Chinese Journal of Chemical Engineering 70 (2024) 20–32

Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

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

Full Length Article

Causal temporal graph attention network for fault diagnosis of chemical processes

Jiaojiao Luo [1], Zhehao Jin [1], Heping Jin [2], Qian Li [2], Xu Ji [1], Yiyang Dai [1,*]

[1] School of Chemical Engineering, Sichuan University, Chengdu 610065, China
[2] China Three Gorges Corporation, Beijing 100038, China

ARTICLE INFO

Article history:
Received 3 November 2023
Received in revised form
23 January 2024
Accepted 24 January 2024
Available online 6 March 2024

Keywords:
Chemical processes
Safety
Fault diagnosis
Causal discovery
Attention mechanism
Explainability

ABSTRACT

Fault detection and diagnosis (FDD) plays a significant role in ensuring the safety and stability of chemical processes. With the development of artificial intelligence (AI) and big data technologies, data-driven approaches with excellent performance are widely used for FDD in chemical processes. However, improved predictive accuracy has often been achieved through increased model complexity, which turns models into black-box methods and causes uncertainty regarding their decisions. In this study, a causal temporal graph attention network (CTGAN) is proposed for fault diagnosis of chemical processes. A chemical causal graph is built by causal inference to represent the propagation path of faults. The attention mechanism and chemical causal graph were combined to help us notice the key variables relating to fault fluctuations. Experiments in the Tennessee Eastman (TE) process and the green ammonia (GA) process showed that CTGAN achieved high performance and good explainability.

1. Introduction

With the rapid development of technology, the modern chemical industry has a good opportunity to move in a large-scale and highly automated direction [1]. Chemical unit operations become more tightly coupled to each other, which will make it easy for a tiny fault to develop into a disastrous accident at a chemical plant. In these cases, safety management is still faced with challenges from increasingly complex chemical processes. Fault detection and diagnosis (FDD), which can identify the abnormal state in time, is the key technology to ensure the safe and stable operation of chemical processes [2,3]. In the Industry 4.0 era, more and more researchers have been focusing on FDD methods.

FDD methods can be classified into three categories: model-based methods, knowledge-based methods, and data-based methods. In the model-based methods, a mathematical model of chemical processes was established based on the first principles [4,5]. The knowledge-based methods depend on expert knowledge of fault states and require a deep understanding of the chemical process [4–6]. In both methods, FDD models are established based on an insight into chemical processes, which have a huge time cost and a high labor cost [7]. It is difficult to build complex fault diagnosis models for large-scale chemical processes. These methods may be preferable to experts with engineering experience. With the rapid development of AI and big data technology, the number of data-based methods that model different chemical processes using the same approach, has increased considerably.

The data-driven methods depend on data, which can be derived from the simulation data, the on-site data, or the generated data by transfer learning. And they require minimal process knowledge [8], which can extract effective information from input data containing multidimensional features. With the increasing of data acquisition capacity, the quantity and quality of data will be improved constantly. This data-driven method can be divided into statistics-based methods and machine learning methods [8,9]. Statistics-based methods include partial least square (PLS) [10,11], principal component analysis (PCA) [11–14], independent component analysis (ICA) [15,16], linear discriminant analysis (LDA) [17], and Fisher discriminant analysis (FDA) [11,17,18]. These methods are proposed to project the original data into a lower-dimensional feature space. Based on this feature space, the key information can be extracted more easily [19]. However, statistics-based methods that focus only on the problems of dimensionality curse combined with

* Corresponding author. Tel.: +86 180 1060 5143.
E-mail address: daiyy@scu.edu.cn (Y. Dai).

https://doi.org/10.1016/j.cjche.2024.01.019
1004-9541/

classification methods to detect and diagnose faults should be the next step.

After feature extraction, different fault diagnosis methods, primarily traditional machine learning methods, are proposed to find the root causes of faults. These methods are called classifiers. Classifiers based on machine learning include support vector machines (SVM) [20–22], k-nearest neighbor (KNN) [23], random forest (RF) [24], Bayesian networks (BN) [19,25,26], artificial neural networks (ANN), artificial immune system (AIS) [27,28], and their variants. Traditional machine learning methods, which can model complex systems using the same simple framework, have nice advantages in the big data era. However, these models with shallow structures have a serious disadvantage, which is their low capability to learn the inherent rules of complex chemical processes [29,30]. With the development of industrial sensors and the increase of industrial data, the demand for data mining outstrips the capability of machine learning models.

Compared with traditional machine learning methods, deep learning methods can extract data features more effectively. In recent years, more and more researchers have been focusing on deep learning-based FDD. Wu *et al.* [31] proposed a fault diagnosis method based on convolutional neural network (CNN), which can extract features in both spatial and temporal domains. Zhang *et al.* [32] used a bidirectional recurrent neural network (BiRNN) to construct FDD models with sophisticated recurrent neural network (RNN) cells. The BiRNN architecture can learn long-term dependencies in chemical process data and achieve a higher fault diagnosis rate (FDR). Wei *et al.* [29] proposed a modified transformer model called the target transformer, which can capture long-term fluctuations using a target-attention mechanism and detect faults earlier. There are other deep learning methods as well, including one-dimensional CNN [20], the hybrid method integrating CNN and BiRNN [33], the attention-based long short-term memory (LSTM) [34], the attention-based CNN [35], and so on. With the development of the computing capabilities of computers, deep learning models have become more complex, and FDR is getting higher and higher. However, this surge in performance has often been achieved through increased model complexity, which turns such methods into "black box" approaches and causes the uncertainty that deep learning models bring to decisions [36].

There is a clear trade-off between the performance of a deep learning model and its ability to produce explainable and interpretable predictions. This need for trustworthy, fair, robust, and high-performing models for industrial applications led to the revival of the field of explainable AI. Wei *et al.* [29] used attention mechanisms to improve the interpretability and performance of deep learning models. Wu *et al.* [37] proposed a FDD structure integrating chemical process topology as extra knowledge, called process topology convolutional network, which uses centrality encoding and spatial encoding. This method can also be combined with the self-attention mechanism to have higher interpretability [38], and it is important for its application in high-risk scenarios. However, the complex chemical process has a large number of nonlinear subsystems interacting with each other in a complex way, and it is not enough to represent the fault propagation information using physical topology between chemical unit operations.

To improve the transparency of model prediction, a novel fault diagnosis method is proposed in this study, with a new structure of causal temporal graph attention network (CTGAN) integrating causal discovery with attention mechanisms. The major contributions of this work are summarized as follows:

(1) A novel model, CTGAN, a causality-aided graph attention network, is proposed for fault diagnosis of chemical processes.
(2) A chemical causal graph is built by causal inference to represent the propagation path of faults.
(3) attention mechanism and chemical causal graph were combined to achieve high interpretability.

The rest of the study is organized as follows: Section 2 introduces the fundamental theories about causal inference and graph attention networks. Section 3 introduces the FDD framework based on CTGAN. Section 4 presents the experimental results of fault diagnoses on the TE process and the GA process. Finally, Section 5 provides a summary of this study.

2. Methods

2.1. Causal discovery by linear non-Gaussian acyclic model

In chemical processes, unit operations and streams are physically connected with pipes. Control loops link up the process variables and manipulated variables with sensors, controllers, actuators, etc. To fully utilize the information on topology, it is more accurate to represent the data of chemical processes by a graph. Different from the graph in process topology convolutional network (PTCN) [37], causal diagrams model the exact dependency between different variables rather than only describing the physical relationships between variables.

A causal diagram is a causal structure representing the causal relationships between the observed variables. Causal relationships have a greater advantage than statistical associations, where the former can drive a knowledge discovery process to discover the essence lying behind the phenomenon [39]. The linear non-gaussian acyclic model (LiNGAM) [40] is a well-known causal discovery method widely used to identify causal relationships between the observed variables from data.

For the temporal causal discovery problem shown in Fig. 1, the observational data with 4 continuous time series variables can be defined by $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4\} \in \boldsymbol{R}^{4\times T}$, and the goal is to discover the causality between all variables and to model a temporal causal graph. In chemical processes, the $\boldsymbol{x}_i \in \boldsymbol{R}^T$ may represent a variable such as temperature, pressure, and so on. Here, T is the length of observations for each variable. The causal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is learned from observational multivariate time-series data $\boldsymbol{x}$ by the LiNGAM methods, where the nodes $v_i \in \mathcal{V}$ represents a time-series variable $\boldsymbol{x}_i$ and each directed edge $e_{ij} = (v_i, v_j) \in \mathcal{E}$ from node v_i to v_j denotes a causal relationship that variable $\boldsymbol{x}_i$ causes an effect in $\boldsymbol{x}_j$. The neighborhood of a node v_i is defined as $\mathcal{N}(v_i) = \{u \in \mathcal{V} | (v_i, u) \in \mathcal{E}\}$. The degree of a node v_i is the number of directed edges in which it participates. The adjacency matrix $\boldsymbol{A}$ is a $n \times n$ matrix with $\boldsymbol{A}_{ij} = 1$ if $e_{ij} \in \mathcal{E}$ and $\boldsymbol{A}_{ij} = 0$ is $e_{ij} \notin \mathcal{E}$. In this case, $n = 4$.

