Reliability Engineering and System Safety 252 (2024) 110468

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Information-based Gradient enhanced Causal Learning Graph Neural Network for fault diagnosis of complex industrial processes

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

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

ARTICLE INFO

Keywords:
Complex industrial processes
Fault diagnosis
Causal intervention
Gradient reactivation
Graph neural networks (GNN)

ABSTRACT

By representing the embedded components and their interactions in industrial systems as nodes and edges in a graph, Graph Neural Networks (GNNs) have achieved outstanding results due to their ability to model statistical correlations. However, these correlations may not capture the true causal relationships within the data, thereby impairing the model's performance in fault diagnosis.

To address this issue, an Information-based Gradient enhanced Causal Learning Graph Neural Network (IGCL-GNN) is proposed for fault diagnosis of complex industrial processes. First, the information theory in graph representations is theoretically analyzed and the optimization objectives are derived separately for the causal and non-causal parts of the graph neural network, which decouple it into a multi-objective optimization problem. Then, to optimize such problem, a causal disentanglement layer is designed in the graph network that could effectively separate causal and non-causal information in graph representations. Thirdly, a novel gradient reactivation method is proposed to dynamically filter features from the disentangled layers, thereby capture the causal representations of graph data more accurately. For robust and efficient optimization, the multi-objective gradient descent algorithm is employed in this paper. Finally, comparative experiments were conducted on the three-phase flow facility (TFF) dataset, achieving a fault diagnosis accuracy of 98.42% for our proposed method.

1. Introduction

With the rapid advancement of technology in the industrial field, the costs and complexity of operating industrial systems and associated equipment in factories have accelerated rapidly. Consequently, there is an urgent need for an effective and reliable diagnostic system to monitor the operation of industrial systems, replacing manual inspection techniques, diminishing the expense of maintenance services, and ensuring the safe operation of industrial systems in an intelligent and efficient manner [1,2]. Characterized by vast scale and high complexity, modern industrial systems require the diagnosis of systems that encompass a variety of measurement devices, with readings characterized by high dimensionality and complex interactions [3–5]. Furthermore, due to the interplay between different equipment units, when a fault occurs, readings from multiple sensors will deviate from their normal state. At the same time, certain sensor signals respond to different types of faults, indicating that a single category of fault is associated with multiple sensor signals, which are irregularly distributed.

Consequently, the traditional manual fault identification methods are no longer suitable [6–8]. In the realm of deep learning algorithms [9], it is imperative to employ deep learning architectures that incorporate multiple layers of nonlinear data processing units for advanced feature learning [10–12]. Research indicates that the identification of faults within complex industrial processes is now recognized as a classification issue that leverages signals from multiple time series, which is gaining increasing attention in both academia and industry [13–15]. Hence, mining the interplay of multi-sensor measurements and integrating information, accurately identifying the nonlinear relationships, correlations, and control rules of the process, are crucial for fault diagnosis in large-scale industrial processes.

Existing methods mostly take grid data as input, neglecting the topological structure of the process and the interactions between monitoring variables, while in practice, graph-domain data with topological structures are far more abundant than grid data. Since graphs can reasonably describe real-world systems, convert some unstructured scenarios into

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

https://doi.org/10.1016/j.ress.2024.110468

Available online 28 August 2024

Acronyms and Abbreviations

S	Non-causal Features
Z	Data Representation of the Graph
Y	Predicted Attributes
X	Node Features
V	Node Set
E	Edge Set
G	Input Graph
A	Adjacency Matrix
H	Node Representation Matrix
M_{edge}	Edge Attention Matrix
M_{node}	Node Attention Matrix
$\mathcal{L}_Y$	Supervised Classification Loss
$\mathcal{L}_C$	Causal intervention Loss
$\mathcal{L}_S$	Uniform Classification Loss
$\mathcal{L}_{total}$	Total Loss of IGCL-GNN Model

graphs and use graph-based methods can achieve better performance, which can be applied to areas such as multi-agent systems [16,17]. Due to the complex interactions between sensor measurements, structural property graphs are suitable data structures for describing the characteristics and relationships of sensor data, where each sensor measurement corresponds to a node, and these nodes are connected by implicit edges that represent interactions. Moreover, fault information such as faults and fault propagation can be represented by this graph. A key task in fault diagnosis is to identify the category of faults [18], therefore, this task is formulated as a graph-level classification problem in this study.

By propagating node features across the graph topology, Graph Neural Networks (GNNs) can learn expressive embeddings useful for node and graph-level predictive tasks [19,20], which makes GNNs become a powerful technique for graph-structured data representation learning [21,22]. Recent research [23,24] have shown that in graph classification tasks, the salient properties that determine a graph's label often originate from specific causal substructures in the graph. Contemporary graph neural networks (GNNs) learn via end-to-end backpropagation on structure-rich graph inputs and predominantly rely on exploiting statistical correlations between graph features and outputs. As such, GNNs exhibit a tendency to utilize potentially spurious non-causal features for making predictions, as long as they are associated with the target labels. However, non-causal features that are correlated with labels but not causally related tend to vary significantly across domains. Overfitting to one domain may increase spurious correlations, thereby compromising the generalization and reliability of graph neural networks [25,26].

Analyzing the decision-making process of GNN graph classification from the perspective of mutual information can provide a better understanding of how the presence of non-causal features affects the generalization process of graph neural network learning. According to research on causal assumptions, non-causal features act as confounding factors that open backdoor paths [27,28], and falsely associate causal features with predictions [29]. By modeling the causal features as C and the non-causal features as S, the mutual information between the input graph G and the predicted label Y can be decomposed as:

$$I(Y;G) = I(Y;C) + I(Y;S|C) \tag{1}$$

where $I(Y;C)$ represents the mutual information between the causal features C and the prediction Y, capturing the invariant explanatory mechanism. $I(Y;S|C)$ represents the mutual information between the non-causal features S and the prediction Y given the causal features C, indicating spurious correlations that do not generalize across distributions. As GNNs tend to exploit any statistical association, they tend to maximize $I(Y;S)$ while ignoring the underlying $I(Y;C)$. Therefore, reducing the model's extraction and prediction of non-causal information during the learning process, and enhancing causal information, will enable the model to extract more valuable relevant information, thereby reducing hindrance from irrelevant information and improving the performance of fault diagnosis.

To address this issue, an Information-based Gradient enhanced Causal Learning in GNN (IGCL-GNN) framework is proposed in this paper. In accordance with the chain rule of mutual information, this method breaks down the maximization of mutual information between the graph and the prediction into two processes: non-causal information learning and causal information learning. This approach restricts the extraction of non-causal information during training while reinforcing the extraction of causal information. Specifically, a composite objective function has been devised that integrates a term for enhancing causal features and a regularization term for non-causal features. The causal term aims to maximize the mutual information $I(Y;C)$ between causal features and the prediction target, thereby extracting invariant and interpretable causal information. Concurrently, the non-causal term seeks to minimize $I(Y;S)$, reducing dependency on non-causal information within the correlations. By optimizing this composite objective, the learning of GNNs can be dynamically steered towards the stable $I(Y;C)$ while avoiding trivial $I(Y;S)$. To further refine these components, an attention-based optimization framework has been proposed to explicitly and dynamically differentiate between causal and non-causal subgraphs. A novel gradient reactivation module has been introduced to ensure the reliability of causal subgraph extraction. Through joint learning of both information objectives and achieving Pareto optimality, the model is capable of distilling genuine relevant information while maximally mitigating irrelevant information. The key contributions of our research are outlined as follows:

(1) Addressing issues in fault diagnosis leveraging Graph Neural Networks (GNNs), this study conducts an exploration of causal feature learning in the context of fault diagnosis (graph classification) from an information-theoretic perspective, and explicates the influence of these two types of features on the classification performance.

