Neural Networks 161 (2023) 397–417

Contents lists available at ScienceDirect

Neural Networks

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

MPGE and RootRank: A sufficient root cause characterization and quantification framework for industrial process faults

Pengyu Song [a], Chunhui Zhao [a,*], Biao Huang [b]

[a] State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, China
[b] Department of Chemical and Materials Engineering, University of Alberta, Edmonton, AB T6G 2G6, Canada

ARTICLE INFO

Article history:
Received 22 April 2022
Received in revised form 18 January 2023
Accepted 23 January 2023
Available online 4 February 2023

Dataset link: https://github.com/chunhuiz/MPGE-RootRank-for-root-cause-diagnosis

Keywords:
Root cause diagnosis
Multi-level predictive graph extraction
RootRank scoring
Hierarchical adjacency pruning
Multi-level Granger causality

ABSTRACT

Root cause diagnosis can locate abnormalities of industrial processes, ensuring production safety and manufacturing efficiency. However, existing root cause diagnosis models only consider pairwise direct causality and ignore the multi-level fault propagation, which may lead to incomplete root cause descriptions and ambiguous root cause candidates. To address the above issue, a novel framework, named multi-level predictive graph extraction (MPGE) and RootRank scoring, is proposed and applied to the root cause diagnosis for industrial processes. In this framework, both direct and indirect Granger causalities are characterized by multi-level predictive relationships to provide a sufficient characterization of root cause variables. First, a predictive graph structure with a sparse constrained adjacency matrix is constructed to describe the information transmission between variables. The information of variables is deeply fused according to the adjacency matrix to consider multi-level fault propagation. Then, a hierarchical adjacency pruning (HAP) mechanism is designed to automatically capture vital predictive relationships through adjacency redistribution. In this way, the multi-level causalities between variables are extracted to fully describe both direct and indirect fault propagation and highlight the root cause. Further, a RootRank scoring algorithm is proposed to analyze the predictive graph and quantify the fault propagation contribution of each variable, thereby giving definite root cause identification results. Three examples are adopted to verify the diagnostic performance of the proposed framework, including a numerical example, the Tennessee Eastman benchmark process, and a real cut-made process of cigarette. Both theoretical analysis and experimental verification show the high interpretability and reliability of the proposed framework.

1. Introduction

With the development of intelligent manufacturing and the progress of big data, data-driven fault detection and diagnosis have been widely studied and implemented in the process industry. The fault detection algorithm is used to capture the abnormal behavior of the process. Subsequently, the diagnosis algorithm is applied to isolate the faulty variables and identify the root cause of the fault. Compared with detection and isolation methods, root cause diagnosis technology can analyze the interaction between faulty variables, determine the fault propagation paths, and unearth the root cause of the fault, further helping engineers eliminate process abnormalities. However, despite the extensive studies of fault detection (Peng, Lu, Kang, & Kai, 2020; Song & Zhao, 2022; Yu, Zhao, & Huang, 2021; Zhao, 2022; Zhao, Chen, & Jing, 2020; Zhao, Lai, & Chen, 2019) and isolation methods (He, Qin, & Wang, 2005; Van den Kerkhof, Vanlaer, Gins, & Van Impe, 2013; Yan, Yao, Huang, & Wong, 2018; Yu & Zhao, 2017), research on root cause diagnosis in industrial processes has just emerged in recent years.

The root cause diagnosis of process faults can be regarded as a causal inference task for multivariate time series, one of the most active research areas in data mining (Runge et al., 2019). The process data collected within a period after the fault occurrence are used to infer the causal relationships between faulty variables so that the fault propagation pathways can be determined. There are three most typical causal inference methods for root cause diagnosis, including Granger causality (GC) (Granger, 1969), transfer entropy (TE) (Schreiber, 2000), and dynamic Bayesian network (DBN) (Murphy, 2002). Compared with TE and DBN, GC methods can be more readily extended to multivariate scenarios and integrated with deep learning frameworks for end-to-end training (Bellot, Branson, & van der Schaar, 2021; Montalto et al., 2015; Tank, Covert, Foti, Shojaie, & Fox, 2021) to better handle multivariate, nonlinear root cause diagnosis tasks. Therefore, GC is mainly focused on in this study. Nonetheless, few studies have been reported on the industrial applications of GC methods based

* Corresponding author.
E-mail address: chhzhao@zju.edu.cn (C. Zhao).

https://doi.org/10.1016/j.neunet.2023.01.030
0893-6080/

Table 1
Comparison of the existing studies and the proposed framework on fault propagation characterization ability and root cause identifiability.

Methods	Techniques	Research gaps
Off-the-shelf causal inference methods: (1) GC (Yuan & Qin, 2014); (2) TE (Duan, Yang, Chen, & Shah, 2013; Duan, Zhao, & Wu, 2022; Lindner, Auret, & Bauer, 2019); (3) DBN (Gharahbagheri, Imtiaz, & Khan, 2017; Yu & Rashid, 2013)	Use single direct causalities to describe fault propagation.	Gap #1: ×
	Build a causal map with pairwise causal relations to observe the root cause variables.	Gap #2: ×
Causal inference + Reachabilitymatrix (Chen & Zhao, 2022; Jiang, Patwardhan, & Shah, 2009)	Approximate multi-level causality using reachability between variables in the causal map composed of direct causal relationships.	Gap #1: ×
	Build a causal map with pairwise causal relations to observe the root cause variables.	Gap #2: ×
Causal inference + Simplified causal map (He, Wang, & Fan, 2019; Liu et al., 2020)	Use single direct causalities to describe fault propagation.	Gap #1: ×
	Enforce the removal of numerous causal relations to simplify the causal map.	Gap #2: ×
Proposed framework	**Multi-level information fusion:** characterize multi-level fault propagation through multiple times of information fusion among variables; **Sparsely constrained adjacency matrix:** guide information fusion and extract salient direct information transfer paths; **Hierarchical adjacency pruning:** automatically select significant multi-level predictive paths to capture multi-level Granger causality.	Gap #1: ✓
	RootRank algorithm: quantify the predictive information amount provided by each variable as the root cause score, and variables with the highest score can be definitely identified as the root causes.	Gap #2: ✓

(Gap #1: whether the method can extract significant multi-level causalities to fully characterize fault propagation;
Gap #2: whether the method can give definite root cause variables from complex causal structures;
✓: the method can address the gap; ×: the method cannot address the gap.)

on deep learning. Chen and Zhao (2022) designed a nonlinear and multivariate GC method based on neural networks. However, this method required an independent network structure for each variable and did not fully extract multivariate interactions.

Notably, the current root cause diagnosis methods are mainly based on causal inference models. Researchers use the causal inference model to extract the causal relationships between faulty variables to construct a causal map to reveal the fault propagation paths. The root cause is generally determined in the following way: if a variable is the cause of other variables but not the effect of any other variable, it is determined to be a root cause. However, such a universal approach brings inevitable shortcomings. First of all, the fundamental goals of causal inference and root cause diagnosis tasks are different. Causal inference aims to determine the direct causal relationship between each pair of variables, whereas root cause diagnosis focuses on identifying root cause variables. Even a causal inference model with high accuracy cannot guarantee root cause identifiability due to the interference of bidirectional and circular causality (Liu et al., 2020). In addition, the information contained in root cause variables may affect multiple downstream variables along fault propagation paths in both direct and indirect ways. Therefore, direct causality is far from enough for root cause identification, and multi-level causality (both direct and indirect causality) should be sufficiently considered. Previous studies (Chen & Zhao, 2022; Jiang et al., 2009) attempted to account for multi-level fault propagation with multi-order reachability matrices of direct causal graphs. However, such approaches are simple extensions based on direct causal relations, which cannot accurately capture multi-level information transfer paths and ensure the significance of indirect causalities. Furthermore, most causal inference models describe causalities in the form of pairwise relationships but lack quantitative representations of the root cause possibility. This makes it difficult to determine the root cause from complex causal structures. Existing studies (He et al., 2019; Liu et al., 2020) designed simplified approaches for causal maps to find root causes from complex pairwise causalities. However, these methods enforce the removal of numerous causal relations, which may cause information loss and reduce diagnosis performance. In short, two research gaps can be identified, including sufficient characterization of multi-level fault propagation and root cause identifiability. We summarize existing studies and their technical responses to these two issues in Table 1. It can be seen that, although some existing methods (Chen & Zhao, 2022; He et al., 2019; Jiang et al., 2009; Liu et al., 2020) have identified these two gaps, these methods still have theoretical flaws and fail to address them well. So far, these gaps have not been well resolved to the best of the authors' knowledge.

To address the above issues, a novel framework for sufficient root cause characterization and quantification, named multi-level predictive graph extraction (MPGE) and RootRank scoring, is proposed and applied to industrial processes. Fig. 1 shows the essential differences between the proposed framework and existing research. As shown in the upper left part of the figure, given variables $\{a, b, c\}$, if there is no direct causality between a and c, then when a causes b, and there is bidirectional causality between b and c, existing studies focusing on direct causalities cannot determine whether the root cause variable is a or both a and c since the indirect effect between a and c is not considered. To solve this problem, fault propagation is modeled as

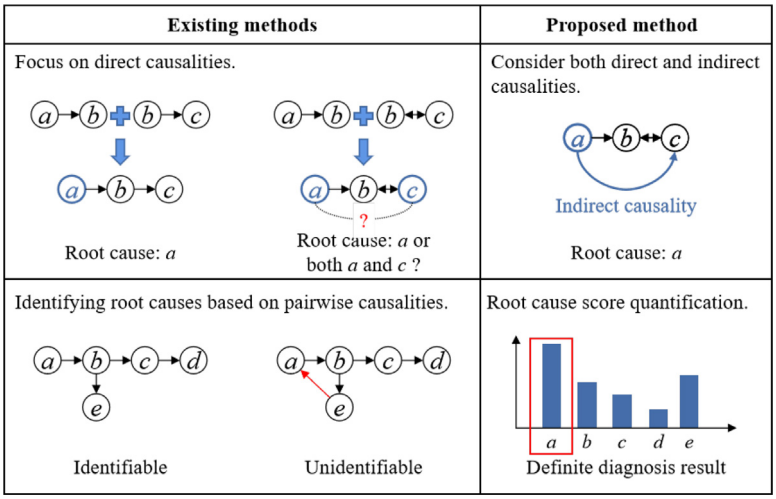

Fig. 1. Comparison of existing root cause diagnosis studies and the proposed method.

multi-level information transmission rather than single direct causality to fully characterize the root cause in this study. In addition, as shown in the lower left part of the figure, the root cause variable cannot be identified by using existing methods based on pairwise causalities when it is affected by the circular causal structure among *a*, *b*, and *e*. To address this issue, we provide definite diagnosis results with the quantitative representations of root cause scores to ensure the identifiability of root cause variables. In the proposed MPGE method, we design a multi-level information fusion (MLIF) module in the multivariate time series forecasting task, where the information of each variable is fused multiple times according to a sparsely constrained adjacency matrix to simulate the fault propagation paths. Additionally, we propose a hierarchical adjacency pruning (HAP) mechanism to automatically select key information transmission paths through adjacency redistribution. Theoretical explanations illustrate that the proposed framework can realize the sparse predictive relationship extraction, thereby characterizing multi-level GC between variables. Moreover, we propose a RootRank scoring algorithm to quantify the root cause score by measuring the predictive information amount provided by each variable. The root causes can be definitely determined as the variables with the highest scores to avoid ambiguous diagnosis results. The solvability of RootRank is mathematically proven to ensure reliability. As shown in Table 1, in comparison with the existing methods, the proposed framework raises effective solutions to the aforementioned two research gaps, providing a sufficient characterization of fault propagation and ensuring the root cause identifiability. The main contributions of this paper are summarized as follows.

(1) We clarify the essential differences between root cause analysis as well as causal inference tasks and point out the weakness of existing root cause diagnosis methods based solely on causal inference models, i.e., the insufficient and inappropriate characterization of root causes using pairwise and direct causalities. For this issue, we propose a novel root cause diagnosis framework considering multi-level fault propagation and providing quantitative root cause characterization.
(2) We propose the MPGE method, in which variables undergo multiple information fusions in the MLIF module to model multi-level fault propagation. Moreover, we design the sparse adjacency constraint and the HAP mechanism to ensure the significance of direct and indirect predictive information transfer paths, thus discovering multi-level Granger causality instead of single direct relations in conventional causal analysis.
(3) A RootRank scoring algorithm is designed to quantify the fault propagation contribution of each variable, thereby giving definite diagnosis results rather than ambiguous root cause candidates. Moreover, the reliability of RootRank is mathematically verified.

The rest of this paper is organized as follows. In Section 2, we review classical causal inference methods and their applications to industrial root cause diagnosis. The preliminary method is briefly revisited in Section 3. Section 4 presents the proposed MPGE method and RootRank scoring algorithm. In Section 5, experimental results are presented to illustrate the performance of the proposed method. Finally, the conclusion is given in Section 6.

2. Related work

In this section, three classical causal inference methods for time series, including GC, TE, and DBN, as well as their variants are revisited. Also, their applications to industrial root cause diagnosis are reviewed.

GC (Granger, 1969) is one of the most common causal inference methods for time series. The basic idea is: given time series **x** and **y**, if the introduction of previous information of **x** can significantly enhance the forecasting accuracy of **y**, then **x** is said to be the Granger cause of **y**. The traditional GC method is limited to the bivariate situation. When dealing with multivariate time series, it needs to determine the causality between each pair of variables, which omits the interactions among multiple variables. In order to solve this problem, conditional Granger (CG) (Geweke, 1984), Lasso Granger (LG) (Arnold, Liu, & Abe, 2007), vector autoregressive Granger (Siggiridou & Kugiumtzis, 2015), and LAPPS penalized sparse Granger (Bore et al., 2020) were successively proposed to consider multivariate effects. Further, GC cannot tackle nonlinearity and nonstationarity due to the use of simple linear autoregressive (AR) models. Kernel-based GC was proposed by Marinazzo, Pellicoro, and Stramaglia (2008), which mapped the original data to a high-dimensional kernel space to cover nonlinearity. Some improved GC methods adopt advanced regression models without linearity assumptions to replace AR, such as GPR-Granger (Chen, Yan, Yao, Huang, & Wong, 2018) and neural network-based Granger (Bellot et al., 2021; Montalto et al., 2015; Tank et al., 2021). Also, time-varying parameters are introduced into GC to tackle the nonstationarity (Raveendran, Huang, & Mitchell, 2020; Schäck, Muma, Feng, Guan, & Zoubir, 2017; Song, Zhao, Huang, & Wu, 2022).

TE (Schreiber, 2000) is a nonlinear temporal causal inference method based on information theory, which describes the direction of information transfer between a pair of time series to determine cause and effect. It has been proven that TE and GC are equivalent for variables obeying the Gaussian distribution (Barnett, Barrett, & Seth, 2009). Staniek and Lehnertz (2008) proposed symbolic transfer entropy (STE) to describe the short-term fluctuations of time series to deal with nonstationarity. Moreover, partial symbolic transfer entropy (PSTE) (Kugiumtzis, 2013) was proposed as a multivariate extension of STE to consider the interaction of multiple variables. However, the performance of TE is significantly affected by the time delay hyperparameter. Recently, Amornbunchornvej, Zheleva, and Berger-Wolf (2021) proposed a variable-lag TE (VLTE) method, which could overcome this problem by aligning the time series.

DBN (Murphy, 2002) is a probabilistic graphical model that describes the statistical dependencies of multivariate time series. In a DBN, each node represents a variable, and the arcs denote the causal relations between variables within the same time slice and across time slices. Constructing a DBN model requires two main steps, including structure learning (Chen, Anantha, & Lin, 2008; Colombo & Maathuis, 2014; Tsamardinos, Brown, & Aliferis, 2006) and parameter learning (Liao & Ji, 2009; Su, Zhang, Ling, & Matwin, 2008). Structure learning focuses on how to build a dependency network between variables. The network parameters can then be learned to represent the probability distribution of variables when historical events are sufficient. The parameters and network structure in DBN are time-invariant, which makes it unable to deal with nonstationarity. Robinson and Hartemink (2008), Robinson, Hartemink, and Ghahramani (2010) proposed a nonstationary DBN to capture the time-varying dependency structure. Besides, the number of parameters of DBN increases exponentially with the number of variables, resulting in substantial computational complexity. Zhang (2015) developed a dynamic uncertain causality graph (DUCG) on the basis of DBN to solve this problem and improved the interpretability of causality inference by introducing logic gate structures.