Given the observed time-series data $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4\}$, LiNGAM assumes that the observed variables $\boldsymbol{x}_i$ can be arranged in a causal order where no later variable causes any earlier variable, and the causal relationships between variables are linear and can be defined as:

Fig. 1. A causal graph was learned from multivariate observational time series data.

$$\boldsymbol{x}_i = \sum_{k(i)>k(j)} b_{ij}\boldsymbol{x}_j + e_i + c_i \tag{1}$$

Here e_i represents a non-Gaussian noise term that is independent of each other. And c_i is an optional constant; b_{ij} is the coefficient expressing causal strength. In this way, the nonlinear chemical systems are simplified as linear causal relationships with noise, which can facilitate modeling and accelerate calculation. The results of experiments by Parida *et al.* [41] show that the basic model of LiNGAM can be used to analyze nonlinear data for effective error minimization and to reduce noises.

The LiNGAM model in Fig. 1 can be written as:

$$\boldsymbol{x}_1 = e_1 \tag{2}$$

$$\boldsymbol{x}_2 = b_{12}\boldsymbol{x}_1 + e_2 \tag{3}$$

$$\boldsymbol{x}_3 = b_{23}\boldsymbol{x}_2 + e_3 \tag{4}$$

$$\boldsymbol{x}_4 = b_{24}\boldsymbol{x}_2 + b_{34}\boldsymbol{x}_3 + e_4 \tag{5}$$

These linear equations can be represented in matrix form as:

$$\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\\ \boldsymbol{x}_3\\ \boldsymbol{x}_4\end{bmatrix} = \begin{bmatrix}0 & 0 & 0 & 0\\ b_{12} & 0 & 0 & 0\\ 0 & b_{23} & 0 & 0\\ 0 & b_{24} & b_{34} & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\\ \boldsymbol{x}_3\\ \boldsymbol{x}_4\end{bmatrix} + \begin{bmatrix}e_1\\ e_2\\ e_3\\ e_4\end{bmatrix} \tag{6}$$

Eq. (6) is the expansion of $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{B}$ is a matrix that could be permuted to strict lower triangularity if the causal ordering $k(i)$ of features is known. Solve for $\boldsymbol{x}$ as follows:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{e} \tag{7}$$

$$\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{B})^{-1} \tag{8}$$

Here, $\boldsymbol{A}$ could also be permuted to a lower triangularity. Finally, applying the independent component analysis (ICA) algorithm to decompose the matrix $\boldsymbol{x}$, we can obtain the LiNGAM matrix $\boldsymbol{B}$, which describes the causal relationships among the features.

2.2. Graph attention network

As structured data in non-Euclidean space, graphs are irregular in that a graph may have a variable size of unordered nodes and a node may have a different number of neighbors. Each node in graphs is related to others by links of various types, such as citations, friendships, and interactions. The irregularity of graphs resulted in some important operations (*e.g.* convolutions) being easy to compute in the image domain but difficult to apply to the graph domain. A graph neural network (GNN) is a type of deep learning model that can process data in the graph domain or non-Euclidean space and has developed by leaps and bounds in recent years.

The GNN model tends to learn a featured embedding that is composed of nodes' neighbor information. For a graph classification problem, the individual graph $\mathcal{G}$ is defined by its node feature $\boldsymbol{X}$ and adjacency matrix $\boldsymbol{A}$. The GNN targets to predict the labels of the graph. Mathematically, it can be represented as:

$$\boldsymbol{h}_v = f(\boldsymbol{X}, \boldsymbol{A}) \tag{9}$$

$$\boldsymbol{z}_v = \text{flatten}(\boldsymbol{h}_v) \tag{10}$$

$$\overline{\boldsymbol{y}} = g(\boldsymbol{z}_v) \tag{11}$$

The f function defines the transition function that transforms the input vector into a d dimensional space. The hidden state matrix $\boldsymbol{h}_v$ is unstacked as a vector $\boldsymbol{z}_v$ by the flatten function. Then the vector $\boldsymbol{z}_v$ will be the input of the next classifier, g. There is an important segment, called message passing or neighborhood aggregation, to get a particular solution for $\boldsymbol{h}_v$. Many types of message-passing mechanisms are proposed to adapt to different application scenarios. As an important variant of GNN models, the graph convolutional network (GCN) aggregates messages from the neighborhood and itself. Its message-passing mechanism can be expressed as:

$$\boldsymbol{H}^{l+1} = \text{RELU}\left(\tilde{\boldsymbol{D}}^{-1/2}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-1/2}\boldsymbol{H}^l\boldsymbol{W}^l\right) \tag{12}$$

$$\text{RELU}(\boldsymbol{x}) = \max(0, \boldsymbol{x}) \tag{13}$$

$$\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I} \tag{14}$$

Here, $\boldsymbol{H}^l$ is the feature matrix in the l^{th} layer. $\boldsymbol{I}$ is the identity matrix, and $\tilde{\boldsymbol{A}}$ is the adjacency matrix of the graph $\mathcal{G}$ with added self-connections. $\tilde{\boldsymbol{D}}$ is the degree matrix of $\tilde{\boldsymbol{A}}$. GCN uses this mechanism to achieve normalized information aggregation. After multiple such layers, each node feature vector will contain information about its neighbors and further away nodes.

Unlike GCN's message-passing mechanism, the graph attention network (GAN) [42] assumes that there are differences in the contributions of neighbor nodes to the central node. GAN adopts attention mechanisms to adaptively learn the weight of importance for each neighbor node. The attention-based message passing of GAN is defined as:

$$\boldsymbol{h}_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)\cup i}\alpha_{i,j}\boldsymbol{W}\boldsymbol{h}_j^l\right) \tag{15}$$

$$\alpha_{i,j} = \text{softmax}\left(a\left(\boldsymbol{W}\boldsymbol{h}_i^l, \boldsymbol{W}\boldsymbol{h}_j^l\right)\right) \tag{16}$$

In Eq. (15), $\boldsymbol{h}_i$ is the feature vector of the central node, σ is an activation function, $\alpha_{i,j}$ is the attention weight between the node i and its neighbor j, $\boldsymbol{W}$ is a shared weight matrix for linear transformation to every node, and a is a shared attention mechanism. GCN learns dynamic neighbor features containing weighted sharing properties to achieve higher expression abilities.

3. Causal Temporal Graph Attention Network Based Fault Diagnosis Method

Because of the increasing requirements of fairness, accountability, and transparency, the CTGAN is proposed to achieve both high performance and good explainability. The structure of CTGAN is shown in Fig. 2.

The historical time-series data of chemical processes can be transformed into directed acyclic graphs (DAG) with causality by the LiNGAM model, where each node represents a specific variable and each edge denotes a causal relationship between two variables. Based on the high-order Markov process (HOMP) [43] that the current system state depends on the multiple-steps-before states, the data were segmented by a sliding window with size t into

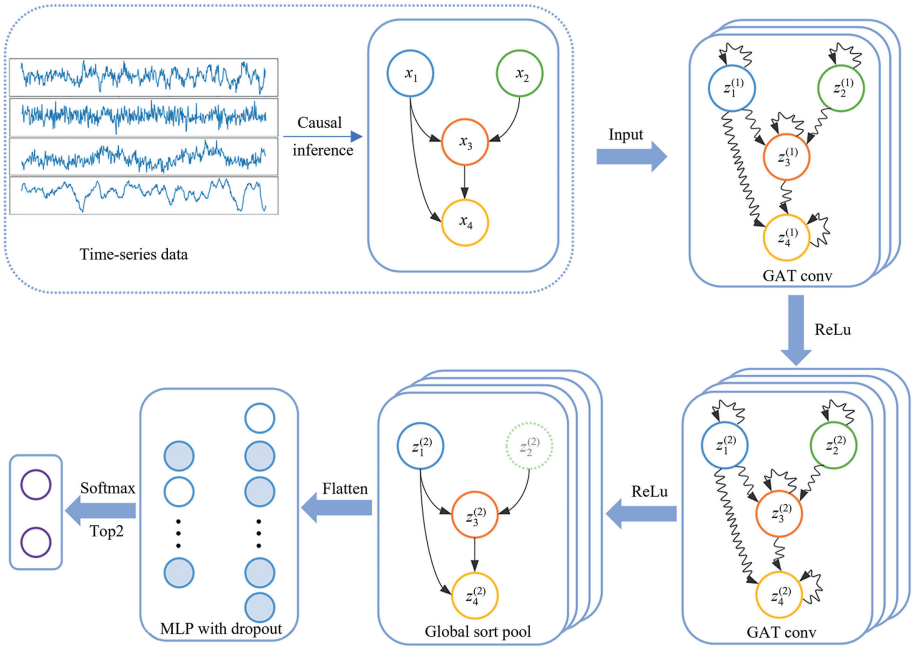


Fig. 2. The structure of CTGAN.

