Reliability Engineering and System Safety 251 (2024) 110328

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Causal intervention graph neural network for fault diagnosis of complex industrial processes

Ruonan Liu [a], Quanhu Zhang [a], Di Lin [a,*], Weidong Zhang [b,c], Steven X. Ding [d]

[a] College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China
[b] School of Information and Communication Engineering, Hainan University, Haikou, 570228, China
[c] Department of Automation, Shanghai Jiao Tong University, Shanghai, 200240, China
[d] School of Automation, University of Duisburg–Essen, Duisburg, 47057, Germany

ARTICLE INFO

Keywords:
Complex industrial processes
Fault diagnosis
Graph neural networks
Causal intervention
Instrumental variable

ABSTRACT

With the development of industry and manufacturing, the mechanical structures of equipment have become intricate and complex. Due to the interaction between components, once a failure occurs, it will propagate through the industrial processes, resulting in multiple sensor anomalies. Identifying the root causes of faults and eliminating interference from irrelevant sensor signals are critical issues in enhancing the stability and reliability of intelligent fault diagnosis. The components of industrial processes and their interactions can be represented by a structural attribute graph. The causal subgraph formed by fault signals determines the fault mode, while irrelevant sensor signals constitute a non-causal subgraph. The structure of non-causal subgraphs is relatively simple, and graph neural networks tend to use this part as a shortcut for prediction, leading to a significant decrease in prediction accuracy. To address this issue, a causal intervention graph neural network (CIGNN) framework is proposed. First, the sensor signals are constructed into structural attribute graphs using an attention mechanism. Due to causal and confounding features are highly coupled in graphs, explicitly decoupling them is almost impossible. Then, we design an instrumental variable to implement causal intervention to mitigate the confounding effect. Experimental results on two complex industrial datasets demonstrate the reliability and effectiveness of the proposed method in fault diagnosis.

1. Introduction

Fault diagnosis is an essential element of prognostics and health management, which is crucial for the safe and reliable operation of complex industrial systems and extending the lifespan of machines [1–3]. Compared to traditional algorithms, deep learning-based methods can adaptively extract features from vibration signals through model training, significantly enhancing diagnosis efficiency and accuracy [4]. Various deep learning models have been applied in fault diagnosis, including convolutional neural networks [5], recurrent neural networks [6,7], and auto-encoder [8].

Multi-source heterogeneous data provides comprehensive equipment information, enhancing the accuracy and reliability of fault diagnosis. However, the effective integration and alignment of such data remains a persistent challenge. Miao et al. proposed a deep feature interaction network for mechanical fault diagnosis, aiming to achieve adaptive feature fusion of multi-source heterogeneous data [9]. Class imbalance caused by the lack of data information is a common problem in data-driven methods. Tian et al. proposed a weighted modified conditional variational autoencoder as a data augmentation technique to solve this issue [10]. Su et al. fused prior knowledge with key health information extracted from raw monitoring data, enhancing the interpretability and robustness of intelligent fault diagnosis [11].

These methods mostly use data from Euclidean space, ignoring the topological structure of complex industrial processes and the interactions between components. Industrial components and their relationships can be represented as a structural attribute graph [12]. Graph Neural Networks (GNN) exploit inductive biases related to functional dependencies and non-Euclidean representations to effectively model and analyze graph data [13,14], demonstrating outstanding performance in fault diagnosis [15–17].

Fault diagnosis is a classification task, and GNN-based fault diagnosis is a graph classification task [12]. Graph classification is typically determined by its causal substructures rather than the entire graph. For instance, the properties of chemical molecules depend on chemical bonds and functional groups, rather than non-causal substructures like ion pairs or hydrogen bonds [18]. The placement of sensors is determined by the functional causal relationships among industrial

* Corresponding author.
E-mail address: Ande.lin1988@gmail.com (D. Lin).

https://doi.org/10.1016/j.ress.2024.110328

Available online 9 July 2024

Acronyms and Abbreviations

G	Structural Attribute Graph
G'	Causal Intervention Graph
V	Node Set
E	Edge Set
A	Adjacency Matrix
X	Node Features
$\mathbf{X}^e$	Edge Attributes
S	Structure Information of Graph
Y	Fault Mode
C	Causal Variable
B	Confounding Variable
Z	Instrumental Variable
$\mathcal{L}_C$	Fault Diagnosis Loss
$\mathcal{L}_{IV}$	Causal Intervention Loss
$\mathcal{L}_R$	Node Feature Reconstruction Loss
$\mathcal{L}_{TOTAL}$	Total Loss of Causal Intervention Graph Neural Network Model
SCM	Structural Causal Model
TFF	Three-phase Flow Facility
NPS	Nuclear Power System
CIGNN	Causal Intervention Graph Neural Network
KNN	K-Nearest Neighbor Classification
GAT	Graph Attention Networks
IAGNN	Interaction-Aware Graph Neural Networks
PKT-MCNN	Progressive Knowledge Transfer-based Multitask Convolutional Neural Network
ROC	Receiver Operating Characteristic
t-SNE	t-distributed Stochastic Neighborhood Embedding

equipment, structures, and components. Once a failure occurs, it propagates through the industrial system [19,20], affecting the normal operation of other components and causing multiple sensors to emit abnormal signals [21]. These anomalous sensor signals constitute causal subgraphs of the structural attribute graph that determine the failure modes, while the other sensor signals constitute the non-causal subgraphs. Due to the non-causal subgraph has a simpler structure, GNN tends to use this part as a shortcut for prediction, resulting in a significant decrease in prediction accuracy [22]. Therefore, GNN-based fault diagnosis methods need to address the interference of the non-causal subgraphs.

Correlation is particularly important for process monitoring and fault diagnosis as it reveals predictive relationships that can be used in practice. However, correlation analysis can only assess the correlation of data and fails to reveal the causal relationships inherent in industrial processes. Causal learning analyzes the deterministic relationships in the prediction based on causal theory, which is rapidly developing in image recognition [23], natural language processing [24], and out-of-distribution (OOD) [25]. Fault diagnosis based on causal learning is also becoming a research hotspot. Li et al. proposed a causal consistency network to address data bias caused by changes in machine working conditions [26]. Sensors are susceptible to measurement noise and process noise, which constrains the capability of data-driven methods to perform causal analysis. Wang et al. proposed causal-trivial attention graph neural network based on the backdoor adjustment formula to mitigate noise interference [27]. Zhang et al. proposed an anti-causal learning approach to estimate transfer effects between source and target domains thereby improving the cross-domain adaptation of the model [28]. However, these methods primarily address uncertainties arising from data bias, noise, and domain variation, without further analyzing the causal relationships between sensor signals and fault.

To address this problem, GNN-based fault diagnosis needs to face two challenges: (1) Considering that non-causal subgraphs are more likely to be learned by GNN and thus interfere with prediction, identifying causal and confounding features in graph is the first challenge to be addressed. Sensor signals that constitute causal subgraphs are causal features, while irrelevant sensor signals are considered as confounding features. (2) Causal and confounding features are highly coupled in graph, making it a challenge to disentangle them. Instrumental variable provide a causal intervention method that eliminate spurious correlations between non-causal subgraphs and fault without the need for direct observation of the confounding features.

Based on the above analysis, we propose an intelligent fault diagnosis method based on causal intervention graph neural network (CIGNN). Specifically, we analyze the causality in GNN-based fault diagnosis process based on the causal theory, and identify irrelevant sensor signals as confounding features. Given the difficulty that confounding features cannot be directly observed, we design an instrumental variable to mitigate the confounding effect. The contributions are as follows.

(1) This study formulates fault diagnosis in complex industrial processes as a graph classification task. The causality between sensor signals and faults in GNN-based fault diagnosis is analyzed based on causal theory, and a CIGNN fault diagnosis framework is proposed.