(2) A novel attention-based causal subgraph extraction module is proposed to separate causal and non-causal subgraphs, along with a new gradient reactivation module to ensure the reliability of causal subgraph extraction.

(3) To address the challenge of handling shortcut features and causal features, a IGCL-GNN strategy has been introduced, which simultaneously enhances the extraction of genuine relevant information and suppresses the extraction of irrelevant information during the learning process, and ensures the maximal extraction of genuine relevant information through the Pareto optimality of two objectives.

(4) Further visualization and in-depth analysis of the IGCL-GNN on the TFF dataset have demonstrated the interpretability and rationality of the IGCL-GNN. Comparisons with existing fault diagnosis algorithms have shown that the IGCL-GNN model can more accurately extract causal and non-causal features, perform more stable classification.

2. Problem formulation and preliminaries

2.1. Problem formulation

(1) Component Signal Fragment: Signal extraction is generally obtained and composed of its constituent elements. Different components are located at various positions within the industrial system, each generating its corresponding signal, thereby forming n raw measurement variables. Over a period of time t, the signal segment generated by the ith component is $s_i = (s_i^1, s_i^2, \ldots, s_i^t)$. However, due to the long operation times of industrial systems, the obtained signal segments span a large range and are difficult to handle. Therefore, it is necessary to obtain multiple signal segments through window sliding, which can be represented as $w_j = (s_i^t, s_i^{t-1}, s_i^{t-2}, \ldots, s_i^{t-m+1}) \in \Omega$. Since signal

segments are stable over short periods, they can serve as inputs for graph-structured modeling.

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

2.2. Attention mechanism in GNN

The attention mechanism excels at focusing on key details and filtering out irrelevant information. Within the context of Graph Neural Networks (GNNs), this mechanism can be adeptly applied to either nodes or edges. It aids in identifying the critical segments of the graph that are most pertinent to the task at hand. These salient subgraphs, in turn, enhance our ability to achieve the desired objectives more effectively.

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

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

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

$$H' = GConv(A, H \odot M_{node}) \tag{3}$$

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

$$h_G = f_{readout}(h_i^{out} | i \in V) \tag{4}$$

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

$$z_G = \Phi(h_G) \tag{5}$$

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

$$\mathcal{L}_{CE} = -\frac{1}{|D|} \sum_{G\in D} \mathbf{y}_G^{\top} \log\left(z_G\right) \tag{6}$$

where L_{CE} is the cross-entropy loss [30] computed on the training data D. y_G is the ground-truth label. Given that this empirical loss hinges on the distributional traits and statistical interdependencies present in the training data, this learning approach captures predictive shortcut features rather than identifying the pivotal causal features.

3. Proposed fault diagnosis method

For fault diagnosis in complex industrial systems, an IGCL-GNN framework is proposed based on multivariate signal segment information, which is introduced in four main parts. Firstly, a theoretical analysis of graph representation learning is presented from an information theory perspective, leading to the proposal of an optimization objective grounded in mutual information. This objective is designed to enhance the association between prediction outcomes and causal features, concurrently diminishing the reliance on non-causal information. Secondly, the disentangle algorithm of causal and non-causal information is described within the graph representation and optimize the proposed objective through variational approximation. Then, how to combine and balance multiple optimization objectives to achieve optimal predictive performance, robustness, and generalization is discussed. Lastly, a novel gradient reactivation module that filters features with greater impact on predictions is introduced to ensure the reliability of causal subgraph extraction.

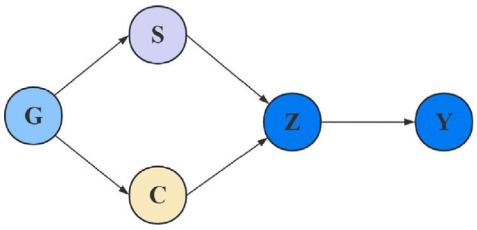

Fig. 1. Structural causal model.

3.1. Mutual information optimization

In this section, graph representation learning is analyzed from an information-theoretic perspective and decompose the information extraction process in neural networks based on causal assumptions and the chain rule of mutual information, proposing an optimization objective grounded in mutual information.

We engaged in a causal examination of GNN modeling and formulated a Structural Causal Model (SCM), as illustrated in Fig. 1. This model delineates the causal relationships among five key variables: graph data G, causal features C, shortcut features S, graph representation Z, and prediction Y. And the arrows denote the causal relationships between the variables. he following causal relationships exist in SCM for graph representation learning:

- $C \leftarrow G \rightarrow S$: C represents the causal features of the graph, which truly reflect the intrinsic properties of the graph data. S represents the shortcut features, which are typically non-causal features caused by data noise. Given that S and C are concurrently present in the graph data, the causal relationships are inherently established.
- $C \rightarrow Z \leftarrow S$: Z is the data representation of the graph. Traditional learning strategies use shortcut features S and causal features C together as inputs to extract discriminative information to obtain the graph representation Z.
- $Z \rightarrow Y$: The classifier makes predictions based on the graph representation Z, obtaining the predicted attributes of the input graph Y.

From the perspective of mutual information, the optimization objective is equivalent to maximizing the mutual information between the representation and the prediction target $I(Z;Y)$ (Objective I), where Z is the learned representation and Y is the prediction target. Nevertheless, because mutual information captures the correlation between variables without assessing causality, it can only assimilate the statistical associations between the input features and the labels present in the training data. Consequently, the optimization of predictive relationships may not be solely driven by the causal features we are interested in; instead, it could be predominantly influenced by non-causal features and their indirect effects on prediction. Therefore, the optimization objective I is equivalent to maximizing the information flow $Z \rightarrow Y$, but it cannot distinguish whether the correlation between S and C is a causal relationship. To address this issue, the mutual information objective is decomposed using the chain rule of mutual information and the aforementioned causal relationships:

$$I(Z;Y) = I(C,S;Y) = I(C;Y) + I(S;Y|C) \tag{7}$$

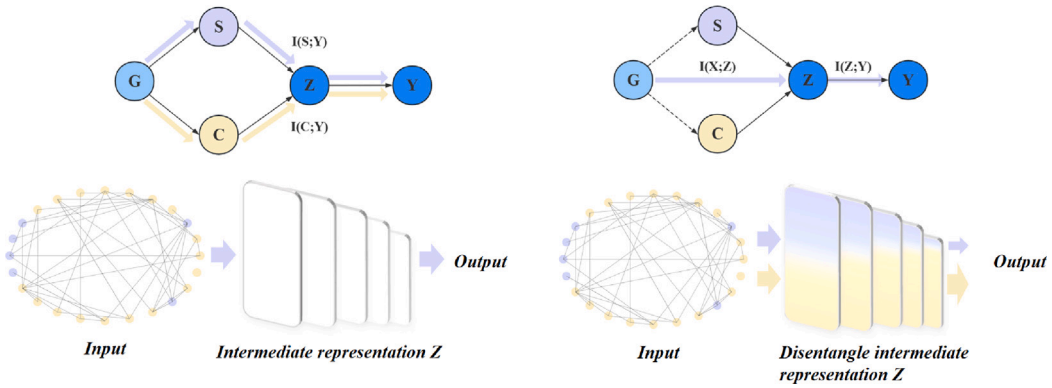

Fig. 2. Differences between Objective II and Objective I.

where $I(C;Y)$ is the mutual information between causal features C and prediction Y, representing true information and causal dependencies. $I(S;Y|C)$ is the mutual information between non-causal features S and prediction Y given C, representing noise and non-causal dependencies that C cannot explain. Due to the entanglement of C and S, optimizing $I(Z;Y)$ will increase $I(C;Y)$ but also increase $I(S;Y|C)$, leading to non-causal dependencies and noise, which negatively impact model generalization. To better optimize causal dependencies, the optimization objective I is replaced with Objective II:

$$\max I(Z;Y) \ \ \& \ \ \max I(C;Y) \ \ \& \ \ \min I(S;Y|C) \tag{8}$$

Fig. 2 vividly illustrates the optimization effect of Objective II. This means that while improving prediction accuracy, it can use stable causal dependencies to enhance robustness and generalization, and reduce fitting to noise and non-causal dependencies.