matrixes possessing two dimensions: time and variables. Then the dataset can be represented as $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_m\}$, $\boldsymbol{X}_i \in \boldsymbol{R}^{n\times t}$, where n is the number of observed variables and t is the length of the time window. A causal graph as an input sample is defined as $\boldsymbol{X}_i^{\mathrm{G}} = (\boldsymbol{X}_i, \boldsymbol{A})$, where $\boldsymbol{A}$ is the adjacency matrix of the process causal graph. The graph data integrates more information about the process than the pure data matrix.

To transform the input data into high-dimensional features with higher expressive power, two graph attentional layers are required. As an initial step, each node undergoes a linear transformation with a shared weight matrix $\boldsymbol{W}$. The attention coefficient e_{ij} is calculated using the transformed node features:

$$e_{ij} = a(\boldsymbol{W}\boldsymbol{h}_i, \boldsymbol{W}\boldsymbol{h}_j) \tag{17}$$

Here, e_{ij} indicates the importance of node j to node i. $\boldsymbol{h}_i$ is the hidden feature of node i, which can be initialized to a time-series variable $\boldsymbol{x}_i$. The attention mechanism a is a single-layer feedforward neural network that applies the nonlinear activation ReLU function, and its parameters can be represented by α:

$$e_{ij} = \mathrm{ReLU}(\alpha[\boldsymbol{W}\boldsymbol{h}_i \| \boldsymbol{W}\boldsymbol{h}_j]) \tag{18}$$

$$\alpha_{ij} = \mathrm{softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k\in N_i}\exp(e_{ik})} \tag{19}$$

Here, || represents the concatenation operation. To make the importance of different nodes more intuitive, the softmax function is used here to normalize the attention coefficients, obtaining the weight coefficients α_{ij} of neighboring nodes. This is the process of calculating the attention coefficients in the graph attention layer. After obtaining the weight coefficients of the neighboring nodes of node i, the information of the neighboring nodes and itself is weighted and aggregated:

$$\boldsymbol{h}_i^{'} = \mathrm{ReLU}\left(\sum_{j\in\mathcal{N}_i\cup i}\alpha_{ij}\boldsymbol{W}\boldsymbol{h}_j\right) \tag{20}$$

The weight coefficients are used to compute a linear combination of the features corresponding to them as the final output feature for every node.

After being calculated by two graph attention layers, the hidden features enter a global sorting pooling layer. The purpose of this pooling layer is to select the k most important nodes through a sorting strategy in order to achieve data downsampling and reduce overfitting. The data from the pooling layer will be flattened into one-dimensional data and input into a fully connected layer with a random dropout. Finally, the output of the fully connected layer will be normalized by the softmax function to obtain a confidence score representing whether it is a certain fault.

The CTGAN-based framework consisting of an offline stage and an online stage for fault diagnosis is shown in Fig. 3. Its diagnosis procedures are described in Table 1.

4. Case Study

4.1. Tennessee Eastman process

4.1.1. Dataset

The Tennessee Eastman process is often used as a benchmark for chemical process fault diagnosis. It is a complex chemical process that involves multiple steps such as esterification, hydrogenation, separation, and regeneration of ethylene glycol. The TE process is shown in Fig. 4.

The experimental data in this study were generated based on the simulator proposed by Bathelt *et al.* [44]. For this experiment, 20 fault types listed in Table 2 were selected for simulation.

In the simulation, a total of 50 variables were considered, comprising nine process manipulation variables, 22 continuous process measurements, and 19 composition analysis measurements. The sampling period was set to 3 min (20 samples per

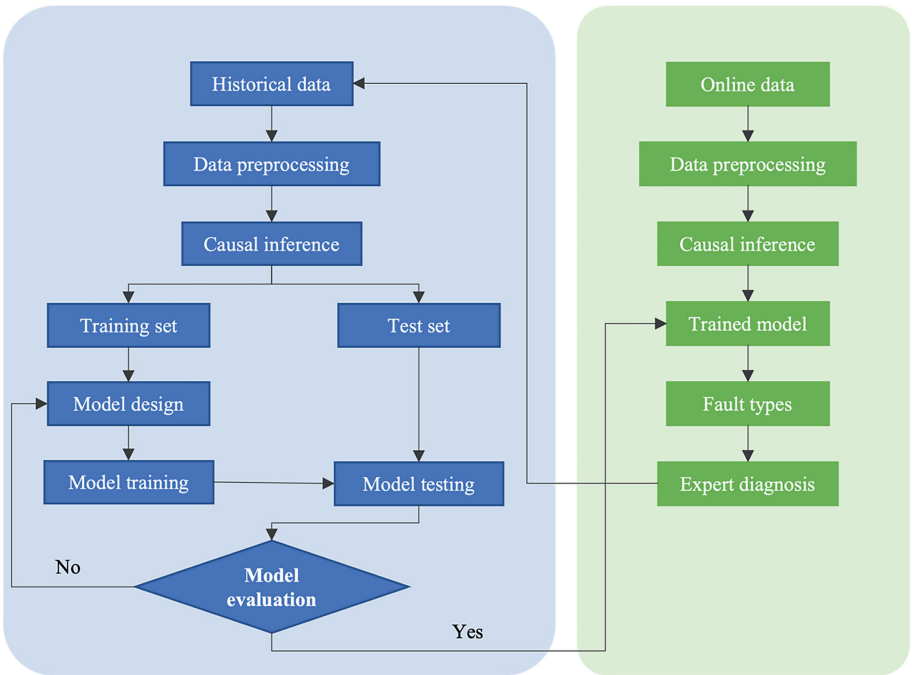


Fig. 3. The CTGAN-based framework.

hour). The normal state simulation was performed for 500 h to obtain 10000 normal state samples. Each fault condition was simulated for 40 h after introducing a fault disturbance at a stable status. The initial condition of the system was changed to repeat each fault simulation 10 times, thus obtaining 8000 samples per abnormal state. However, for the fault six simulation, the simulation was terminated after 7 h because reactor pressure exceeded the limit. 80% of all data under each condition was randomly selected as the training set, and the remaining 20% of samples as the testing set. The simulation time and sample sizes are shown in Table 3. The number of samples was similar across the 21 simulations, mainly due to ensuring that the proposed method is trained on a balanced dataset in the offline stage.

4.1.2. Result and discussion

The causal graph of the TE process calculated by the LiNGAM algorithm is shown in Fig. 5. In graph theory, the node is the basic unit of a graph, and the degree is an important attribute of the node because the degree distribution of a graph can represent the importance of nodes. The degree of each node in Fig. 5 is displayed in Table 4.

The degrees of nodes 1, 11, 18, and 21 are all greater than 7, which means the neighbors of these nodes are abundant. Node 1 represents the A feed in stream 1. In the TE process, A is an important reactant that participates in the generation of all products and by-products. Node 11 is the temperature of the product separator. In a production process, there may be many factors affecting the temperature of the product separator. Node 18 represents the temperature of the stripper. Once a fault occurs in the process, the temperature of the stripper will produce significant fluctuations. Node 21 is the reactor cooling water outlet temperature. This variable is also easily disturbed in the TE process. These nodes play pivotal roles in the message passing of CTGAN.