Some successful applications of the aforementioned causal inference methods have been reported in the process industry (Raveendran & Huang, 2018). Existing root cause diagnosis methods for industrial processes can be divided into linear and nonlinear causal inference methods. The linear causal inference method mainly refers to diagnostic strategies based on traditional GC. Yuan and Qin (2014) used GC to discover the root causes of plant-wide oscillations. Liu et al. (2020) used a maximum spanning tree to simplify the causal inference results of GC, highlighting the root cause variable. He et al. (2019) introduced comparative GC to avoid misleading results, which determined the root cause by comparing the difference in causal relations between variables under normal operating conditions and fault conditions. To address the common nonlinearity in real scenarios, nonlinear causal inference methods, including TE and DBN, have also been applied to root cause diagnosis (Duan et al., 2013, 2022; Gharahbagheri et al., 2017; Lindner et al., 2019; Yu & Rashid, 2013).

3. Revisit of Lasso Granger

As mentioned earlier, GC methods can be extended to neural network frameworks to consider multivariate and nonlinear causal relations in an end-to-end manner. Here, we introduce a linear multivariate GC method, LG (Arnold et al., 2007), as the theoretical basis of our method. LG applies sparse constraints to extract the causality for multivariate processes. Before introducing LG, conventional GC is reviewed first. Given two time series, $\mathbf{x}$ and $\mathbf{y}$, GC believes that if $\mathbf{x}$ is the cause of $\mathbf{y}$, then the past information of $\mathbf{x}$ can assist in predicting $\mathbf{y}$ and vice versa. More specifically, $\mathbf{x}$ is considered the Granger cause of $\mathbf{y}$ if and only if forecasting for $\mathbf{y}$ based on both past values of $\mathbf{y}$ and $\mathbf{x}$ is significantly more accurate than doing so with past values of $\mathbf{y}$ solely. Mathematically, two different autoregressive (AR) models are established, one of which is a bivariate model:

$$y(t) = \sum_{p=1}^{h} a_{1,p} x(t-p) + \sum_{p=1}^{h} a_{2,p} y(t-p) + \varepsilon(t) \tag{1}$$

and the other is the reduced model:

$$y(t) = \sum_{p=1}^{h} b_p y(t-p) + \varepsilon_r(t) \tag{2}$$

where $a_{i,p}$ and b_p are the parameters of the two AR models; ε and ε_r represent the prediction errors of the bivariate model and the reduced model, respectively; $x(t)$ and $y(t)$ are the samplings of $\mathbf{x}$ and $\mathbf{y}$ at the tth time point, respectively; h denotes the time lag.

If the prediction residual of the bivariate model (ε) is significantly less than that of the reduced model (ε_r), it means that the past value of $\mathbf{x}$ contributes to the prediction of $\mathbf{y}$. The following F statistic is constructed:

$$F = \frac{(RSS_0 - RSS_1)/h}{RSS_1/(N-2h-1)} \sim F(h, N-2h-1) \tag{3}$$

where RSS_0 and RSS_1 are the sums of squared residuals of the reduced and the bivariate models, and N is the sample size. The null hypothesis is that the introduction of $\mathbf{x}$ cannot improve the prediction accuracy of $\mathbf{y}$. If the null hypothesis is rejected with a confidence level of α, then $\mathbf{x}$ is the Granger cause of $\mathbf{y}$.

Despite its simplicity, conventional GC cannot well tackle the multivariate time series. For a multivariate time series with J variables, if the GC test is performed for each pair of variables, the total number of operations is as high as $J^2 - J$. Also, the multivariate interactions are ignored. LG method overcomes this disadvantage through Lasso regression. Given a multivariate time series $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_J]$, where $\mathbf{x}_i$ denotes the ith variable, LG builds a prediction model for each variable $\mathbf{x}_i$ relating all variables, and the optimization objective can be written as:

$$\min \left[x_i(t) - \sum_{j=1}^{J} \sum_{p=1}^{h} \omega_{ji,p} x_j(t-p) \right]^2 + \lambda \sum_{i=1}^{J} \sum_{p=1}^{h} \left| \omega_{ji,p} \right| \tag{4}$$

where $\omega_{ji,p}$ denotes the regression coefficient of $x_j(t-p)$ on $x_i(t)$, and λ is the L1 penalty factor.

The L1 penalty method has the effect of sparse variable selection (Tibshirani, 1996), tending to give larger regression coefficients to variables that have significant contributions, whereas the regression coefficients of other variables are shrunk to 0. Hence, if $\omega_{ji,p}$ is equal to zero for each p, it means that $\mathbf{x}_j$ has no prediction contribution to $\mathbf{x}_i$. Thus, $\mathbf{x}_j$ is not a Granger cause of $\mathbf{x}_i$. Conversely, if there exists a p, such that $\omega_{ji,p}$ is not equal to zero, then $\mathbf{x}_j$ "Granger causes" $\mathbf{x}_i$. LG establishes a regression model for each variable rather than for each pair of variables, improving computation efficiency.

4. Methodology

4.1. Motivation

In existing research, root cause diagnosis is generally based on the pairwise causality between variables. Given J variables $\{\mathbf{x}_1, \ldots, \mathbf{x}_J\}$, the main diagnostic steps include:

(1) Infer the direct causal relationship between each pair of variables.

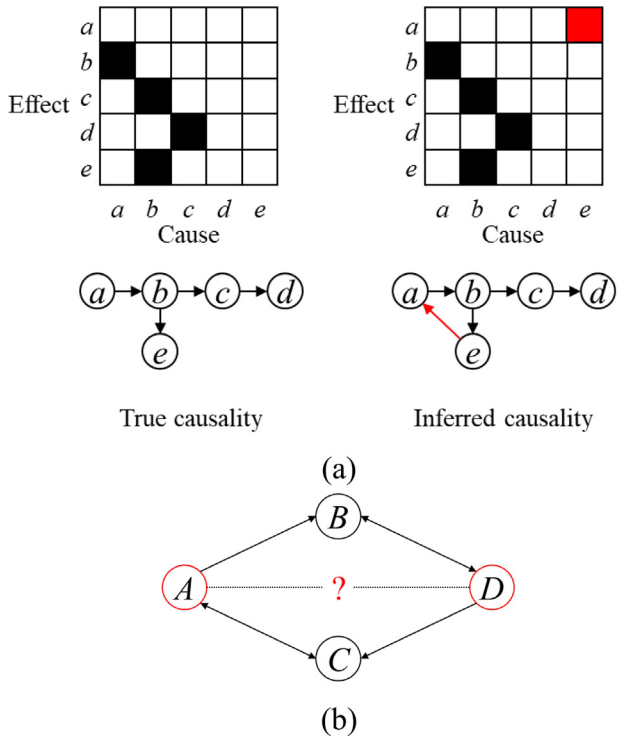

Fig. 2. Two schematic diagrams to illustrate the difference between causal inference and root cause diagnosis. (a) A high-accuracy inference result is compared with true causality, in which the red square and arrow indicate the false causality. (b) A causal map showing the direct causality between four variables from which the root cause cannot be determined. The dashed line represents the unobserved indirect causality.

(2) Combine the obtained causal relationships into a causal matrix, where the element in the ith row and jth column indicates whether $\mathbf{x}_j$ is the cause of $\mathbf{x}_i$.

(3) Find the fault source according to the causal matrix to determine the root cause. In general, if a variable is a cause rather than the effect of any other variable, it is determined to be the root cause.

Here, cause and effect are defined by temporal causality. Specifically, given variables $\mathbf{x}$ and $\mathbf{y}$, if the current and past observations of $\mathbf{x}$ significantly affect the future observations of $\mathbf{y}$, then $\mathbf{x}$ is defined as the cause of $\mathbf{y}$, and $\mathbf{y}$ is the effect of $\mathbf{x}$. In the context of root cause diagnosis, the temporal causality between faulty variables reflects the transmission direction of fault information.

However, causal inference models cannot guarantee the identifiability of the root cause for the following reasons:

(1) The fundamental goals of causal inference and root cause diagnosis tasks are different. Even a causal inference model with high accuracy cannot ensure that the root cause can be identified. Fig. 2(a) shows a comparison between an inference result with high causal inference accuracy (95%) and real causal relations. Here, causal inference accuracy refers to the proportion of correctly inferred causal relationships to all inferred relations. Among the twenty inferred relationships, only one is incorrect. Thus, the causal inference accuracy is calculated as: $(20-1)/20 = 95\%$. Nevertheless, we still cannot determine the root cause from the inference result because of the interfering relationship between e and a.

(2) Causal inference methods only focus on direct causality, but root cause diagnosis also needs to consider indirect effects. The root cause variable is the information source for the entire fault propagation process. All faulty variables will be directly or indirectly affected by the root cause. A reasonable root cause diagnosis model should consider both direct and indirect causality to highlight the location of the root cause. For example, Fig. 2(b) presents a causal map showing the direct causality between four variables. Since there are paths from variables A and D to all other variables, we can regard A and D as root cause candidates. However, the root cause still cannot be determined because of the bidirectional causality. Since there is no direct causality between A and D, the direction of information propagation is unobservable. If we can know the relative strength of the two information flows, $A \to B \to D$ and $D \to C \to A$, then the indirect causal relationship between A and D can be determined to identify the root cause.

(3) The existing methods present causality in causal matrices, which only describe pairwise relationships and cannot directly determine the root cause variable. There are complex couplings and feedback control between variables in industrial processes. Therefore, even the root cause variables may be the effect of other variables. In these cases, the causal matrix cannot give a definite diagnosis result but can only provide ambiguous root cause candidates. Both of the above examples confirm this problem.

This study proposes a novel framework of root cause diagnosis to overcome the shortcomings of traditional causal inference methods. First, we model the multi-level information transmission between variables to consider both the direct and indirect causalities. Second, we design a quantitative approach to describe the causal strength between variables to improve the interpretability of diagnostic results. Moreover, we propose a scoring algorithm to quantify the contribution of each variable to the fault propagation, thereby giving definite root cause identification results rather than causal matrices.

4.2. Multi-level predictive graph extraction

MPGE is a neural network structure designed for end-to-end multi-level predictive relationship extraction. The schematic of MPGE is shown in Fig. 3. The input of MPGE is a multivariate time series with J variables $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_J]^{\mathrm{T}}$, where $\mathbf{x}_i$ denotes the ith variable. Define the forecasting period as T, and variable $\mathbf{x}_i$ can be expressed as $\mathbf{x}_i = [x_i(1), \ldots, x_i(T)]$, where $x_i(t)$ is the sampling of $\mathbf{x}_i$ at the tth time point. First, MPGE uses convolutional neural networks (CNN) (Krizhevsky, Sutskever, & Hinton, 2012) to extract the temporal features of each variable without interactions between different variables. Then MPGE constructs an adjacency matrix with sparse constraints. The feature vectors of variables are deeply fused according to the adjacency matrix to realize multi-level information transmission. Simultaneously, key adjacency relationships are screened layer by layer through the designed HAP mechanism. The forecasted value of each variable is output at the end of the network. The adjacency matrix can capture the predictive relationships between variables to describe multi-level GC by minimizing the forecasting error. Detailed information is presented below.

4.2.1. Temporal feature extraction

In the proposed MPGE, the CNN model is used for temporal feature extraction. CNN can extract nonlinear features and capture temporal correlations.

Before identifying predictive relationships between variables, it is required that there is no interaction between variables in feature extraction. To this end, J CNNs can be used to process all the J variables of the input data, respectively. Notably, such repeated operations can be simplified to reduce complexity. Here we use a two-dimensional CNN with a kernel size of $(1 \times n)$ to extract temporal features from the input multivariate time series $\mathbf{X}$ with a dimension of $(J \times T)$. Fig. 4 shows the operation of the designed CNN structure (taking four variables as an example).

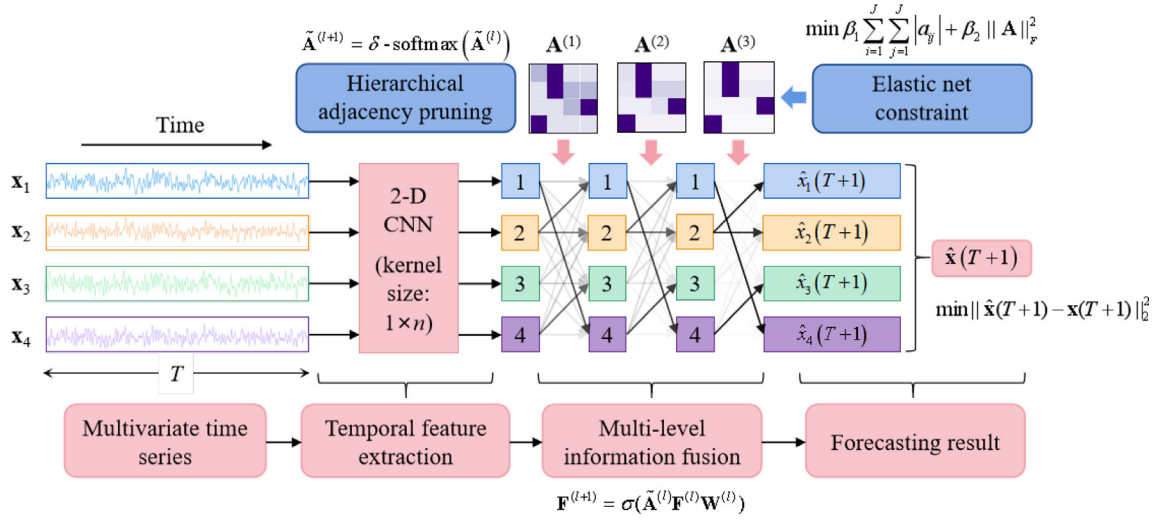

Fig. 3. Schematic of the proposed MPGE. (The input to the model is a multivariate time series with J variables. Here, we take the case of $J = 4$ as an example. Denote the time series of the ith variable as $\mathbf{x}_i$, $i = 1, \ldots, 4$, where $\mathbf{x}_i = [x_i(1), \ldots, x_i(T)]$, $x_i(t)$ is the sampling of $\mathbf{x}_i$ at the tth time point, and T is the forecasting period. The output of the model is the predicted value $\hat{\mathbf{x}}(T+1)$ for the value of the multivariate time series at $T+1$ (denoted by $\mathbf{x}(T+1)$). The temporal features of each variable are extracted by a two-dimensional CNN, whose kernel size is one in the variable dimension, and then input into the multi-level information fusion (MILF) module. The MILF module performs weighted fusions of the features of each variable according to the adjacency matrix $\mathbf{A}$. The gray arrows indicate the direction and intensity of information transfer between variables. A darker grayscale indicates a stronger adjacency relationship. The adjacency matrix $\mathbf{A}$ is constrained by the elastic net to ensure sparsity while preserving useful variable groups (see Section 4.5). Additionally, the HAP mechanism performs the δ-softmax transformation on the $\mathbf{A}$ matrix layer by layer to discard insignificant information transmission paths. These operations aim to extract the critical multi-level (both direct and indirect) predictive relations while minimizing the prediction error, i.e., $\left\|\hat{\mathbf{x}}(T+1) - \mathbf{x}(T+1)\right\|_2^2$).