3.2. Disentanglement and optimization

The overview of IGCL-GNN is given in Fig. 3.

To optimize this objective, the representation's causal and non-causal information is first disentangled. To achieve this, two attention layers are proposed in this paper that disentangle causal and non-causal information at the edge and node levels, respectively. Specifically, given a GNN encoder $f(\cdot)$ and graph $G = \{A, X\}$, the encoded representation can be obtained:

$$H = f(A, X); \tag{9}$$

where H contains both causal and non-causal information from G. To separate them, the causal information extractor Att_C and the non-causal information extraction layer Att_S are proposed in our approach, which learn attention over nodes and edges that are causally and non-causally relevant from the representation, respectively:

$$\begin{cases} \alpha_x, \alpha_a = \sigma\left(Att_C(H, A)\right) \\ \beta_x, \beta_a = \sigma\left(Att_S(H, A)\right) \end{cases} \tag{10}$$

where α_x, α_a represent the node and graph level causal attentions, emphasizing the significance of nodes and edges in establishing causal relationships. β_x, β_a represent the corresponding non-causal attentions, emphasizing the significance in non-causal dependencies. The original graph G is disentangled into causal graph Z_C and non-causal graph Z_S based on the two attentions:

$$\begin{cases} Z_C = GConv_C\left(A \odot \alpha_a, X \odot \alpha_x\right) \\ Z_S = GConv_S\left(A \odot \beta_a, X \odot \beta_x\right) \end{cases} \tag{11}$$

Utilizing disentangled causal representations Z_C and non-causal representations Z_S, Objective II can be optimized. As detailed in the discussion of Section 3.1, to enhance the learning of true causal dependencies, the maximization of $I(C;Y)$ is aimed for. However, due to the high complexity of mutual information optimization, a variational lower bound is derived and minimized.:

$$I(C;Y) \geq E_p(c,y)[log\ q(y|c)] - H(Y) \tag{12}$$

where $q(y|c)$ is a conditional probability distribution that can be modeled by a classifier $f(\cdot)$; $H(Y)$ is the entropy of Y, which is a constant; $E_p(c,y)[log\, q(y|c)]$ is the expectation over the classification results of the overall causal features. The lower bound of $I(C;Y)$ can be optimized by maximizing $E_p(c,y)[log\ q(y|c)]$, the causal intervention loss $\mathcal{L}_C$ is optimized to enhance the learning of causal dependencies in this paper:

$$\mathcal{L}_C = -\sum_{c\in C}\sum_{y\in Y} p(c,y) log\ q(y|c) \tag{13}$$

Similarly, to maximize $I(Z;Y)$ and improve overall predictive performance, the cross-entropy loss of the global representation, utilizing both causal and non-causal information, is optimized:

$$\mathcal{L}_Y = -\sum_{z\in Z}\sum_{y\in Y} p(z,y) log\ q(y|z) \tag{14}$$

Then, to minimize $I(S;Y|C)$ and reduce the impact of non-causal features such as data noise on predictions, S is made independent of Y by minimizing the KL divergence between Z_S and a uniform distribution:

$$\mathcal{L}_S = KL(S \parallel u(S)) = -\sum_{s\in S} p(s) log \frac{p(s)}{u(s)} \tag{15}$$

where KL is the KL divergence, $p(s)$ is the distribution of the non-causal representation S, and $u(S)$ is the uniform distribution over S. Since the uniform distribution has the maximum entropy, minimizing L_S encourages S to not contain information about Y, thereby minimizing the dependency between the non-causal representation S and the prediction Y. By optimizing the aforementioned loss functions, overall predictive accuracy and reliance on causal features can be improved, while reducing reliance on non-causal features. However, despite achieving the transition from mutual information correlations to causal dependencies, there are still complex interactions between these objectives, and there may even be conflicts between reducing non-causal dependencies and enhancing overall predictive accuracy.

3.3. Combinations and tradeoff

Given the pronounced correlation between the complexity of causal dependencies and optimization objectives, optimization scheme is conceptualized as a multi-objective optimization problem in this paper. Mathematically, our problem can be articulated as:

$$\min_{\theta\in\Theta} \mathcal{L}(\theta) = \min_{\theta\in\Theta}(\mathcal{L}_Y(\theta), \mathcal{L}_C(\theta), \mathcal{L}_S(\theta)) \tag{16}$$

The resolution of multi-objective optimization problems is typically geared towards achieving a holistic optimum, known as Pareto

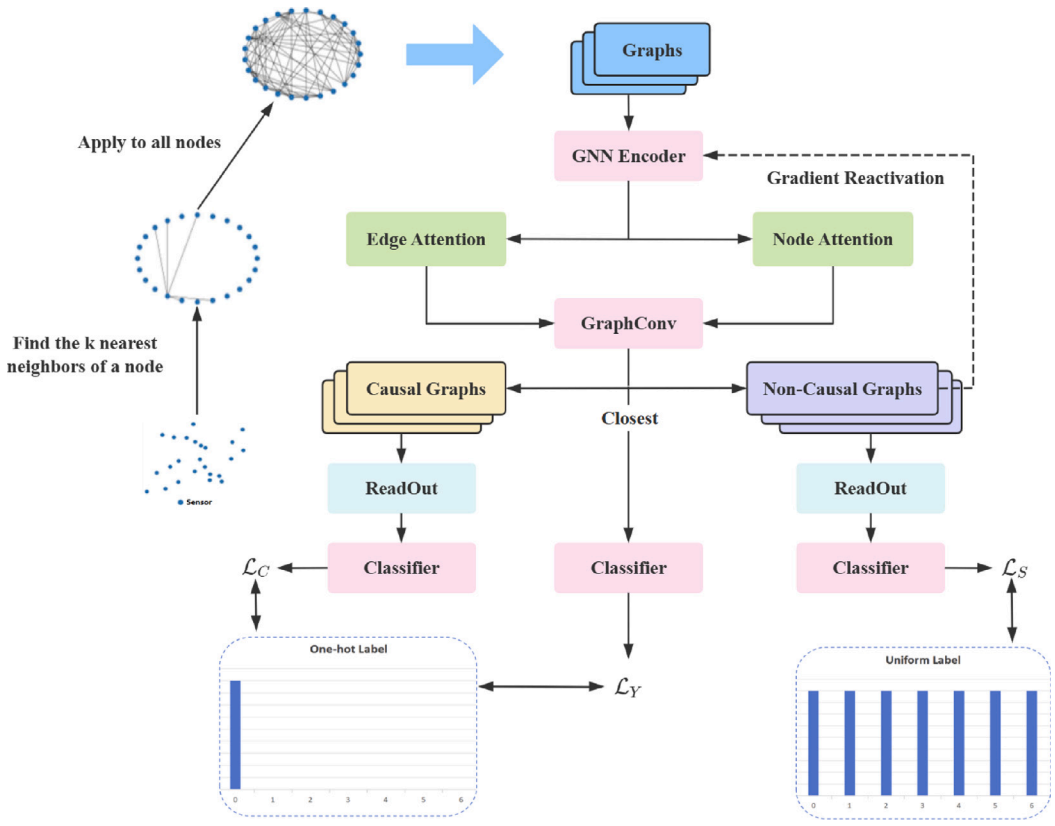

Fig. 3. Overview of IGCL-GNN.