Some nodes' degrees are equal to 0, such as nodes 2, 3, 4, 8, 14, and 15, which indicates that these nodes have no obvious causal

Table 1
Algorithm of fault diagnosis based on the CTGAN

Offline Stage:
Step 1: Historical data (X, Y) of chemical processes is collected.
Step 2: Data preprocessing. Based on the HOMP, the time-series data is transformed into $n \times t$ matrices.
Step 3: A causal graph is established by the LiNGAM method, and the matrices are represented as graphs ($\boldsymbol{X}^{\mathrm{G}}, \boldsymbol{Y}$).
Step 4: Divide data into train set ($\boldsymbol{X}^{\mathrm{G}}_{\mathrm{train}}, \boldsymbol{Y}_{\mathrm{train}}$) and test set ($\boldsymbol{X}^{\mathrm{G}}_{\mathrm{test}}, \boldsymbol{Y}_{\mathrm{test}}$).
Repeat:
Step 5: The CTGAN is designed for chemical processes.
Step 6: The CTGAN is trained with train set ($\boldsymbol{X}^{\mathrm{G}}_{\mathrm{train}}, \boldsymbol{Y}_{\mathrm{train}}$).
Step 7: The CTGAN is tested with test set ($\boldsymbol{X}^{\mathrm{G}}_{\mathrm{test}}, \boldsymbol{Y}_{\mathrm{test}}$) and judged.
Until:
The prediction on test set is satisfactory.
Online Stage:
Step 1: Collect online data $\boldsymbol{X}$, which may consist of some new faults that are not included in historical data.
Step 2: Data preprocessing is applied to transform online data into a graph $\boldsymbol{X}^{\mathrm{G}}$.
Step 3: Online samples are inputted into the trained CTGAN, which generates a prediction $\boldsymbol{Y}_{\mathrm{pred}}$ for each data.
Step 4: Evaluate and determine the type of fault by experts.
Step 5: The new data ($\boldsymbol{X}, \boldsymbol{Y}_{\mathrm{true}}$) will be used to retrain or fine-tuning the model in the offline stage.

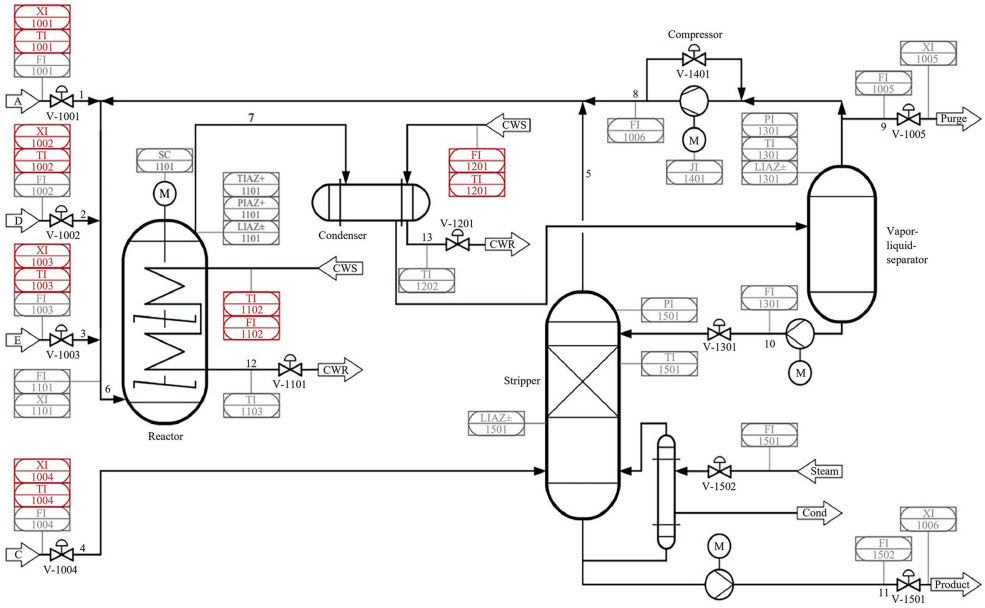


Fig. 4. The TE process.

relationship with others. In the message-passing mechanism of the graph attention layer, the information of neighboring nodes is aggregated in each iteration. These nodes with a degree of 0 are isolated, which means that these nodes cannot learn additional effective features from other nodes, and their information also cannot be passed to others. In this study, isolated nodes with their represented variables are removed from the causal graph. On the one hand, the fluctuations of variables represented by isolated nodes in a causal graph have no effect on other nodes, and these variables are not caused by others. However, once the fault occurs in a chemical process, a chain reaction will occur between variables. Thus, isolated nodes can be seen as redundant variables unrelated to the fault diagnosis task. On the other hand, the elimination of redundant variables can reduce the difficulty for models to extract valid information from data and improve the efficiency of the model training process.

After data processing and causal discovery, the CTGAN is trained, where the batch size is set to 128, the number of epochs is set to 10, and the learning rate is equal to 0.0001. The categorical cross-entropy loss function is used to train the model, and the Adam optimizer is used to accelerate the CTGAN's convergence. After the model training, the testing dataset is inputted into the proposed model to test the fault diagnosis effect. The confusion matrix with details of the diagnosis results is shown in Fig. 6, where the horizontal axis represents the predicted labels and the vertical axis represents the true labels. The color depth of the background is related to the proportion of the predicted samples in the total sample size. Most of the samples are distributed on the diagonal of the confusion matrix, indicating that most labels predicted by the CTGAN match the true labels. However, there is still difficulty in classifying faults 15 and 16. In future work, in-depth data exploration can be carried out to identify the root causes of the difficulty in classifying these two faults.

Table 2
Fault types in the TE process

Fault number	Fault description	Variation
1	A/C feeding ratio, B component unchanged (flow 4)	Step
2	B component, A/C feeding ratio unchanged (flow 4)	Step
3	Feed temperature of D (flow 2)	Step
4	Inlet temperature of reactor cooling water	Step
5	Condenser cooling water inlet temperature	Step
6	Feed loss of A (flow 1)	Step
7	Resistance loss of C (flow 4)	Step
8	Feed components of A, B, C (flow 4)	Random variables
9	Feed temperature of D (flow 2)	Random variables
10	Feed temperature of C (flow 2)	Random variables
11	Inlet temperature of reactor cooling water	Random variables
12	Condenser cooling water inlet temperature	Random variables
13	Reaction dynamics	Slow drift
14	Reactor cooling water valve	Stick
15	Condenser cooling water valve	Stick
16	Unknown	Unknown
17	Unknown	Unknown
18	Unknown	Unknown
19	Unknown	Unknown
20	Unknown	Unknown

Table 3
Simulation time and sample sizes

Status	Train simulation time/h	Test simulation time/h	Train sample size	Test sample size
IDV 01–05, 07–20	40 × 10 × 19 × 0.8	40 × 10 × 19 × 0.2	121600	30400
IDV 06	7 × 10 × 0.8	7 × 10 × 0.2	1120	280
Total	6536	1634	130720	32680

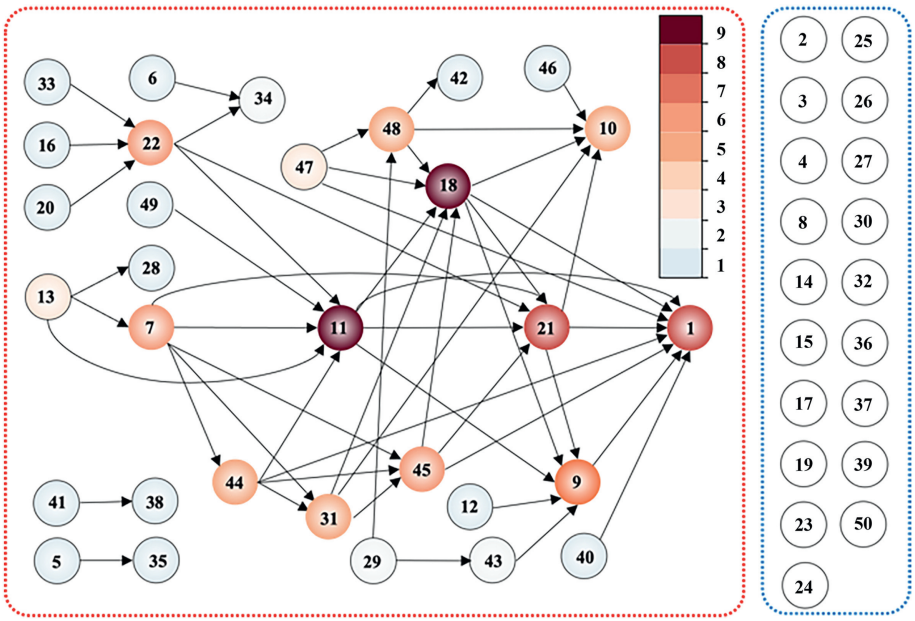


Fig. 5. Causal graph of the TE process.