(2) A graph construction module is proposed that uses an attention mechanism to construct structural attribute graphs based on sensor signals to initially learn the topology and interactions between components.

(3) An instrumental variable is designed to implement causal intervention to enhances model's ability to extract causal features, thereby mitigating confounding effects caused by confounding features.

(4) Extensive experiments conducted on two complex industrial processes data. Experimental results and analysis demonstrate the superiority of CIGNN compared to state-of-art methods, offering a viable direction for developing models with enhanced interpretability and robustness.

The rest of the article is organized as follows. Section 2 first introduces background and problem formulation, and then elaborates on the construction and details of CIGNN . In Section 3, the effectiveness of the proposed method is validated on two complex industrial process datasets. Finally, Section 4 concludes this article.

2. Causal Intervention Graph Neural Network

In this study, CIGNN is developed to address the confounding effects caused by irrelevant features in fault diagnosis. As depicted in Fig. 1, the framework comprises two parts: (1) Constructing industrial process sensor signals as graph data through an attention mechanism. (2) Designing an instrumental variable to perform causal intervention on the input graphs to obtain causal intervention graphs.

2.1. GNN-based fault diagnosis

In GNN-based fault diagnosis, each sensor is considered a node, and the correlations between them are viewed as edges. Thus, the industrial process components and their relationships can be represented as a graph [29]. A simple graph can be defined as $G = G(V, E)$. where V and E are the sets of nodes and edges, respectively. Node $v_i \in V$, and $e_{ij} \in E$

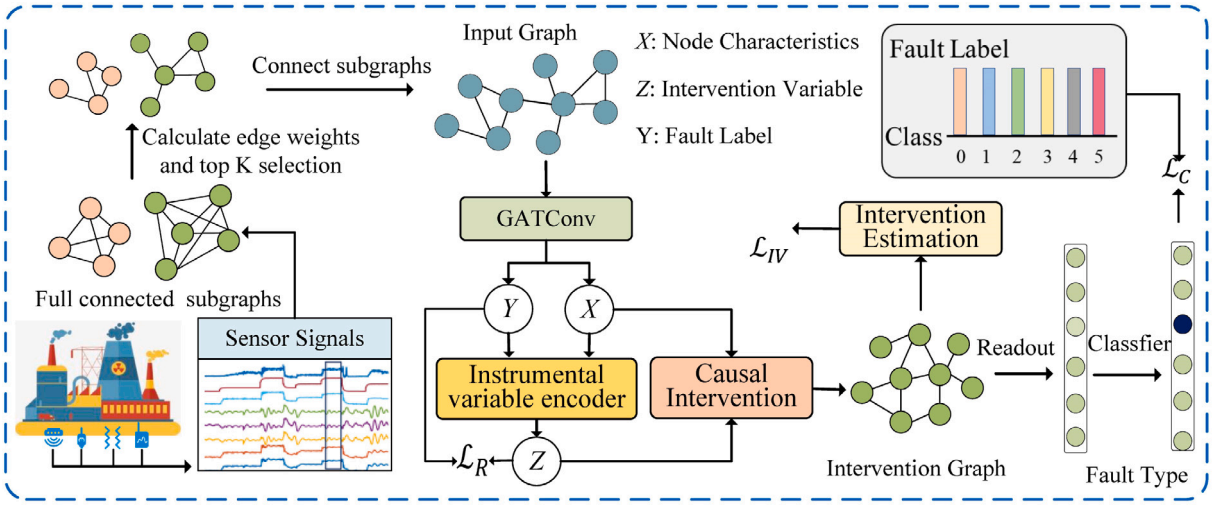

Fig. 1. The overview of the Causal Intervention Graph Neural Network framework. (1) The multi-sensor signals are sliced and each segment is used as a feature input to a node, using the attention mechanism to construct input graph. (2) The input graph is subjected to causal intervention learning, including instrumental variable construction and intervention implementation, to obtain a causal intervention graph. (3) Evaluating the causal intervention effect and utilizing the causal intervention graph for fault diagnosis.

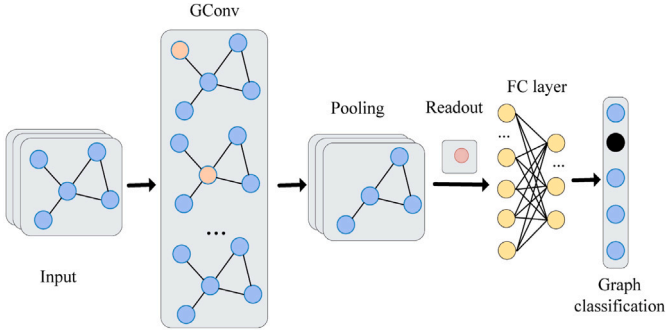

Fig. 2. The architecture for GNN-based fault diagnosis.

donates an edge between v_i and v_j. In general, to describe the topology of a graph using an adjacency matrix $\mathbf{A} \in \mathcal{R}^{N\times N}$, where $N = |V|$, indicating the number of nodes. $\mathbf{A}_{ij}$ stands for an edge link between nodes v_i and v_j. $\mathbf{X} \in \mathbb{R}^{n\times d_n}$ and $\mathbf{X}^e \in \mathbb{R}^{c\times d_c}$ denote node features and edge attributes, respectively. The GNN-based fault diagnosis is a graph classification task where given a set of graphs $\{G_1, G_2, \ldots, G_N\} \in \mathcal{G}$ to identify fault types $\{y_1, y_2, \ldots, y_N\} \in Y$.

A typical architecture for graph-level fault diagnosis is shown as Fig. 2. In this architecture, input graphs are constructed using three general methods, i.e., KNNGraph, RadiusGraph, and PathGraph. A GConv layer follows by a graph pooling layer to coarsen the input graphs into subgraphs, which reduces the dimensionality of the input graphs and speeds up the computation. After that, a readout layer collapses the node embedding of the sub-graphs into a graph representation by using the sum/max/mean operation. Finally, the learned graph representation is inputted to the FC layer for realizing graph-level fault diagnosis.

2.2. Structural Attribute Graph Construction

Inspired by the interpretability of the attention mechanism and insights from related work [30], we propose an automatic graph construction method to capture the interdependencies among sensors in complex industrial processes. If two segments of sensor signals exhibit significant correlation, they should be linked when constructing the structural attribute graph. The graph construction process is shown as Fig. 3.

Graph attention networks (GAT) [31] improves neighbor aggregation by automatically learning the attention coefficients of neighboring nodes. Considering a graph $G(V, E)$, the attention coefficient of $edge(i, j)$ is:

$$e_{ij} = \text{LeakyReLU}\left(\vec{\mathbf{a}}^{\mathrm{T}}\left[\mathbf{W}\vec{h}_i \parallel \mathbf{W}\vec{h}_j\right]\right), \quad \forall (i,j) \in E \tag{1}$$

where $\vec{h}_i$ and $\vec{h}_j$ are the features of nodes v_i and v_j, respectively. $\mathbf{W}$ is an attention vector used for node features embedding, while $\vec{\mathbf{a}}$ is a weight matrix used for calculating the correlation between two embeddings. To facilitate calculation and comparison, the attention coefficients are normalized using the *softmax* function:

$$\alpha_{ij} = \text{softmax}\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k\in N_i}\exp\left(e_{ik}\right)} \tag{2}$$

where $N_i = \{k \in N : (i,k) \in E\} \cup \{i\}$ is the self-containing neighboring set of node V_i. The new feature $\vec{h}'_i$ of node V_i can be aggregated by:

$$\vec{h}'_i = \sigma\Big(\sum_{j\in N_i} \alpha_{ij}\mathbf{W}\vec{h}_j\Big) \tag{3}$$

where $\sigma(\cdot)$ is ReLU activation function.

