcal{G}_v$ for different categories on TE dataset.

4.20% in terms of F1-Score. Furthermore, Fig. 20 depicts the comparison results in the form of violin plots. CIE-GNNs exhibit a higher median value and a smaller variance of fault diagnosis accuracies compared to their corresponding GNN-based methods. Remarkably, unlike GCN, CIE-GCN substantially diminishes the accuracy fluctuation, with the experimental results predominantly concentrated around the median value. This result substantiates the superior efficacy of our method in industrial rotating machinery fault diagnosis. It is evident that the proposed CIE-GNN exhibits significant advantages over these GNN-based methods, underscoring the critical role of causal inference to elevate the fault diagnosis performance.

Notably, Table 7 demonstrates that CLFormer and Convformer-NSE display high diagnostic performance on XJTUSpurgear dataset, with F1-Score values of 0.9654 and 0.9648, respectively. This can be attributed to the fact that CLFormer and Convformer-NSE are specifically designed for fault diagnosis in industrial rotating machinery. Conversely, in contrast to their impact on TE dataset and TFF dataset, CIE-GNNs only demonstrate a marginal advantage over CLFormer

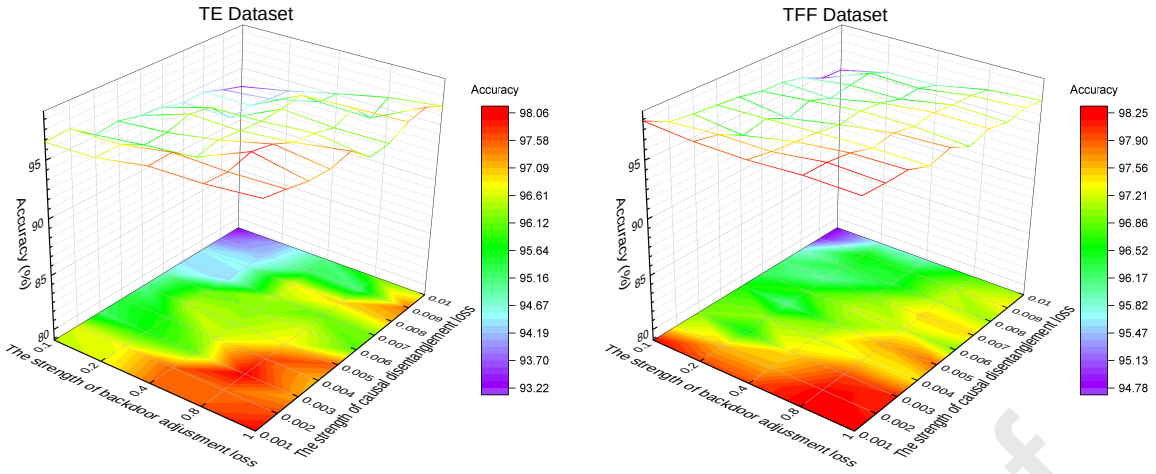


Figure 18: Parameter sensitivity analysis of α and β on TE dataset and TFF dataset.

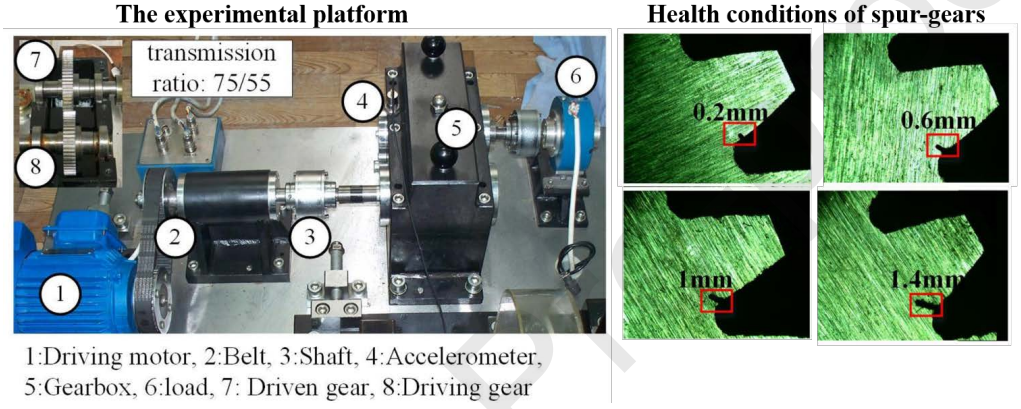


Figure 19: The experimental platform of XJTUSpurgear Dataset.