We set the stride of the convolution kernel to 1 for each moving in both temporal and variable dimensions. In addition, the CNN is designed to have multiple filters to fully extract temporal features. In the temporal dimension, the convolution kernel can be regarded as a time window of length n, which sweeps the time series of each variable to extract temporal features. For the ith variable $\mathbf{x}_i$ in $\mathbf{X}$, the kth filter sweeps through $\mathbf{x}_i$ and calculates the output $\mathbf{y}_{i,k}$ as follows:

$$\mathbf{y}_{i,k} = \sigma(\mathbf{W}_k * \mathbf{x}_i + \mathbf{b}_k), i = 1, 2, \ldots, J \tag{5}$$

where $*$ denotes the convolution operation, $\mathbf{W}_k$ is the convolution kernel parameter matrix, and $\mathbf{b}_k$ is the bias vector.

In the variable dimension, since the size of the convolution kernel in the variable dimension is 1, although the designed CNN is two-dimensional, it can avoid the interactions between variables and achieve independent feature extraction for each variable.

More than one CNN layer can be stacked to extract deep-level features. Finally, the feature vector $\mathbf{f}_i$ of $\mathbf{x}_i$ is extracted as:

$$\mathbf{f}_i = \left[\mathbf{y}_{i,1}, \mathbf{y}_{i,2}, \ldots, \mathbf{y}_{i,P}\right] \tag{6}$$

where $[\cdot]$ means concatenation of vectors, and P is the number of filters.

Notably, the feature extraction of each variable in the above operation is independent and has no interaction with other variables, which is the premise of subsequent causality extraction. The designed CNN structure here can achieve this requirement without repeated modeling for each variable, thus simplifying the operation. In contrast, if a recurrent network such as long short-term memory (Hochreiter & Schmidhuber, 1997) is used, it needs to process the time series of each variable separately, leading to considerable computational complexity. After obtaining the temporal features, combine the features of each variable into a feature matrix $\mathbf{F}$, the ith row of which is the feature vector of $\mathbf{x}_i$.

4.2.2. Multi-level information fusion (MLIF) module

Under the GC framework, the information transmission between variables can be embodied by predictive contributions. The information from the root cause variable is transmitted through multiple steps, directly or indirectly affecting the future value of downstream variables. Inspired by this, multiple information fusion layers are designed in the MLIF module to capture the multi-level causality. In each layer, the information transfer between variables is fused according to a predictive graph $\mathcal{G}$. $\mathcal{G}$ is a directed weighted graph structure, which is an effective representation of the information transfer relations between multiple variables. In $\mathcal{G}$, each node can represent a variable, and the weights and directions of edges between every two nodes represent the directions and strengths of predictive relations between variables. In the proposed MPGE method, the information fusion between variables can be readily guided by the inter-node adjacencies described by $\mathcal{G}$. Through the multi-layer stacking of the information fusion layer, multi-level information transfer between variables can be realized. At the end of the MLIF module, the fused information is used for multivariate time series forecasting. While minimizing the prediction error, MLIF finds the most suitable graph structure $\mathcal{G}$ for the prediction to characterize the multi-level GC.

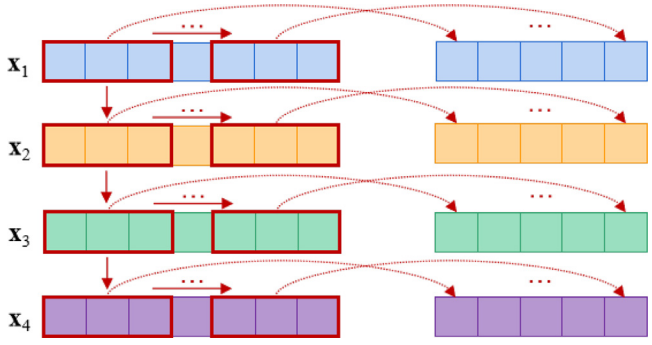

Fig. 4. Schematic of convolution operation in temporal feature extraction (The red box represents the convolution kernel).

Specifically, $\mathcal{G}$ is composed of two parts: the feature matrix $\mathbf{F}$ obtained in the previous step and the adjacency matrix $\mathbf{A}$, where $\mathbf{A} \in \Re^{J\times J}$. The element a_{ij} in the ith row and jth column of matrix $\mathbf{A}$ is a positive number that represents the adjacency strength from the jth node to the ith node. Each node can also be linked to itself to represent self-predictability. For a single information fusion layer, denote its input matrix as $\mathbf{F}_{\text{in}}$, its output $\mathbf{F}_{\text{out}}$ can be calculated as:

$$\mathbf{F}_{\text{out}} = \sigma(\mathbf{D}^{-1}\mathbf{A}\mathbf{F}_{\text{in}}\mathbf{W}) = \sigma(\widetilde{\mathbf{A}}\mathbf{F}_{\text{in}}\mathbf{W}) \tag{7}$$

where $\mathbf{D} = \text{diag}\left\{d_1, d_2, \ldots, d_J\right\}$ is the indegree matrix of $\mathcal{G}$, which is a diagonal matrix, and the ith element of its diagonal d_i $(i = 1, \ldots, J)$ is calculated as follows:

$$d_i = \sum_{j=1}^{J} a_{ij} \tag{8}$$

$\widetilde{\mathbf{A}}$ is the standardized adjacency matrix, $\mathbf{W}$ represents the parameter matrix of the information fusion layer, and $\sigma(\cdot)$ is a nonlinear activation function such as $\text{ReLU}(x) = \max(0, x)$.

The form of the information fusion layer is similar to the graph neural network (GNN) (Wu et al., 2020), which enables adjacent nodes in the graph structure to realize information transfer. The difference is that the inter-node adjacencies in GNN are usually deterministic, whereas the adjacency matrix $\mathbf{A}$ in the proposed method is a learnable structure. According to Eq. (7), for each graph node generated by variable $\mathbf{x}_i$ $(i = 1, \ldots, J)$, its feature output by the information fusion layer $\mathbf{f}_i^{(\text{out})}$ can be calculated as:

$$\mathbf{f}_i^{(\text{out})} = \sigma(\sum_{j=1}^{J} \widetilde{a}_{ij}\mathbf{f}_j\mathbf{W}) \tag{9}$$

where $\widetilde{a}_{ij}$ is the element in the ith row and the jth column of matrix $\widetilde{\mathbf{A}}$, denoting the standardized adjacent weight from the jth node to the ith node.

Notably, for any $i = 1, \ldots, J$, the following equation holds:

$$\sum_{j=1}^{J} \widetilde{a}_{ij} = 1 \tag{10}$$

Hence, for each variable node, the information fusion layer essentially performs a weighted summation on its own feature vector and the features of its adjacent nodes according to matrix $\mathbf{A}$. Element $\widetilde{a}_{ij}$ can be regarded as a valve from variable j to the variable i. The larger its value, the more information of variable j can be transferred to variable i. In other words, $\mathbf{A}$ controls the direct information transmission path between variables.

However, as mentioned earlier, direct information transmission is far from enough for root cause analysis tasks. In order to allow the information of the root cause variable to propagate to the downstream variable through intermediate nodes, multiple information fusion layers are stacked. For the lth information fusion layer, denote its input matrix as $\mathbf{F}^{(l)}$, its output $\mathbf{F}^{(l+1)}$ can be calculated as:

$$\begin{aligned} \mathbf{F}^{(l+1)} &= \sigma\left(\mathbf{D}^{-1}\mathbf{A}\mathbf{F}^{(l)}\mathbf{W}^{(l)}\right) = \sigma(\widetilde{\mathbf{A}}^{(l)}\mathbf{F}^{(l)}\mathbf{W}^{(l)}) \\ \mathbf{F}^{(1)} &= \mathbf{F} \end{aligned} \tag{11}$$

where $\mathbf{W}^{(l)}$ represents the parameter matrix of the lth information fusion layer; $\mathbf{A}^{(l)}$ and $\mathbf{D}^{(l)}$ are the adjacency matrix and the indegree matrix of the graph structure in the lth layer, respectively.

Through multiple information fusion layers, the fused features are acquired for forecasting. Denote the total number of information fusion layers as L, then the predicted value of variable $\mathbf{x}_i$ is calculated as:

$$\hat{\mathbf{x}}_i(T+1) = \mathbf{f}_j^{(L)}\mathbf{w}_i = \sigma(\sum_{j=1}^{J} \widetilde{a}_{ij}^{(L)}\mathbf{f}_j^{(L-1)}\mathbf{W}^{(L)})\mathbf{w}_i \tag{12}$$

where $\mathbf{w}_i$ is the weight vector for predicting the ith variable.

So far, the forecasting result $\hat{\mathbf{x}}(T+1)$ is obtained, which is required to be as close as possible to the actual value $\mathbf{x}(T+1)$. It can be found that the adjacency matrix $\mathbf{A}$ guides the entire feature extraction and forecasting process. If the adjacency from the jth node to the ith node is relatively strong, namely the parameter a_{ij} is relatively large, $\mathbf{x}_j$ will have a significant contribution on the prediction of $\mathbf{x}_i$, and vice versa. As mentioned earlier, a significant predictive relationship represents the existence of GC, which can be used to describe fault propagation. To capture significant predictive relationships, an elastic network (EN) sparsity constraint (Zou & Hastie, 2005) integrating both the L1 and L2 norms is imposed on the adjacency matrix $\mathbf{A}$. Then the overall optimization objective function of MPGE is designed as follows:

$$\min_{\mathbf{A},\mathbf{W}^{(l)}} \mathbb{E}\left\{\| \hat{\mathbf{x}}(T+1) - \mathbf{x}(T+1) \|_2^2\right\} + \beta_1 \sum_{i=1}^{J}\sum_{j=1}^{J} \left|a_{ij}\right| + \beta_2 \| \mathbf{A} \|_F^2 \tag{13}$$

where β_1 and β_2 are the L1 and L2 penalty factors of the EN method, respectively, and $\|\cdot\|_F$ is the Frobenius matrix norm. In general, β_1 is much greater than β_2 to achieve sparsity.

MPGE method extracts crucial predictive relationships while minimizing the forecasting error. EN can play a role in variable selection. Since the EN constraint is added to $\mathbf{A}$, it will tend to assign larger adjacency weights to key variables that are more useful for prediction, ignoring those variables that have no or less effect. In this way, matrix $\mathbf{A}$ will show a certain degree of sparsity (most elements approximate zero). The problem in Eq. (13) can be solved by a gradient descent algorithm such as Adam (Kingma & Ba, 2014) so that the adjacency matrix $\mathbf{A}$ carrying the predictive relationship can be obtained. According to Eqs. (10)–(12), the value of $\widetilde{a}_{ij}$ represents the contribution of variable $\mathbf{x}_j$ to the prediction of $\mathbf{x}_i$ in each information fusion layer. Since Eq. (10) holds for each variable, $\widetilde{a}_{ij}$ can be regarded as a direct prediction contribution rate of $\mathbf{x}_j$ to $\mathbf{x}_i$.

After obtaining the direct predictive relationships described by $\mathbf{A}$, we can derive the multi-level prediction contribution rate between variables through L $(L > 1)$ information fusion layers. Define the multi-level prediction contribution rate of $\mathbf{x}_j$ to $\mathbf{x}_i$ after passing through l $(l < L)$ information fusion layers as $p_{ij}^{(l)}$. The prediction contribution matrix in the lth information fusion layer is defined as $\mathbf{P}^{(l)}$, and the element in the ith row and jth column of $\mathbf{P}^{(l)}$ equals to $p_{ij}^{(l)}$. For any $k = 1, \ldots, J$, the direct prediction contribution rate of $\mathbf{x}_k$ to $\mathbf{x}_i$ variable is $\widetilde{a}_{ik}$. Hence, $p_{ij}^{(l+1)}$ can be calculated as:

$$p_{ij}^{(l+1)} = \sum_{k=1}^{J} \widetilde{a}_{ik}p_{kj}^{(l)} = \left[\widetilde{\mathbf{A}}\mathbf{P}^{(l)}\right]_{ij} \tag{14}$$

where $\left[\widetilde{\mathbf{A}}\mathbf{P}^{(l)}\right]_{ij}$ denotes the element in the ith row and jth column of $\widetilde{\mathbf{A}}\mathbf{P}^{(l)}$.

Eq. (14) can be rewritten in matrix form:

$$\mathbf{P}^{(l+1)} = \widetilde{\mathbf{A}}\mathbf{P}^{(l)} \tag{15}$$

In addition, since $\mathbf{P}^{(1)}$ happens to be equal to $\widetilde{\mathbf{A}}$, the following conclusions can be drawn:

$$\mathbf{P} = \mathbf{P}^{(L)} = \widetilde{\mathbf{A}}\mathbf{P}^{(L-1)} = \widetilde{\mathbf{A}}^2\mathbf{P}^{(L-1)} = \cdots = \widetilde{\mathbf{A}}^{L-1}\mathbf{P}^{(1)} = \widetilde{\mathbf{A}}^L \tag{16}$$

and

$$p_{ij} = \left[\widetilde{\mathbf{A}}^L\right]_{ij} \tag{17}$$

where matrix $\mathbf{P}$ is the final prediction contribution matrix, and p_{ij} is the element in the ith row and jth column of $\mathbf{P}$.

It can be seen that p_{ij} quantifies the total prediction contribution rate of $\mathbf{x}_j$ to $\mathbf{x}_i$ in the L-order information transmission process. All possible information transmission paths from variable $\mathbf{x}_j$ to $\mathbf{x}_i$ are covered. Therefore, $\mathbf{P}$ measures the multi-level predictive relationship between variables under the GC framework and can more fully reflect the predictive contribution of root causes than direct causality.

In this subsection, we present the proposed MILF module. Compared with conventional multi-layer perceptron (MLP) (Rosenblatt, 1958), MLIF has various advantages. First, MLIF can process two-dimensional time series data directly to capture both the inter-variable predictive relationships and the intra-variable temporal features, whereas MLP can only deal with one-dimensional data to achieve feature extraction. Second, MLIF can characterize the predictive contributions between variables through the normalized adjacency matrix $\widetilde{\mathbf{A}}$, which is more suitable for root cause diagnosis tasks. Third, we design a novel layer-by-layer regularization method for MLIF to screen significant multi-level predictive relationships. In contrast, regularization in MLP acts on each layer independently and thus cannot select multi-level predictive paths. This regularization method will be presented in the following subsection.

4.2.3. Hierarchical adjacency pruning (HAP)

In the above subsections, we present the main architecture and optimization objective of the MPGE model. Multi-level GC can be captured through the key predictive relationship extraction and the stacking of information fusion layers. However, despite the EN constraint imposed on the adjacency matrix $\mathbf{A}$, the multi-level causality between variables is represented by the $\mathbf{P}$ matrix instead of $\mathbf{A}$. Ideally, we expect that the matrix $\mathbf{P}$ can also show sparseness to select the critical multi-level causality. A naive idea is to directly impose EN constraints on matrix $\mathbf{P}$ instead of $\mathbf{A}$. The sparseness effect is due to the L1 norm penalty in the EN constraint. According to Eqs. (16) and (17), the L1 norm penalty of the $\mathbf{P}$ matrix is calculated as follows:

$$\sum_{i=1}^{J}\sum_{j=1}^{J}|p_{ij}| = \sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{k_1=1}^{J}\cdots\sum_{k_{L-1}=1}^{J}\left|\widetilde{a}_{ik_1}\widetilde{a}_{k_1k_2}\widetilde{a}_{k_2k_3}\ldots\widetilde{a}_{k_{L-1}j}\right| \quad (18)$$

where $|p_{ij}| = \sum_{k_1=1}^{J}\cdots\sum_{k_{L-1}=1}^{J}\left|\widetilde{a}_{ik_1}\widetilde{a}_{k_1k_2}\widetilde{a}_{k_2k_3}\ldots\widetilde{a}_{k_{L-1}j}\right|$.

It can be found that each component $(\widetilde{a}_{ik_1}\widetilde{a}_{k_1k_2}\widetilde{a}_{k_2k_3}\ldots\widetilde{a}_{k_{L-1}j})$ in p_{ij} represents a possible predictive path from variable node j to i ($j \to k_{L-1} \to \ldots \to k_1 \to i$). Thus, the L1 norm penalty for the matrix $\mathbf{P}$ is essentially equivalent to selecting the key predictive paths. It is not easy to directly implement the sparsity constraint on the $\mathbf{P}$ matrix due to the limitation of the optimization algorithm. Therefore, we hope to start with the model structure to simplify information transmission and select critical predictive paths. To this end, we propose a hierarchical adjacency pruning (HAP) mechanism. HAP increases the difficulty of information transmission layer by layer, and only the most critical adjacencies are retained in the deep layers. In this way, most information transmission paths are blocked to facilitate the sparseness of the predictive contribution matrix $\mathbf{P}$.