The structural attribute of graphs is related to fault modes, where strong influences should be represented as edges. Therefore, we sort all α_{ij} select the top $D \times N$ scores to construct the edge set E, where D is used to regulate the degree of each node. Simultaneously, the structural information S of the sensor association graph $G(V, E)$ can be derived intuitively.

According to the study by [30], reducing the receptive field of the attention mechanism results in a stronger sense of dependency. Following this principle, we initially use the k-nearest neighbors [32] algorithm to cluster all industrial process sensors into $K = \sqrt{N}$ groups. Subsequently, the aforementioned method is used to calculate edge attributes between nodes within each group. Finally, the same procedure is performed on the K center nodes to construct the complete structural attribute graph G.

2.3. Causal view in GNN-based fault diagnosis

We analyze the causal relationships between sensor signals and fault in complex industrial processes to analyze the identify the causality of GNN-based fault diagnosis. A structural causal model (SCM) [33] is constructed by examining causal relationships among seven variables:

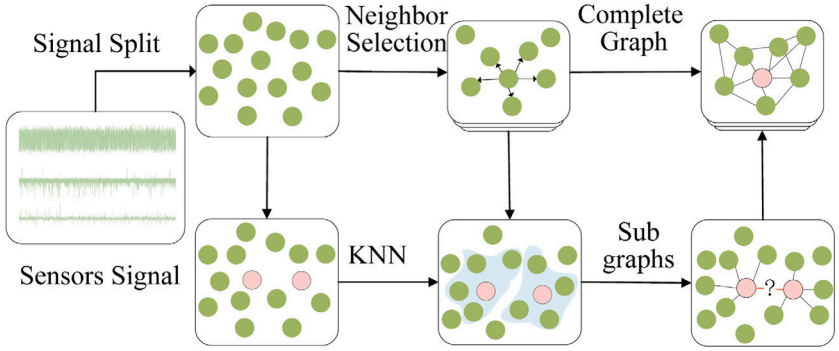

Fig. 3. Structural Attribute Graph Construction process.

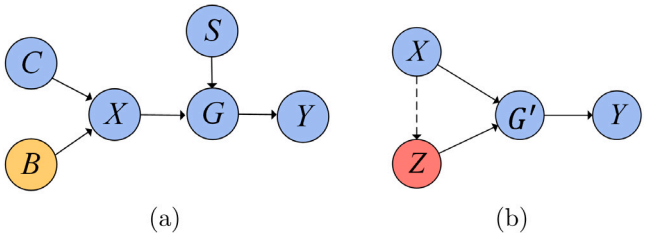

Fig. 4. (a) SCM of GNN-based fault diagnosis, (b) Causal intervention for SCM through instrumental variable.

causal variable $\boldsymbol{C}$, confounding variable $\boldsymbol{B}$, nodes features $\boldsymbol{X}$, instrumental variable $\boldsymbol{Z}$, structure information $\boldsymbol{S}$, graph data $\boldsymbol{G}$, and fault mode $\boldsymbol{Y}$. The SCM is illustrated in Fig. 4.

- $\boldsymbol{C}\rightarrow \boldsymbol{X}\leftarrow\boldsymbol{B}$. Node features $\boldsymbol{X}$ are generated from two highly coupled latent variables: the causal variable $\boldsymbol{C}$, which represents sensor signals determining the fault mode, and the confounding variable $\boldsymbol{B}$, which represents irrelevant sensor signals.
- $\boldsymbol{X}\rightarrow \boldsymbol{S}\leftarrow\boldsymbol{Z}$. $\boldsymbol{S}$ represents the structural information obtained from node features and instrumental variable. $\boldsymbol{S}$ is an adjacency matrix, describing the graph structure and reflects the interrelationships between sensors.
- $\boldsymbol{X}\dashrightarrow\boldsymbol{Z}$. The instrumental variable $\boldsymbol{Z}$ is randomly generated based on the characteristics spatial of $\boldsymbol{X}$.
- $\boldsymbol{X}\rightarrow \boldsymbol{G}'\leftarrow\boldsymbol{Z}$. Causal intervention is performed on the input graph to obtain a causal graph.
- $\boldsymbol{G}\rightarrow\boldsymbol{Y}$. GNN-based fault diagnosis is to predict the properties of the input graphs. The classifier will make prediction $\boldsymbol{Y}$ based on the graph $\boldsymbol{G}$.

Based on the SCM, it is obvious total effects between $\boldsymbol{X}$ and $\boldsymbol{Y}$ denoted as $P(Y|X)$ is different from causal effects of $\boldsymbol{X}\rightarrow\boldsymbol{Y}$, donated as $P(Y|do(X=x))$. Causal effects only involve the direct path from $\boldsymbol{X}$ to $\boldsymbol{Y}$, while total effects involve all paths connecting $\boldsymbol{X}$ and $\boldsymbol{Y}$. Thus, irrelevant sensor signals act as error terms to obscure causality.

2.4. Instrument variable for causal intervention

To bridge the gap between total effects and causal effects, we need to adjust potential confounding variable $\boldsymbol{B}$. Causal intervention is a viable approach [34]. There are four causal interventions: front-door adjustment, back-door adjustment, randomized controlled trial, and instrumental variable estimation. Additional observable variable is needed for both front-door and back-door adjustments, while in complex industrial processes where fault and noise signals are highly coupled, confounding variables cannot be directly observed. Randomized controlled trial are dynamic and uncertain. Instead of directly observing confounding factors, we use instrumental variable $\boldsymbol{Z}$ to eliminate spurious correlations.

Instrumental variable $\boldsymbol{Z}$ must fulfill two requirements: (1) The influence of error terms on $\boldsymbol{Y}$ is independent of $\boldsymbol{X}$, and $\boldsymbol{Z}$ is independent of error terms. (2) $\boldsymbol{Z}$ is strongly correlated with $\boldsymbol{X}$. In other words, all correlations between $\boldsymbol{Z}$ and $\boldsymbol{Y}$ require $\boldsymbol{X}$ as a mediator.

Based on the above analysis, we implement a two stages approach to construct instrumental variable [35]: In the first stage, a coefficient α is obtained by regression estimation of $\boldsymbol{X}$ and $\boldsymbol{Z}$, denoted as $Cov(X,Z)$. In the first stage, the expression for $\boldsymbol{Z}$ is replaced with an expression including $\boldsymbol{Y}$, and then regress $\boldsymbol{Y}$ on $\boldsymbol{Z}$, denoted as $Cov(Y,Z)$. Due to the restriction in the definition of $\boldsymbol{Z}$, there is confounding bias between $\boldsymbol{Y}$ and $\boldsymbol{Z}$.

$$X=\alpha Z+\varepsilon_X; Y=\omega X+\varepsilon_Y \tag{4}$$

where ε_X and ε_Y are error terms including confounding variable $\boldsymbol{B}$. Following the adjustments, the effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ is asymptotic unbiased.

$$\omega=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(z_i-\bar{z}\right)\left(y_i-\bar{y}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(z_i-\bar{z}\right)\left(x_i-\bar{x}\right)}=\frac{\mathrm{Cov}(Z,Y)}{\mathrm{Cov}(Z,X)} \tag{5}$$

2.5. Causal intervention learning

2.5.1. Instrumental variable generation

The limitations of instrumental variable can be summarized in two points: (1) the instrumental variable $\boldsymbol{Z}$ does not affect the prediction through any other path than $\boldsymbol{X}$, (2) the instrumental variable $\boldsymbol{Z}$ affects the characteristics $\boldsymbol{X}$. Based on these two basic points, and inspired by the work on data augmentation [36], we choose random perturbations as instrumental variable. Specifically, the instrumental variable $\boldsymbol{Z}$ is randomly generated from the characteristics of nodes $\boldsymbol{X}$. It satisfies well the requirements of instrumental variables: (1) Stochastic perturbations clearly have no independent effect on the prediction, (2) Stochastic perturbation alter the node characteristics and structural information of the graph, thereby affecting the prediction results.