optimality. Pareto optimality denotes the optimal trade-off among multiple objectives. Within the context of our proposed problem, Pareto optimality can be delineated as follows:

Definition 1 (*Pareto Optimality*)**.** For the multi-objective optimization problem $\min_{\theta\in\Theta}\left(\mathcal{L}_Y(\theta),\mathcal{L}_C(\theta),\mathcal{L}_S(\theta)\right)$, a solution $\theta^* \in \Theta$ is Pareto optimal if there does not exist another solution $\theta \in \Theta$ dominates it, i.e.,

$$\begin{aligned}&\theta\in\Theta \text{ s.t. } \mathcal{L}_i(\theta)\le\mathcal{L}_i\left(\theta^*\right) \text{ for all } i=Y,C,S\\&\text{and } \mathcal{L}_j(\theta)<\mathcal{L}_j\left(\theta^*\right) \text{ for some } j=Y,C,S\end{aligned} \tag{17}$$

Here, Θ represents the set of feasible solutions. This implies that a Pareto optimal solution cannot be improved in one objective without at least one other objective deteriorating. The collection of all Pareto optimal solutions is referred to as the **Pareto Frontier**.

Although Pareto optimality [31,32] is regarded as possessing many desirable qualities and serves as the ultimate goal for numerous multi-objective problems, it is not applicable to our problem with causal assumptions. If Objective II has achieved Pareto optimality, and further improvements in $\max I(Z;Y)$ are possible, then the optimization of $\max I(Z;Y)$ must be halted to avoid undermining $\max I(C;Y)$ and $\min I(S;Y|C)$. In other words, enhancing causal dependencies and reducing non-causal dependencies may impede the improvement of overall prediction, which is clearly contrary to our motivations and causal assumptions. Therefore, the objectives are reformulated as:

$$\theta\in\Theta \ \max I(Z;Y) \ s.t. \ \theta\in P(I(C;Y),-I(S;Y|C)) \tag{18}$$

This implies that the predictive relevance of the representation is maximized under the premise of achieving optimal causal dependencies for predictive information. To attain Pareto optimality that maximizes causal dependencies and minimizes non-causal dependencies, a Multi-Objective Gradient Descent Algorithm (MGDA) [33] is employed in optimizing $\mathcal{L}_C(\theta)$ and $\mathcal{L}_S(\theta)$. MGDA is a gradient-based multi-objective optimization algorithm that progressively balances the gradients of multiple objective functions at each iteration.

The crux of MGDA lies in identifying a direction d at each iteration such that a small step η along d can improve all objective functions. In our context, given that there are only two sub-objectives $\mathcal{L}_C(\theta)$ and $\mathcal{L}_S(\theta)$, the MGDA solution approach can be simplified. First, the gradients of the two sub-objectives $\nabla\mathcal{L}_C(\theta)$ and $\nabla\mathcal{L}_S(\theta)$ are computed, and then the angle θ between the two gradients is calculated:

$$\alpha=arccos\frac{\nabla\mathcal{L}_C(\theta)^T\,\nabla\mathcal{L}_S(\theta)}{\nabla\mathcal{L}_C(\theta)\nabla\mathcal{L}_S(\theta)} \tag{19}$$

By calculating α, the direction in which both objectives can descend concurrently is ascertained. For the Pareto optimal solution θ^*, it can be utilized as a constraint to optimize $I(Z;Y)$ This corresponds to finding a solution on the Pareto frontier that maximizes $I(Z;Y)$ to obtain the final representation Z^* for the predictive task. Through this multi-objective optimization process, a stable trade-off between predictive relevance and causal dependency is achieved.

3.4. Gradient reactivation

In this paper, a gradient reactivation method is further proposed to filter non-causal subgraphs, thereby enhancing the precision of separating causal and non-causal subgraphs. After obtaining the non-causal subgraph, nodes and edges with substantial gradients are reactivated back into the causal subgraph, indicating that they significantly contribute to the prediction. For the derived non-causal subgraph Z_S, the objective is to extract the erroneously assigned causal components from it. The loss incurred from predicting the true labels using Z_S is computed:

$$\mathcal{L}_t=-\sum_{s\in S}\sum_{y\in Y}p(s,y)log\ q(y|c) \tag{20}$$

The cross-entropy loss L_t does not participate in backpropagation and is solely utilized for the computation of gradients $\overline{M}_{node}$ and $\overline{M}_{edge}$.

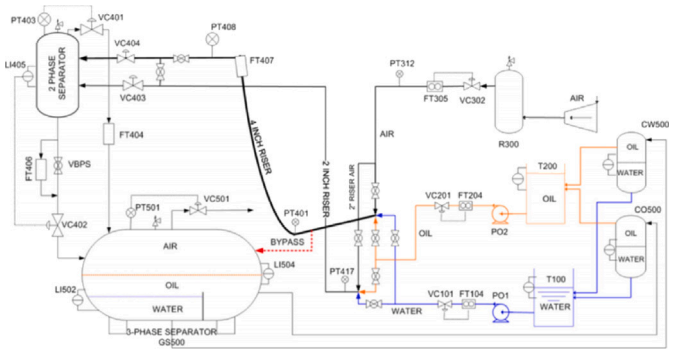

Fig. 4. Overall structure and sensor distribution of the three-phase flow facility.

Elements with substantial gradients within $\overline{M}_{node}$ and $\overline{M}_{edge}$ should not be included in the non-causal subgraph. They are removed from the non-causal subgraph, with the new non-causal subgraph masks denoted as $\overline{M}'_{node}$ and $\overline{M}'_{edge}$. Subsequently, two additional losses are introduced to ensure that elements with significant gradients are not encapsulated within the non-causal subgraph.

$$\begin{aligned} Loss_{node} &= (M_{node} - M'_{node})^2 \\ Loss_{edge} &= (M_{edge} - M'_{edge})^2 \end{aligned} \tag{21}$$

So, the final loss framework becomes:

$$Loss_{total} = \mathcal{L}_Y + \lambda_1 \cdot \mathcal{L}_S + \lambda_2 \cdot \mathcal{L}_C + \lambda_3 \cdot Loss_{node} + \lambda_4 \cdot Loss_{edge} \tag{22}$$

This loss framework not only automatically optimizes λ_1, λ_2 through Multi-Objective Gradient Descent Algorithm (MGDA), but also offers flexibility through the tuning of hyperparameters λ_3, λ_4. This allows for a nuanced approach to enhancing causal relationships and minimizing non-causal dependencies, all aimed at maximizing predictive accuracy and model robustness.