Denote the row vector corresponding to the variable $\mathbf{x}_i$ in $\widetilde{\mathbf{A}}$ matrix be $\widetilde{\mathbf{a}}_i = [a_1, a_2, \ldots, a_J]$, where a_i represents the adjacency strength, i.e., the first-order prediction contribution rate of the ith variable, and $\sum_{j=1}^{J} a_j = 1$. A pruning function, termed δ-softmax, is designed to select key prediction relationships and disconnect less effective adjacencies in the HAP mechanism. For the vector $\widetilde{\mathbf{a}}_i$, let

$$\delta - \text{soft}\max(\widetilde{\mathbf{a}}_i) = \left[a_1^{(s)}, \ldots, a_J^{(s)}\right] \quad (19)$$

where

$$\forall i = 1, \ldots, J, \; a_j^{(s)} = \frac{\exp(\delta a_j)}{\sum_{i=1}^{J}\exp(\delta a_j)} \quad (20)$$

and δ is a hyperparameter called the pruning coefficient.

In each information fusion layer, HAP performs a δ-softmax transformation on the $\widetilde{\mathbf{a}}_i$ vector. The transfer function of the lth information fusion layer in MPGE with HAP can be written as:

$$\begin{aligned}&\mathbf{F}^{(l+1)} = \sigma(\widetilde{\mathbf{A}}^{(l)}\mathbf{F}^{(l)}\mathbf{W}^{(l)}), l = 1, \ldots, L\\ &\widetilde{\mathbf{A}}^{(l+1)} = \delta - \text{soft}\max\left(\widetilde{\mathbf{A}}^{(l)}\right), l = 1, \ldots, L\\ &\mathbf{F}^{(1)} = \mathbf{F}, \widetilde{\mathbf{A}}^{(1)} = \widetilde{\mathbf{A}} = \mathbf{D}^{-1}\mathbf{A}\end{aligned} \quad (21)$$

where δ-softmax $\left(\widetilde{\mathbf{A}}^{(l)}\right)$ means to perform δ-softmax transformation on each row vector $\widetilde{\mathbf{a}}_i$ $(i = 1, \ldots, J)$ in $\widetilde{\mathbf{A}}^{(l)}$.

It can be found that, when the value of δ is large enough, δ-softmax transformation will highlight the large elements in the vector $\widetilde{\mathbf{a}}_i$ to select the vital predictive relationships for each variable $\mathbf{x}_i$. As the number of layers deepens, only the most significant predictive paths can be passed. After adding the HAP mechanism to the model, the calculation of the $\mathbf{P}$ matrix should also be changed accordingly instead of using Eqs. (16) and (17). According to Eqs. (15) and (16), the multi-level prediction contribution matrix $\mathbf{P}$ between variables can be calculated as:

$$\mathbf{P} = \prod_{i=0}^{L-1}\widetilde{\mathbf{A}}^{(L-i)} = \prod_{i=1}^{L}\delta - \text{softmax}^{(L-i)}\left(\widetilde{\mathbf{A}}\right) \quad (22)$$

where $\delta - \text{soft}\max^{(k)}\left(\widetilde{\mathbf{A}}\right)$ means to iteratively perform k times of δ-softmax transformations on $\widetilde{\mathbf{A}}$.

Since the sum of the elements of each row in matrix $\widetilde{\mathbf{A}}$ is one, it is not difficult to verify that the sum of the elements in each row of matrix $\mathbf{P}$ is still equal to one.

HAP allows automatic path selection while ensuring the necessary information transmission. Hence, the significant multi-level causality is captured by the cooperation of the EN constraint on matrix $\mathbf{A}$ and the HAP model, which focus on extracting direct and multi-level paths, respectively.

4.3. Multi-level predictive information quantification

The MPGE method learns the multi-level GC between variables from a multivariate time series, which can be represented by the prediction contribution matrix $\mathbf{P}$. However, like the traditional root cause diagnosis methods, the $\mathbf{P}$ matrix describes the pairwise causality but does not directly determine the root cause variable. In the GC framework, causality is embodied in predictive relationships (Arnold et al., 2007; Bore et al., 2020; Granger, 1969). Therefore, among all variables, the root cause variables should provide the most effective information in the entire multivariate forecasting process. Accordingly, a RootRank scoring algorithm is proposed to quantify the predictive information provided by each variable, thereby characterizing the root cause score and determining the root cause variable.

For each variable $\mathbf{x}_i$, denote its root cause score as s_i. The root cause score is specified to be non-negative, indicating the predictive information amount provided by the variable. Therefore, a variable with a higher score is more likely to be the root cause. According to the nature of root cause, the evaluation of the root cause score of $\mathbf{x}_i$ follows two principles: (1) The more significant the prediction contribution of $\mathbf{x}_i$, namely, the greater the value of p_{ji} $(j = 1, \ldots, J)$, the more predictive information $\mathbf{x}_i$ provides, so s_i should be higher. (2) If $\mathbf{x}_i$ has a non-zero prediction contribution rate p_{ji} to $\mathbf{x}_j$, s_j should be positively correlated with s_i. This is because part of the predictive information provided by $\mathbf{x}_j$ comes from $\mathbf{x}_i$.

The RootRank scoring algorithm models the predictive information transmission as follows. The predictive information of each variable is transferred to other variables in accordance with the corresponding prediction contribution rate. Accordingly, the following equation is established:

$$s_i = \sum_{j=1}^{J} p_{ji} s_j \tag{23}$$

where the product of s_j and p_{ji} represents the predictive information transmitted from $\mathbf{x}_i$ to $\mathbf{x}_j$.

RootRank satisfies the two principles mentioned above. It covers both cases when a variable has high predictive contribution rates and can forecast variables with high scores. Under this mechanism, the following equations are established:

$$\begin{cases} s_1 = \sum_{j=1}^{J} p_{j1} s_j = p_{11} s_1 + p_{21} s_2 + \cdots + p_{J1} s_J \\ s_2 = \sum_{j=1}^{J} p_{j2} s_j = p_{12} s_1 + p_{22} s_2 + \cdots + p_{J2} s_J \\ \qquad \cdots \\ s_J = \sum_{j=1}^{J} p_{jJ} s_J = p_{1J} s_1 + p_{2J} s_2 + \cdots + p_{JJ} s_J \end{cases} \tag{24}$$

which is equal to:

$$\mathbf{s} = \mathbf{P}^{\mathrm{T}} \mathbf{s} \tag{25}$$

where $\mathbf{s} = \left[s_1, s_2, \ldots, s_J\right]^{\mathrm{T}}$ is the score vector composed of the scores of all variables.

Therefore, the quantization of root cause scores is transformed into the eigenvalue decomposition problem of matrix $\mathbf{P}^{\mathrm{T}}$. Mathematical proof of the solvability of RootRank can be found in Section 4.5. When the prediction contribution matrix $\mathbf{P}$ is obtained through the MPGE method, RootRank can give scores for each variable. Variables with the highest scores are automatically identified as the root causes, providing definite diagnosis results.

4.4. Outline of the root cause diagnosis strategy

This subsection summarizes the root cause diagnosis strategy based on MPGE and RootRank into the following steps.

Input Multivariate time series $\mathbf{U}$ composed of faulty variables at all sampling points (collected from fault conditions), forecasting period T, step n for time series segmentation (usually set to 1), L1 and L2 penalty factors β_1 and β_2, the pruning coefficient δ, and structural parameters of MPGE.

Output Root cause variables of the fault.

Prestep: Establish fault detection and isolation models in advance. When a fault is detected in the online stage, the fault isolation method is activated to find the key faulty variables. The multivariate time series $\mathbf{U}$ is then constructed by these faulty variables.

Step 1: Generate training samples. Perform Z-score normalization on $\mathbf{U}$, then split the multivariate time series $\mathbf{U}$ to generate the training samples and the corresponding labels. The sliding window method is used here, where the ith training sample $\mathbf{X}^{(i)}$ is obtained as follow:

$$\mathbf{X}^{(i)} = [\mathbf{U}(ni - n + 1), \mathbf{U}(ni - n + 2), \ldots, \mathbf{U}(ni - n + T)]^{\mathrm{T}} \tag{26}$$

and label (component to be forecasted) of $\mathbf{X}^{(i)}$ can be calculated as:

$$\mathbf{L}^{(i)} = \mathbf{x}(T + 1) = \mathbf{U}(ni - n + T + 1) \tag{27}$$

Step 2: Train the MPGE model. The samples obtained in Step 1 are used for training, namely, to solve the optimization problem in Eq. (13).

Step 3: Calculate the prediction contribution matrix P. Obtain the adjacency matrix $\mathbf{A}$ from the MPGE model trained in the previous step. Then, $\mathbf{P}$ can be calculated by using Eq. (22).

Step 4: Determine the root causes. The RootRank scoring algorithm is used to analyze the matrix $\mathbf{P}$ and solve for the root cause score of each variable. Variables with the highest scores are determined as the root causes.

For convenience, the overall pipeline of the proposed root cause diagnosis framework is presented in Algorithm 1.

4.5. Theoretical verification and discussions

In this subsection, we provide discussions and theoretical explanations of important components of the proposed method.

(1) EN sparse constraint

In the proposed MPGE method, we use EN instead of Lasso (Tibshirani, 1996) used in the LG method to achieve sparsity because EN has been proven to have better variable selection performance than Lasso. Specifically, the Lasso method using only the L1 norm was proven to have a disadvantage: if there is a variable group in which variables are highly correlated with each other, then Lasso tends to select only one variable from the group and does not care which one is selected (Zou & Hastie, 2005). In other words, if the variables in this group all have significant predictive contributions to the target to be predicted, Lasso will miss most of these predictive relations. This makes Lasso unable to fully capture predictive paths in the proposed MPGE method, resulting in information loss. By introducing the L2 norm, EN can preserve useful variable groups while discarding insignificant predictive relationships, i.e., encouraging the group effect. This has been thoroughly verified by theoretical analysis and experiments in the previous study (Zou & Hastie, 2005). Thus, we adopt both L1 and L2 norms here to avoid the insufficient extraction of predictive relations.

(2) Discussions on the designed HAP mechanism

The designed HAP mechanism is a critical part of the proposed method, which is achieved by blocking most information transmission paths. We discuss the rationality of this operation here. Under the HAP mechanism, the information can be transmitted in shallow layers, and redundant paths are pruned in deep layers. Because only a few predictive relationships are retained in the deep information fusion layer, most multi-level information transmission paths are invalidated so that the majority of the components in Eq. (18) will be close to zero. In this way, the HAP mechanism is beneficial to the approximate sparseness of matrix $\mathbf{P}$ to extract significant multi-level predictive relationships. In addition, since δ-softmax transformations are carried out layer by layer, necessary information transmissions can be implemented at shallow layers instead of being discarded. In particular, δ-softmax transformation is not performed in the first information fusion layer. Thus, although only the most critical paths are preserved in the deeper layers, much of the necessary information transfer has been considered in shallower ones.

(3) Solvability of the proposed RootRank algorithm

A mathematical proof of RootRank's solvability is given here. We verify that the equation for solving root cause scores, namely Eq. (24), must have a real number solution where all the elements are non-negative, thus illustrating its theoretical reliability. The detailed proof is presented below.

Define vector $\mathbf{1}$ as:

$$\mathbf{1} = [1, 1, \ldots, 1]^{\mathrm{T}} \tag{28}$$

where $\mathbf{1} \in \Re^J$.

We can find that:

$$\mathbf{P1} = \left[\sum_{i=1}^{J} p_{1i}, \sum_{i=1}^{J} p_{2i}, \ldots, \sum_{i=1}^{J} p_{Ji}\right]^{\mathrm{T}} = [1, 1, \ldots, 1]^{\mathrm{T}} = \mathbf{1} \tag{29}$$

This shows that the vector $\mathbf{1}$ must be an eigenvector of $\mathbf{P}$, and the corresponding eigenvalue equals 1. Since the eigenvalues of

Algorithm 1 The Proposed Root Cause Diagnosis Framework

Data Acquisition Stage

1 Establish fault detection and isolation models in advance. When a fault is detected in the online stage, the fault isolation method is activated to find the key faulty variables.

2 Collect the multivariate time series of faulty variables **U** over a period of time after the fault occurs.

Model Training Stage

3 Perform *Z*-score normalization on **U**.

4 Split the multivariate time series **U** to generate the training samples and the corresponding labels using Eqs. (26) and (27).

5 Train the MPGE model with the optimization objective in Eq. (13) based on the gradient descent algorithm.

6 Obtain the adjacency matrix **A** from the MPGE model and calculate the **P** matrix using Eq. (22).

Root Cause Variable Determination Stage

7 Use the RootRank scoring algorithm to analyze the matrix **P** and solve for the root cause score of each variable.

8 Identify the variables with the highest scores as the root causes.

an arbitrary matrix are not changed after transposing, $\mathbf{P}^{\mathrm{T}}$ must have an eigenvalue equal to 1 same as **P**, which means that the problem in Eq. (29) must have a real number solution. Next, we need to prove that RootRank must have a non-negative real number solution.

Define matrix **Q**, where $\mathbf{Q} \in \Re^{J\times J}$, and

$$q_{ij} = [\mathbf{Q}]_{ij} = \begin{cases} p_{ij} & i \neq j \\ p_{ij} - 1 & i = j \end{cases} \tag{30}$$

Based on Eq. (24), for any $i = 1, \ldots, J$, the following equation holds:

$$\sum_{k=1}^{J} q_{ki} s_k = 0 \tag{31}$$

Assume that both non-negative numbers and negative numbers exist in the solution of Eq. (24). Let H be the set of indices of all negative numbers in the solution, namely,

$$H = \{i | s_i < 0, i = 1, \ldots, J\} \tag{32}$$

According to Eq. (31), for any $k \notin H$,

$$\sum_{i\in H}\left(\sum_{k=1}^{J} q_{ki} s_k\right) = \sum_{i\in H}\sum_{k\in H} q_{ki} s_k + \sum_{i\in H}\sum_{k\notin H} q_{ki} s_k = 0 \tag{33}$$

where

$$\sum_{i\in H}\sum_{k\in H} q_{ki} s_k = \sum_{k\in H} s_k \left(\sum_{i\in H} p_{ki} - 1\right) \tag{34}$$

Additionally,

$$\sum_{i\in H} p_{ki} - 1 < \sum_{i=1}^{J} p_{ki} - 1 = 0 \tag{35}$$

and because for any $k \in H$, $s_k < 0$, we can get:

$$\sum_{i\in H}\sum_{k\in H} q_{ki} s_k = \sum_{k\in H} s_k \left(\sum_{i\in H} p_{ki} - 1\right) > 0 \tag{36}$$

Moreover, if $i \in H$ and $k \notin H$, we know that $q_{ki} > 0$ and $s_k \geq 0$. Thus,

$$\sum_{i\in H}\sum_{k\notin H} q_{ki} s_k \geq 0 \tag{37}$$

So far, we can obtain the following equation:

$$\sum_{i\in H}\sum_{k\in H} q_{ki} s_k + \sum_{i\in H}\sum_{k\notin H} q_{ki} s_k > 0 \tag{38}$$

which is contrary to Eq. (33). In summary, we prove by contradiction, verifying that RootRank must have a solution vector where the elements are all non-negative.