Table 4
The degree distribution of TE causal graph

Node	Degree	Node	Degree	Node	Degree	Node	Degree	Node	Degree
1	8	11	9	21	8	31	5	41	1
2	0	12	1	22	6	32	0	42	1
3	0	13	3	23	0	33	1	43	2
4	0	14	0	24	0	34	2	44	5
5	1	15	0	25	0	35	1	45	6
6	1	16	1	26	0	36	0	46	1
7	6	17	0	27	0	37	0	47	3
8	0	18	9	28	1	38	1	48	5
9	6	19	0	29	2	39	0	49	1
10	5	20	1	30	0	40	1	50	0

Other deep learning models were trained using the same dataset and hyperparameters. The fault diagnosis rates (FDRs) and false positive rates (FPRs) are compared with the CTGAN, as shown in Table 5. The proposed method achieves the best results in detecting 20 types of faults, where the FDRs of three faults are greater than 99%. Moreover, the FDR for faults 15 and 16 is significantly higher than the other three deep learning methods.

Assume that experts can accurately diagnose which fault it is. If the proposed method provides two possible faults, then the fault diagnosis rate will be greatly improved. In the actual production process, it requires manual intervention to solve faults that occur during the runtime. Therefore, the CTGAN was set to return two possible types of faults. The accuracy in this case is called the Topk accuracy, where k represents the output number of fault types. The confusion matrix of Top2 accuracy in Fig. 7 shows that the expert-assisted model achieves outstanding performance in fault diagnosis tasks.

The proposed method performs well in the TE process, which is mainly due to causal discovery and attention mechanisms. The time-series data in the TE process is restructured as causal graphs by the LiNGAM method, which integrates the topology of chemical processes occurring in faults into the data. Moreover, a causal graph without isolated nodes can be seen as a feature selection method, which removes redundant variables from the data and benefits model training. The dimension-reduction algorithm t-distributed Stochastic Neighbor Embedding (*t*-SNE) [45] is used to visualize the raw data and the causal data, as shown in Fig. 8.

Randomly selecting 1000 samples from 21 types of faults, the t-SNE algorithm is used to transform the 20 × 50 sample matrixes into vectors of length 2. In Fig. 8, each point represents a sample,

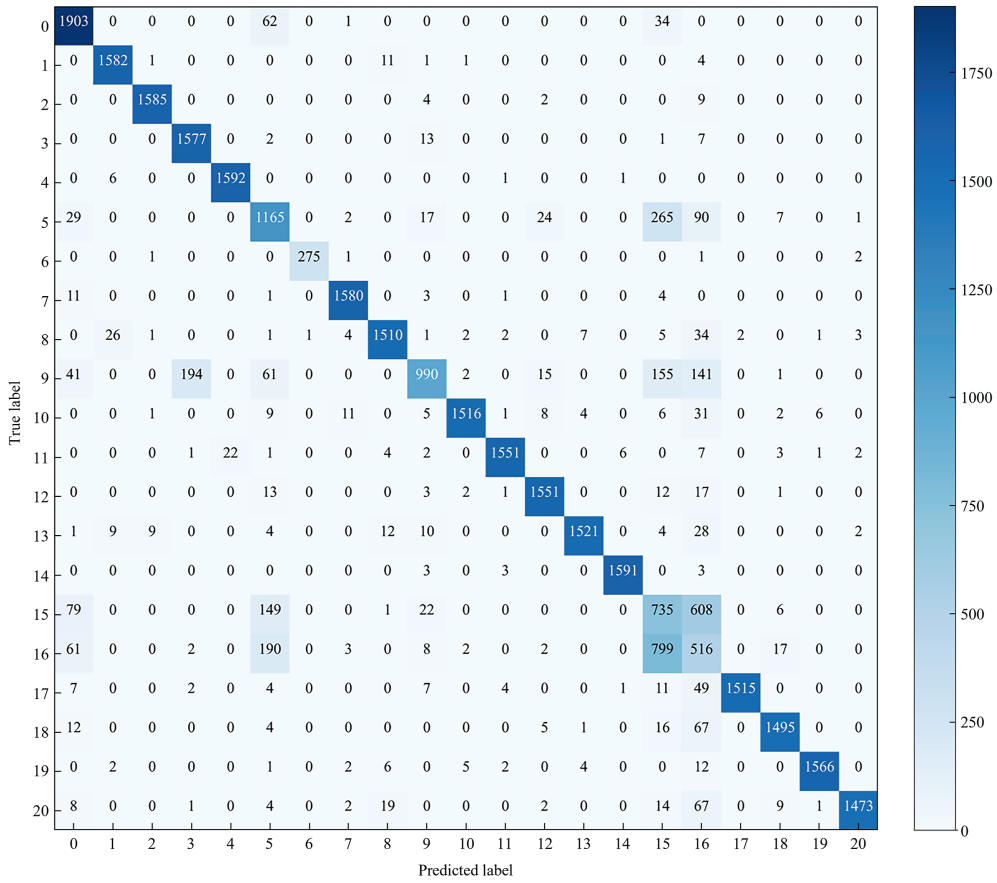


Fig. 6. The confusion matrix of the proposed method in TE process.

Table 5
The FDR and FPR in the TE process

	CTGAN		Causal GCN		CNN		LSTM		Transformer	
	FDR	FPR	FDR	FPR	FDR	FPR	FDR	FPR	FDR	FPR
Fault 0	0.966	0.005	0.883	0.003	0.988	0.018	0.94	0.018	0.967	0.006
Fault 1	0.989	0.001	0.991	0.0	0.992	0.002	0.996	0.001	0.971	0.002
Fault 2	0.986	0.0	0.990	0.0	0.983	0.0	0.986	0.0	0.986	0.0
Fault 3	0.977	0.008	0.951	0.004	0.879	0.01	0.685	0.003	0.636	0.006
Fault 4	0.992	0.001	0.997	0.0	0.976	0.0	0.999	0.0	0.999	0.001
Fault 5	0.902	0.002	0.898	0.002	0.892	0.005	0.897	0.01	0.755	0.001
Fault 6	0.982	0.0	0.957	0.0	0.971	0.0	0.989	0.0	0.996	0.0
Fault 7	0.999	0.0	1.0	0.0	0.998	0.0	1.0	0.0	1.0	0.0
Fault 8	0.949	0.001	0.960	0.002	0.934	0.001	0.956	0.001	0.916	0.002
Fault 9	0.601	0.004	0.767	0.023	0.559	0.029	0.61	0.024	0.326	0.011
Fault 10	0.936	0.0	0.968	0.001	0.966	0.0	0.969	0.0	0.973	0.0
Fault 11	0.981	0.001	0.976	0.0	0.984	0.001	0.984	0.0	0.964	0.0
Fault 12	0.976	0.002	0.979	0.003	0.941	0.001	0.946	0.001	0.974	0.004
Fault 13	0.966	0.001	0.939	0.001	0.951	0.0	0.946	0.0	0.963	0.0
Fault 14	0.992	0.0	0.997	0.0	0.985	0.0	0.988	0.0	0.991	0.0
Fault 15	0.471	0.035	0.393	0.032	0.3	0.033	0.385	0.035	0.393	0.059
Fault 16	0.461	0.042	0.439	0.036	0.318	0.031	0.387	0.036	0.459	0.059
Fault 17	0.949	0.0	0.947	0.0	0.942	0.0	0.948	0.0	0.942	0.0
Fault 18	0.936	0.004	0.921	0.0	0.937	0.001	0.938	0.001	0.935	0.001
Fault 19	0.978	0.001	0.989	0.001	0.984	0.0	0.987	0.0	0.986	0.0
Fault 20	0.934	0.0	0.931	0.0	0.931	0.0	0.934	0.0	0.929	0.0
Average	0.901	0.005	0.898	0.005	0.877	0.006	0.880	0.006	0.8610	0.007

and the different fault types of samples are represented by different colors in the graph. It can be observed that the samples will cluster after reducing dimension by t-SNE. Theoretically, the samples with more obvious clustering effects could be recognized and classified more easily. In the original data, most of the samples were mixed and not separated. The clustering performance of the data was greatly improved after the removal of isolated nodes by causal discovery.