The first step of instrumental variable generation is to establish a causal relationship between $\boldsymbol{Z}$ and $\boldsymbol{X}$, denoted as $\boldsymbol{Z}\rightarrow\boldsymbol{X}$. To enhance the comprehensiveness of the information, we align the one-hot label $\boldsymbol{Y}$ with $\boldsymbol{X}$ by a graph attention layer and subsequently concatenate them to get new the $\boldsymbol{X}$. The instrumental variable $\boldsymbol{Z}$ is then generated from $\boldsymbol{X}$ and $\boldsymbol{Y}$ by an encoder, which contains a multi-layer perceptions (MLP) and a pair of graph attention layer. Specifically, we firstly utilize MLP to derive the mean μ and standard deviation σ of the latent space for the intervention variables.

$$\begin{aligned} X_{att}&=GAT(X,Y)\\ \mu,\sigma&=\mathrm{encoder}(X_{att},Y)\end{aligned} \tag{6}$$

Instrumental variable $\boldsymbol{Z}$ is a random matrix that obeys a normal distribution with mean μ and standard deviation σ.

2.5.2. Causal intervention

X and Z are each spliced after one layer of the graph attention network to obtain the reconstruction X_{recon}.

$$\begin{aligned} Z_{att} &= GAT(Z, Y) \\ X_{recon} &= GAT(X_{att}, Z_{att}) \\ X_{recon} &= \sigma(f(X_{recon})) \end{aligned} \tag{7}$$

where $\sigma(\cdot)$ is sigmoid activation function, and $f(\cdot)$ is a liner layer. We select binary cross-entropy (BCE) function to compute the reconstruction loss:

$$\mathcal{L}_R = f_{BCE}(X, X_{recon}) \tag{8}$$

The GNN-based fault diagnosis is a classification task, as detailed in Section 1. Given a set of graphs $\{G_1, G_2, \ldots, G_N\} \in \mathcal{G}$ with M fault types, we use negative log-likelihood (NLL) function to compute the classification loss.

$$\mathcal{L}_C = -\sum_{i=1}^{N}\sum_{j=1}^{M} y_j \ \log \ q_j(w|G_i) \tag{9}$$

w are the hyper-parameters of the model, y_j is the ground-truth label, and q_j represents the probability of belonging to the fault type j.

It is necessary to note that the model cannot completely ignore the raw data, which is incompatible with standard fault diagnosis process. We set the hyperparameters λ_1 and λ_2 to achieve equilibrium and combine the reconstruction loss $\mathcal{L}_R$, the causal intervention loss $\mathcal{L}_{IV}$, and the fault diagnosis loss $\mathcal{L}_C$. The total loss of CIGNN is denoted as $\mathcal{L}_{TOTAL}$, and the detailed implementation of CIGNN in Algorithm 1.

$$\mathcal{L}_{TOTAL} = \mathcal{L}_C + \lambda_1\mathcal{L}_{IV} + \lambda_2\mathcal{L}_R \tag{10}$$

Algorithm 1 CIGNN

Input: The multiple sensor signals $\mathbf{X} \in \mathbb{R}^{n\times d_n}$, hyperparameters λ_1 and λ_2
Output: The parameters of the trained model.
1: Initialize the node features, $\mathbf{H}^{(0)} \leftarrow \mathbf{X}$;
2: Cluster all nodes V into K groups, $K = \sqrt{N}, N = |V|$;
3: Compute attention coefficient e_{ij} of node v_i and $v_j \leftarrow$ Eq. 3;
4: Compute attention coefficient e_{k_1,k_2} between central nodes;
5: Obtain input graph G with node feature x_i and adjacency matrix $A_i = \{e_{11}, e_{12}, ..., e_{nn}\}$
6: **for** N input graphs $\{G_i = (A_i, X_i)\}_{i=1}^{N}$ **do**
7: compute mean μ and standard deviation $\sigma \leftarrow$ Eq.6
8: instrumental variable $\boldsymbol{Z} = \mu + \sigma \times \mathcal{N}(0, 1)$
9: reconstruction of node feature $X_{recon} \leftarrow$ Eq.7
10: compute reconstruction loss $\mathcal{L}_R \leftarrow$ Eq.8
11: compute proportionality coefficient $\alpha \leftarrow$ Eq.9
12: compute instrumental variable estimation loss $\mathcal{L}_{IV} \leftarrow$ Eq.10
13: compute fault diagnosis loss $\mathcal{L}_C \leftarrow$ Eq.11
14: total loss $\mathcal{L}_{TOTAL} = \mathcal{L}_C + \lambda_1\mathcal{L}_{IV} + \lambda_2\mathcal{L}_R$
15: **end for**

For a specific stochastic perturbation z_i, we have $z_i = f(x, z_i) \approx \alpha_i \cdot x$. α is a self-learning parameter, which is used to represent the relationship between x and z. Substituting the relation above between original feature x and augmentation feature x_{z_i} into Eq. (1), we will get the $y|_{x=z_i}$ with different proportionality coefficient α.

$$\begin{aligned} y|_{x=z_i} &= W_{xy} \cdot z_i + W_{cy} \cdot c + \hat{\varepsilon_y} \\ &= \alpha_i \cdot W_{xy} \cdot x + W_{cy} \cdot c + \hat{\varepsilon_y} \\ &= \alpha_i \cdot y|_{do(x=x)} + W_{cy} \cdot c + \hat{\varepsilon_y} \end{aligned} \tag{11}$$

where $y|_{do(x=x)}$ denotes causal path $\boldsymbol{X} \rightarrow \boldsymbol{Y}$. Substituting the $y|_{x=z_i}$ with Eq. (9) obtained by learning, we obtain the learning object the instrumental variable estimation. The benefit is obviously that we can suppress the confounding effect without directly observing the confounding variable $\boldsymbol{B}$. The goals of causal intervention learning can be summarized as:

$$\mathcal{L}_{IV} = \min \sum_{i\neq j} \left\| \left(\alpha_i - \alpha_j\right) \cdot \left(W_{by} \cdot b + \hat{\varepsilon_y}\right) \right\| \tag{12}$$

3. Case study

In this section, two case studies are conducted to verify the validity of the proposed CIGNN model. The first case study is the fault diagnosis of a three-phase flow device. The second case study is the fault diagnosis of a gas turbine in a nuclear power system. Experiments were conducted on NVIDIA GeForce RTX 3090 GPU using Pytorch-1.11 and DGL-1.0.2 frameworks.

3.1. Data description

3.1.1. Three-phase flow facility data

The three-phase flow facility (TFF) [37] is a complex industrial simulation system designed by Cranfield University to provide measurable flow rates of water, oil and air flows to pressurized systems. Fig. 5(a) illustrates the simplified sketch of TFF facility. The main components comprise several pipes of varying orifice sizes and geometries and a gas–liquid two-phase separator. The facility is capable of transporting single-phase air, water and oil, as well as mixtures of these fluids, at specific speeds. The system has a total of 24 sensors as input variables to the model system for measuring pressure, flow, density and temperature at different locations in the system. Data is captured at a sampling rate of 1 Hz for all three-phase flow facilities. The system has four air flows and five water flows, which can be combined in 20 different ways. The TFF facility simulates three datasets under normal operating conditions and six typical faults that can occur during actual operation using 20 process inputs. Note that the first 23 variables are used for all the fault study, whereas variable 24 is only used in fault 6. After running for a certain period of time, the simulated system introduces faults, which are automatically resolved once they reach a certain level, resulting in the system returning to its normal state. As a result, normal operation data and faulty data alternate. To better extract the fault data, we cut the data into 50-second segments as samples. The data is available at TFF data.