4. Experiment results and comparisons

4.1. Datasets

The TFF [34], designed by Cranfield University, is one of the typical industrial systems. This system is a genuine and complex industrial-scale test bench used for controlling a pressurized system, equipped with 24 component sensors. These sensors are strategically positioned at various critical locations within the system to detect the density, temperature, pressure, and flow rate at different key points, thereby measuring the flow rates of water, oil, and air. An illustration of this system is shown in Fig. 3. Furthermore, the system is capable of functioning across various operating conditions and offers experimental data. Details of the sensors can be found in [34], and the TFF dataset is available for download at the following link TFF (see Fig. 4). To obtain data under different operating conditions, the data acquisition setup included 20 sets of process inputs, which were composed of combinations from the system's four air flows and five water flows, and three sets of data were obtained through simulation. For the fault dataset, the system simulated a total of six typical faults that might occur in practice. During the fault data collection process, the equipment initially operated in a normal state, then faults were injected. Once the faults developed to a certain extent, the fault injection was stopped, and the system gradually returned to a healthy state. Therefore, the data generated for each fault type contains transition information from the initial state to the fault state, including the process of the fault developing from minor to severe. Moreover, to ensure a comprehensive and rich dataset, the data collection considered the equipment under both steady-state and varying conditions, hence each type of fault includes multiple datasets. The data sampling frequency is 1 Hz. Additionally, this paper employs the min–max normalization method for each component's data, where $x = (x - x_{min})/(x_{max} - x_{min})$. To better extract fault feature information, normal data is removed from the fault data, with each sample containing 50 s of temporal information. The samples are randomly divided, with 0.9 allocated to the training set and 0.1 reserved for the test set. The fault and normal types in the dataset, along with their corresponding sample counts, are shown in Table 1.

Table 1
Fault categories and corresponding sample sizes in the TFF dataset.

Fault class	Fault type	Numbers of samples
0	Normal	667
1	Air line blockage	194
2	Water line blockage	172
3	Top separator input blockage	433
4	Open direct bypass	226
5	Slugging conditions	105
6	Pressurization of the 2″line	96

Table 2
The bold values indicate that the IGCL-GNN model outperforms other comparison methods in terms of Accuracy, Micro F1, and Macro F1.

	Accuracy	Micro F1	Macro F1
GAT	77.25	86.39	75.97
BHGNN	94.51	92.82	92.56
CTA-GNN	96.30	96.72	94.73
IGCL-GNN	**98.42**	**98.39**	**97.94**

4.2. Experimental settings

4.2.1. Current baseline methods

Current Baseline Methods: To demonstrate the performance of our IGCL-GNN method in real industrial systems, the IGCL-GNN method is compared with existing baseline methods: Graph attention networks (GAT) [35], Bayesian hierarchical graph neural network (BHGNN) [36], causal-trivial attention graph neural network (CTA-GNN) [37].

(1) GAT: GAT learns based on node features to obtain edge weights and measure the importance of edges. When unstructured multivariate time series data is constructed into a fully connected graph, GAT can be directly applied to multivariate time series data.

(2) BHGNN: BHGNN captures epistemic and aleatoric uncertainties by employing a variational dropout approach and adjusts the strength of temporal consistency constraints using the uncertainty information of each sample. Additionally, the BHGNN method models process data as a hierarchical graph, integrating data with domain knowledge by fully leveraging interaction-aware modules and the physical topology of industrial processes.

(3) CTA-GNN: CTA-GNN first generates representations of nodes and edges by estimating soft masks, then obtains causal and shortcut features from the graph through disentanglement, and finally combines each causal feature with various shortcut features through the adjustment of the backdoor criterion parameterized by causal theory.

4.2.2. Evaluation indicators and parameter settings

To demonstrate the superiority of IGCL-GNN, Accuracy, Micro F1, Macro F1, and the confusion matrix are employed as metrics for comparison with other methods. Additionally, extensive experiments are conducted on various baseline methods and the IGCL-GNN model, comparing them to select the hyperparameters that exhibit the best performance and testing outcomes. For the TFF dataset, the input graph comprises 24 nodes, corresponding to 24 sensors. The feature size

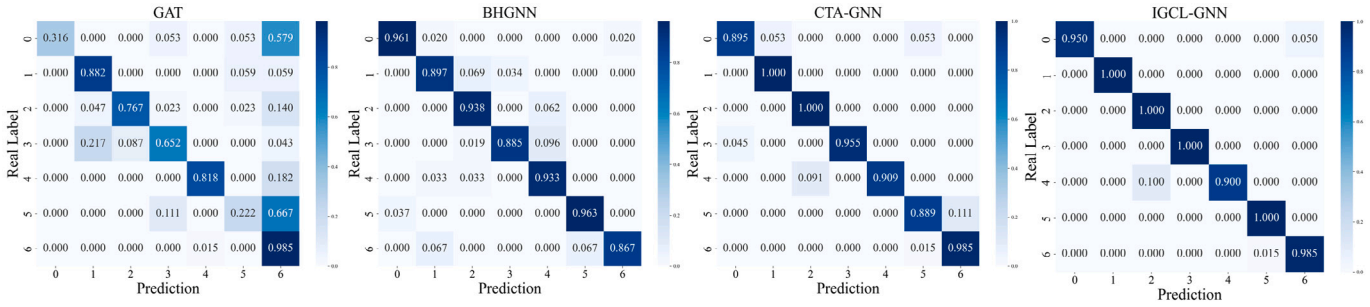

Fig. 5. Confusion matrix comparison. (a) GAT. (b) BHGNN. (c) CTA-GNN. (d) IGCL-GNN.

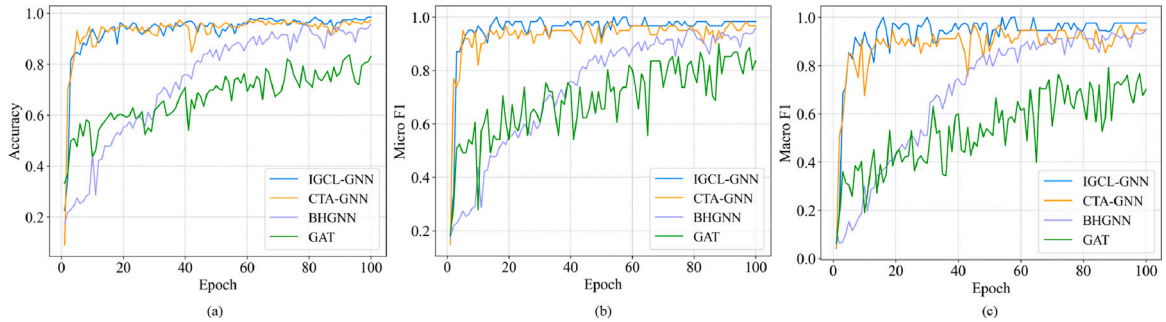

Fig. 6. Accuracy and F1 score comparison. (a) Accuracy. (b) Micro F1 score. (c) Macro F1 score.

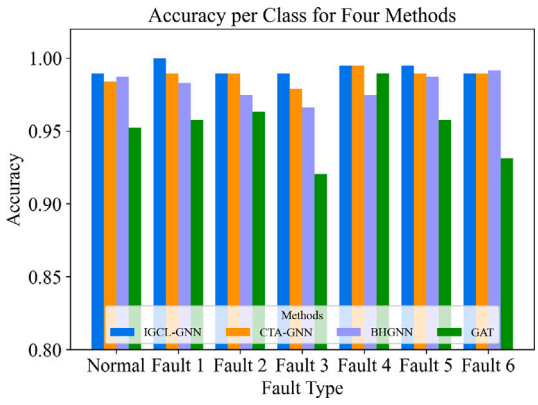

Fig. 7. Accuracy per class for four methods.

of each node is 50, representing the captured 50-s signal segments. Additionally, the maximum number of epochs is set to 100, with a learning rate of 0.001. In the loss function, $\lambda 3$ is assigned a value of 0.5, and $\lambda 4$ is also assigned a value of 0.5.