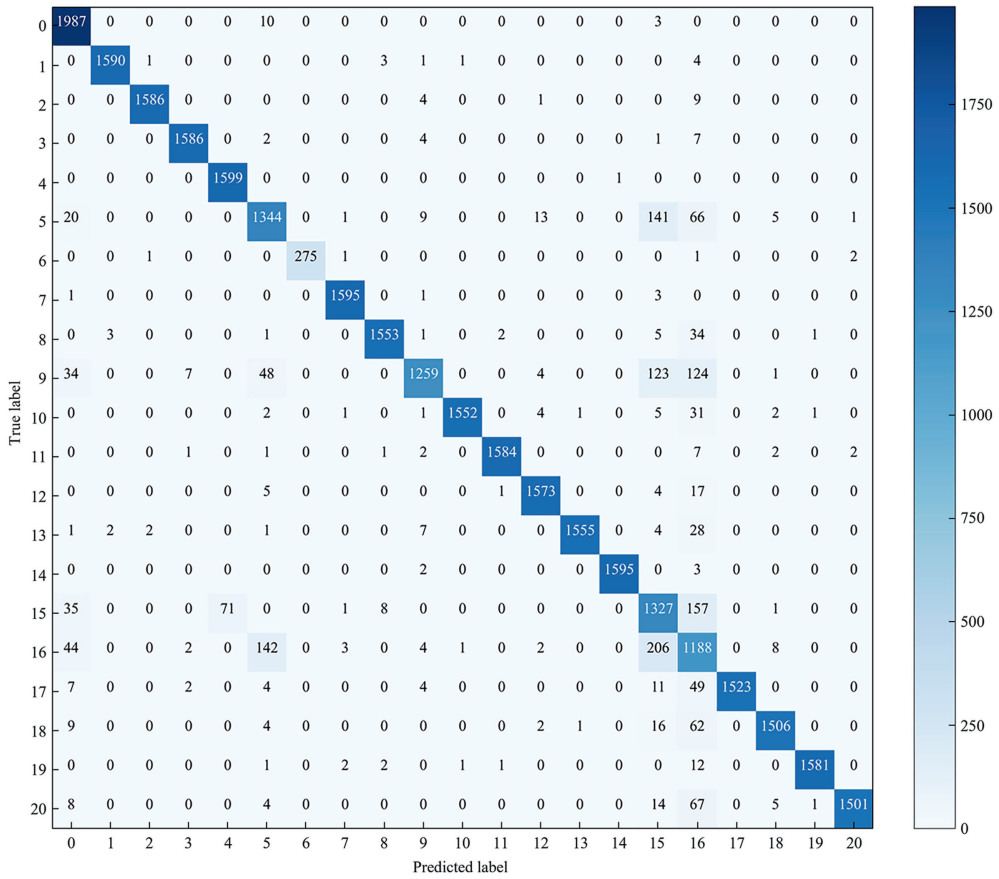
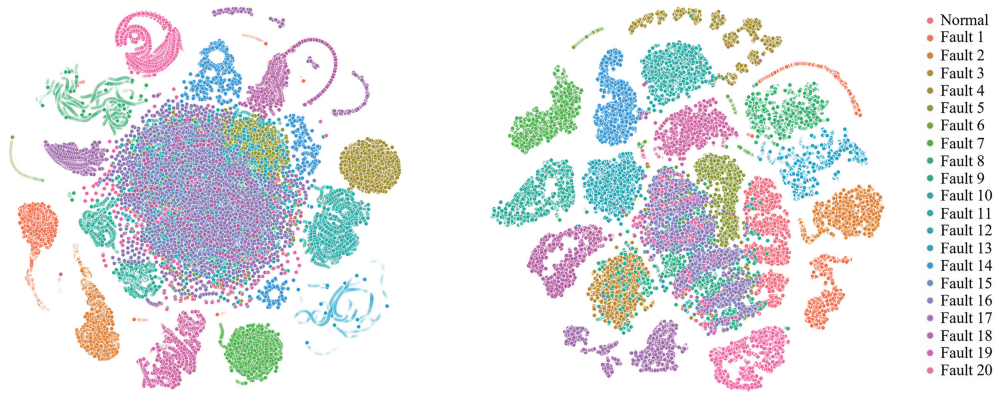


Fig. 7. Top two confusion matrix in the TE process.



(a) Origin data (b) Causal data

Fig. 8. The t-SNE algorithm for origin data and causal data in the TE process.

The attention coefficients of attention layers visualized in Fig. 9 show the weights for edges in the causal graph during message passing, which indicates the importance of different edges or nodes. In each subgraph, the *x*-axis represents the sequence numbers of edges, and the *y*-axis represents the attention weights of different batches. The figure shows the relative values of the attention weights, where the darker the color of the grid, the greater the attention weight. It can be observed that the model has different attention weights for different edges in the causal graph. The edges 1 (4 → 35), 13 (17 → 13), 16 (17 → 27), 24 (20 → 48), 31 (30 → 47), and 40 (49 → 23) have large attention weights in the first attention layer, indicating that the aggregation of information between these nodes is more important to the CTGAN. The same goes for the second attention layer. In chemical processes, it is required to pay more attention to the fluctuations of these variables. The attention mechanism enhances the aggregation of effective information while reducing the interference of unimportant nodes

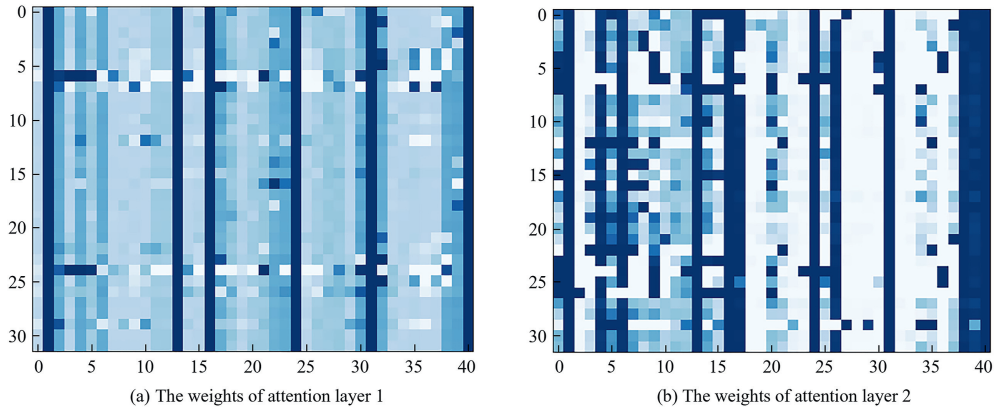


Fig. 9. The attention coefficients of CTGAN trained by TE dataset.

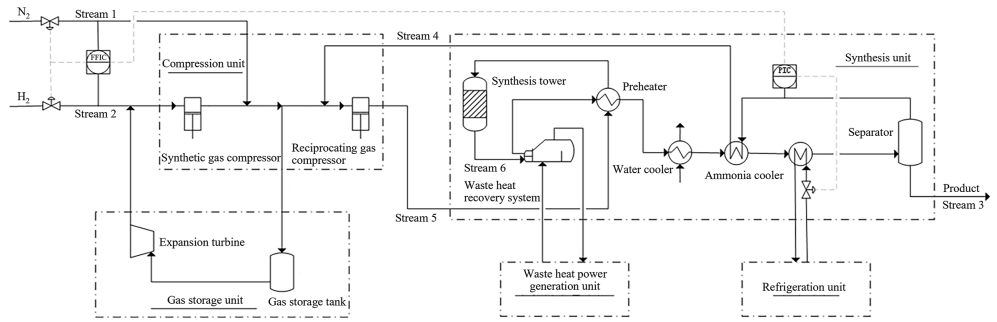


Fig. 10. The GA process.

in the entire flow of information. This targeted information propagation gradually enhances the dissimilarity of node features, thereby avoiding excessive smoothing caused by similar deep node features. The attention mechanism on the graph is more conducive to the fault diagnosis of chemical processes.