3.1.2. Nuclear power system data

The nuclear power system (NPS) [38] dataset consists of encrypted operational data collected by large sensors in the nuclear power system. The system is a power plant energy generation module consisting of steam generators, turbines, generators, condensers, pumps, valves and associated equipment, as shown in Fig. 5(b). The NPS which has 121 sensors collecting system monitoring data, recording pressures, temperatures, and associated valve opening and closing variables for the major components. For each simulation, the system starts in a normal state, operates for a period of time, and then introduces a fault at some point during operation, so that the fault data alternates with the normal data. The time series multivariate signals are sampled at 4 Hz, and we split the NPS data into segments consists of 60 signals. In actual monitoring, multiple sensors are installed at the same location, resulting in data redundancy. To remove redundant and irrelevant measurements and to reduce computational costs, we selected the first 64 sensor variables, as well as 22 typical fault types and normal conditions.

The TFF and NPS datasets were preprocessed by max–min normalization and randomly select 80% of the samples as the training set and the rest as the test set, and the specific fault types and samples are shown in Table 1.

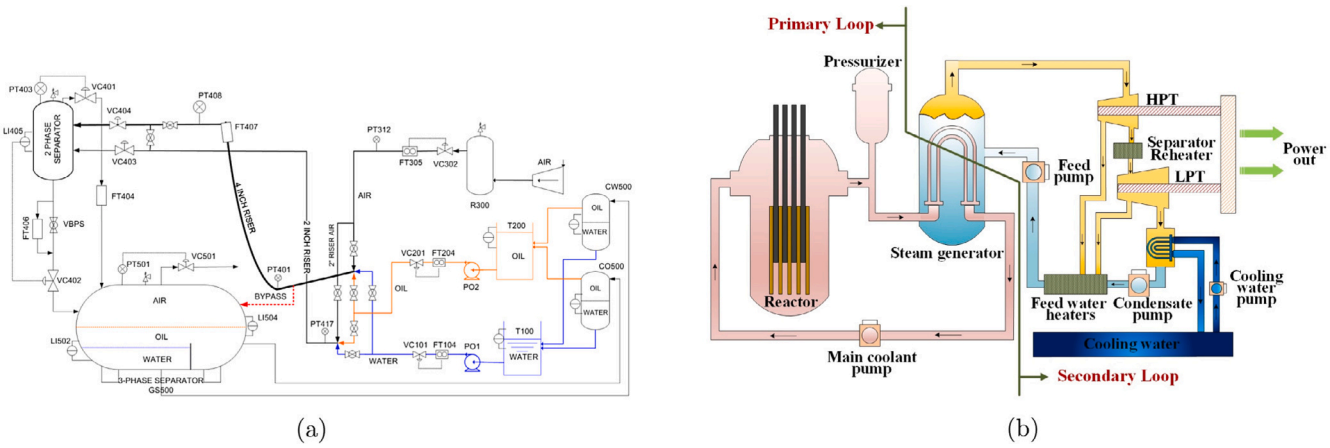

Fig. 5. Simplified sketch (a) three-phase flow facility, (b) nuclear power system.

Table 1
Fault types in datasets.

Dataset	Fault type	Samples of training set	Samples of testing set
TFF	Fault case(1-6)	1386	315
	Normal	534	133
NPS	Fault case(1-21)	3557	890
	Normal	1563	390

3.2. Experimental setup

3.2.1. Baseline methods

To validate the effectiveness and superiority of CIGNN, we compare CIGNN with four different fault diagnosis methods, including machine learning methods, deep learning-based methods, and GNN-based methods.

(1) KNN: KNN [32] is a common classification machine learning algorithm that can be used for fault diagnosis.

(2) GAT: GAT [31] utilizes the attention mechanism to capture the importance of neighboring nodes, focusing on relevant information during feature extraction and learning complex relationships in graph data.

(3) IAGNN: IAGNN [12] creates feature subgraphs from uncovering correlations between sensor signals through interactive sensing. The final graph embedding is created by fusing each subgraph feature using a weighted summation function.

(4) PKT-MCNN: PKT-MCNN [38] develops a coarse-to-fine progressive knowledge transfer structure. Specifically, the MCNN module learns coarse/fine-grained tasks and extracts generic fault information. PKT migrates the coarse-grained knowledge to the fine-grained tasks.

3.2.2. Implementation details and evaluation metrics

Extensive experiments are conducted on the CIGNN model to select the hyperparameters that produce the best results. The learning rate is selected from {0.0001, 0.0005, 0.001}, and the batch size is selected from {32, 64}. When applying the CIGNN model to the TFF dataset, input graph consists of 24 nodes corresponding to 24 sensors. For the NPS dataset, the number of nodes is 64. Metrics such as accuracy, receiver operating characteristic (ROC) curve, and confusion matrix are used to evaluate the performance of different fault diagnosis methods.

Table 2
Accuracy of different fault diagnosis methods.

Dataset \ Model	KNN	GAT	IAGNN	PKT-MCNN	CIGNN
TFF	0.8197	0.8768	0.9229	0.8858	**0.9634**
NPS	0.7547	0.7886	0.8734	0.8650	**0.8929**

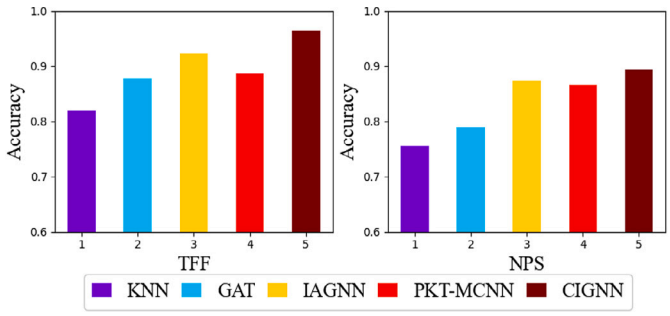

Fig. 6. Accuracy of different fault diagnosis methods.

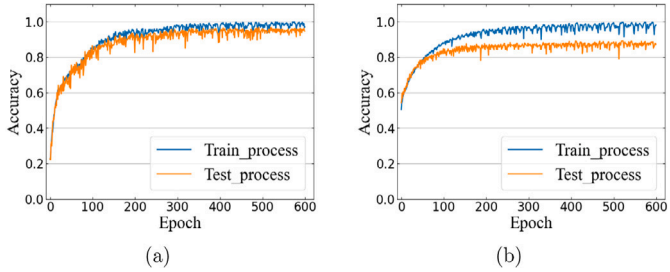

Fig. 7. Convergence of the CIGNN on (a) TFF, (b) NPS dataset.

Industrial equipment usually stops working before a fault occurs, so the fault data is significantly less than the normal data. Following this principle, a large number of experiments have been carried out and the results of different fault diagnosis methods are presented in Table 2 and Fig. 6 provides a visualization of the results. Fig. 7 depicts the training and testing accuracy trajectories of the CIGNN model on the TFF and NPS datasets.