5. Experiments

This section verifies the performance of the proposed method with three examples, including a numerical example, the Tennessee Eastman process (TEP) example, and a real cut-made process of cigarette example. The main source code of the proposed method is available on GitHub.[1] Three classical causal inference methods, i.e., LG (Arnold et al., 2007), neural Granger (NG) (Tank et al., 2021), and STE (Staniek & Lehnertz, 2008), are utilized to provide comparisons. Among them, LG and NG methods are linear and nonlinear representations of sparse GC, respectively. STE is an improved TE model that can well handle causal inference tasks of nonlinear and nonstationary time series. In addition, two state-of-the-art methods, neural graphical modeling (NGM) (Bellot et al., 2021) as well as VLTE (Amornbunchornvej et al., 2021), are considered as comparative models. NGM is one of the most advanced Granger causality models, which considers continuous-time causality in the form of differential equations under the neural network framework to adapt to process nonlinearity and different time scales. It can be regarded as an improved version of NG. VLTE is a recently published improved transfer entropy method. An optimal alignment mechanism for time series is designed in VLTE to consider the changing time delay, which shows better performance than the transfer entropy-based methods with fixed time delays.

Moreover, hyperparameter selection is discussed. In the proposed method, the main hyperparameters include pruning coefficient δ, penalty factor of L1 norm β_1, penalty factor of L2 norm β_2, number of information fusion layers, and forecasting period T. For the number of information fusion layers, we set it by the rule of thumb. The number of information fusion layers represents the maximum fault propagation path length considered in MPGE. In general, fault propagation paths in industrial processes are not very long. Previous research in graph theory (Fronczak, Fronczak, & Hołyst, 2004) suggests that the average path length between two nodes in a graph structure does not exceed 3 when the number of nodes does not exceed 100. Industrial fault data also agree with this rule. For example, in the study of fault mecha-

[1] https://github.com/chunhuiz/MPGE-RootRank-for-root-cause-diagnosis

Table 2
Parameters adopted in the illustration cases.

Case	δ	β_1	β_2	Number of information fusion layers	Forecasting period T
Numerical example	5	0.15	0.001	3	10
TEP IDV(1)	10	0.2	0.001	3	10
TEP IDV(8)	10	0.2	0.001	3	10
Cut-made process of cigarette	5	0.9	0.001	3	12

nisms for TE processes (Duan, Chen, Shah, & Yang, 2014; Gao, Xu, & Zhu, 2016), it can be found that most fault propagation path lengths do not exceed 3. Thus, we set the number of information fusion layers to 3 here. This means that at most three-hop fault propagation is considered. For other parameters (δ, β_1, β_2, T), we set them through grid search and cross-validation. We use the prediction error on the validation set to select the best group of parameters because a lower prediction error usually leads to more accurate predictive relationships captured by the model. For convenience, the key parameters of MPGE in the experiments are listed in Table 2. For the comparative models based on GC, similar to the proposed MPGE, the hyperparameters are set according to cross-validation. For TE-based comparative methods, we follow the guideline for parameter settings in the existing research on permutation entropy (Bandt & Pompe, 2002).

Next, we present the detailed experiment settings:

Training data generation In the numerical example, we constructed a multivariate time series with six variables. The causal relations between these variables were known. We generated 500 samples for model training. In the TEP and cut-made process of cigarette examples, we used 200 and 100 faulty samples for model training, respectively. In addition, as mentioned earlier, the root cause diagnosis task needs to use the fault isolation model to determine the faulty variables in advance so that the information transfer between the fault variables can be used to characterize the fault propagation. For the fault cases of TEP, we adopted the fault isolation results from previous studies (Chen et al., 2018; Li, Qin, & Yuan, 2016), respectively. The faulty variables for the cigarette example were determined from our previous work (Yu & Zhao, 2017).

Data preprocessing: In order to avoid the influence of data scale on model training, the input data are standardized by the Z-score normalization method.

Root cause determination criteria: For the comparative methods, we obtained the pairwise causal relationships inferred by them. If a variable acts only as the cause rather than an effect of other variables, it is determined to be the root cause. This is a common criterion in previous root cause diagnosis studies. For the proposed method, since we propose a novel diagnosis framework, the diagnosis results are presented in terms of root cause scores calculated by RootRank rather than pairwise causalities. The variables with the highest root cause scores in the proposed method were directly identified as root causes.

Evaluation metrics: For the root cause diagnosis task, we only focus on whether the root cause variables are correctly identified. We use the root cause diagnosis accuracy to measure the experimental results. Here, root cause diagnosis accuracy refers to the proportion of root cause variables that were correctly identified among all true root cause variables of the fault cases. Additionally, the causal inference accuracy is also introduced for the comparative methods, which is defined in Section 4.1. We use this metric not to measure diagnosis performance but to illustrate the difference between root cause diagnosis and causal inference tasks.

Multiple times of experiments: For the proposed method, NG, and NGM, considering the randomness of neural network training, we carried out multiple times of experiments. We found that some different diagnosis results might appear when the number of experiments was large enough. For online industrial applications, one cannot run multiple experiments to select the best result since the real root cause variable is unknown. Therefore, to simulate the online diagnosis scenario, we randomly select a set of experimental results to present in Sections 5.1–5.3. To further discuss the impact of randomness, we analyze the average root cause diagnosis accuracy in multiple times of experiments where the process data or the initialization of trainable parameters are changed and perform statistical tests to illustrate the advantages of the proposed method in Section 5.4. For the LG, STE, and VLTE methods, they are not affected by random factors, so there is no need to run experiments for them repeatedly.

5.1. Numerical example

A nonlinear process including six variables with known causal relations was constructed in this case to illustrate the difference between traditional causal inference and root cause diagnosis tasks. Also, we discussed the impact of hyperparameters in the proposed framework on diagnostic performance.

First, two root cause variables t_1 and t_2 were generated as follows:

$$t_1(t) = \begin{cases} 0 & t = 0 \\ 0.5t_1(t-1) + 0.8\varepsilon_1(t) & t > 0 \end{cases} \tag{39}$$

$$t_2(t) = \begin{cases} 0 & t = 0 \\ 0.5t_2(t-1) + 0.8\varepsilon_2(t) & t > 0 \end{cases} \tag{40}$$

where $\varepsilon_1(t)$ and $\varepsilon_2(t)$ were two independent random perturbations sampled from the standard uniform distribution $U(0,1)$ at the tth sampling point.

Based on the two root cause variables, the other four variables were generated:

$$\begin{aligned} x_1(t) &= 2t_1^2(t-3) - t_1(t-1) \\ x_2(t) &= -x_1^2(t-2) + 2x_1(t-2) + 0.5t_1(t-1) \\ x_3(t) &= \sin(t_2 - 2) \\ x_4(t) &= x_2(t-1)x_3(t-1) \end{aligned} \tag{41}$$

The direct causal relationships between these variables are shown in Fig. 5(e). The darker parts in the matrix indicate the existence of direct causal relationships. As shown, the first two rows of the matrix in Fig. 5(e) do not have a darker part, indicating that t_1 and t_2 are not the effect of other variables and thus are root cause variables.

According to Eqs. (39)–(41), 500 samples were generated to train the MPGE model. The root cause diagnosis results of MPGE, LG, NG, and STE methods are shown in Fig. 5. We visualized the predicted contribution matrix **P** for MPGE in the plot of Fig. 5(a). Unlike the direct relations in Fig. 5(g), the diagnosis result of MPGE presented multi-level predictive contributions between

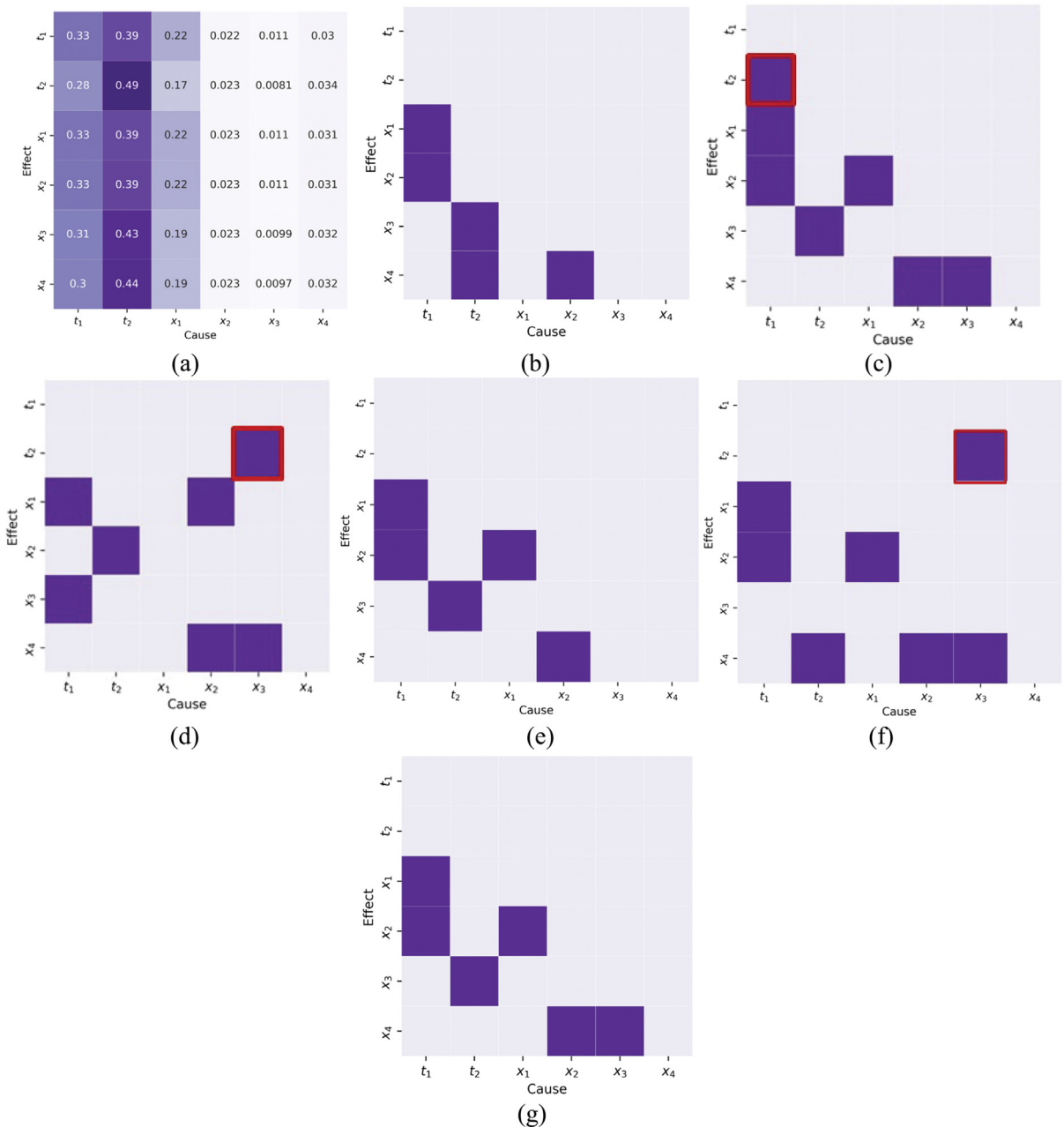

Fig. 5. Root cause diagnosis results of different methods for the numerical example. (a) The prediction contribution matrix (**P** matrix) for MPGE. (b) LG (λ was set to 0.1). (c) NG. (d) STE. (e) NGM. (f) VLTE. (g) The truth. (For subfigure (a), the higher prediction contribution rates are represented by darker colors; for subfigures (b) through (g), the darker part in the matrix indicates the existence of a direct causal relationship. The red boxes indicate the minor mistakes that cause the root cause diagnosis to fail.) .

variables, which considered both direct and indirect causalities to highlight the root causes. It could be found that MPGE was the only method that could quantitatively describe the predictive strength, and the higher prediction contribution rates are represented by darker colors. The first two columns of the prediction contribution matrix of MPGE had darker colors, indicating that both t_1 and t_2 had relatively significant predictive contributions to other variables. Compared with t_1 and t_2, downstream variables had lower prediction contribution rates. These results mean the proposed MPGE could highlight and obtain the true root cause variables (t_1 and t_2) correctly, which is agreed with the root cause information shown in Fig. 5(g).

Fig. 6 presents the root cause scores given by RootRank. We normalized all the root cause scores to provide a fair comparison, namely:

$$\widetilde{s}_i = \frac{s_i}{\sum_{i=1}^{J} s_i} \tag{42}$$

where $\widetilde{s}_i$ is the normalized score for the ith variable.

As shown in Fig. 6, the scores of t_1 and t_2 are significantly higher than other variables, indicating that the proposed method correctly determined the root causes. Moreover, the score of x_1 was also relatively high because x_1 was closer to the root cause t_1 and provided more predictive information than x_2, x_3, and x_4. The experimental results above show that the proposed MPGE and RootRank methods can effectively extract predictive relationships and provide reliable quantitative representations.

Among the comparative methods, LG and NGM correctly determined the root causes. NG, STE, and VLTE missed the root

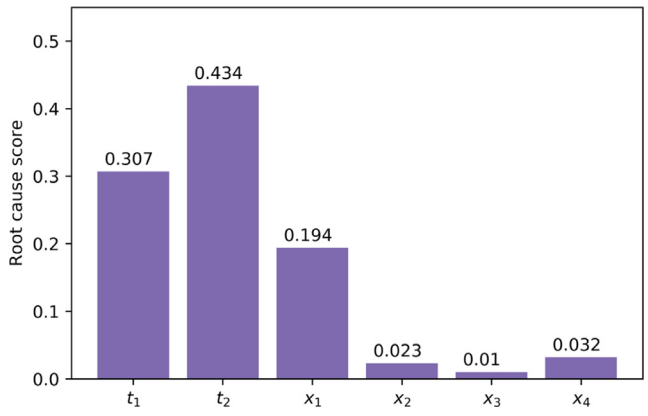

Fig. 6. Root cause scoring result of the proposed RootRank algorithm for the numerical example.

Table 3
Root cause scores given by the MPGE and RootRank models with different δ parameters for the numerical example (β_1 was set to 1).

δ	Root cause scores					
	t_1	t_2	x_1	x_2	x_3	x_4
0.1	0.247	0.381	0.231	0.066	0.038	0.036
1	0.244	0.391	0.195	0.064	0.057	0.049
5	0.255	0.502	0.161	0.022	0.039	0.023
10	0.214	0.541	0.160	0.040	0.030	0.014
20	0.304	0.415	0.124	0.042	0.042	0.073
100	0.272	0.507	0.149	0.013	0.010	0.048

cause variable t_2 because of minor mistakes (marked by the red box in Fig. 5). In other words, NG, STE, and VLTE identified t_2 as a downstream variable of other variables, resulting in t_2 not being correctly diagnosed as the root cause. Notably, as one of the most advanced causal inference models, the NG method achieved high causal inference accuracy (defined in Section 4.1) of 96.7%. Nevertheless, it still failed to give a correct diagnosis result. However, LG's causal inference accuracy (90%) was lower than NG's but successfully identified the root cause. The contradiction suggests the inadequacy of direct causality for root cause representation and the inappropriateness of pairwise causality for root cause diagnosis tasks. Contrariwise, MPGE covered the multi-level predictive relationships related to the root causes, thereby significantly highlighting the root cause locations. Further, RootRank enabled the prediction contribution matrix from MPGE to be converted into root cause scores, improving the reliability and ensuring the root cause identifiability.

In general, a root cause diagnosis model has several hyperparameters that affect the diagnosis performance. In our framework, β_1 and δ are the key hyperparameters that control the sparsity of the diagnosis results. We listed the root cause scores given by the MPGE and RootRank models with different β_1 and δ parameters in Tables 3 and 4, respectively. It could be seen that the root cause variables t_1 and t_2 had the highest root cause scores within a wide parameter range, indicating that the proposed diagnosis framework is not sensitive to hyperparameters. Due to space limitations, we did not visualize MPGE's **P** matrices under different parameter settings. In fact, the **P** matrix under different parameters had certain differences, but they all correctly captured the predictive relationships of the root causes to other variables. In addition, RootRank could give quantitative root cause scores so that different **P** matrices had consistent root cause identification results. These experimental results confirm the stability and practicability of the proposed framework.

Table 4
Root cause scores given by the MPGE and RootRank models with different β_1 parameters for the numerical example (δ was set to 20).