4.3. Fault classification performance

Utilizing the confusion matrix derived from the testing phase, it is possible to evaluate the performance of various fault diagnosis methods. By visualizing the confusion matrices, one can gain a clear insight into the diagnostic capabilities of each model across different fault categories. The visual representations of the confusion matrices for the respective models are depicted in Fig. 5. As the training process progresses, the performance on the test set is also subject to change. The Micro F1 scores, Macro F1 scores, and accuracy metrics are recorded during the testing phase to observe their variations throughout the process. The F1 score is a crucial measure of a model's classification accuracy, offering a balanced assessment by combining precision and recall. The Micro F1 score aggregates predictions across all categories to provide an overall accuracy metric, which is particularly useful for datasets with class imbalances as it mitigates the disproportionate impact of minority classes. In contrast, the Macro F1 score calculates the average F1 score for each category, ensuring a fair evaluation of all categories, including those with fewer samples, and helps pinpoint areas where the model underperforms. The results of the Micro F1 scores, Macro F1, and accuracy ratings are shown in Fig. 6. Simultaneously, we utilized Fig. 7 to compare the accuracy of various methods across different fault types. Table 2 displays a comparison of the classification performance of different models on the TFF dataset.

By comparing its performance with GAT, the IGCL-GNN model demonstrates the ability to comprehensively leverage the raw information of nodes and learn the interrelationships among node information, leading to enhanced performance. In contrast, the GAT model uses an attention mechanism to weigh and aggregate the features of neighboring nodes within the graph, yet it overlooks the graph's underlying structure. Unlike the GAT, which employs fully connected graphs, the IGCL-GNN model utilizes graph structure descriptions specific to various fault types as its input, as illustrated in Fig. 8. Consequently, GAT does not account for the interactions between various components. Given that in industrial systems, intricate interactions among different components constitute a significant aspect of fault characteristics, this limitation hinders GAT's ability to differentiate between distinct fault types effectively.

BHGNN takes into account the correspondence between the model architecture and the actual process system, representing a fault diagnosis method within the Bayesian Deep Learning(BDL) framework. BHGNN constructs a hierarchical graph of the industrial process to extract fault features, primarily consisting of two-tiered structures: the sensor-level graph and the unit-level graph. Additionally, it employs a variational dropout method to capture epistemic and aleatoric uncertainties, and utilizes the uncertainty information of each sample to adjust the strength of the temporal consistency (TC) constraints. However, the BHGNN may exhibit overconfidence in the uncertainty information of the output probabilities and among samples, resulting in classification performance that is inferior to IGCL-GNN. Therefore,

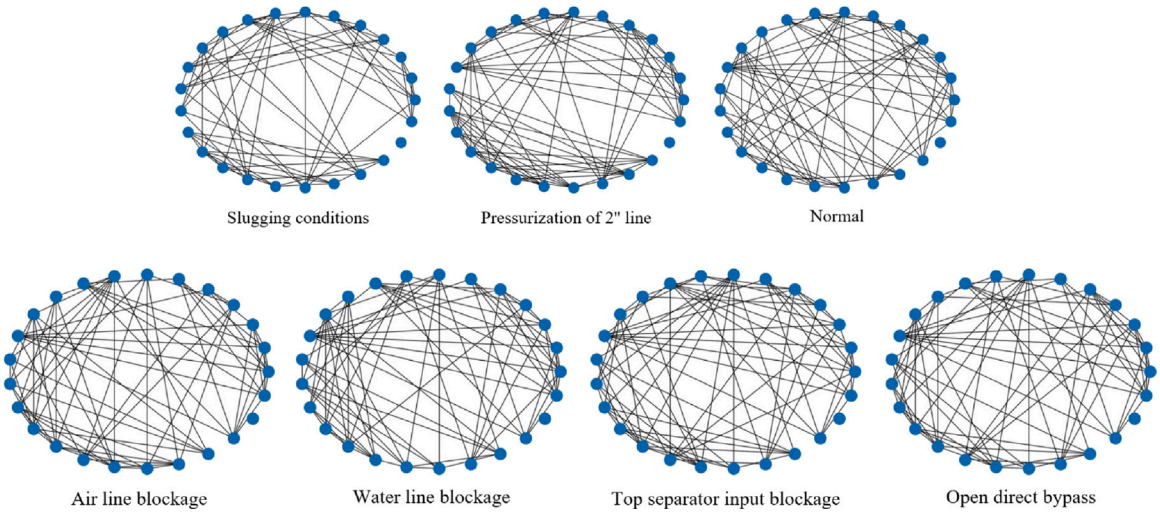

Fig. 8. Graph structure illustration.

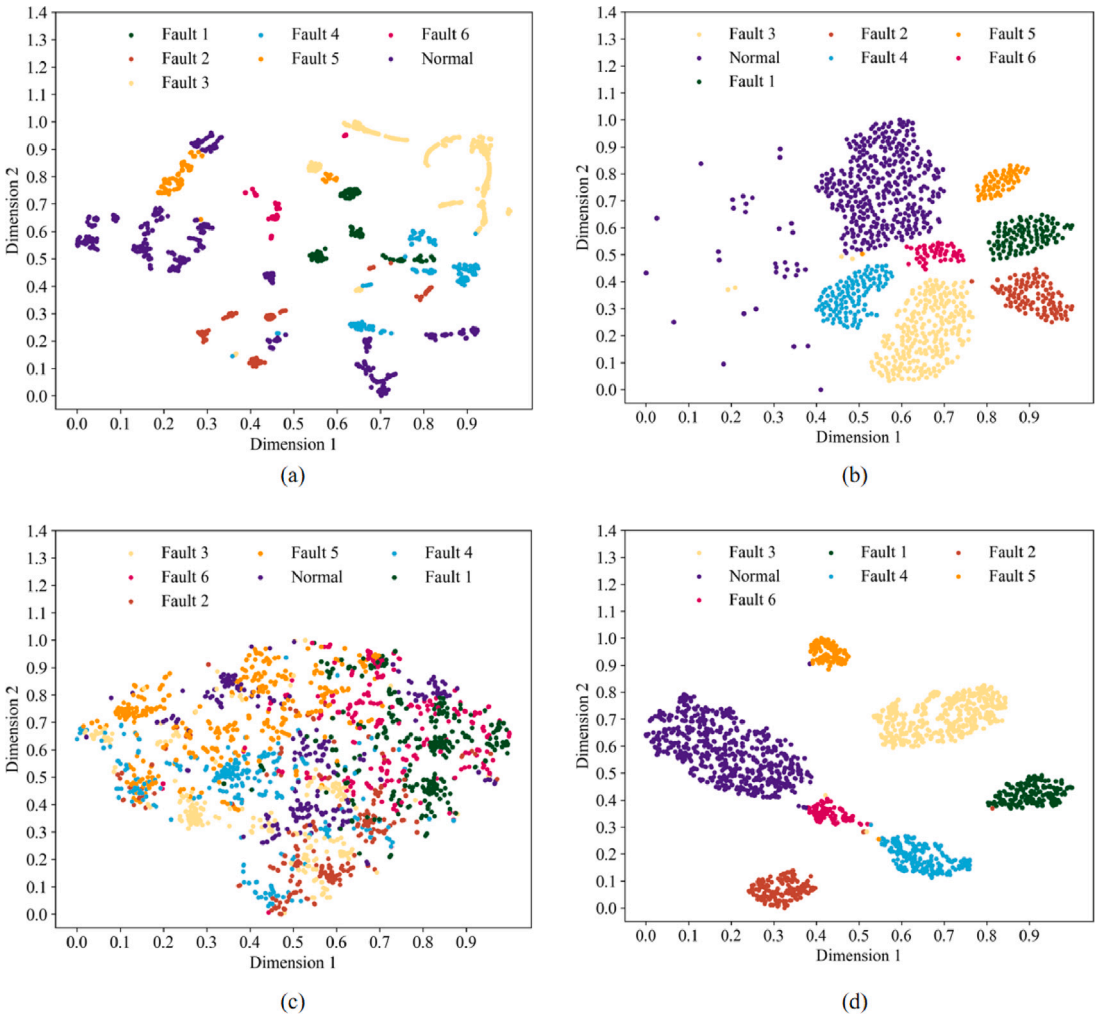

Fig. 9. Use the t-SNE method to visualize the results. (a) Original graph data space. (b) Z_C containing only causal information. (c) Z_S containing only non-causal information. (d) Z containing both causal and non-causal information.

a framework capable of introducing high uncertainty for these samples is necessary for reliable fault diagnosis.