4.2. Green ammonia process

4.2.1. Dataset

The Green ammonia process is a process for preparing ammonia using renewable energy sources, such as wind and solar. It involves many different pieces of equipment, such as electrolytic cells, hydrogen storage tanks, compressors, *etc.* Each step and piece of equipment in the process have the potential to encounter faults, which can result in interruptions and production declines in the process and even cause damage to the equipment. The production of the GA process is influenced by the supply of upstream energy, which requires dynamic load adjustments to be conducted in response to the fluctuations of energy. Faults in the GA process are easily caused by frequent load switching. Therefore, by promptly identifying and resolving faults through FDD methods, production interruptions can be

Table 6
Fault types in GA process

Fault number	Fault description	Variation
1	Flow detection delay	Random variables
2	Controller fault	Step
3	Level display fault	Step
4	Components of H_2, N_2	Random variables

Table 7
The variables in GA process

Variables number	Variables description
1	H_2 feed (stream 1)
2	Temperature of stream 1
3	Pressure of stream 1
4	N_2 feed (stream 2)
5	Temperature of stream 2
6	Pressure of stream 2
7	NH_3 feed (stream 3)
8	Product separator pressure
9	Product separator temperature
10	Product separator level
11	Synthetic gas compressor work
12	Reciprocating gas compressor work
13	Synthesis tower temperature
14	Synthesis tower pressure
15	Stream 4 component H_2
16	Stream 4 component N_2
17	Stream 4 component NH_3
18	Temperature of stream 4
19	Pressure of stream 4
20	Stream 5 component H_2
21	Stream 5 component N_2
22	Temperature of stream 5
23	Pressure of stream 5
24	Stream 6 component H_2
25	Stream 6 component N_2
26	Stream 6 component NH_3
27	Temperature of stream 6
28	Pressure of stream 6

Table 8
The sample sizes of simulations for GA process

Status	Train sample size	Test sample size
Normal	8000	2000
IDV 01-04	$1000 \times 8 \times 4$	$1000 \times 2 \times 4$
Total	40000	10000

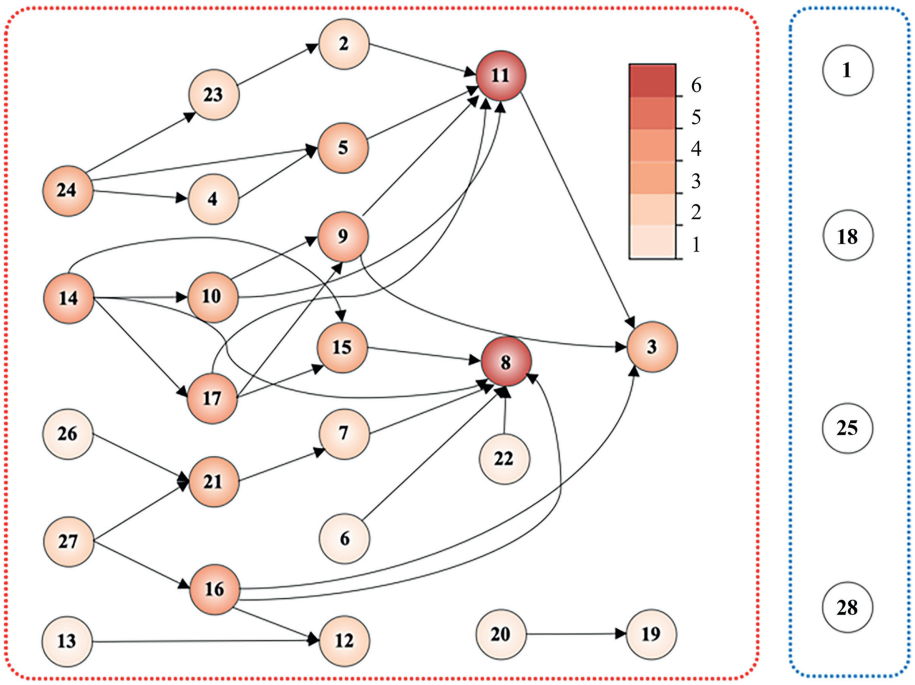


Fig. 11. The causal graph for the GA process.

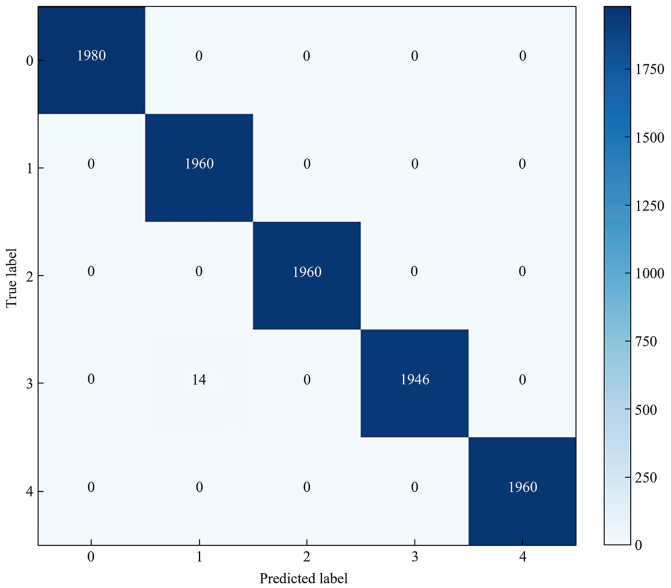


Fig. 12. The confusion matrix of the proposed method in the GA process.

minimized to the greatest extent, and both production efficiency and quality can be improved while reducing unnecessary losses. The fault data were obtained through simulation in UniSim, and the process flowchart is shown as Fig. 10. For this experiment, four types of faults listed in Table 6 were selected for simulation. The collected 28 variables are listed in Table 7.

The sampling period is set to 1 min, which means 60 samples per hour. The normal state simulation was performed to obtain 10000 normal samples. For fault simulations, a normal state simulation ran for 2 h to ensure a stable status and then introduced fault disturbances. Each fault simulation was repeated 10 times in different initial conditions of the system to obtain 10000 abnormal samples. 80% of all data under each condition were randomly selected as the training set, and the remaining 20% of samples as the testing set. The sample sizes of the simulations are shown in Table 8.

4.2.2. Result and discussion

The CTGAN was applied to the GA process dataset, and the model exhibited excellent performance. The causal directed acyclic

Table 9
The FDR and FPR in the TE process

	CTGAN		Causal GCN		CNN		LSTM		Transformer	
	FDR	FPR	FDR	FPR	FDR	FPR	FDR	FPR	FDR	FPR
Fault 0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0
Fault 1	1.0	0.0	1.0	0.091	0.284	0.060	0.0	0.0	1.0	0.249
Fault 2	1.0	0.002	1.0	0.001	1.0	0.0	1.0	0.0	1.0	0.0
Fault 3	0.993	0.0	0.637	0.001	0.761	0.178	1.0	0.252	0.0	0.0
Fault 4	1.0	0.0	0.992	0.0	1.0	0.0	1.0	0.0	1.0	0.0
Average	0.999	0.0004	0.9258	0.019	0.809	0.047	0.800	0.050	0.800	0.050

graph is shown in Fig. 11. In this causal graph, it is observed that some causal relationships that align with our understanding, such as node 8 representing the pressure of the product separator, are influenced by various upstream factors.

After embedding time-series data into the causal graph without isolated nodes, the CTGAN is trained. The performance of the trained model on the testing set is shown in Fig. 12. For normal samples and fault samples in the GA process, the proposed method achieves a fault diagnosis rate of 99%, whichis higher than other deep learning models shown in Table 9.

5. Conclusions

A causality-assisted graph attention network called the causal temporal graph attention network is proposed for complex chemical processes in this study. The attention mechanism and chemical causal graph were combined to achieve high interpretability and wide applicability for different chemical processes. A directed acyclic graph with causality is extracted from the historical data by the causal discovery algorithm LiNGAM, which can infer the information transportation chain when faults occur. Messages in series-time data are passed and aggregated by two graph attention layers in the CTGAN, and the important information is further extracted by the global sorting pooling. The proposed FDD method achieved the optimal FDR and FPR results among all the methods evaluated by the TE process and the GA process. The attention weights combined with the causal graph can help us notice the key variables relating to fault fluctuations. There is a trade-off between the performance and interpretability of deep learning models, but both are important for a practical application. Exploring ways to enhance interpretability while the performance is not declining will be an important and promising direction in future research.

Declaration of 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.

Acknowledgements

The authors are grateful for the support of the National Key Research and Development Program of China (2021YFB4000505).