β_1	Root cause scores					
	t_1	t_2	x_1	x_2	x_3	x_4
0.05	0.254	0.444	0.258	0.009	0.016	0.017
0.1	0.311	0.443	0.176	0.017	0.044	0.009
0.2	0.282	0.446	0.199	0.018	0.039	0.016
0.5	0.272	0.375	0.257	0.048	0.008	0.039
1	0.244	0.391	0.195	0.064	0.057	0.049
2	0.164	0.647	0.095	0.023	0.051	0.020

5.2. Tennessee Eastman process (TEP)

The TEP is a simulation system created based on an actual chemical process of Eastman Chemical Company. It simulates various typical characteristics of the complex industrial process, which is a commonly used benchmark problem in the field of process control. The whole process includes five operating units: a reactor, a product condenser, a vapor–liquid separator, a recycle compressor, and a product stripper, involving 12 control variables and 41 measured variables. The details of the process description can be found in Downs and Vogel (1993). Here, fault cases IDV(1) and IDV(8) collected from the TEP were used to verify the root cause diagnosis performance of the proposed method. The fault in IDV(1) was an abnormal step change, and it did not directly transmit from the root cause variable to the downstream variables but was gradually transmitted through physical coupling and control adjustment. We adopted this case to illustrate the necessity of considering multi-level causality and the effectiveness of the HAP mechanism. Moreover, we used fault case IDV(8), which was a random fault, to illustrate that the proposed method can be applied to different fault types.

5.2.1. Step change fault case IDV(1) in TEP example

In the IDV(1) case, the A/C feed ratio in stream 4 had a step change, resulting in a decrease in the amount of A in the recycling stream. Here, we adopt the root cause analysis and fault isolation results in the previous study (Chen et al., 2018). After the fault occurred, the corresponding valve opening (x_{44}) was adjusted to increase the flow rate of A feed in A stream 1 (x_1). This series of abnormal process behaviors affected many other process variables. The flow rate of stream 4 (x_4) was affected by the action of the level controller, which further changed the stripper pressure (x_{16}), and temperature (x_{18}). After that, the product separator pressure (x_{13}) and the reactor pressure (x_7) were also disturbed. Meanwhile, the temperature control of the stripper affected the stripper steam valve (x_{50}) to adjust the stripper steam flow (x_{19}). Based on the above analysis, x_1 and x_{44} were the root cause variables. The involved faulty variables include $\{x_1, x_4, x_7, x_{13}, x_{16}, x_{18}, x_{19}, x_{44}, x_{50}\}$.

The MPGE model was trained using 200 samples collected after the occurrence of the fault. The prediction contribution matrix of MPGE is presented in Fig. 7(a). It could be found that the two root cause variables were given significantly high prediction contribution rates.

Fig. 8 shows the root cause scores calculated by the RootRank algorithm. The variables x_1 and x_{44} had the highest scores, indicating that MPGE and RootRank correctly determined the root causes.

For comparison, the root cause diagnosis results of five comparative methods are shown in Fig. 7(b)–(f). Overall, the diagnostic performance of these methods was not satisfactory. STE misidentified x_7 and x_{50} as root causes, and VLTE incorrectly diagnosed x_4 as the root cause. In addition, NGM missed one of the root cause variables, x_1. Moreover, Although LG and NG correctly

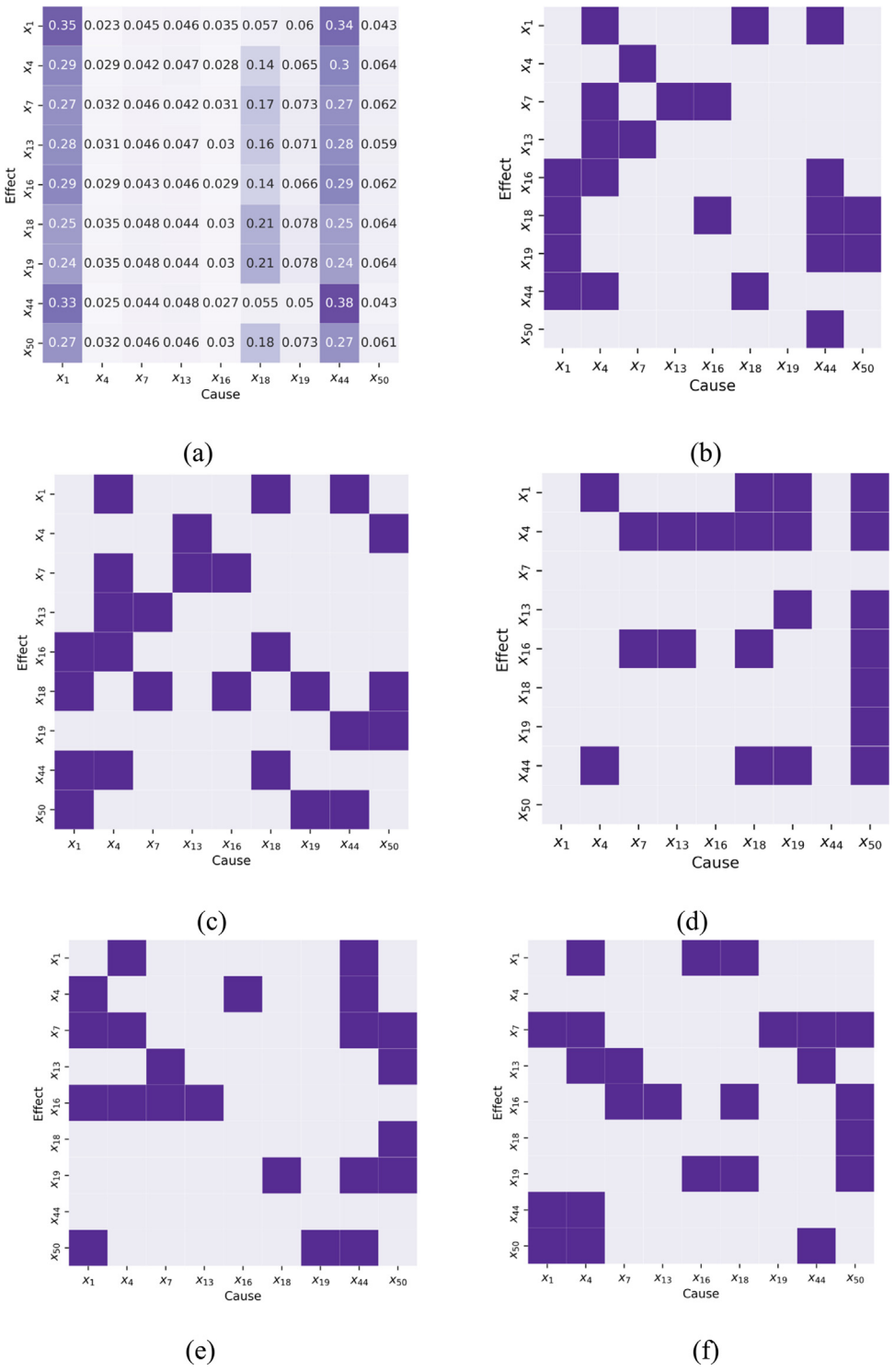

Fig. 7. Root cause diagnosis results of different methods for the IDV(1) case in the TEP example. (a) The prediction contribution matrix (**P** matrix) for MPGE. (b) LG. (c) NG. (d) STE. (e) NGM. (f) VLTE. (For subfigure (a), the higher prediction contribution rates are represented by darker colors; for subfigures (b) through (f), the darker part in the matrix indicates the existence of a direct causal relationship.)

identified some causal relationships, the root causes x_1 and x_{44} were both determined as the effects of other variables, so the definite diagnosis results could not be given. The results suggest that the direct causality cannot fully characterize the root causes due to the multi-level fault propagation caused by the feedback control and complex couplings between process variables. We visualized the **P** matrix to further verify this viewpoint when the number of information fusion layers in MPGE was set to one. As shown in Fig. 9, the **P** matrix of MPGE was also difficult to indicate the root causes when only direct predictive relationships were considered. The experimental phenomena highlight the necessity of considering multi-level causalities in root cause diagnosis tasks.

We also verified the effectiveness of the HAP model. The **P** matrix of MPGE without the HAP mechanism is shown in Fig. 10. It could be found that the **P** matrix was not sparse at all and could

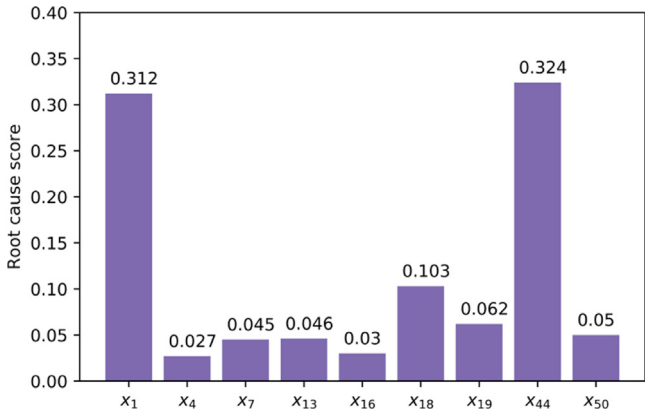

Fig. 8. Root cause scoring result of the proposed RootRank scoring algorithm for the IDV(1) case in the TEP example.

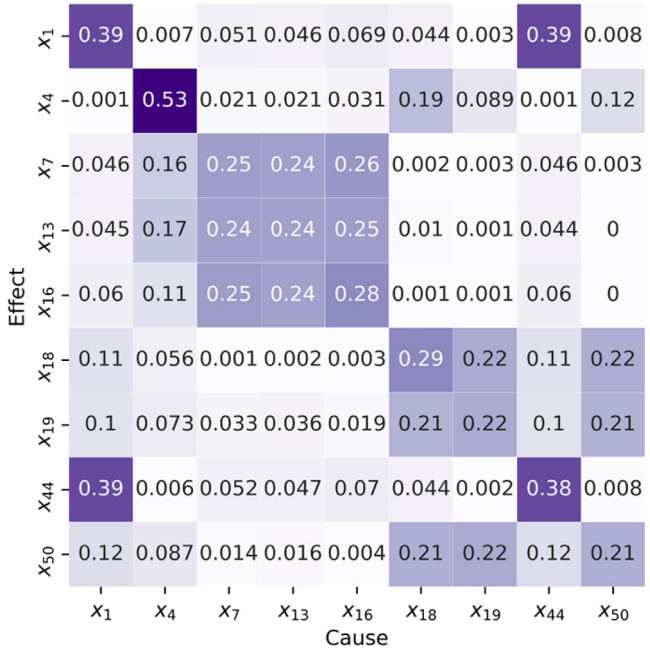

Fig. 9. Prediction contribution matrix of MPGE with only one information fusion layer for the IDV(1) case in the TEP example.

not highlight the root cause correctly. Therefore, this experiment confirms that the HAP mechanism can remove redundant predictive relationships to make the **P** matrix sparse, thereby helping root cause identification.

5.2.2. Random fault case IDV(8) case in TEP example

In the IDV(8) case, the total feed of materials A, B, and C in stream 4 had abnormal variations. Unlike the fault in IDV(1), the variation here was random rather than stepwise. According to the analysis in the previous study (Li et al., 2016), the fault directly affected the variable x_{26} (total feed flow in stream 4). Subsequently, the abnormality in x_{26} was propagated through the control loop to x_4 (total feed in stream 4). Also, downstream variables in stream 3, x_3 (E feed in stream 3) and x_{24} (E feed flow in stream 3), were forced to change to keep pace with stream 4. Eleven faulty variables were isolated, including $\{x_3, x_4, x_6, x_9, x_{14}, x_{21}, x_{23}, x_{24}, x_{26}\ x_{31}\ x_{32}\}$, and the root cause variable was x_{26} (Li et al., 2016).

The diagnosis results of the proposed MPGE method and comparative models are visualized in Fig. 11. As shown in Fig. 11(a),

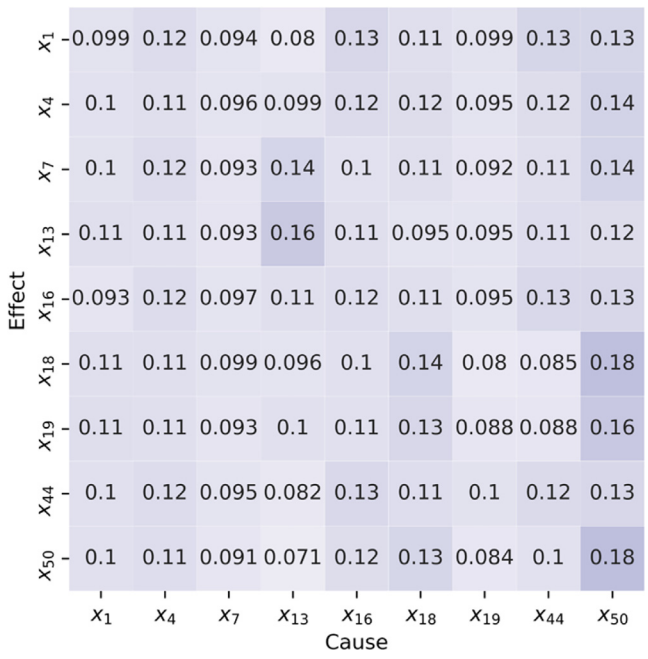

Fig. 10. Prediction contribution matrix of MPGE without HAP mechanism for the IDV(1) case in the TEP example.

the prediction contribution rates associated with the root cause variable x_{26} were assigned relatively high values in the **P** matrix of MPGE. This means that MPEG correctly characterized the multi-level fault propagation and highlighted the root cause. Further, the RootRank algorithm was activated to quantify root cause scores for each variable. As shown in Fig. 12, x_{26} has the highest root cause score among all candidates. Thus, the proposed method successfully diagnosed the root cause in this random fault case.

For comparative methods, only NGM yielded correct diagnosis results. The fault in this case propagated through complex control loops in multiple stages and involved a relatively large number of faulty variables, increasing the diagnosis difficulty. It can be seen from Fig. 11(b) and Fig. 11(c) that the diagnostic results of LG and NG have certain similarities, and both of them mistakenly identified x_{23} as the root cause. The technical principles of the two are similar, and the only difference is whether the nonlinearity is considered, so the results are reasonable. As mentioned earlier, both LG and NG only considered first-order (direct) causality and ignored multi-level fault propagation, making the extracted causal information insufficient and leading to incorrect diagnosis results. Additionally, STE misidentified x_6 as the root cause. Notably, the inference result of STE contained a causal relation from x_6 to x_{26}, and it was this relationship that prevented the correct root cause identification. Similarly, in the experimental results of VLTE, the causality from x_{14} to x_{26} led to an incorrect diagnosis. The phenomena further confirm our view, that is, root cause diagnosis and causal inference are different tasks, and minor errors in causal inference results can lead to a failure of root cause diagnosis.