CTA-GNN introduces a causal GNN framework that filters shortcut features while mining causal features through three steps: estimating soft masks, disentangling, and causal intervention. CTA-GNN employs an attention module to learn both causal and shortcut features from a given graph, although the shortcut features may still be contaminated with causal features. Additionally, manually tuning the hyperparameters, specifically $\lambda 1$ that controls the degree of disentanglement and $\lambda 2$ that governs the extent of causal intervention, is time-consuming and

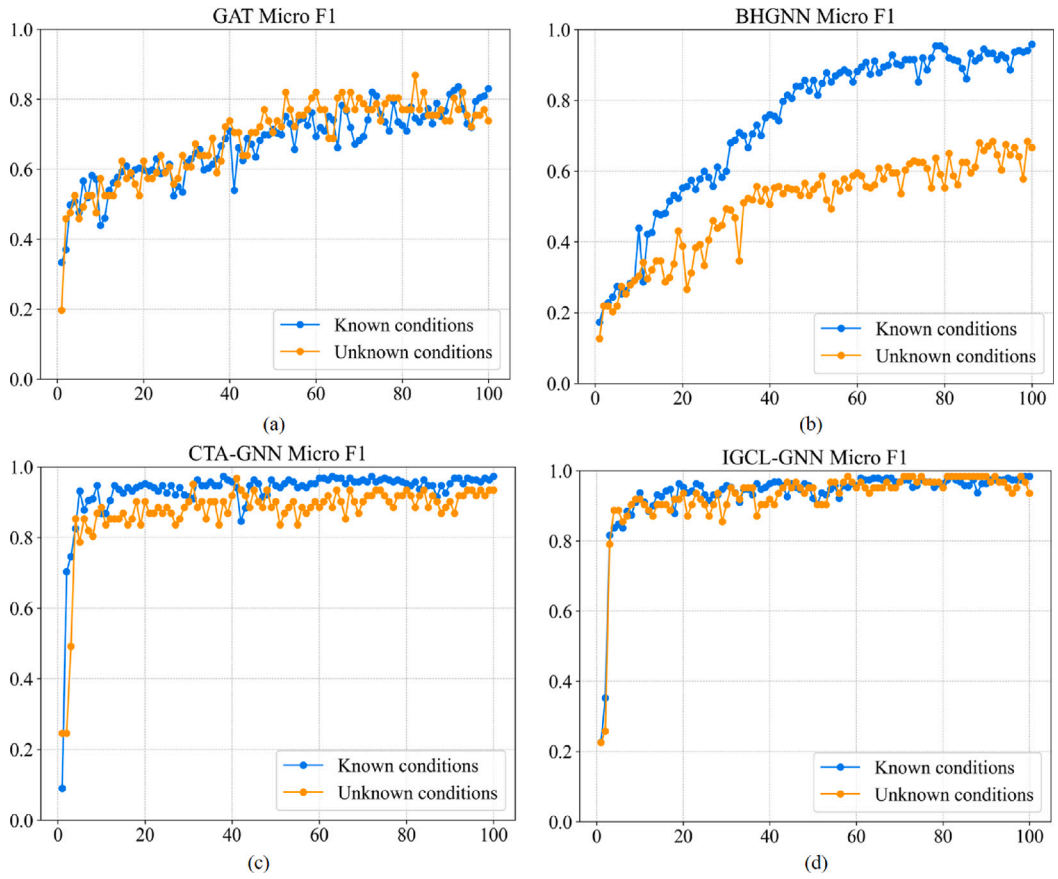

Fig. 10. Micro F1 score change comparison. (a) GAT. (b) BHGNN. (c) CTA-GNN. (d) IGCL-GNN.

labor-intensive, making it challenging to identify the optimal solution (see Fig. 9).

Fig. 6 displays the simplified two-dimensional feature maps and the learned fault characteristics of the IGCL-GNN model obtained through the t-Distributed Stochastic Neighbor Embedding (t-SNE) [38] method. Specifically, (a) represents the data space of the initial graph, (b) depicts the representation Z_C that contains only causal information, (c) illustrates the representation Z_S that contains only non-causal information, and (d) shows the representation Z that encompasses both causal and non-causal information. Apparently, due to the widespread presence of environmental noise and varying degrees of faults, the sample features for the same type of fault are diverse. By maximizing the mutual information $I(C;Y)$, the intra-class relevance and predictive relevance of Z_C after classification are further increased, enhancing the dependence on C. At the same time, the clustering effect of S is reduced, indicating that minimizing $I(S;Y)$ has successfully decreased the correlation between non-causal features and predictions. It can also be seen that the correlation between the joint representation Z and predictions has significantly increased, fully demonstrating the effectiveness of the objective $I(Z;Y)$. After learning through the IGCL-GNN model, features of the same fault are clustered together, enabling efficient classification.

Additionally, given the challenges in estimating mutual information, we have derived a variational bound that enables us to transform the aforementioned objectives into a manageable loss function. By employing the Multi-Objective Gradient Descent Algorithm (MGDA), a stable and effective transition from information correlation to causal dependence is achieved. MGDA, as a multi-objective optimization algorithm, eliminates the need to preset weights for individual objectives, efficiently navigating the solution space using gradient information. It reliably converges to the Pareto optimal front, thereby significantly enhancing the efficiency and adaptability of the optimization process.

According to the outcomes of our experiments, the IGCL-GNN model shows better fault diagnostic performance on the TFF dataset than the baseline methods. The IGCL-GNN model more precisely captures causal and non-causal features, notably strengthens the association between the predictions and causal characteristics, diminishes the dependence on non-causal features, and more accurately identifies the underlying causes of faults in industrial systems, culminating in excellent diagnostic performance.

4.4. Ablation experiment

To demonstrate the model's generalization capability, this paper assumes that two sensors fail, causing all the signal segments they generate to be zeroed out. Under this scenario, the KNN algorithm [39] is applied to the signal segments from the 24 components to identify nodes that are close to each other or functionally similar and categorize them into groups. Subsequently, a new topological graph is computed and fed into the original model. Normal conditions are defined as known conditions, while sensor failure conditions are defined as unknown conditions. The Micro F1 score is used as the metric to evaluate changes in model performance, as shown in Fig. 10. The results indicate that the proposed IGCL-GNN model performs the best under unknown conditions, with diagnostic accuracy essentially consistent with that under known conditions. For the CTA-GNN model, there may be some discrepancy from the results under known conditions, possibly due to inadequate control of the disentanglement and causal intervention degrees. For the BHGNN model, the sensor failure has a significant impact on the model's classification effectiveness, indicating that the BHGNN model is not sufficiently stable. These experiments conclusively illustrate that the IGCL-GNN model has strong stability and adaptability.