and Convformer-NSE on XJTUSpurgear dataset. This is because XJTUSpurgear dataset comprises a smaller number of sensors (12 sensors), and TE dataset and TFF dataset involve a larger number of sensors (51 and 24 sensors). This disparity demonstrates that CIE-GNNs are more suitable for fault diagnosis within large-scale systems effectively by harnessing the complex interactions among multiple sensor measurements. To further elucidate the impact of the number of sensors, we present the performances of CIE-GCN with varying numbers of sensor nodes on XJTUSpurgear dataset. Fig. 21 demonstrates the progressive diminishments in accuracy as the number of sensor nodes decreases. Additionally, Fig. 22 presents the confusion matrices of CIE-GCN with N=3 and N=12, revealing that CIE-GCN with N=3 exhibits a detection accuracy decrease across all fault types compared to N=12. This further confirms that our method may exhibit limitations when applied to systems with a small number of sensor nodes. In summary, these experimental results emphasize the importance of considering the scale and complexity of the system when applying our CIE-GNNs for fault diagnosis within real-world industrial settings.

Table 7
The fault diagnosis performances on XJTUSpurgear dataset.

Method	$\overline{Prec}$	$\overline{Rec}$	$\overline{F1}$
CLFormer	0.9689	0.9681	0.9654
Convformer-NSE	0.9708	0.9673	0.9648
ChebyNet	0.9591	0.9585	0.9585
GraphSage	0.9657	0.9635	0.9637
HoGCN	0.9132	0.9127	0.9125
STAGED	0.9222	0.9153	0.9147
GCN	0.9301	0.9238	0.9236
GIN	0.9351	0.9305	0.9304
GAT	0.9613	0.9593	0.9592
CAL	0.9648	0.9640	0.9639
CIE-GCN	0.9664	0.9656	0.9656
CIE-GIN	0.9691	0.9686	0.9686
CIE-GAT	0.9747*	0.9742*	0.9741*

6. Conclusion

In this paper, we propose Causal Inference-Enabled Graph Neural Networks (CIE-GNN) for generalized fault

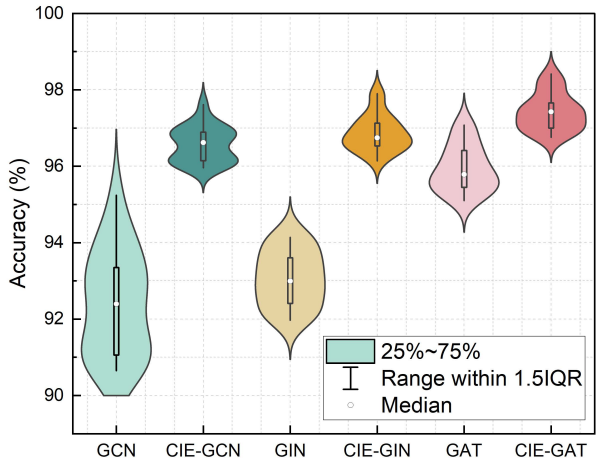


Figure 20: The fault diagnosis accuracy of each category on XJTUSpurgear Dataset.

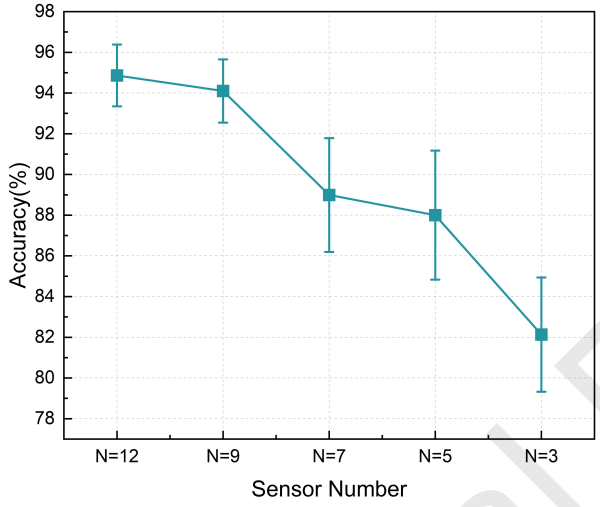


Figure 21: The fault diagnosis accuracy of CIE-GCN with different node numbers on XJTUSpurgear Dataset.

diagnosis in large-scale IIoT systems. We design the disentangled transformation module with the learnable representational masks to generate the fault-causal subgraphs and the non-causal subgraphs. Furthermore, we propose two novel regularization loss functions: the causal disentanglement loss, designed to facilitate the effective decoupling of these subgraphs, and the backdoor adjustment loss, aimed at mitigating the adverse impacts of non-causal factors. Extensive experiments and theoretical analysis verify the superior performance and generalization ability of the proposed CIE-GNNs across diverse working environments. Whereas, the causal inference mechanism augments the computational demands during model training, resulting in an increased computational complexity.