5.3. Cut-made process of cigarette

The cut-made process is a typical procedure in cigarette manufacturing, which is critical to the flavor and style characteristics of cigarettes. The cutting process of cigarettes mainly includes three operating parts: leaf processing, silk processing, and blending and spicing. In this process, twenty-three measured variables were

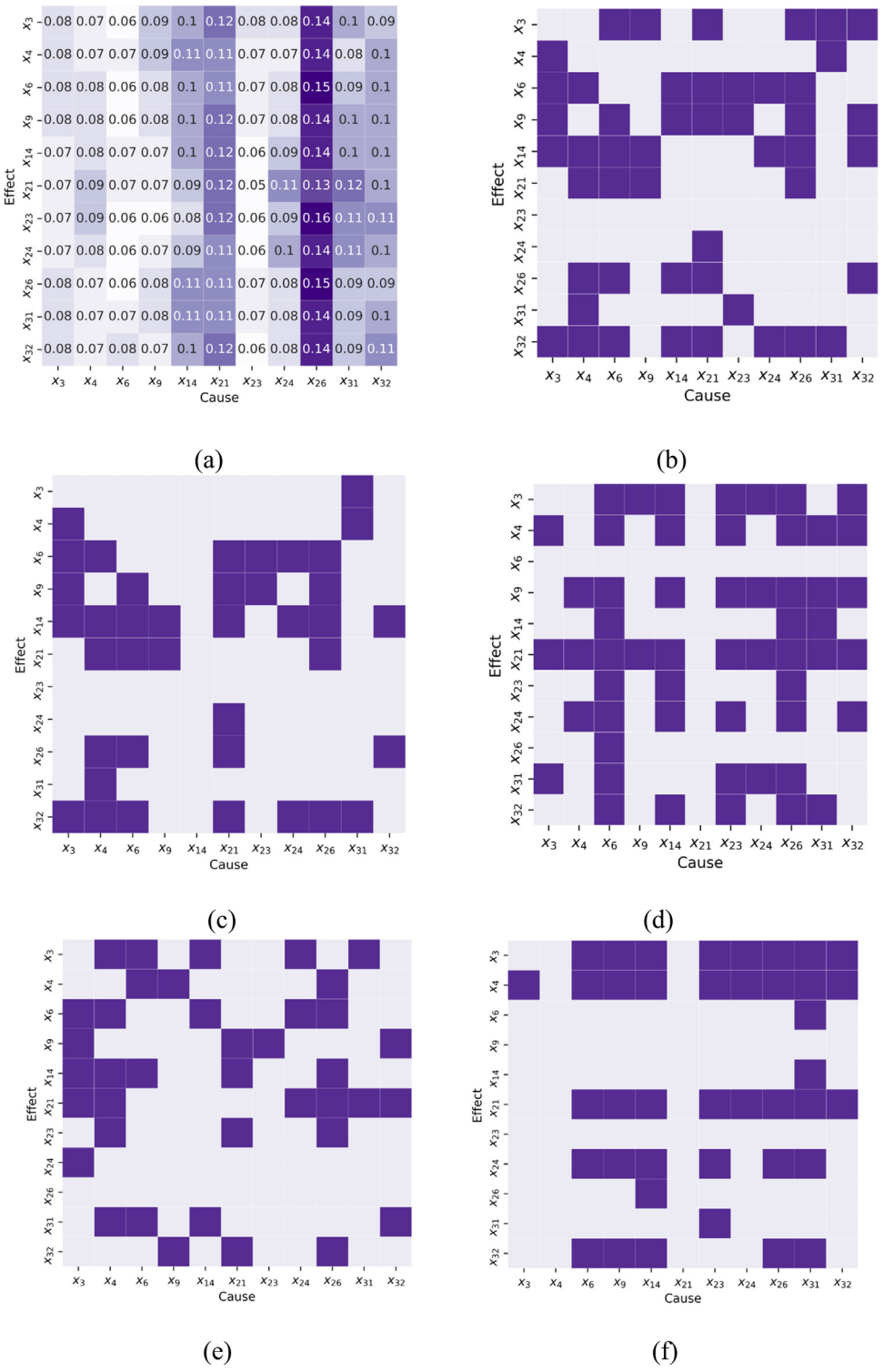

Fig. 11. Root cause diagnosis results of different methods for the IDV(8) case in the TEP example. (a) The prediction contribution matrix (**P** matrix) for MPGE. (b) LG. (c) NG. (d) STE. (e) NGM. (f) VLTE. (For subfigure (a), the higher prediction contribution rates are represented by darker colors; for subfigures (b) through (f), the darker part in the matrix indicates the existence of a direct causal relationship.)

collected with a sampling period of ten seconds, which are listed in Table 5. More detailed information about this process can be found in Zhao and Gao (2016). A real fault case in the cut-made process was used to verify the diagnostic performance of the proposed framework in the actual industrial scene.

In this case, a step change was imposed in the set point of the KLD hot windspeed (x_{18}), further resulting in anomalies in other variables. We gave a detailed propagation mechanism and root cause analysis of this fault case in our previous study (Chen & Zhao, 2022). In addition, according to the previous fault

Table 5
Variables in cut-made process of cigarette used for modeling.

Variable no.	Variable description	Variable no.	Variable description	Variable no.	Variable description
1	Initial cut tobacco flow rate (kg/h)	9	KLD moisture exhaust damper opening (%)	17	KLD hot wind temperature (°C)
2	Initial moisture content (%)	10	KLD vapour pressure (bar)	18	KLD hot wind speed (m/s)
3	SIROX vapour pressure (bar)	11	Region-1 vapour pressure (bar)	19	KLD water removal mass (l/h)
4	SIROX temperature (°C)	12	Region-1 wall temperature	20	KLD dried moisture content (%)
5	SIROX vapour volume flow rate (m^3/h)	13	Region-2 vapour pressure (bar)	21	KLD dried temperature (°C)
6	SIROX vapour mass flow rate (kg/h)	14	Region-2 wall temperature (°C)	22	Cooling moisture content (%)
7	SIROX vapour diaphragm valve opening (%)	15	Region-1 condensed water temperature (°C)	23	Cooling temperature (°C)
8	KLD moisture exhaust negative pressure (ubar)	16	Region-2 condensed water temperature (°C)		

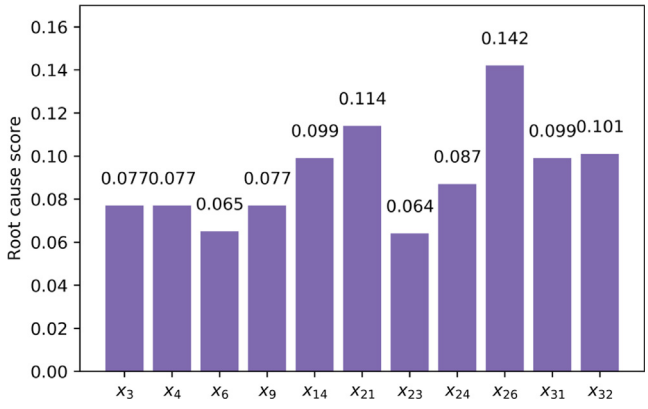

Fig. 12. Root cause scoring result of the proposed RootRank scoring algorithm for the IDV(8) case in the TEP example.

isolation results (Yu & Zhao, 2017), the faulty variables included x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, and x_{23}. One hundred samples collected after the fault occurred were used for fault diagnosis. The diagnostic results of all methods involved are shown in Fig. 13, and the root cause scoring result of RootRank is presented in Fig. 14. The variable x_{18} had the highest score, indicating that MPGE correctly identified the root cause. The score of x_{19} was also relatively high. Nevertheless, observing the **P** matrix of MPGE in Fig. 13(a), we knew that the high score of x_{19} was due to its strong self-predictability. Therefore, we could rule out the possibility of x_{19} as the root cause. In the above analysis, the mutual explanations between the **P** matrix and the root cause scores could help us effectively determine the root cause variable.

As for the comparative methods, only STE correctly determined the root cause. LG and NGM misidentified the root cause as x_{21}, VLTE gave an incorrect root cause variable x_{19}, and NG discovered the redundant root causes x_{19} and x_{21}. According to the **P** matrix of MPGE, we could conclude that the self-predictability of x_{19} and x_{21} were relatively strong, which may interfere with the inference results of LG, NGM, and NG. These results imply that the predictive strength quantification of MPGE is superior to traditional binary causality representation due to its numerical interpretability.

5.4. Random experiments and ablation studies

In this subsection, multiple times of random experiments and statistical tests are carried out to consider the influence of randomness and further illustrate the advancement of the proposed method. In addition, the effectiveness of each component in our model is further verified through ablation studies.

5.4.1. Random experiments

Two sets of random experiments were conducted to consider the randomness and show the effectiveness of the proposed method through the independent samples t-test (Ross & Willson, 2018). One was to randomly change the settings of variables in the dataset of the numerical example, which was used to observe the performance of the proposed method and the comparative methods on different data. The other was to randomly change the initial set point of the trainable parameters in the network, thus considering randomness during model training.

For the first set of experiments, as shown in the following equation, we introduce several changeable factors, $\{\theta, v, r, \pi\}$, into the data generation process in the numerical example, so as to generate different data in each experiment without changing the causal relationships between variables:

$$\begin{aligned} x_1(t) &= \theta t_1^2(t-3) - t_1(t-1) \\ x_2(t) &= v x_1^2(t-2) + r x_1(t-2) + 0.5 t_1(t-1) \\ x_3(t) &= \sin(t_2 - 2) \\ x_4(t) &= x_2(t-1) x_3(t-\pi) \end{aligned} \tag{43}$$

The optional range of each factor is: θ: {1, 2, 3}; v: {−2, −1, 1, 2}; r: {1, 2, 3, 4}; π: {1, 2, 3}. In each experiment, we randomly set these factors and generate different data for model training. The results of 30 times of random experiments are listed in Table 6. It can be seen that the proposed method outperformed all the comparative models, indicating its effectiveness. Notably, the comparative models showed correct results in some of the experiments, but their overall performance of them was not satisfactory. This was because these comparative methods identify the root cause variables through the form of pairwise causalities, which resulted in unstable diagnosis results. In such a form, even if the vast majority of causal relationships were correctly inferred, as long as the model mistakenly identified the root cause variable as a downstream variable of other variables, the diagnosis was invalid. The proposed method used quantitative scores to diagnose the root causes, which focused on the overall

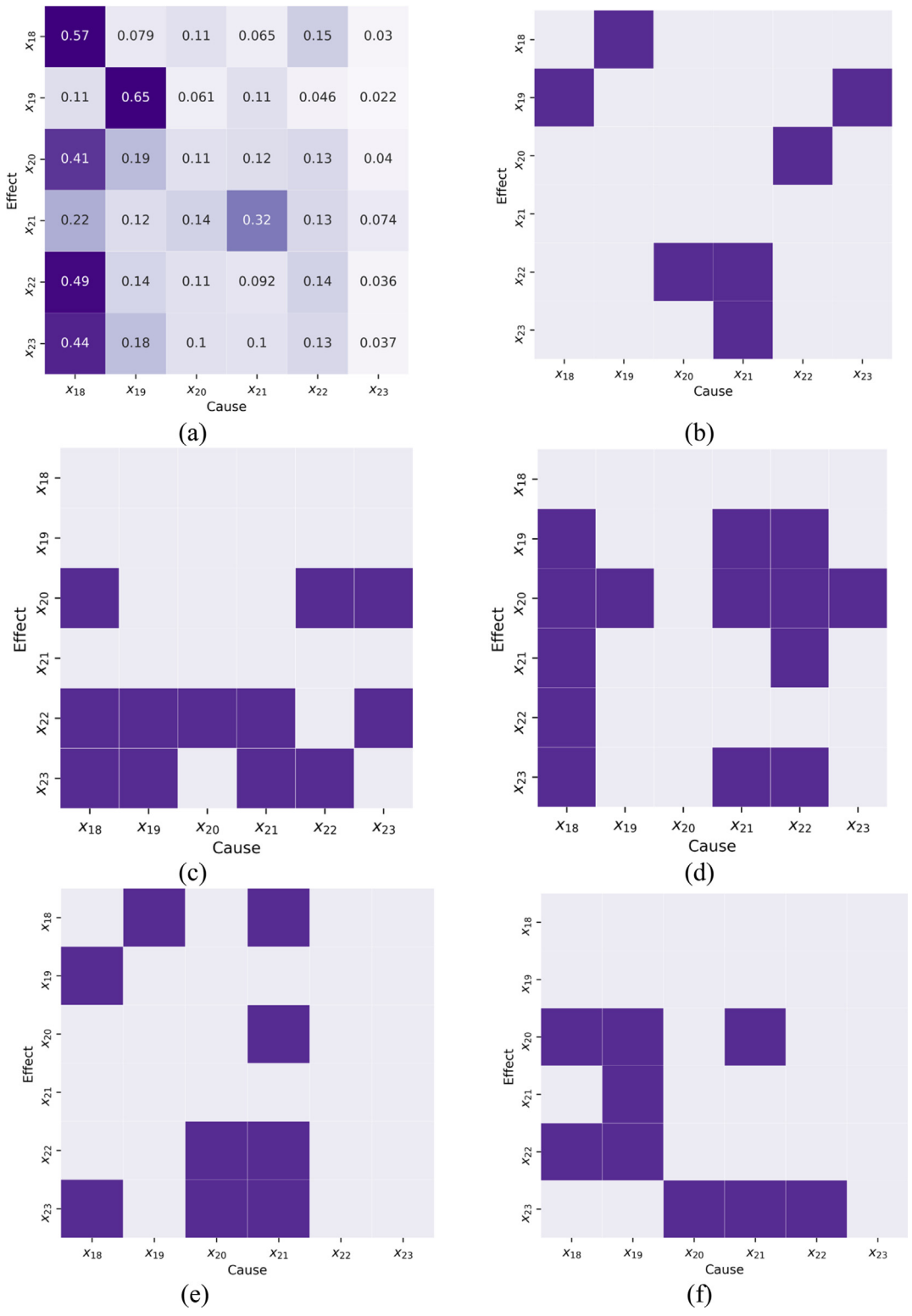

Fig. 13. Root cause diagnosis results of different methods for the cut-made process of cigarette example. (a) The prediction contribution matrix (**P** matrix) for MPGE. (b) LG. (c) NG. (d) STE. (e) NGM. (f) VLTE. (For subfigure (a), the higher prediction contribution rates are represented by darker colors; for subfigures (b) through (f), the darker part in the matrix indicates the existence of a direct causal relationship.)

predictive contribution of each variable rather than the specific pairwise causalities, thus overcoming this problem and improving accuracy and reliability.

For the second set of experiments, each time, we randomly changed the initial set points of the trainable parameters in the network. Note that the comparative methods, LG, STE, and VLTE, were not affected by the randomness during training, and thus we did not consider them here. The results of 30 times of random experiments are listed in Table 7. It can be seen that the proposed method achieved high average accuracy in these experiments, indicating that the randomness did not significantly affect its performance. Also, through the independent samples t-test, we

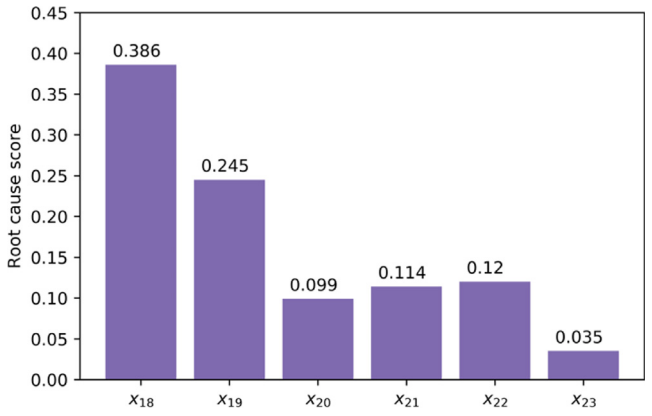

Fig. 14. Root cause scoring result of the proposed RootRank algorithm for the cut-made process of cigarette example.

could conclude that the proposed method achieved higher root cause diagnosis accuracy than NG and NGM.

Notably, although the results given by the comparative methods seem to be unsatisfactory, this does not mean that they perform poorly because they actually achieved high causal inference accuracy in some cases. This reveals the unreasonableness of previous root cause diagnosis methods based solely on causal inference and validates the essential difference between root cause diagnosis and causal inference tasks described in Section 4.1. Through the experimental results for the numerical example, we can find that even using state-of-the-art causal inference methods such as NGM, direct causality alone cannot guarantee the successful identification of root causes due to the multi-level propagation of faults. This reflects the necessity and correctness of considering multi-level information transmission to extract both direct and indirect causal relations. Also, pairwise causality-based methods may lead to complex causal matrices, and thus root cause variables may not be directly accessible from such complex causal matrices. In contrast, the proposed method quantitatively characterizes the root cause variables through the root cause score calculated by the RootRank algorithm, providing a more intuitive form of diagnosis results. This ensures that the root cause variables must be available, as the variables with the highest root cause scores can be directly identified as the root causes.