References

[1] P.D. Christofides, J.F. Davis, N.H. El-Farra, D. Clark, K.R.D. Harris, J.N. Gipson, Smart plant operations: vision, progress and challenges, *AIChE J.* 53 (11) (2007) 2734–2741.
[2] S. Yin, S.X. Ding, X.C. Xie, H. Luo, A review on basic data-driven approaches for industrial process monitoring, *IEEE Trans. Ind. Electron.* 61 (11) (2014) 6418–6428.
[3] S. Yin, Z.H. Huang, Performance monitoring for vehicle suspension system *via* fuzzy positivistic C-means clustering based on accelerometer measurements, *IEEE ASME Trans. Mechatron.* 20 (5) (2015) 2613–2620.
[4] C. Nan, F. Khan, M.T. Iqbal, Real-time fault diagnosis using knowledge-based expert system, *Process. Saf. Environ. Prot.* 86 (1) (2008) 55–71.
[5] P.M. Frank, Analytical and qualitative model-based fault diagnosis – a survey and some new results, *Eur. J. Contr.* 2 (1) (1996) 6–28.
[6] D. Peng, Y. Xu, Q.X. Zhu, Study and application of case-based extension fault diagnosis for chemical process, *Chin. J. Chem. Eng.* 21 (4) (2013) 366–375.
[7] X.Y. Chen, X.F. Yan, Fault diagnosis in chemical process based on self-organizing map integrated with Fisher discriminant analysis, *Chin. J. Chem. Eng.* 21 (4) (2013) 382–387.
[8] M. Alauddin, F. Khan, S. Imtiaz, S. Ahmed, A bibliometric review and analysis of data-driven fault detection and diagnosis methods for process systems, *Ind. Eng. Chem. Res.* 57 (32) (2018) 10719–10735.
[9] N. Md Nor, C.R. Che Hassan, M.A. Hussain, A review of data-driven fault detection and diagnosis methods: applications in chemical process systems, *Rev. Chem. Eng.* 36 (4) (2020) 513–553.
[10] J.V. Kresta, J.F. MacGregor, T.E. Marlin, Multivariate statistical monitoring of process operating performance, *Can. J. Chem. Eng.* 69 (1) (1991) 35–47.
[11] L.H. Chiang, E.L. Russell, R.D. Braatz, Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis, *Chemom. Intell. Lab. Syst.* 50 (2) (2000) 243–252.
[12] B. Xiao, Y.G. Li, B. Sun, C.H. Yang, K.K. Huang, H.Q. Zhu, Decentralized PCA modeling based on relevance and redundancy variable selection and its application to large-scale dynamic process monitoring, *Process. Saf. Environ. Prot.* 151 (2021) 85–100.
[13] J.X. Zhang, W.J. Luo, Y.Y. Dai, Y.M. Yao, Cycle temporal algorithm-based multivariate statistical methods for fault diagnosis in chemical processes, *Chin. J. Chem. Eng.* 47 (2022) 54–70.
[14] W.D. Tian, Y.J. Ren, Y.X. Dong, S.G. Wang, L.Z. Bu, Fault monitoring based on mutual information feature engineering modeling in chemical process, *Chin. J. Chem. Eng.* 27 (10) (2019) 2491–2497.
[15] J.M. Lee, S.J. Qin, I.B. Lee, Fault detection of non-linear processes using kernel independent component analysis, *Can. J. Chem. Eng.* 85 (4) (2007) 526–536.
[16] L.F. Cai, X.M. Tian, S. Chen, A process monitoring method based on noisy independent component analysis, *Neurocomputing* 127 (2014) 231–246.
[17] C.H. Zhao, F.R. Gao, A nested-loop Fisher discriminant analysis algorithm, *Chemom. Intell. Lab. Syst.* 146 (2015) 396–406.
[18] L.Y. Jiang, L. Xie, S.Q. Wang, Fault diagnosis for batch processes by improved multi-model Fisher discriminant analysis, *Chin. J. Chem. Eng.* 14 (3) (2006) 343–348.
[19] G. Yang, X.H. Gu, Fault diagnosis of complex chemical processes based on enhanced naive Bayesian method, *IEEE Trans. Instrum. Meas.* 69 (7) (2020) 4649–4658.
[20] M. Onel, C.A. Kieslich, E.N. Pistikopoulos, A nonlinear support vector machine-based feature selection approach for fault detection and diagnosis: application to the Tennessee Eastman process, *AIChE J.* 65 (3) (2019) 992–1005.
[21] S. Mahadevan, S.L. Shah, Fault detection and diagnosis in process data using one-class support vector machines, *J. Process. Contr.* 19 (10) (2009) 1627–1639.
[22] Y. Mao, Z. Xia, Z. Yin, Y.X. Sun, Z. Wan, Fault diagnosis based on fuzzy support vector machine with parameter tuning and feature selection, *Chin. J. Chem. Eng.* 15 (2) (2007) 233–239.
[23] G.Z. Wang, J.C. Liu, Y. Li, Fault diagnosis using kNN reconstruction on MRI variables, *J. Chemom.* 29 (7) (2015) 399–410.
[24] C. Li, R.V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, R.E. Vásquez, Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals, *Mech. Syst. Signal Process.* 76–77 (2016) 283–293.
[25] W.K. Sun, A.R.C. Paiva, P. Xu, A. Sundaram, R.D. Braatz, Fault detection and identification using Bayesian recurrent neural networks, *Comput. Chem. Eng.* 141 (2020) 106991.
[26] C.Y. Lou, X.S. Li, M.A. Atoui, Bayesian network based on an adaptive threshold scheme for fault detection and classification, *Ind. Eng. Chem. Res.* 59 (34) (2020) 15155–15164.
[27] Y.Y. Dai, J.S. Zhao, Fault diagnosis of batch chemical processes using a dynamic time warping (DTW)-based artificial immune system, *Ind. Eng. Chem. Res.* 50 (8) (2011) 4534–4544.
[28] L. Ming, J.S. Zhao, Feature selection for chemical process fault diagnosis by artificial immune systems, *Chin. J. Chem. Eng.* 26 (8) (2018) 1599–1604.
[29] Z.C. Wei, X. Ji, L. Zhou, Y.G. Dang, Y.Y. Dai, A novel deep learning model based on target transformer for fault diagnosis of chemical process, *Process. Saf. Environ. Prot.* 167 (2022) 480–492.
[30] L. Deng, Y. Zhang, Y.Y. Dai, X. Ji, L. Zhou, Y.G. Dang, Integrating feature optimization using a dynamic convolutional neural network for chemical process supervised fault classification, *Process. Saf. Environ. Prot.* 155 (2021) 473–485.
[31] H. Wu, J.S. Zhao, Deep convolutional neural network model based chemical process fault diagnosis, *Comput. Chem. Eng.* 115 (2018) 185–197.
[32] S.Y. Zhang, K.X. Bi, T. Qiu, Bidirectional recurrent neural network-based chemical process fault diagnosis, *Ind. Eng. Chem. Res.* 59 (2) (2020) 824–834.
[33] A. Kopbayev, F. Khan, M. Yang, S.Z. Halim, Gas leakage detection using spatial and temporal neural network model, *Process. Saf. Environ. Prot.* 160 (2022) 968–975.
[34] L. Zhang, Z.H. Song, Q.H. Zhang, Z.P. Peng, Generalized transformer in fault diagnosis of Tennessee Eastman process, *Neural Comput. Appl.* 34 (11) (2022) 8575–8585.
[35] S.W. Xiong, L. Zhou, Y.Y. Dai, X. Ji, Attention-based long short-term memory fully convolutional network for chemical process fault diagnosis, *Chin. J. Chem. Eng.* 56 (2023) 1–14.
[36] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, A.I. Explainable, A review of machine learning interpretability methods, *Entropy* 23 (1) (2020) 18.
[37] D.Y. Wu, J.S. Zhao, Process topology convolutional network model for chemical process fault diagnosis, *Process. Saf. Environ. Prot.* 150 (2021) 93–109.

[38] D.Y. Wu, X.T. Bi, J.S. Zhao, ProTopormer: toward understandable fault diagnosis combining process topology for chemical processes, *Ind. Eng. Chem. Res.* 62 (21) (2023) 8350–8361.
[39] M. Vuković, S. Thalmann, Causal discovery in manufacturing: a structured literature review, *J. Manuf. Mater. Process.* 6 (1) (2022) 10.
[40] S. Shimizu, P.O. Hoyer, A. Hyvärinen, A. Kerminen, A linear non-Gaussian acyclic model for causal discovery, *J. Mach. Learn. Res.* 7 (2006) 2003–2030.
[41] P.K. Parida, T. Marwala, S. Chakraverty, Altered-LiNGAM (ALiNGAM) for solving nonlinear causal models when data is nonlinear and noisy, *Commun. Nonlinear Sci. Numer. Simul.* 52 (2017) 190–202.
[42] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, In: International Conference on Learning Representations, Vancouver, BC, Canada, 2018.
[43] J.B. Yu, Machine health prognostics using the Bayesian-inference-based probabilistic indication and high-order particle filtering framework, *J. Sound Vib.* 358 (2015) 97–110.
[44] A. Bathelt, N.L. Ricker, M. Jelali, Revision of the Tennessee Eastman process model, *IFAC-PapersOnLine* 48 (8) (2015) 309–314.
[45] L. Maaten, G.E. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* 9 (2008) 2579–2605.