In future work, we aim to design a more lightweight CIE-GNN network architecture to strike a balance between performance and computational efficiency, thereby improving the scalability and feasibility of CIE-GNN for real-world applications. Furthermore, our future research will investigate the adaptability and robustness of our method when confronted with adversarial samples and unknown fault types, which can potentially invalidate and compromise the fault diagnosis system.

References

[1] Pivoto D G S, de Almeida L F F, da Rosa Righi R, et al. Cyber-physical systems architectures for industrial internet of things applications in Industry 4.0: A literature review[J]. Journal of manufacturing systems, 2021, 58: 176-192.

[2] Wang S, Liu Z, Jia Z, et al. Intermittent fault diagnosis for electronics-rich analog circuit systems based on multi-scale enhanced convolution transformer network with novel token fusion strategy[J]. Expert Systems with Applications, 2024, 238: 121964.

[3] Liu H, Xu Q, Han X, et al. Attention on the key modes: Machinery fault diagnosis transformers through variational mode decomposition[J]. Knowledge-Based Systems, 2024, 289: 111479.

[4] Yin M, Li J, Shi Y, et al. Fusing logic rule-based hybrid variable graph neural network approaches to fault diagnosis of industrial processes[J]. Expert Systems with Applications, 2024, 238: 121753.

[5] Calabrese F, Regattieri A, Bortolini M, et al. Data-Driven Fault Detection and Diagnosis: Challenges and Opportunities in Real-World Scenarios[J]. Applied Sciences, 2022, 12(18): 9212.

[6] Djenouri Y, Belhadi A, Srivastava G, et al. Fast and accurate deep learning framework for secure fault diagnosis in the industrial internet of things[J]. IEEE Internet of Things Journal, 2021.

[7] Kumar D, Ujjan S M, Dev K, et al. Towards soft real-time fault diagnosis for edge devices in industrial IoT using deep domain adaptation training strategy[J]. Journal of Parallel and Distributed Computing, 2022, 160: 90-99.

[8] Zhang S, Tong H, Xu J, et al. Graph convolutional networks: a comprehensive review[J]. Computational Social Networks, 2019, 6(1): 1-23.

[9] Chen D, Liu R, Hu Q, et al. Interaction-aware graph neural networks for fault diagnosis of complex industrial processes[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.

[10] Li S, Meng W, He S, et al. STAGED: A Spatial-Temporal Aware Graph Encoder-Decoder for Fault Diagnosis in Industrial Processes[J]. IEEE Transactions on Industrial Informatics, 2023.

[11] Jia M, Liu Y, Xu D, et al. Topology-Informed Graph Convolutional Network for Fault Diagnosis[C]//2022 IEEE 11th Data Driven Control and Learning Systems Conference (DDCLS). IEEE, 2022: 595-599.

[12] Yin P, Nie J, Liang X, et al. A multi-scale graph convolutional neural network framework for fault diagnosis of rolling bearing[J]. IEEE Transactions on Instrumentation and Measurement, 2023.

[13] Chen D, Liu R, Yu W, et al. Fault Diagnosis of Industrial Control System With Graph Attention Network on Multi-view Graph[C]//2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT). IEEE, 2021: 617-623.

[14] Liu K, Nie G, Jiao S, et al. Research on fault diagnosis method of vehicle cable terminal based on time series segmentation for graph neural network model[J]. Measurement, 2024: 114999.

[15] Ni P, Zhang Y, Xiong X, et al. MSGAFN: Multi-scale graph attention fusion network for machine fault diagnosis[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2024: 09544062241230217.

[16] Li C, Mo L, Yan R. Fault diagnosis of rolling bearing based on WHVG and GCN[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-11.