5. Conclusion

This paper firstly transforms the complex industrial fault diagnosis problem into a graph recognition task, followed by a theoretical analysis of graph representation learning from an information-theoretic perspective, aiming to enhance the generalization of graph neural networks by optimizing causal dependencies. In industrial fault diagnosis, the final prediction results should primarily rely on causal features, but in reality, they are often disturbed by non-causal factors. Therefore, through information-theoretic analysis, the limitations of relying on spurious correlations are identified in this paper, and an objective is introduced that maximizes causal mutual information while minimizing non-causal terms. The optimization objective is achieved through a causal disentanglement module and multi-objective optimization. To ensure the reliability of the separation between causal and non-causal subgraphs, a gradient reactivation module is proposed to constrain the correlation between non-causal subgraphs and actual labels. The IGCL-GNN model, as evaluated on the TFF dataset, has shown commendable results in fault diagnosis tasks. It has made significant advancements in fault diagnosis, exhibiting enhanced stability and adaptability compared to other benchmark methods.

Fault diagnosis often entails open-set recognition challenges, given the inherent unpredictability of machinery and operational environments. Looking ahead, research should focus on leveraging IGCL-GNN for open-set recognition tasks. This involves not only pinpointing known faults with precision through causal features but also adeptly detecting previously unseen faults, thereby preventing them from remaining undetected and potentially disrupting industrial operations.

CRediT authorship contribution statement

Ruonan Liu: Methodology, Data curation. **Yunfei Xie:** Writing – original draft, Methodology, Data curation. **Di Lin:** Methodology, Data curation. **Weidong Zhang:** Investigation. **Steven X. Ding:** Investigation.

Declaration of competing interest

The authors declare that they have no relevant financial or non-financial interests that could be perceived as influencing the results or discussion reported in this paper. The research was conducted independently, and no commercial party had any involvement in the study design, data collection, analysis, interpretation, writing, or the decision to submit the paper for publication.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors confirm that all authors have read and approved the final manuscript and that there are no known conflicts of interest associated with this publication.

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 under Grant No. 22015810300, Tianjin Applied Basic Research Project under Grant No. 22JCQNJC00410, Young Elite Scientist Sponsorship Program under Grant No. YESS20220409, Alexander von Humboldt Foundation Grant No. 1226831 and State Key Laboratory of Reliability and Intelligence of Electrical Equipment No. EERI-KF2022001.

References

[1] 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.
[2] 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.
[3] Fedullo T, Morato A, Tramarin F, Rovati L, Vitturi S. A comprehensive review on time sensitive networks with a special focus on its applicability to industrial smart and distributed measurement systems. Sensors 2022;22(4):1638.
[4] 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.
[5] 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.
[6] Theissler A, Pérez-Velázquez J, Kettelgerdes M, Elger G. Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. Reliab Eng Syst Saf 2021;215:107864.
[7] Xu Z, Saleh JH. Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. Reliab Eng Syst Saf 2021;211:107530.
[8] Xia M, Shao H, Williams D, Lu S, Shu L, de Silva CW. Intelligent fault diagnosis of machinery using digital twin-assisted deep transfer learning. Reliab Eng Syst Saf 2021;215:107938.
[9] Liu R, Xiao D, Lin D, Zhang W. Intelligent bearing anomaly detection for industrial internet of things based on auto-encoder Wasserstein generative adversarial network. IEEE Internet Things J 2024.
[10] Wu J, Zhao Z, Sun C, Yan R, Chen X. Learning from class-imbalanced data with a model-agnostic framework for machine intelligent diagnosis. Reliab Eng Syst Saf 2021;216:107934.
[11] Wu D, Zhao J. Process topology convolutional network model for chemical process fault diagnosis. Process Saf Environ Prot 2021;150:93–109.
[12] 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.
[13] Uddin MP, Mamun MA, Hossain MA. PCA-based feature reduction for hyperspectral remote sensing image classification. IETE Tech Rev 2021;38(4):377–96.
[14] Li G, Choi B, Xu J, Bhowmick SS, Chun K-P, Wong GL-H. Shapenet: A shapelet-neural network approach for multivariate time series classification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, 2021, p. 8375–83.
[15] Benhaddi M, Ouarzazi J. Multivariate time series forecasting with dilated residual convolutional neural networks for urban air quality prediction. Arab J Sci Eng 2021;46:3423–42.
[16] Berahmand K, Nasiri E, Li Y, et al. Spectral clustering on protein-protein interaction networks via constructing affinity matrix using attributed graph embedding. Comput Biol Med 2021;138:104933.
[17] Huang T, Zhang Q, Tang X, Zhao S, Lu X. A novel fault diagnosis method based on CNN and LSTM and its application in fault diagnosis for complex systems. Artif Intell Rev 2022;55(2):1289–315.
[18] Wang N, Li H, Wu F, Zhang R, Gao F. Fault diagnosis of complex chemical processes using feature fusion of a convolutional network. Ind Eng Chem Res 2021;60(5):2232–48.
[19] Varbella A, Gjorgiev B, Sansavini G. Geometric deep learning for online prediction of cascading failures in power grids. Reliab Eng Syst Saf 2023;237:109341.
[20] Zhu J, Wang J, Han W, Xu D. Neural relational inference to learn long-range allosteric interactions in proteins from molecular dynamics simulations. Nat Commun 2022;13(1):1661.
[21] Trivedi R, Yang J, Zha H. Graphopt: Learning optimization models of graph formation. In: International conference on machine learning. PMLR; 2020, p. 9603–13.
[22] Jin W, Derr T, Liu H, Wang Y, Wang S, Liu Z, Tang J. Self-supervised learning on graphs: Deep insights and new direction. 2020, arXiv preprint arXiv:2006.10141.
[23] Lv F, Liang J, Li S, Zang B, Liu CH, Wang Z, Liu D. Causality inspired representation learning for domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 8046–56.
[24] Marcinkevičs R, Vogt JE. Interpretable models for granger causality using self-explaining neural networks. 2021, arXiv preprint arXiv:2101.07600.
[25] Bahng H, Chun S, Yun S, Choo J, Oh SJ. Learning de-biased representations with biased representations. In: International conference on machine learning. PMLR; 2020, p. 528–39.
[26] Liu J, Zheng S, Wang C. Causal graph attention network with disentangled representations for complex systems fault detection. Reliab Eng Syst Saf 2023;235:109232.
[27] Pearl J. Interpretation and identification of causal mediation. Psychol Methods 2014;19(4):459.

[28] Neuberg LG. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. Econometric Theory 2003;19(4):675–85.
[29] Liu R, Zhang Q, Lin D, Zhang W, Ding SX. Causal intervention graph neural network for fault diagnosis of complex industrial processes. Reliab Eng Syst Saf 2024;110328.
[30] Lyakhov P, Lyakhova U, Kalita D. Multimodal neural network system for skin cancer recognition with a modified cross-entropy loss function. 2023.
[31] Censor Y. Pareto optimality in multiobjective problems. Appl Math Optim 1977;4(1):41–59.
[32] Stiglitz JE. Pareto optimality and competition. J Finance 1981;36(2):235–51.
[33] Désidéri J-A. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. C R Math 2012;350(5–6):313–8.
[34] 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.
[35] Casanova P, Lio ARP, Bengio Y. Graph attention networks. ICLR. Petar Velickovic Guillem Cucurull Arantxa Casanova Adriana Romero Pietro Liò and Yoshua Bengio 2018.
[36] Chen D, Xie Z, Liu R, Yu W, Hu Q, Li X, Ding SX. Bayesian hierarchical graph neural networks with uncertainty feedback for trustworthy fault diagnosis of industrial processes. IEEE Trans Neural Netw Learn Syst 2023.
[37] 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 2023.
[38] Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(11).
[39] Boutet A, Kermarrec A-M, Mittal N, Taïani F. Being prepared in a sparse world: the case of KNN graph construction. In: 2016 IEEE 32nd international conference on data engineering. ICDE, IEEE; 2016, p. 241–52.