5.4.2. Ablation studies

There are several key parameters in our model, including β_1, β_2, δ, and the number of information fusion layers, which represent the roles of the L1 norm penalty, the L2 norm penalty, the HAP mechanism, and the MLIF module. To verify their effectiveness, ablation studies are conducted here. For β_1 and β_2, we set them to 0 to eliminate their effects. For δ, considering it represents the strength of hierarchical adjacency pruning (HAP) in our model, we verify its effect by removing the HAP mechanism in the ablation experiment. For the number of information fusion layers, we set it to one in the ablation study to avoid multi-level information fusion, thereby verifying the effectiveness of considering multi-level predictive relations. Also, we conducted multiple times of experiments (30 times here) to ensure the generalizability of results by randomly changing the initial set points of trainable parameters in the model. The mean and standard deviation of the root cause diagnosis accuracy of these experimental results are used for statistical tests to illustrate the effectiveness of each component.

The results of ablation experiments are listed in Table 8. We used the independent samples t-test (Ross & Willson, 2018) to measure the significance of differences between experimental results. It could be seen that the root cause diagnosis accuracy significantly decreased in most cases of ablation experiments for each parameter, indicating the effectiveness of the corresponding modules in the proposed method. The performance degradation of the model is most significant in both cases with $\beta_1 = 0$ and without the HAP mechanism. This is because the determination of causality requires the significance of the predictive relationship. The L1 norm penalty term controlled by β_1 and the HAP mechanism controlled by δ guarantee the significance of direct and multi-level predictive pathways, respectively, which are crucial for root cause diagnosis. In addition, model performance dropped in three examples when only one information fusion layer was used, illustrating the effectiveness of considering multi-level predictive relationships. Also, in the TEP IDV(8) and the cut-made process of cigarette examples, the performance degraded when the L2 norm constraint was not used, confirming that the used elastic net constraint has advantages in multivariate root cause diagnosis compared to the single L1 norm constraint.

Table 6
Results of multiple times (30 times) experiments for the numerical example.

Method		Performance
The proposed model		0.867 (0.288)
LG	Accuracy	0.583 (0.389)
	Statistical test	✓
NG	Accuracy	0.717 (0.334)
	Statistical test	✓
STE	Accuracy	0.633 (0.407)
	Statistical test	✓
NGM	Accuracy	0.733 (0.281)
	Statistical test	✓
VLTE	Accuracy	0.617 (0.499)
	Statistical test	✓

(✓: the result is different from that of the proposed model with statistical significance (90% confidence level);
×: the result is not significantly different from that of the proposed model statistically;
the values in brackets are the standard deviations.)

6. Conclusion

In this study, a multi-level GC extraction and quantification framework, called MPGE and RootRank scoring, is proposed and applied to the root cause diagnosis of industrial process faults. Both direct and indirect causality can be considered to sufficiently characterize fault propagation. Moreover, the designed RootRank scoring algorithm can provide definite diagnosis results rather than ambiguous root cause candidates by transforming complex causal structures into understandable root cause scores. Experiments on a numerical example, a benchmark process, and a real industrial process verified the performance of the proposed framework. Compared with counterpart methods, the proposed framework clearly identified one or two root causes in each example, which were all in line with the truth. Additionally, the scores assigned to the root cause variables were significantly higher than that of downstream variables. Also, the experimental results indicate that the proposed framework can be effectively applied to actual industrial scenarios. Furthermore, random experiments and ablation studies confirmed the reliability and advancement of the proposed method.

Table 7
Results of multiple times (30 times) experiments for different examples and methods with randomly initialized trainable parameters.

Method		Numerical example	TEP IDV(1)	TEP IDV(8)	Cut-made process of cigarette
The proposed model		0.9 (0.238)	0.917 (0.227)	0.833 (0.373)	0.967 (0.18)
NG	Accuracy	0.733 (0.309)	0.417 (0.41)	0.267 (0.442)	0.767 (0.423)
	Statistical test	✓	✓	✓	✓
NGM	Accuracy	0.75 (0.382)	0.533 (0.34)	0.633 (0.482)	0.533 (0.499)
	Statistical test	✓	✓	✓	✓

(✓: the result is different from that of the proposed model with statistical significance (90% confidence level);
×: the result is not significantly different from that of the proposed model statistically;
the values in brackets are the standard deviations.)

Table 8
Ablation study results of multiple times (30 times) experiments for different examples.

Case		Numerical example	TEP IDV(1)	TEP IDV(8)	Cut-made process of cigarette
Complete model		0.9 (0.238)	0.917 (0.227)	0.833 (0.373)	0.967 (0.18)
$\beta_1 = 0$	Accuracy	0.717 (0.38)	0.767 (0.359)	0.633 (0.482)	0.533 (0.499)
	Statistical test	✓	✓	✓	✓
$\beta_2 = 0$	Accuracy	0.783 (0.247)	0.8 (0.279)	0.767 (0.423)	0.933 (0.249)
	Statistical test	✓	✓	×	×
Without the HAP mechanism	Accuracy	0.617 (0.402)	0.5 (0.316)	0.6 (0.49)	0.5 (0.5)
	Statistical test	✓	✓	✓	✓
With only one information fusion layer	Accuracy	0.783 (0.308)	0.783 (0.334)	0.633 (0.482)	0.833 (0.372)
	Statistical test	×	✓	✓	✓

(✓: the result is different from that of the complete model with statistical significance (90% confidence level);
×: the result is not significantly different from that of the complete model statistically;
the values in brackets are the standard deviations.)

Declaration of competing interest

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

Data availability

The link to the code and numerical case data can be found in the manuscript. Data availability link given as https://github.com/chunhuiz/MPGE-RootRank-for-root-cause-diagnosis. For data on the actual industrial case, the authors are not authorized to share for privacy reasons.

Acknowledgment

This work was supported by the National Natural Science Foundation of China[http://dx.doi.org/10.13039/501100001809] (No. 62125306).

References

Amornbunchornvej, C., Zheleva, E., & Berger-Wolf, T. (2021). Variable-lag granger causality and transfer entropy for time series analysis. *ACM Transactions on Knowledge Discovery from Data*, *15*(4), 1–30.

Arnold, A., Liu, Y., & Abe, N. (2007). Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 66–75).

Bandt, C., & Pompe, B. (2002). Permutation entropy: a natural complexity measure for time series. *Physical Review Letters*, *88*(17), Article 174102.

Barnett, L., Barrett, A. B., & Seth, A. K. (2009). Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical Review Letters*, *103*(23), Article 238701.

Bellot, A., Branson, K., & van der Schaar, M. (2021). Neural graphical modelling in continuous-time: consistency guarantees and algorithms. In *International conference on learning representations*.

Bore, J. C., Li, P., Harmah, D. J., Li, F., Yao, D., & Xu, P. (2020). Directed EEG neural network analysis by LAPPS ($p \leq 1$) penalized sparse Granger approach. *Neural Networks*, *124*, 213–222.

Chen, X. W., Anantha, G., & Lin, X. (2008). Improving Bayesian network structure learning with mutual information-based node ordering in the K2 algorithm. *IEEE Transactions on Knowledge and Data Engineering*, *20*(5), 628–640.

Chen, H. S., Yan, Z., Yao, Y., Huang, T. B., & Wong, Y. S. (2018). Systematic procedure for Granger-causality-based root cause diagnosis of chemical process faults. *Industrial and Engineering Chemistry Research*, *57*(29), 9500–9512.

Chen, J., & Zhao, C. (2022). Multi-lag and multi-type temporal causality inference and analysis for industrial process fault diagnosis. *Control Engineering Practice*, *124*, Article 105174.

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, *15*(1), 3741–3782.

Downs, J. J., & Vogel, E. F. (1993). A plant-wide industrial process control problem. *Computers & Chemical Engineering*, *17*(3), 245–255.

Duan, P., Chen, T., Shah, S. L., & Yang, F. (2014). Methods for root cause diagnosis of plant-wide oscillations. *AIChE Journal*, *60*(6), 2019–2034.

Duan, P., Yang, F., Chen, T., & Shah, S. L. (2013). Direct causality detection via the transfer entropy approach. *IEEE Transactions on Control Systems Technology*, *21*(6), 2052–2066.

Duan, S., Zhao, C., & Wu, M. (2022). Multiscale partial symbolic transfer entropy for time-delay root cause diagnosis in nonstationary industrial processes. *IEEE Transactions on Industrial Electronics*, http://dx.doi.org/10.1109/TIE.2022.3161761.

Fronczak, A., Fronczak, P., & Hołyst, J. A. (2004). Average path length in random networks. *Physical Review E*, *70*(5), Article 056110.
Gao, H., Xu, Y., & Zhu, Q. (2016). Spatial interpretive structural model identification and AHP-based multimodule fusion for alarm root-cause diagnosis in chemical processes. *Industrial and Engineering Chemistry Research*, *55*(12), 3641–3658.
Geweke, J. F. (1984). Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, *79*(388), 907–915.
Gharahbagheri, H., Imtiaz, S. A., & Khan, F. (2017). Root cause diagnosis of process fault using KPCA and Bayesian network. *Industrial and Engineering Chemistry Research*, *56*(8), 2054–2070.
Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 424–438.
He, Q. P., Qin, S. J., & Wang, J. (2005). A new fault diagnosis method using fault directions in Fisher discriminant analysis. *AIChE Journal*, *51*(2), 555–571.
He, F., Wang, C., & Fan, S. S. (2019). Fault detection and root cause analysis of a batch process via novel nonlinear dissimilarity and comparative granger causality analysis. *Industrial and Engineering Chemistry Research*, *58*(47), 21842–21854.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.
Jiang, H., Patwardhan, R., & Shah, S. L. (2009). Root cause diagnosis of plant-wide oscillations using the concept of adjacency matrix. *Journal of Process Control*, *19*(8), 1347–1354.
Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.
Kugiumtzis, D. (2013). Partial transfer entropy on rank vectors. *The European Physical Journal Special Topics*, *222*(2), 401–420.
Li, G., Qin, S. J., & Yuan, T. (2016). Data-driven root cause diagnosis of faults in process industries. *Chemometrics and Intelligent Laboratory Systems*, *159*, 1–11.
Liao, W., & Ji, Q. (2009). Learning Bayesian network parameters under incomplete data with domain knowledge. *Pattern Recognition*, *42*(11), 3046–3056.
Lindner, B., Auret, L., & Bauer, M. (2019). A systematic workflow for oscillation diagnosis using transfer entropy. *IEEE Transactions on Control Systems Technology*, *28*(3), 908–919.
Liu, Y., Chen, H. S., Wu, H., Dai, Y., Yao, Y., & Yan, Z. (2020). Simplified Granger causality map for data-driven root cause diagnosis of process disturbances. *Journal of Process Control*, *95*, 45–54.
Marinazzo, D., Pellicoro, M., & Stramaglia, S. (2008). Kernel method for nonlinear Granger causality. *Physical Review Letters*, *100*(14), Article 144103.
Montalto, A., Stramaglia, S., Faes, L., Tessitore, G., Prevete, R., & Marinazzo, D. (2015). Neural networks with non-uniform embedding and explicit validation phase to assess Granger causality. *Neural Networks*, *71*, 159–171.
Murphy, K. P. (2002). *Dynamic bayesian networks: representation, inference and learning*. Berkeley: University of California.
Peng, C., Lu, R., Kang, O., & Kai, W. (2020). Batch process fault detection for multi-stage broad learning system. *Neural Networks*, *129*, 298–312.
Raveendran, R., & Huang, B. (2018). Variational Bayesian approach for causality and contemporaneous correlation features inference in industrial process data. *IEEE Transactions on Cybernetics*, *49*(7), 2580–2590.
Raveendran, R., Huang, B., & Mitchell, W. (2020). A variational Bayesian causal analysis approach for time-varying systems. *IEEE Transactions on Control Systems Technology*, *29*(3), 1191–1202.
Robinson, J., & Hartemink, A. (2008). Non-stationary dynamic Bayesian networks. *Advances in Neural Information Processing Systems*, *22*, 1369–1376.
Robinson, J. W., Hartemink, A. J., & Ghahramani, Z. (2010). Learning non-stationary dynamic Bayesian networks. *Journal of Machine Learning Research*, *11*(12).
Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386.
Ross, A., & Willson, V. L. (2018). *Basic and advanced statistical tests: writing results sections and creating tables and figures*. Springer.
Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., ... Zscheischler, J. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, *10*(1), 1–13.
Schäck, T., Muma, M., Feng, M., Guan, C., & Zoubir, A. M. (2017). Robust nonlinear causality analysis of nonstationary multivariate physiological time series. *IEEE Transactions on Biomedical Engineering*, *65*(6), 1213–1225.
Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, *85*(2), 461.
Siggiridou, E., & Kugiumtzis, D. (2015). Granger causality in multivariate time series using a time-ordered restricted vector autoregressive model. *IEEE Transactions on Signal Processing*, *64*(7), 1759–1773.
Song, P., & Zhao, C. (2022). Slow down to go better: A survey on slow feature analysis. *IEEE Transactions on Neural Networks and Learning Systems*, http://dx.doi.org/10.1109/TNNLS.2022.3201621.
Song, P., Zhao, C., Huang, B., & Wu, M. (2022). Sparse and time-varying predictive relation extraction for root cause quantification of nonstationary process faults. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–13.
Staniek, M., & Lehnertz, K. (2008). Symbolic transfer entropy. *Physical Review Letters*, *100*(15), Article 158101.
Su, J., Zhang, H., Ling, C. X., & Matwin, S. (2008). Discriminative parameter learning for bayesian networks. In *Proceedings of the 25th international conference on machine learning* (pp. 1016–1023).
Tank, A., Covert, I., Foti, N., Shojaie, A., & Fox, E. B. (2021). Neural Granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, http://dx.doi.org/10.1109/TPAMI.2021.3065601.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.*, *58*(1), 267–288.
Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max–min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, *65*(1), 31–78.
Van den Kerkhof, P., Vanlaer, J., Gins, G., & Van Impe, J. F. (2013). Analysis of smearing-out in contribution plot based fault isolation for statistical process control. *Chemical Engineering Science*, *104*, 285–293.
Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 4–24.
Yan, Z., Yao, Y., Huang, T. B., & Wong, Y. S. (2018). Reconstruction-based multivariate process fault isolation using Bayesian Lasso. *Industrial and Engineering Chemistry Research*, *57*(30), 9779–9787.
Yu, J., & Rashid, M. M. (2013). A novel dynamic bayesian network-based networked process monitoring approach for fault detection, propagation identification, and root cause diagnosis. *AIChE Journal*, *59*(7), 2348–2365.
Yu, W., & Zhao, C. (2017). Sparse exponential discriminant analysis and its application to fault diagnosis. *IEEE Transactions on Industrial Electronics*, *65*(7), 5931–5940.
Yu, W., Zhao, C., & Huang, B. (2021). MoniNet with concurrent analytics of temporal and spatial information for fault detection in industrial processes. *IEEE Transactions on Cybernetics*, *52*(8), 8340–8351.
Yuan, T., & Qin, S. J. (2014). Root cause diagnosis of plant-wide oscillations using Granger causality. *Journal of Process Control*, *24*(2), 450–459.
Zhang, Q. (2015). Dynamic uncertain causality graph for knowledge representation and reasoning: continuous variable, uncertain evidence, and failure forecast. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *45*(7), 990–1003.
Zhao, C. (2022). Perspectives on nonstationary process monitoring in the era of industrial artificial intelligence. *Journal of Process Control*, *116*, 255–272.
Zhao, C., Chen, J., & Jing, H. (2020). Condition-driven data analytics and monitoring for wide-range nonstationary and transient continuous processes. *IEEE Transactions on Automation Science and Engineering*, *18*(4), 1563–1574.
Zhao, C., & Gao, F. (2016). Critical-to-fault-degradation variable analysis and direction extraction for online fault prognostic. *IEEE Transactions on Control Systems Technology*, *25*(3), 842–854.
Zhao, H., Lai, Z., & Chen, Y. (2019). Global-and-local-structure-based neural network for fault detection. *Neural Networks*, *118*, 43–53.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B.*, *67*(2), 301–320.