[17] Gao S, Li Y, Zhao D. A Novel Directed Graph Convolutional Neural Network for Rolling Bearings in Fault Diagnosis[C]//2023 IEEE International Conference on Unmanned Systems (ICUS). IEEE, 2023: 1207-1212.

[18] Li T, Zhao Z, Sun C, et al. Multireceptive field graph convolutional networks for machine fault diagnosis[J]. IEEE Transactions on Industrial Electronics, 2020, 68(12): 12739-12749.

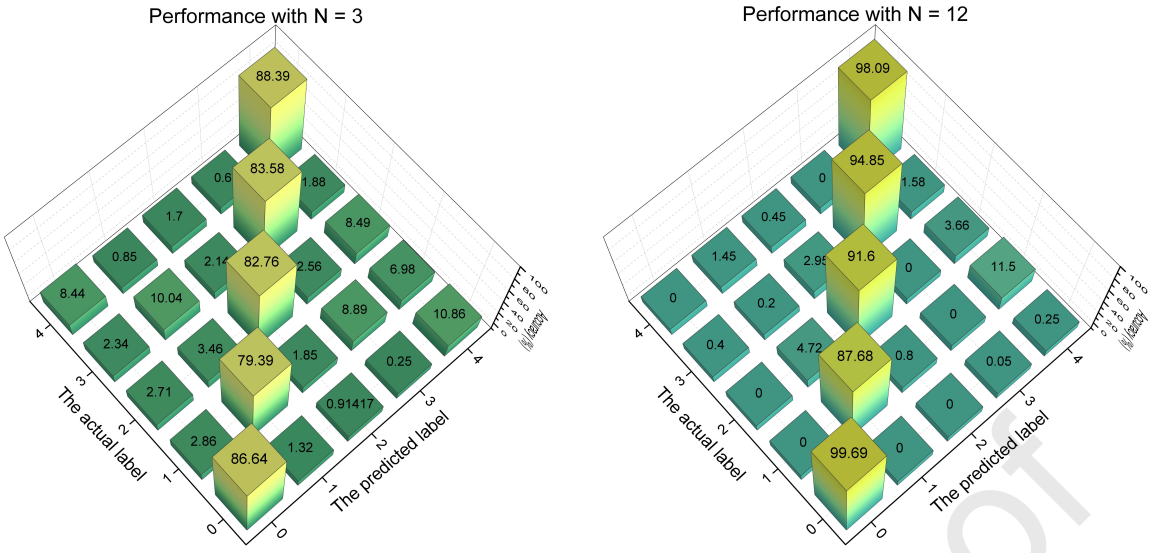


Figure 22: Performances of CIE-GCN with N=3 and N=12 on XJTUSpurgear Dataset.

[19] Zhang H, Wang Z, Qiu L, et al. Polynomial Improved Convolution Kernel Graph Network For Fault Diagnosis[C]//2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE). IEEE, 2023: 15-19.

[20] Wang H, Zhang Z, Xiong H, et al. GRAND: A Graph Neural Network Framework for Improved Diagnosis[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023.

[21] Wang L, Xie F, Zhang X, et al. Spatial-temporal graph feature learning driven by time–frequency similarity assessment for robust fault diagnosis of rotating machinery[J]. Advanced Engineering Informatics, 2024, 62: 102711.

[22] Li X, Xie L, Deng B, et al. Deep dynamic high-order graph convolutional network for wear fault diagnosis of hydrodynamic mechanical seal[J]. Reliability Engineering & System Safety, 2024, 247: 110117.

[23] Li T, Zhao Z, Sun C, et al. Domain adversarial graph convolutional network for fault diagnosis under variable working conditions[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-10.

[24] Jarwar M A, Khowaja S A, Dev K, et al. NEAT: A resilient deep representational learning for fault detection using acoustic signals in IIoT environment[J]. IEEE Internet of Things Journal, 2021.

[25] Pourkeshavarz M, Zhang J, Rasouli A. CaDeT: a Causal Disentanglement Approach for Robust Trajectory Prediction in Autonomous Driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 14874-14884.

[26] Jia L, Chow T W S, Yuan Y. Causal disentanglement domain generalization for time-series signal fault diagnosis[J]. Neural Networks, 2024, 172: 106099.

[27] Mooij J M, Janzing D, Schölkopf B. From ordinary differential equations to structural causal models: the deterministic case[J]. arXiv preprint arXiv:1304.7920, 2013.