3.3. Fault diagnosis performance

3.3.1. TFF experiment results analysis

Firstly, the CIGNN model exhibits superior performance on the TFF dataset compared to the baseline methods. Experiments demonstrate

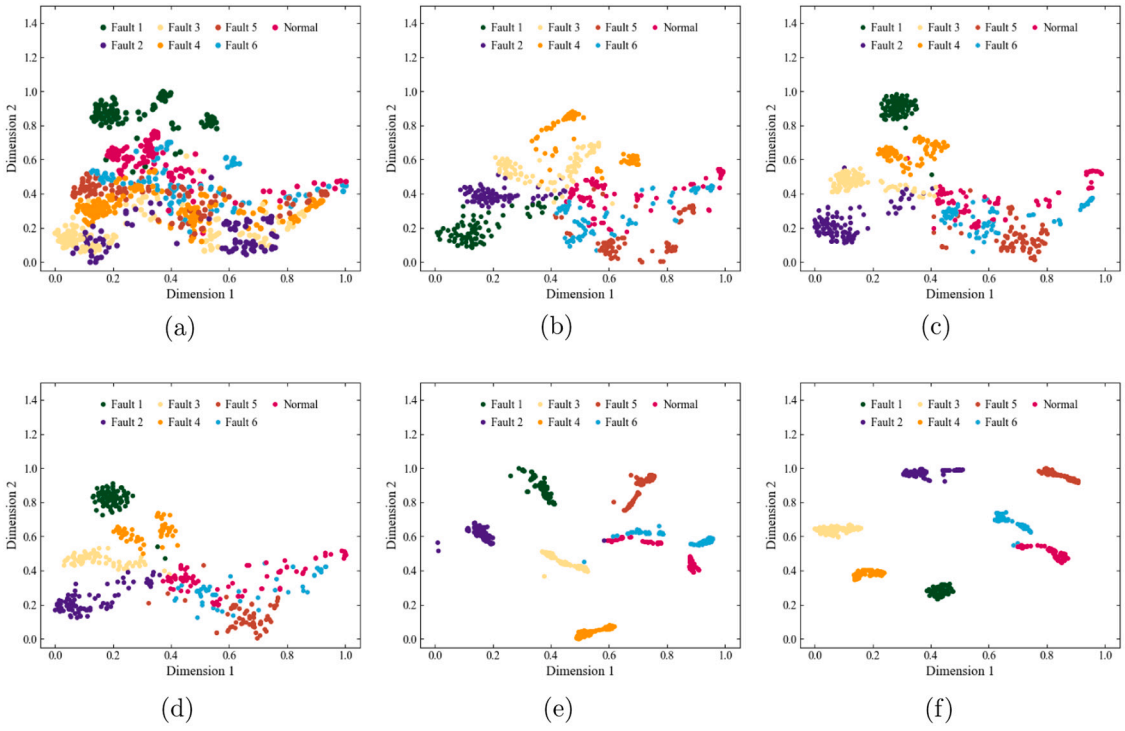

Fig. 8. TFF features visualization via t-SNE of :(a) Raw data, (b) KNN, (c) GAT, (d) PKT-MCNN, (e) IAGNN, (f) CIGNN.

that the multivariate time series embeddings learned by CIGNN effectively reveal the fault characteristics of complex industrial processes. Secondly, it can be observed that GAT outperforms KNN because GAT is better at fusing information from multiple sensor signals, while KNN can only cluster based on data. PKT-MCNN decomposes numerous fault diagnosis problems in complex industrial processes into multiple sub-problems with fewer faults by constructing fault trees, and thus the diagnostic performance is better than that of KNN and GAT. In practice, PKT-MCNN lacks versatility due to the small size of the target diagnostic task and the imprecise granularity structure. IAGNN utilizes an interaction-aware module that topologically takes into account the differences between fault types, and thus the IAGNN module has the best diagnosis performance among the baseline methods. However, it does not take into account that irrelevant sensor signals will interfere with prediction, and thus IAGNN is not as effective as CIGNN.

We analyze the causality between sensor signals and fault based on causal theory. Fault signals play a solely deterministic role in prediction, referred to as causal features. Irrelevant sensor signals are referred to as confounding features. Moreover, causal features and confounding features are highly coupled in sensor signals, making it impossible to explicitly decouple them. To mitigate the confounding effects caused by confounding features, CIGNN designs an instrumental variable to implement causal intervention on graph.

To visually analyze the causal intervention effect of CIGNN, we extract the raw data features of the TFF dataset and the features of last layer of different models training and visualized them using t-distributed stochastic neighborhood embedding (t-SNE) [39]. The visualization results show that the sample features of TFF faults are separated according to different machine states and operating conditions, as shown in Fig. 8(a). The features extracted by CIGNN show better clustering performance due to the fact that CIGNN optimizes the feature and structural information of the graph through intervening variable, which addresses the bias caused by confounding features. CIGNN is able to accurately identify fault types even though many fault types have similar spatial distributions.

Table 3
Computational time comparison (seconds)

Dataset \ Model	KNN	GAT	IAGNN	PKT-MCNN	CIGNN
TFF	6.536	1.893	1.901	1.183	**1.017**
NPS	64.332	11.789	15.437	**3.115**	6.326

3.3.2. NPS experiment results analysis

CIGNN also achieves optimal results on large-scale fault diagnosis tasks for complex industrial processes. We use the confusion matrices to describe the ability of the CIGNN to handle unbalanced data, as shown in Fig. 9. CIGNN demonstrates superior accuracy in recognizing all 22 fault types, surpassing the competing methods, when handling large-scale imbalanced datasets such as NPS. The CIGNN can effectively perform causal intervention and mitigate confounding effects even in the face of large-scale dataset. To evaluate the sensitivity of CIGNN, we compare the ROC curve of different fault diagnosis methods on TFF and NPS dataset, as shown in Fig. 10. It can be seen that the ROC curve of CIGNN achieve optimal results, which are significantly higher than the baseline methods. In addition, the ROC curve of CIGNN is smooth on the NPS, whereas baseline methods show significant tremble, which further proves the superiority of CIGNN.

In terms of computational time, measuring the testing time yielded the results shown in Table 3. CIGNN achieves the shortest computation time on the TFF, indicating that it can produce classification results more quickly. PKT-MCNN records the shortest computation time on the NPS. Due to the NPS containing a large variety of fault types, PKT-MCNN is designed with a hierarchical fault tree from coarse to fine, segmenting fault diagnosis into multiple subtasks. It only needs to compare with similar faults according to the fault tree, significantly improving computational efficiency. However, PKT-MCNN cannot guarantee accuracy with each comparison, as its accuracy is not high and it is unable to construct a fault tree on the TFF with few fault types. Compared with PKT-MCNN, CIGNN sacrifices a small amount of computational time to significantly enhance the accuracy. Moreover, the computation time of CIGNN is shorter than other baseline methods.

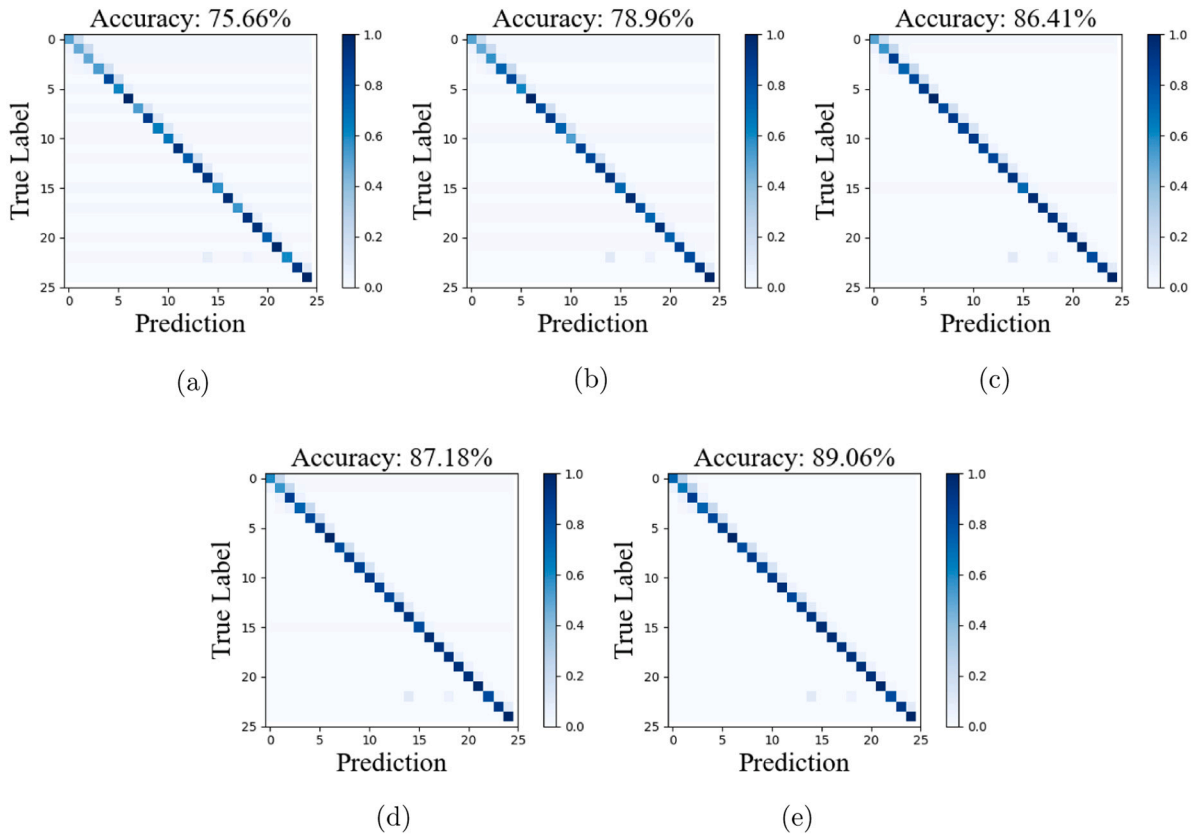

Fig. 9. Confusion matrices of the NPS. (a) KNN, (b) GAT, (c) PKT-MCNN, (d) IAGNN, and (e) CIGNN.

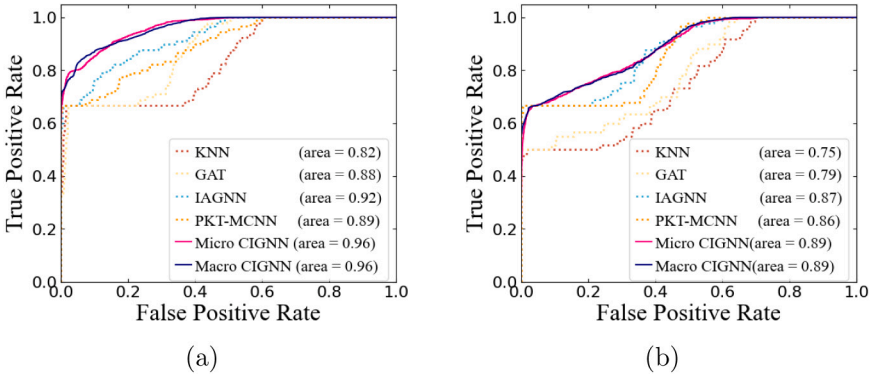

Fig. 10. ROC curve comparison. (a) TFF, (b) NPS.

3.4. Analysis on causality validity

The GAT, IAGNN, and CIGNN models are all GNN-based methods. GAT introduces an attention mechanism that allows each node to dynamically adjust weights based on its relationship with neighboring nodes, allocating more attention to important neighboring nodes. However, this attention mechanism only analyzes at the feature level and cannot reflect the true interaction between components in complex industrial processes. In contrast, IAGNN utilizes an interaction-aware module to analyze the interaction between components from a topological perspective and explores the differences between different fault types, thus achieving higher fault diagnosis accuracy than other baseline methods. However, IAGNN does not consider the influence of irrelevant sensor signals. To address this issue, CIGNN identifies causal features that play a decisive role in fault diagnosis based on causal theory and designs an instrumental variable to implement causal intervention on input graph to mitigate the interference effects. Obviously, the CIGNN achieves best results, which indicates that the causal analysis is correct and the causal intervention is effective from the GNN perspective.

Table 4
Ablation study results.

Attention	Intervention	TFF	NPS
—	—	0.8751	0.7886
✓	—	0.8923	0.8013
—	✓	0.9437	0.8628
✓	✓	**0.9634**	**0.8928**

Ablation studies on: (1) attention mechanism for graph construction, (2) causal intervention learning.

To evaluate the effectiveness and contribution of each module in the CDGNN model, we conducted ablation studies. This involved creating variations of the CIGNN model, with the specific configurations and results detailed in Table 4. It is evident that each module in CIGNN

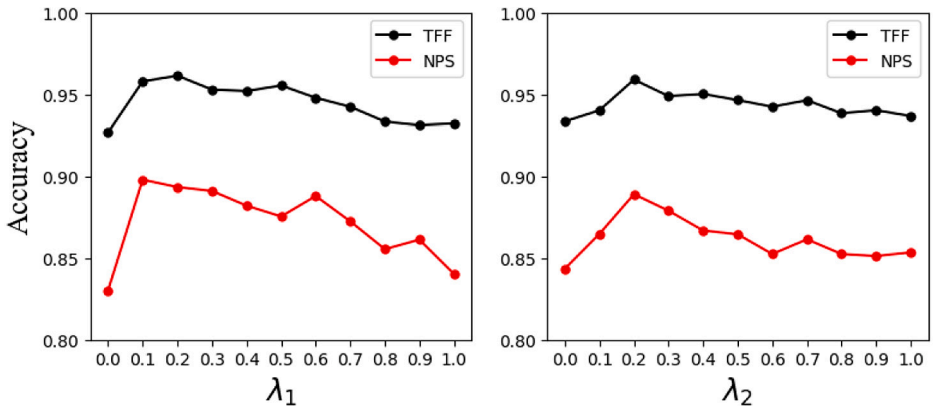

Fig. 11. Hyper-parameter analysis on CIGNN.

plays a crucial role. Constructing a graph using attention mechanisms is beneficial for learning the physical topology relationships among components in complex industrial processes. The causal intervention learning module has the greatest impact on classification performance, highlighting its core role within CIGNN. According to the ablation results, when these two modules are removed, CIGNN becomes a standard GNN model, and its classification performance approaches that of GAT. This indicate that the causal analysis is correct and the causal intervention is effective from the GNN perspective

3.5. Hyper-parameter analysis

We analyze the sensitivity of λ_1 and λ_2 and plotted the classification performance in Fig. 11. For λ_1 and λ_2, there is a specific range that maximizes the test performance across all datasets. CIGNN performs best when λ_1 is 0.2 and λ_2 is 0.1. Notably, we observe that the NPS dataset shows higher sensitivity to change in λ_1 and λ_2, which may be due to the fact that the NPS dataset is more complex and unbalanced.

4. Conclusion

In this study, we analyze the causality between sensor signals and fault based on causal theory, and then propose a causal intervention graph neural network (CIGNN) model. This model considers fault diagnosis task from a causal theory perspective and transforms it into graph classification tasks. Initially, we construct the sensor model into a structural attribute graph through the attention mechanism, preliminary determining the topological relationships among multiple sensors. To mitigate the confounding effect caused by irrelevant sensor signals, we designed an instrumental variable to implement causal interventions on input graphs. The prediction derived from the causal intervention graphs is closer to the actual results. By reducing the confounding effect, CIGNN can improve the robustness and interpretability of intelligent fault diagnosis. Fault diagnosis is usually an open-set recognition task due to the volatility of mechanical equipment and operating conditions. Future research directions include applying CIGNN to open-set recognition, which requires not only accurate diagnosis of known faults using causal features, but also effective identification of unknown faults to prevent new faults from hiding and affecting industrial production.

CRediT authorship contribution statement

Ruonan Liu: Writing – review & editing, Project administration, Funding acquisition. **Quanhu Zhang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Data curation. **Di Lin:** Writing – review & editing, Formal analysis. **Weidong Zhang:** Writing – review & editing. **Steven X. Ding:** Writing – review & editing, Project administration.

Declaration of competing interest

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

Acknowledgments

This research was partly supported by the National Key R&D Program of China under Grant No. 2022ZD0119900, the National Natural Science Foundation of China under Grant Nos. 62206199 and U2141234, Shanghai Science and Technology Program, China under Grant No. 22015810300, Tianjin Applied Basic Research Project, China under Grant No. 22JCQNJC00410, Young Elite Scientist Sponsorship Program, China under Grant No. YESS20220409, Alexander von Humboldt Foundation, Germany Grant No. 1226831 and State Key Laboratory of Reliability and Intelligence of Electrical Equipment, China No. EERI-KF2022001.

References

[1] Hu Y, Miao X, Si Y, Pan E, Zio E. Prognostics and health management: A review from the perspectives of design, development and decision. Reliab Eng Syst Saf 2022;217:108063.
[2] Han T, Tian J, Chung C, Wei Y-M. Challenges and opportunities for battery health estimation: Bridging laboratory research and real-world applications. J Energy Chem 2024;89:434–6.
[3] Xie W, Han T, Pei Z, Xie M. A unified out-of-distribution detection framework for trustworthy prognostics and health management in renewable energy systems. Eng Appl Artif Intell 2023;125:106707.
[4] Lei Y, Yang B, Jiang X, Jia F, Li N, Nandi AK. Applications of machine learning to machine fault diagnosis: A review and roadmap. Mech Syst Signal Process 2020;138:106587.
[5] Shin JH, Bae J, Kim JM, Lee SJ. An interpretable convolutional neural network for nuclear power plant abnormal events. Appl Soft Comput 2023;132:109792.
[6] Meng H, Geng M, Han T. Long short-term memory network with Bayesian optimization for health prognostics of lithium-ion batteries based on partial incremental capacity analysis. Reliab Eng Syst Saf 2023;236:109288.
[7] Jeong Y. Fault detection with confidence level evaluation for perception module of autonomous vehicles based on long short term memory and Gaussian mixture model. Appl Soft Comput 2023;149:111010.
[8] Deng C, Deng Z, Miao J. Semi-supervised ensemble fault diagnosis method based on adversarial decoupled auto-encoder with extremely limited labels. Reliab Eng Syst Saf 2024;242:109740.
[9] Miao M, Yu J. Deep feature interactive network for machinery fault diagnosis using multi-source heterogeneous data. Reliab Eng Syst Saf 2024;242:109795.
[10] Tian J, Jiang Y, Zhang J, Luo H, Yin S. A novel data augmentation approach to fault diagnosis with class-imbalance problem. Reliab Eng Syst Saf 2024;243:109832.
[11] Su Y, Shi L, Zhou K, Bai G, Wang Z. Knowledge-informed deep networks for robust fault diagnosis of rolling bearings. Reliab Eng Syst Saf 2024;244:109863.
[12] Chen D, Liu R, Hu Q, Ding SX. Interaction-aware graph neural networks for fault diagnosis of complex industrial processes. IEEE Trans Neural Netw Learn Syst 2023;34(9):6015–28.

[13] Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 2021;32(1):4–24.
[14] Ju W, Fang Z, Gu Y, Liu Z, Long Q, Qiao Z, et al. A comprehensive survey on deep graph representation learning. Neural Netw 2024;173:106207.
[15] Xia L, Liang Y, Leng J, Zheng P. Maintenance planning recommendation of complex industrial equipment based on knowledge graph and graph neural network. Reliab Eng Syst Saf 2023;232:109068.
[16] Li T, Zhou Z, Li S, Sun C, Yan R, Chen X. The emerging graph neural networks for intelligent fault diagnostics and prognostics: A guideline and a benchmark study. Mech Syst Signal Process 2022;168:108653.
[17] Zheng S, Wang C, Zio E, Liu J. Fault detection in complex mechatronic systems by a hierarchical graph convolution attention network based on causal paths. Reliab Eng Syst Saf 2024;243:109872.
[18] Luo D, Cheng W, Xu D, Yu W, Zong B, Chen H, et al. Parameterized explainer for graph neural network. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H, editors. In: Advances in neural information processing systems, vol. 33, Curran Associates, Inc; 2020, p. 19620–31.
[19] Yin S, Ding SX, Xie X, Luo H. A review on basic data-driven approaches for industrial process monitoring. IEEE Trans Ind Electron 2014;61(11):6418–28.
[20] Anjaiah K, Pattnaik SR, Dash P, Bisoi R. A real-time DC faults diagnosis in a DC ring microgrid by using derivative current based optimal weighted broad learning system. Appl Soft Comput 2023;142:110334.
[21] Li X, Wang Y, Yao J, Li M, Gao Z. Multi-sensor fusion fault diagnosis method of wind turbine bearing based on adaptive convergent viewable neural networks. Reliab Eng Syst Saf 2024;245:109980.
[22] Fan S, Wang X, Mo Y, Shi C, Tang J. Debiasing graph neural networks via learning disentangled causal substructure. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. In: Advances in neural information processing systems, vol. 35, Curran Associates, Inc; 2022, p. 24934–46.
[23] Sui Y, Wang X, Wu J, Lin M, He X, Chua T-S. Causal attention for interpretable and generalizable graph classification. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery; 2022, p. 1696–705.
[24] Wang S, Zhou J, Sun C, Ye J, Gui T, Zhang Q, et al. Causal intervention improves implicit sentiment analysis. In: International conference on computational linguistics. 2022, p. 6966–77.
[25] Fan S, Wang X, Shi C, Cui P, Wang B. Generalizing graph neural networks on out-of-distribution graphs. IEEE Trans Pattern Anal Mach Intell 2024;46(1):322–37.
[26] Li J, Wang Y, Zi Y, Zhang H, Li C. Causal consistency network: A collaborative multimachine generalization method for bearing fault diagnosis. IEEE Trans Ind Inf 2023;19(4):5915–24.
[27] Wang H, Liu R, Ding SX, Hu Q, Li Z, Zhou H. Causal-trivial attention graph neural network for fault diagnosis of complex industrial processes. IEEE Trans Ind Inf 2024;20(2):1987–96.
[28] Zhang G, Kong X, Wang Q, Du J, Wang J, Ma H. Single domain generalization method based on anti-causal learning for rotating machinery fault diagnosis. Reliab Eng Syst Saf 2024;250:110252.
[29] Chen Z, Xu J, Alippi C, Ding SX, Shardt Y, Peng T. Graph neural network-based fault diagnosis: A review. 2021, arXiv e-prints.
[30] Yu JJQ. Graph construction for traffic prediction: A data-driven approach. IEEE Trans Intell Transp Syst 2022;23(9):15015–27.
[31] Velickovic P, Cucurull G, Casanova. Graph attention networks. In: 6th international conference on learning representations. 2017, p. 1–15.
[32] Elshenawy LM, Chakour C, Mahmoud TA. Fault detection and diagnosis strategy based on k-nearest neighbors and fuzzy C-means clustering algorithm for industrial processes. J Franklin Inst 2022;359(13):7115–39.
[33] Pearl J, Glymour M, Jewell NP. Causal inference in statistics: A primer. John Wiley & Sons; 2016.
[34] Pearl J. Causality. Cambridge University Press; 2009.
[35] Angrist JD, Pischk J-S. Mostly harmless econometrics. In: Mostly harmless econometrics. 2008.
[36] Guohang L, Shibin Z, Haozhe T, Lu Y, Lu J, Yuanyuan H. Easy data augmentation method for classification tasks. In: 2020 17th international computer conference on wavelet active media technology and information processing. 2020, p. 166–9.
[37] Ruiz-Cárcel C, Cao Y, Mba D, Lao L, Samuel R. Statistical process monitoring of a multiphase flow facility. Control Eng Pract 2015;42:74–88.
[38] Wang Y, Liu R, Lin D, Chen D, Li P, Hu Q, et al. Coarse-to-fine: Progressive knowledge transfer-based multitask convolutional neural network for intelligent large-scale fault diagnosis. IEEE Trans Neural Netw Learn Syst 2023;34(2):761–74.
[39] van der Maaten L, Hinton GE. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.