[28] Fang H, Deng J, Bai Y, et al. CLFormer: A lightweight transformer based on convolutional embedding and linear self-attention with strong robustness for bearing fault diagnosis under limited sample conditions[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 71: 1-8.

[29] Han S, Shao H, Cheng J, et al. Convformer-NSE: A novel end-to-end gearbox fault diagnosis framework under heavy noise using joint global and local information[J]. IEEE/ASME Transactions on Mechatronics, 2022, 28(1): 340-349.

[30] Rao S, Zou G, Yang S, et al. A feature selection and ensemble learning based methodology for transformer fault diagnosis[J]. Applied Soft Computing, 2024, 150: 111072.

[31] Zhang K, Sun W, Ba Y, et al. Transformer Fault Diagnosis Method Based on SCA-VMD and Improved GoogLeNet[J]. Applied Sciences, 2024, 14(2): 861.

[32] Wu X, Peng H, Cui X, et al. Multi-Channel vibration signals fusion based on rolling bearings and MRST-Transformer fault diagnosis model[J]. IEEE Sensors Journal, 2024.

[33] Hanif A, Ali S, Ahmed A. A framework for fault diagnosis using continuous bayesian network and causal inference[C]//2021 IEEE 19th International Conference on Industrial Informatics (INDIN). IEEE, 2021: 1-8.

[34] Chen J, Zhao C. Multi-lag and multi-type temporal causality inference and analysis for industrial process fault diagnosis[J]. Control Engineering Practice, 2022, 124: 105174.

[35] Jia S, Li Y, Wang X, et al. Deep causal factorization network: A novel domain generalization method for cross-machine bearing fault diagnosis[J]. Mechanical Systems and Signal Processing, 2023, 192: 110228.

[36] Uchida Y, Fujiwara K, Saito T, et al. Causal Plot: Causal-Based Fault Diagnosis Method Based on Causal Analysis[J]. Processes, 2022, 10(11): 2269.

[37] Li J, Wang Y, Zi Y, et al. Causal consistency network: A collaborative multimachine generalization method for bearing fault diagnosis[J]. IEEE Transactions on Industrial Informatics, 2023, 19(4): 5915-5924.

[38] Sui Y, Mao W, Wang S, et al. Enhancing Out-of-distribution Generalization on Graphs via Causal Attention Learning[J]. ACM Transactions on Knowledge Discovery from Data, 2024, 18(5): 1-24.

[39] Liu R, Zhang Q, Lin D, et al. Causal intervention graph neural network for fault diagnosis of complex industrial processes[J]. Reliability Engineering & System Safety, 2024, 251: 110328.

[40] Pearce N, Lawlor D A. Causal inference—so much more than statistics[J]. International journal of epidemiology, 2016, 45(6): 1895-1903.

[41] Hagmayer Y, Sloman S A, Lagnado D A, et al. Causal reasoning through intervention[J]. Causal learning: Psychology, philosophy, and computation, 2007: 86-100.

[42] Cheng P, Hao W, Dai S, et al. Club: A contrastive log-ratio upper bound of mutual information[C]//International conference on machine learning. PMLR, 2020: 1779-1788.

[43] Yue Z, Zhang H, Sun Q, et al. Interventional few-shot learning[J]. Advances in neural information processing systems, 2020, 33: 2734-2746.

[44] Xiao D, Qin C, Yu H, et al. Unsupervised machine fault diagnosis for noisy domain adaptation using marginal denoising autoencoder based

on acoustic signals[J]. Measurement, 2021, 176: 109186.
[45] Rieth C A, Amsel B D, Tran R, et al. Additional tennessee eastman process simulation data for anomaly detection evaluation[J]. Harvard Dataverse, 2017, 1: 2017.
[46] Ruiz-Cárcel C, Cao Y, Mba D, et al. Statistical process monitoring of a multiphase flow facility[J]. Control Engineering Practice, 2015, 42: 74-88.
[47] Anitha C, Rajesh Kumar T, Balamanigandan R, et al. Fault Diagnosis of Tenessee Eastman Process with Detection Quality Using IMVOA with Hybrid DL Technique in IIOT[J]. SN Computer Science, 2023, 4(5): 458.
[48] Schiff Y, Quanz B, Das P, et al. Predicting deep neural network generalization with perturbation response curves[J]. Advances in Neural Information Processing Systems, 2021, 34: 21176-21188.

Declaration of interests

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

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: