Counterfactual Data Generation method for Fault Diagnosis of complex Electromechanical systems

Chong Wang, Jie Liu, Junwei Cao, Xi Chen, Lihua Chen, Yindong Ji

***Abstract*—The high reliability and maintainability of complex electromechanical systems make the collection of representative fault data difficult in routine operations. This issue poses significant challenges to data-driven models, hindering their ability to accurately capture system fault mechanisms. Generating simulated fault samples is a popular and effective approach to augment fault data. However, the state-of-art data generation methods can only produce synthetic data with respect to the distribution of the collected data, and may violate the actual fault mechanisms. To address this, this paper proposes a counterfactual data generation method grounded in causality, aiming to simulate the intrinsic generation process of monitoring data and thereby obtain counterfactual data that conforms to the system's fault mechanisms. In the proposed method, a priori-constrained causal discovery method is designed to uncover the causalities among the monitoring variables in complex electromechanical systems. A graph decoupling network is designed to disentangle causal mechanisms and extract decoupled features representing different uncertainty sources of the monitoring data. Finally, a causality-based generative adversarial network is proposed to generate counterfactual monitoring data that satisfy the mined causalities. The experimental results show that the generalizability and stability of fault diagnosis models can be enhanced with the counterfactual data.**

***Index Terms*—Causality; Counterfactual data generation; Decoupling causal representation; Electromechanical systems; Fault diagnosis**

Symbols

$\boldsymbol{D}$	the collected monitoring data of a complex system
$\boldsymbol{X}$	the monitoring variables
$\boldsymbol{Y}$	the system status variable
$\boldsymbol{C}$	the system monitoring parameters
$\boldsymbol{U}$	the component-level degradation status representation
$\boldsymbol{\varepsilon}$	the environmental factors
$\boldsymbol{M}$	the system operation mode
$\boldsymbol{A}^{\mathrm{C}}$	the causal adjacency matrix of the collected variables (the monitoring variables and the system status variable)
$\boldsymbol{A}^{\mathrm{S}}$	the causal strength matrix of the collected variables
$\boldsymbol{A}'$	the causal strength matrix of the system monitoring parameters $\boldsymbol{C}$
$\boldsymbol{D}'$	the generated counterfactual data $\boldsymbol{D}'=(\boldsymbol{C}',\boldsymbol{M},\boldsymbol{\varepsilon},\boldsymbol{Y}')$
$\boldsymbol{A}''$	the causal adjacency matrix of the generated counterfactual data

This work was supported by the National Key Research and Development Program of China under Grant 2022YFB4300500 and the Natural Science Foundation of China under Grants 62233012. (*Corresponding author: Jie Liu*).

Chong Wang is now with the Department of Automation in Tsinghua University, No. 30 Shuangqing Road, Haidian, Beijing, China (e-mail: chongzi@buaa.edu.cn)

Jie Liu is now with the School of Reliability and System Engineering in Beihang University, 37 Xueyuan Road, Haidian, Beijing, China (e-mail: liujie805@buaa.edu.cn).

Junwei Cao is now with the Beijing National Research Center for Information Science and Technology, Tsinghua University, No. 30 Shuangqing Road, Haidian, Beijing, China (e-mail: junweicao@gmail.com).

Xi Chen is now with the School of Information Science and Technology, Southwest Jiaotong University, 111 Second Ring Road, Chengdu, Sichuan Province, China (e-mail: cx@grci.com.cn).

Lihua Chen is now with the School of Information Science and Technology, Southwest Jiaotong University, 111 Second Ring Road Chengdu, Sichuan Province, China (e-mail: chenlihua5678@qq.com).

Yindong Ji is now with the Department of Automation in Tsinghua University, No. 30 Shuangqing Road, Haidian, Beijing, China (e-mail: jyd@tsinghua.edu.cn).

I. INTRODUCTION

COMPLEX electromechanical systems refer to industrial systems composed of multiple interacted mechanical and electronic components, which are highly intelligent and automated. Nowadays, such complex electromechanical systems have been widely used in aerospace, rail transit, national defense, intelligent manufacturing, etc. [1]. Considering the high economic and reputational losses brought by severe failures, fault diagnosis has attracted widespread attention [2].

The large amount of operation data has led to the widespread application of data-driven methods in fault diagnosis of electromechanical systems [3][4]. Among these methods, supervised learning models with outstanding fitting capabilities, such as support vector machines (SVMs) [5], convolutional neural networks (CNNs) [6], deep neural networks (DNNs) [7], and artificial neural networks (ANNs) [8], have garnered significant attention and positive outcomes have been reported in the published works.

However, the prerequisite for achieving high-precision of the previous models is the sufficient and reliable failure data, which is often difficult to satisfy in practical applications. For highly reliable complex electromechanical systems, limited failure data are available and supervised methods can hardly capture the actual failure mechanisms, ultimately resulting in poor model generalization performance.

To overcome this limitation, generating diverse and representative simulated fault data has emerged as a popular and effective approach. These simulated data closely mimic real data but may represent different fault scenarios, thus can provide a more comprehensive coverage of system faults. By incorporating these fault samples into the training process, the models can learn to recognize a wider range of faults, thereby enhancing their generalization performance [9].

In recent years, different synthetic data generation methods have been proposed and adopted in fault diagnosis. These methods can be roughly classified into statistical methods and generative methods. Synthetic minority over-sampling technique (SMOTE) is a representative statistical method to

generate synthetic data based on the Euclidean spatial distribution of the monitored real data [9]. The variants of SMOTE, such as BorderlineSMOTE1 and BorderlineSMOTE2, focus solely on over-sampling around the borderline fault sample to enhance the classification decision boundary [10]. Furthermore, adaptive synthetic (ADASYN) can adaptively shift the classification decision boundary by employing a weighted distribution for different fault samples [11]. A classical generative model is the generative adversarial network (GAN) which is capable of generating simulation data that conforms to the distribution of real fault data through adversarial learning [12][13]. Moreover, certain modified GAN approaches have further improved the efficiency and effectiveness of data generation process [14][15][16]. However, most of the existing studies aimed at obtaining synthetic samples similar to the distributions of the collected fault samples. The generated data are similarly distributed to the real data, but may inadvertently represent spurious failure mechanisms. Consequently, this would render it more challenging for fault diagnosis models to capture the actual failure mechanisms.

Recently, some researchers have explored the capability of causality in promoting synthetic data generation process, i.e. counterfactual data generation. For example, in [17], the image-data generation process is decomposed into independent causal representations (latent factors), including object shape, object texture and background. Based on the assumption that the intrinsic generation mechanism of images remains constant, counterfactual data is generated by altering these causal representations. Specifically, counterfactual images are obtained by changing the textures or backgrounds of original images. By adding the counterfactual images to the model training process, the image classifier can more effectively learn the key factor (i.e. the object shape). These works inspire the authors to integrate causality into synthetic fault data generation. The existing methods focus on image recognition. For complex systems fault diagnosis, the monitoring data are generally numerical, making the current approaches for image recognition unsuitable for this purpose.

Based on the assumption that the system causality and the failure mechanisms are invariant, a novel counterfactual data generation method is proposed in this paper. This method incorporates a priori-constrained causal discovery method, which integrates prior knowledge into the causal discovery to mine the causal network of complex electromechanical systems. Based on the causal network, a graph decoupling network is designed to decouple latent factors, i.e. the component-level degradation state representations. The decoupling operation is the inverse process of the information propagation, aiming to intuitively extract independent sources of system faults. Finally, a causality-based generative adversarial network (CGAN) is presented to generate counterfactual data consistent to system causality. The generator creates simulation data by selectively manipulating the degradation state representations, while the discriminator assesses the consistency of the generated data with the system's causality. The effectiveness of the proposed method is demonstrated through experiments using simulation data from aviation software and real-world data from a high-speed train braking system. Since direct evaluation of the numerical data is challenging, the generated data are integrated into the training process of fault diagnostic models to indirectly assess their quality. The generalization performance and model stability are employed to evaluate the enhancing effect of the generated data on the monitoring data.

The rest of this paper is organized as follows. The proposed counterfactual data generation method is introduced in Section 2. Section 3 illustrates the experiment results on simulation data and real-world data. In Section 4, conclusions are drawn and further research directions are discussed.

II. Counterfactual Data Generation Based on Causality

As shown in Fig.1, the proposed counterfactual data generation method includes three main parts: ① priori-constrained causal discovery for complex electromechanical systems; ② graph decoupling network for causal mechanisms decoupling; ③ causal-based generative adversarial network (CGAN) for counterfactual data generation. By adding prior constraints to the causal discovery process, prior knowledge of the complex electromechanical system is integrated into the data generated by the proposed method.

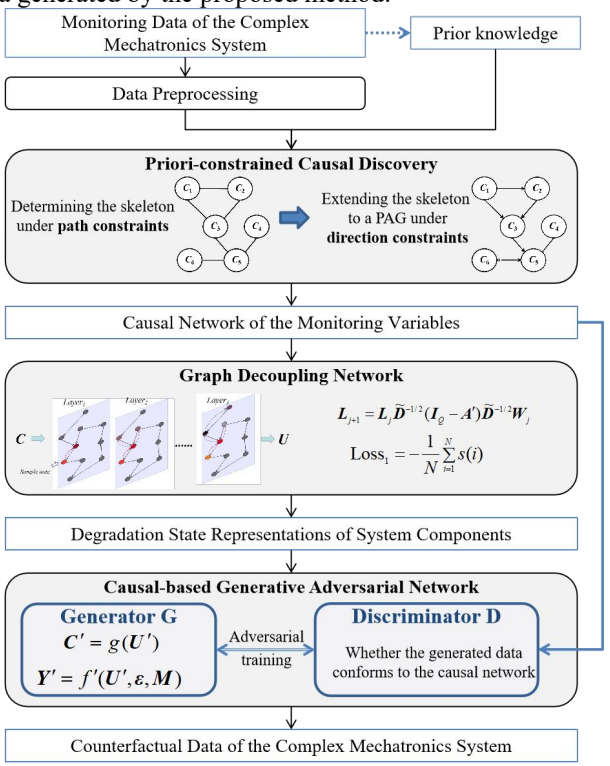


Fig. 1. The proposed counterfactual data generation method.

A. Priori-constrained causal discovery for complex electromechanical systems

Causality can reveal the most fundamental relationship between objects, such as the factors that affect the system failure [18]. Nowadays, there have been a variety of causal discovery algorithms, which can be separated into two types: constraint-based and score-based [19][20]. The constraint-based algorithms explore causality based on conditional independence constraints, while the score-based algorithms

construct the causal structure by maximizing a predefined global score. In this paper, the constraint-based algorithms are considered because of their strong mathematical foundation and relatively stable output [21].

PC (Peter and Clark) algorithm [22] and fast causal inference (FCI) algorithm[23] are two commonly used constraint-based algorithms. The PC algorithm assumes that there are no unobserved confounders, i.e. there are no unobserved common causes for two connected variables. The FCI algorithm, as a modification of the PC algorithm, is used in the presence of latent confounders and selection bias [23][24]. Considering that it is impossible to identify confounders and selection bias in practical applications, especially for fault diagnosis of complex electromechanical systems, the FCI algorithm is used as the basic causal discovery method. Thus, the priori-constrained causal discovery proposed in this work refers to the application of prior constraints to the FCI algorithm.

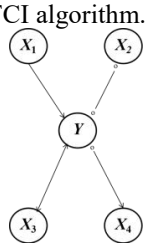


Fig. 2. An example causal PAG.

The output of the FCI algorithm is normally a partial ancestral graph (PAG), as shown in Fig.2. The possible relationships between two nodes (e.g. A and B) in a PAG are shown in Fig.3.

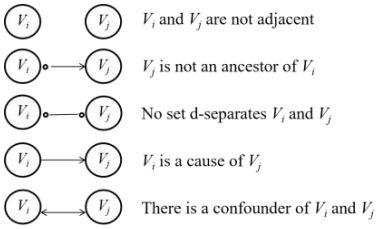


Fig. 3. The possible relationships between two nodes in a PAG.

The considered priori constraints in this work are divided into the following two types:

① **Path constraint.** Constrain that there is no direct causal relationship between variable A and variable B, thus there are no edges between node A and node B in the causal network.

② **Direction constraint.** Constrain that variable A is the cause of variable B, thus there are causal paths from A to B in the causal network. The causal path is the combination of linked edges from cause node to result node, which includes direct causal paths and indirect causal paths.

The FCI algorithm can be divided into two parts: the skeleton determination and the edge orientation. The details can be seen in [24]. The priori-constrained FCI is achieved by the following two steps:

① **Determining the skeleton under path constraints.** By deleting the edges corresponding to the path constraints in the first part of the FCI algorithm, this step can be implemented as shown in Fig.4.

② **Extending the skeleton to a PAG under direction constraints.** This step is implemented by adding orientation rules as shown in Fig.5.

Algorithm 1 Determining the skeleton under path constraints

INPUT: Node (variable) set V, Conditional independence information
OUTPUT: Estimated skeleton C, separation sets S (only needed when directing the skeleton afterwards)
From the complete undirected graph $\widetilde{C}$, in which there is an edge $\circ\!\!-\!\!\circ$ between every pair of nodes in the node set V.
Delete all the path constraint edges $i\circ\!\!-\!\!\circ j$ in the undirected graph $\widetilde{C}$
Denote this new undirected graph by $\widetilde{C}$
$l = -1$; $C = \widetilde{C}$
repeat
  $l = l+1$
  repeat
    Select a new ordered pair of nodes i, j that are adjacent in C such that $|adj(C,i)\setminus\{j\}| \geq l$
    repeat
      Choose new $\boldsymbol{k} \in adj(C,i)\setminus\{j\}$ with $|\boldsymbol{k}| = l$.
      if i and j are conditionally independent given $\boldsymbol{k}$ **then**
        Delete edge $i\circ\!\!-\!\!\circ j$
        Denote this new graph by C
        Save $\boldsymbol{k}$ in $S(i,j)$ and $S(j,i)$
      end if
    until edge $i\circ\!\!-\!\!\circ j$ is deleted or all $\boldsymbol{k} \in adj(C,i)\setminus\{j\}$ with $|\boldsymbol{k}| = l$ have been chosen
    until all ordered pairs of adjacent nodes i and j such that $|adj(C,i)\setminus\{j\}| \geq l$ and $\boldsymbol{k} \in adj(C,i)\setminus\{j\}$ with $|\boldsymbol{k}| = l$ have been tested for conditional independence
until for each ordered pairs of adjacent nodes i, j: $|adj(C,i)\setminus\{j\}| < l$.

Fig. 4. The process of determining the skeleton under path constraints.

Algorithm 2 Extending the skeleton to a PAG under direction constraints

INPUT: Skeleton G_{skel}, separation sets S
OUTPUT: PAG G
for all pairs of non-adjacent nodes i, j with common neighbour k **do**
  if $k \notin S(i,j)$ **then**
    Replace $i\circ\!\!-\!\!\circ k\circ\!\!-\!\!\circ j$ in G_{skel} by $i\circ\!\!\rightarrow k\leftarrow\!\!\circ j$
  end if
end for
for all pairs of direction constraint nodes i, j and path $p = \langle i, \ldots, j\rangle$ **do**
    Orient $\circ\!\!-\!\!\circ$ into $\circ\!\!\rightarrow$ whenever there are only $\circ\!\!-\!\!\circ$, $\circ\!\!\rightarrow$ and $\rightarrow$ in path p.
    Orient $\leftarrow\!\!\circ$ into $\leftrightarrow$ and orient $\circ\!\!-\!\!\circ$ into $\circ\!\!\rightarrow$ whenever there is $\leftarrow\!\!\circ$ in path p.
    Delete the edge with the lowest ACE in p whenever there is $\leftarrow$ in path p.
  end if
end for
In the resulting PAG, try to orient edges as many as possible by repeated application of the orientation rules *R1-R10* in [20].

Fig. 5. The process of extending the skeleton to a PAG under direction constraints.

Assuming that the monitoring data of a complex electromechanical system is $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K, \boldsymbol{Y})$, $\boldsymbol{X}_i \in \boldsymbol{R}^{N\times 1}$ (i= 1, 2, ... , K) and $\boldsymbol{Y} \in \boldsymbol{R}^{N\times 1}$. $X_1, X_2, \ldots, X_K$ are the monitoring variables, and Y is the system status variable. In this work, these above variables (the monitoring variables and the system status variable) are marked as $D_1, D_2, \ldots, D_{K+1}$.

Using the priori-constrained FCI algorithm, the obtained causal network of the complex electromechanical system can be expressed by the causal adjacency matrix $\boldsymbol{A}^{\mathrm{c}}$:

$$\boldsymbol{A}^{\mathrm{c}} = \begin{bmatrix} a^{\mathrm{c}}{}_{11} & a^{\mathrm{c}}{}_{12} & \cdots & a^{\mathrm{c}}{}_{1(K+1)} \\ a^{\mathrm{c}}{}_{21} & a^{\mathrm{c}}{}_{22} & \cdots & a^{\mathrm{c}}{}_{2(K+1)} \\ \vdots & \vdots & \ddots & \vdots \\ a^{\mathrm{c}}{}_{(K+1)1} & a^{\mathrm{c}}{}_{(K+1)2} & \cdots & a^{\mathrm{c}}{}_{(K+1)(K+1)} \end{bmatrix} \quad (1)$$

where, for the relationships between two nodes D_i and D_j, if $D_i \rightarrow D_j$ or $D_i \leftrightarrow D_j$ then $a^c_{ij} = 1$, if $D_i \circ\!\!\rightarrow D_j$ or $D_i \circ\!\!-\!\!\circ D_j$ then $a^c_{ij} = 0.5$, otherwise $a^c_{ij} = 0$.

To quantitatively describe the causality between two variables, the causal strength matrix $\boldsymbol{A}^S$ can be obtained by replacing the values in $\boldsymbol{A}^c$ with the average causal effect (ACE). Currently, there are already some ready-made methods for calculating ACE, and the focus of this article is not on ACE. In this work, ACE are directly estimated based on the *model.estimate_effect* function of the python package *DoWhy*. $\mathrm{ACE}(D_i, D_j)$ is defined as the ACE value from D_i to D_j.

$$\boldsymbol{A}^S = \begin{bmatrix} a^S_{11} & a^S_{12} & \cdots & a^S_{1(K+1)} \\ a^S_{21} & a^S_{22} & \cdots & a^S_{2(K+1)} \\ \vdots & \vdots & \ddots & \vdots \\ a^S_{(K+1)1} & a^S_{(K+1)2} & \cdots & a^S_{(K+1)(K+1)} \end{bmatrix} \tag{2}$$

where, a^S_{ij} is the causal strength from D_i to D_j, if $D_i \rightarrow D_j$ or $D_i \leftrightarrow D_j$ then

$$a^S_{ij} = \mathrm{ACE}(D_i, D_j) \tag{3}$$

if $D_i \circ\!\!\rightarrow D_j$ or $D_i \circ\!\!-\!\!\circ D_j$ then

$$a^S_{ij} = \frac{1}{2}\mathrm{ACE}(D_i, D_j) \tag{4}$$

otherwise $a^S_{ij} = 0$.

In addition, when faced with an abundance of monitoring variables lacking practical significance and prior knowledge, directly extracting causal relationships becomes arduous, and the results are often unverifiable. Considering that complex systems typically possess prior knowledge regarding the relationships between their subsystems, principal component analysis (PCA) [25] is conducted at the subsystem level first, and then the priori constrained FCI algorithm is used to mine the causal relationship between the subsystems. PCA assumes that the data can be represented in a lower-dimensional space while retaining most of the original information, making it particularly effective in dealing with high-dimensional data that may contain redundant and correlated features. In complex systems, the monitoring variables are often redundant and correlated, allowing for a comprehensive monitoring of the system's health status. It is therefore reasonable to utilize PCA to reduce data dimensionality and extract the principal components at the subsystem level. The obtained principal components are independent, which can significantly simplify the causal analysis of the complex electromechanical system. This variable reduction and fusion process based on PCA is further described in the application cases.

B. Graph decoupling network for causal mechanisms decoupling

Generally, the causal information of complex electromechanical systems related to fault diagnosis can be roughly described as shown in Fig.6. The variables in the solid wireframes are usually observable, while the variables in the dashed wireframe (i.e. the degradation status of each component) are unobservable. In Fig.6, $\boldsymbol{C}$ is the component monitoring parameters collected by the set sensors, M is the system operation mode, $\boldsymbol{\varepsilon}$ is the environmental factors, and $\boldsymbol{U}$ represents the degradation status of the system components. Fault diagnosis focuses on the health status of individual components. Unexpected failures are normally excluded from preventive fault diagnosis research. Instantaneous external environmental factors merely cause measurement errors, rather than impacting component health or system operational modes [26]. In this case, the current environmental factors $\boldsymbol{\varepsilon}$ shown in Fig.6 are generally deemed not to exert a direct influence on other variables. Uncertain factors are widespread and often associated with environmental factors. This paper treats environmental factors as representations of those uncertain factors that influence system fault status.

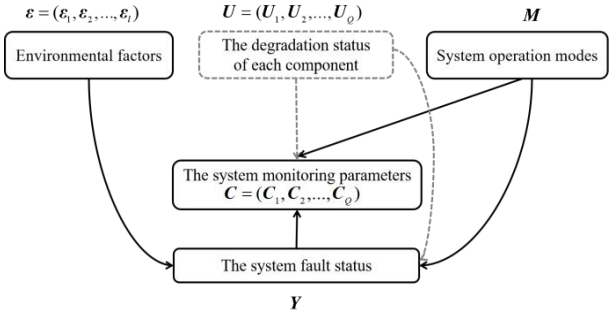


Fig. 6. The general causal network of complex electromechanical systems.

For the monitoring data $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{Y})$, in which $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K) = (\boldsymbol{C}, \boldsymbol{\varepsilon}, \boldsymbol{M})$, the causal network of the system monitoring parameters $\boldsymbol{C}$ can be expressed by (5). Decoupling causal mechanisms for fault diagnosis is essentially to excavate the unobserved variables that directly affect the system fault status. Therefore, the goal of this part is to explore the causal representations of component-level degradation status based on the observation data (i.e. observed variables).

$$\boldsymbol{A}' = \begin{bmatrix} a^S_{11} & a^S_{12} & \cdots & a^S_{1Q} \\ a^S_{21} & a^S_{22} & \cdots & a^S_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ a^S_{Q1} & a^S_{Q2} & \cdots & a^S_{QQ} \end{bmatrix} \tag{5}$$

Indeed, $\boldsymbol{A}'$ is the matrix formed by the first Q rows and the first Q columns of $\boldsymbol{A}^S$.

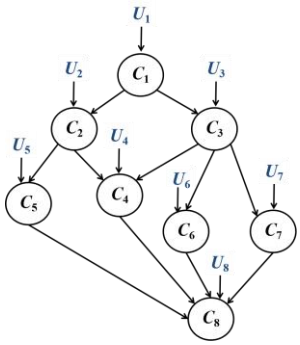


Fig. 7. An example causal network of the system monitoring parameters.

For a complex electromechanical system with single operation model, it is assumed that the causal relationship between the system monitoring parameters (or the subsystem principal components) is shown in Fig.7. According to the structural equation model (SEM) in causal theory [18], the

monitoring value of a node is influenced by its health status and the monitoring values of its father nodes. Thus, the causality in this system can be described as:

$$\begin{cases} C_1 = f_1(U_1) \\ C_2 = f_2(C_1, U_2) \\ C_3 = f_3(C_1, U_3) \\ C_4 = f_4(C_2, C_3, U_4) \\ C_5 = f_5(C_2, U_5) \\ C_6 = f_6(C_3, U_6) \\ C_7 = f_7(C_3, U_7) \\ C_8 = f_8(C_4, C_5, C_6, C_7, U_8) \end{cases} \tag{6}$$

where C_i is the i-th monitoring parameter, U_i is the degradation status representation of the component corresponding C_i.

Under the assumption that the causal mechanisms between the components are all linear, (6) can also be expressed as (7).

$$\begin{cases} C_1 = U_1 \\ C_2 = U_2 + a^{\mathrm{S}}{}_{12} C_1 \\ C_3 = U_3 + a^{\mathrm{S}}{}_{13} C_1 \\ C_4 = U_4 + a^{\mathrm{S}}{}_{24} C_2 + a^{\mathrm{S}}{}_{34} C_3 \\ C_5 = U_5 + a^{\mathrm{S}}{}_{25} C_2 \\ C_6 = U_6 + a^{\mathrm{S}}{}_{36} C_3 \\ C_7 = U_7 + a^{\mathrm{S}}{}_{37} C_3 \\ C_8 = U_8 + a^{\mathrm{S}}{}_{48} C_4 + a^{\mathrm{S}}{}_{58} C_5 + a^{\mathrm{S}}{}_{68} C_6 + a^{\mathrm{S}}{}_{78} C_7 \end{cases} \tag{7}$$

In where, the coefficient of U_i and constant term in linear regression models are merged into U_i, because $U_1, U_2, ..., U_8$ is a group of representations which will not change the actual meanings after linear transformation.

(7) can also be expressed in matrix operation form as:

$$\boldsymbol{C} = \boldsymbol{U} + \boldsymbol{A}'\boldsymbol{C} \tag{8}$$

From (8), the degradation status representation $\boldsymbol{U}$ can be calculated by:

$$\boldsymbol{U} = (C_1, C_2, ..., C_8)(\boldsymbol{I}_Q - \boldsymbol{A}') = \boldsymbol{C}(\boldsymbol{I}_Q - \boldsymbol{A}') \tag{9}$$

where $\boldsymbol{I}_Q$ is the Q-dimensional unit matrix.

Of course, the causal relationships in complex electromechanical systems are predominantly nonlinear. Directly decoupling these nonlinear causal mechanisms is challenging. Inspired by traditional graph convolutional networks, which iteratively learn complex relationships using linear structures [27], this paper repeatedly employs linear causal decoupling to address nonlinear causal mechanisms. By incorporating trainable weights into each layer of linear causal decoupling, the nonlinear causal mechanisms can be approximated and decoupled, yielding the component-level degradation status representation $\boldsymbol{U}$. $\boldsymbol{U}$ is expected to distinguish different faults under various operational modes. Consequently, the optimization objective of $\boldsymbol{U}$ is to make the monitoring samples of different ($\boldsymbol{Y}$, $\boldsymbol{M}$) combinations distinguishable. In this case, referring to the silhouette coefficient used in clustering tasks [28], the loss function for optimizing $\boldsymbol{U}$ is:

$$\mathrm{Loss}_1 = -\frac{1}{N}\sum_{i=1}^{N} s(i) \tag{10}$$

where $s(i)$ is the silhouette coefficient [28] of the i-th sample.

As shown in Fig.8, a graph decoupling network is designed in this work to decouple the degradation status representation, and the j-th layer operation process of the network is:

$$\boldsymbol{L}_{j+1} = \boldsymbol{L}_j \widetilde{\boldsymbol{D}}^{-1/2}(\boldsymbol{I}_Q - \boldsymbol{A}')\widetilde{\boldsymbol{D}}^{-1/2}\boldsymbol{W}_j \tag{11}$$

where $\widetilde{\boldsymbol{D}}^{-1/2} \in \boldsymbol{R}^{Q\times Q}$ is the diagonal matrix for feature normalization; the i-th row and i-th column element of $\widetilde{\boldsymbol{D}}^{-1/2}$ is $1\Big/\sqrt{\left|1-\sum_{t=1}^{Q} a^{\mathrm{S}}{}_{it}\right|}$; $\sum_{t=1}^{Q} a^{\mathrm{S}}{}_{it}$ is the sum of the i-th row elements in $\boldsymbol{A}'$; $\boldsymbol{W}_j \in \boldsymbol{R}^{Q\times Q}$ is a trainable regular matrix for feature coordinate rotation; $\boldsymbol{L}_j \in \boldsymbol{R}^{N\times Q}$ and $\boldsymbol{L}_{j+1} \in \boldsymbol{R}^{N\times Q}$ are respectively the input and output of the j-th layer. Moreover, $\boldsymbol{C}$ is the input of the first layer and $\boldsymbol{U}$ is the output of the last layer.

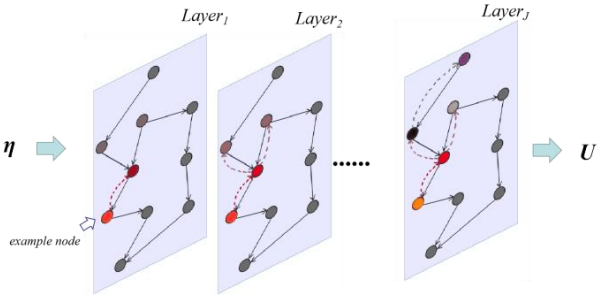


Fig. 8. The structure of the graph decoupling network.

By optimizing the number of layers and the feature weight matrix of each layer to minimize the loss function shown in (10), an interpretable degradation status representation $\boldsymbol{U}$ can be obtained.

In addition, a graph decoder is designed and trained to convert the degradation status representation into the system monitoring parameters. With the same number of layers as the decoupling network above, the i-th layer of the decoder is:

$$\boldsymbol{L}_{i+1} = \boldsymbol{L}_i \widetilde{\boldsymbol{D}}^{1/2}(\boldsymbol{I}_Q - \boldsymbol{A}')^{-1}\widetilde{\boldsymbol{D}}^{1/2}\boldsymbol{W}_i' \tag{12}$$

where $\boldsymbol{W}_i' \in \boldsymbol{R}^{Q\times Q}$ is the trainable weigh matrix, $\boldsymbol{L}_i \in \boldsymbol{R}^{N\times Q}$ and $\boldsymbol{L}_{i+1} \in \boldsymbol{R}^{N\times Q}$ are respectively the input and output of the i-th layer. $\boldsymbol{U}$ is the input of the first layer, and the estimated value of $\boldsymbol{C}$ (denoted by $\boldsymbol{C}'$) is the output of the last layer.

The decoder is trained by optimizing $\boldsymbol{W}_i'$ to minimize the gap between $\boldsymbol{C}'$ and $\boldsymbol{C}$. The loss function of the decode is shown in (13). In this work, the trained decoder is recorded as $\boldsymbol{C}' = g(\boldsymbol{U})$.

$$\mathrm{Loss}_2 = \sum_{i=1}^{Q}\left\|\boldsymbol{C}_i' - \boldsymbol{C}_i\right\|_2 = \sum_{i=1}^{Q}\sum_{j=1}^{N}(c_{ij}' - c_{ij})^2 \tag{13}$$

According to the general causal network shown in Fig.6, under a single operation mode, the system fault status $\boldsymbol{Y}$ is determined by the degradation status representation $\boldsymbol{U}$ and the environmental factors $\boldsymbol{\varepsilon}$, as shown in (14).

$$\boldsymbol{Y} = f(\boldsymbol{U}, \boldsymbol{\varepsilon}) \tag{14}$$

For a system with τ operation modes, the causal mechanism model (CMM) of fault occurrence is recorded as:

$$\boldsymbol{Y} = \begin{cases} f_1(\boldsymbol{U},\boldsymbol{\varepsilon}), \boldsymbol{M} = M_1 \\ f_2(\boldsymbol{U},\boldsymbol{\varepsilon}), \boldsymbol{M} = M_2 \\ \cdots \\ f_\tau(\boldsymbol{U},\boldsymbol{\varepsilon}), \boldsymbol{M} = M_\tau \end{cases} \triangleq f'(\boldsymbol{U},\boldsymbol{\varepsilon},\boldsymbol{M}) \tag{15}$$

Where $f'(*)$ represents the CMM of fault occurrence, i.e. the fault diagnosis model based on the degradation status representation $\boldsymbol{U}$. The CMM can be gained by training a group of classifiers (for example, Logistic regression [29], SVM [5], neural networks [6][7], etc).

C. CGAN for counterfactual data generation

By assuming that the CMM is stable, a causal-based generative adversarial network (CGAN) is designed in this part to generate counterfactual data. The proposed CGAN contains a generator and a discriminator, as shown in Fig.9. The generator consists of the decoder and the CMM, while the discriminator is based on the traditional FCI algorithm to judge whether the generated data conforms to the system causality. This process can make the expanded data, i.e. the new model training data, reflect more practical causal information.

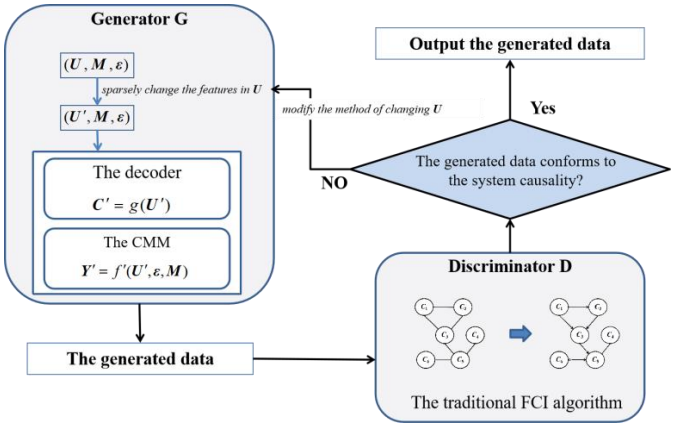


Fig. 9. The structure of the CGAN.

Generally, the counterfactual data are required not to be too different from the actual data, otherwise it is difficult to assess their credibility. Considering that in fault diagnosis tasks, the instantaneous monitoring values of the environmental factors have little impact on system fault status, and the system operation modes are fixed. Therefore, in the generator, the counterfactual data are generated by sparsely changing the features in $\boldsymbol{U}$. Specifically, this entails randomly replacing one or two features in $\boldsymbol{U}$ with white noise. Modifying the method of changing $\boldsymbol{U}$, as presented in Fig.9, means selecting different target features for alteration. This process is inspired by common counterfactual image generation methods, but it differs from simple feature rotation or flipping by using white noise replacement to ensure data diversity.

The features in $\boldsymbol{U}$, which represent the component degradation status, do not exhibit any direct causal relationships among themselves. Thus, changing one feature in $\boldsymbol{U}$ will not affect the values of other features. When one or two features are randomly changed, the difference between the counterfactual features and the original features is very small, thus preserving the similarity between the counterfactual data and the original data. Furthermore, sparsely replacing these features with white noise can obtain diverse and counterfactual health state values. The fault status of the counterfactual data can be determined based on the range of the component-level degradation status representation. To streamline this process, the CMM is employed to determine the fault labels of the counterfactual data.

To conform the generated data to the system causal network, the loss function of the counterfactual generation network is:

$$\text{Loss}_2 = \left\| \boldsymbol{A}'' - \boldsymbol{A}^{\text{C}} \right\|_2 \tag{16}$$

where $\boldsymbol{A}''$ is the causal adjacency matrix of the generated counterfactual data $\boldsymbol{D}' = (\boldsymbol{C}', \boldsymbol{M}, \boldsymbol{\varepsilon}, \boldsymbol{Y}')$ based on the traditional FCI algorithm, $\boldsymbol{A}^{\text{c}}$ is the causal adjacency matrix of the complex electromechanical system.

III. CASE STUDIES

A. Application results on simulation data

In the case of single operation mode, the simulation data of an airborne software system is used to verify the effectiveness of the proposed method. Ideally, only key components are monitored, each component outputs a single monitoring parameter. The causal strength matrix $\boldsymbol{A}'$ of the system key components is assumed to be a sparse upper triangular matrix. This is because the information transmission in the airborne software system is usually sequential according to time, and there are usually no cycle structures in the instantaneous monitoring. For example, the causality in a virtual system with 5 components in Fig.10 can be defined by $\boldsymbol{A}_5$ in (17).

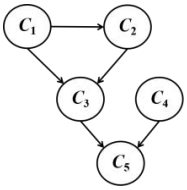


Fig. 10. The causal network of a simulation system with 5 components.

$$\boldsymbol{A}_5 = \begin{bmatrix} 0 & 0.6 & 0.5 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0.7 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{17}$$

Considering that there are usually multitudinous components in complex electromechanical systems, the causal strength matrices of virtual systems are not specifically listed here, but named as $\boldsymbol{A}_5$, $\boldsymbol{A}_{10}$, $\boldsymbol{A}_{20}$, etc. The monitoring value of a component C_i is affected by its parent node components and its own degenerate status representation U_i. When the causal effects between the components are all linear, the monitoring parameters of the virtual complex electromechanical system can be obtained by (18).

$$\boldsymbol{C} = \boldsymbol{U}(\boldsymbol{I}_Q - \boldsymbol{A}')^{-1} \tag{18}$$

In general, when any degradation status representation U_i is greater than a certain threshold ρ_i, system fault occurs and can be marked as Y_i (i.e. component C_i fault). The environmental

factor (uncertainty factor) ε is defined as white noise. Without considering compound faults (i.e. multiple components fail at the same time), the CMM of the system failure is:

$$\boldsymbol{Y}=\begin{cases} Y_1 & ,U_1+\varepsilon\geq\rho_1 \\ Y_2 & ,U_2+\varepsilon\geq\rho_2 \\ \vdots & \vdots \\ Y_Q & ,U_Q+\varepsilon\geq\rho_Q \end{cases} \tag{19}$$

In order to simulate the ubiquitous nonlinear relationship in complex electromechanical systems, some of the influence mechanism between components in (18) are replaced by nonlinear functions as:

$$\boldsymbol{C}=\boldsymbol{U}_{\varphi}(\boldsymbol{I}_Q-\boldsymbol{A}')^{-1} \tag{20}$$

where $\boldsymbol{U}_{\varphi}$ is obtained by replacing U_i of $\boldsymbol{U}$ with $\varphi(U_i)$; i is in a subset L of {1, 2, ..., Q}; $\varphi()$ is a nonlinear mapping which can be exponential function, power function, *sigmoid* function, *Rule* function, etc.

Mathematically, to learn the real impact of each representation U_i on the system health status Y, the values of other representations and environmental factors need to be controlled to remain unchanged. This means that, the features in $\boldsymbol{U}$ should be absolutely random and independent. However, in practice, due to the influence of operation time and maintenance history, it is impossible to guarantee that the degradation status of components represented by the system monitoring data is independent. Based on the above considerations, the process of generating the simulation data is as follows:

1) A series of degradation state representation $\boldsymbol{U}^0$ can be randomly generated, and adjusted to make the Kaiser-Meyer-Olkin (KMO) test [30] statistic M* of the features in $\boldsymbol{U}^0$ between 0.3 and 0.5. The KMO test statistic M* can be calculated by:

$$M^*=\frac{R}{R+O}=\frac{\sum_{i=1}^{Q}\sum_{j=1}^{Q}r_{ij}^{\ 2}}{\sum_{i=1}^{Q}\sum_{j=1}^{Q}r_{ij}^{\ 2}+\sum_{i=1}^{Q}\sum_{j=1}^{Q}q_{ij}^{\ 2}} \tag{21}$$

where R and O are respectively the quadratic sum of Pearson correlation coefficients [31] and partial correlation coefficients [32] for all features in $\boldsymbol{U}^0$.

2) Under the premise of filtering composite faults, data $(\boldsymbol{U}^0,\boldsymbol{Y}^0)$ containing 20,000 samples can be obtained according to (20). In this work, the environmental factor $\boldsymbol{\varepsilon}^0$ is white noise with mean value of 0 and variance value of 0.01. And, each threshold ρ_i (i = 1, 2, ..., Q) in (19) is set as 0. The corresponding $\boldsymbol{C}^0$ is calculated based on (20). The obtained data ($\boldsymbol{C}^0$, $\boldsymbol{\varepsilon}^0$, $\boldsymbol{Y}^0$) is regarded as a basic dataset.

3) The simulation data ($\boldsymbol{C}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{Y}$) with 5000 samples is obtained by biased sampling from the basic data set. Here, the biased sampling means that different selection ratios are sit for different sample classes.

The process of generating counterfactual data in this part is as follows:

1) Based on the simulation data ($\boldsymbol{C}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{Y}$), the causal strength matrix $\boldsymbol{A}^{\mathrm{S}}$ for fault diagnosis can be obtained by the FCI algorithm with priori constraints as shown in Fig.5. And, the causal strength matrix $\boldsymbol{A}'$ of $\boldsymbol{C}$ is formed by the first Q rows and the first Q columns of $\boldsymbol{A}^{\mathrm{S}}$. Based on $\boldsymbol{A}'$, a decoupling network is constructed to make different faults be clustered on the degradation status representation $\boldsymbol{U}$. Since only a few nonlinear causal relationships are considered when generating the simulation data, the decoupling network in this part is a single-layer structure network.

2) A single-layer perceptron with *Relu* activation function is used to establish the estimated CMM model, which is trained by ($\boldsymbol{U}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{Y}$). And, the decoder with single-layer is trained by ($\boldsymbol{U}$, $\boldsymbol{C}$).

3) In order to prevent compound failure, for the fault Y_i, deleting all the fault samples that are not Y_i in the simulation data ($\boldsymbol{C}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{Y}$). The obtained data is recorded as ($\boldsymbol{C}^{(i)}$, $\boldsymbol{\varepsilon}^{(i)}$, $\boldsymbol{Y}^{(i)}$). The corresponding counterfactual degradation state representation is marked as $\boldsymbol{U}^{(i)}$.

4) By replacing the i-th feature in $\boldsymbol{U}^{(i)}$ with white noise with mean value of 0 and variance value of 0.1, the counterfactual degradation state representation $\boldsymbol{U}'^{(i)}$ is obtained. And, the i-th group counterfactual data $(\boldsymbol{C}'^{(i)},\varepsilon^{(i)},\boldsymbol{Y}'^{(i)})$ can be generated based on the estimated CMM and the decoder.

5) Q groups of counterfactual data are obtained by making i traverse the set {1, 2, ..., Q}, and are merged to be the final counterfactual dataset.

The effectiveness of the proposed method is verified based on the biased simulation data with the number of components (i.e. Q) respectively being 5, 10, 20, 30 and 50. With 5-fold cross-validation to reduce the impact of uncertain factors, each biased simulation dataset is randomly divided into 5 training sets and 5 test sets. For assessing the enhancement effect of generated data on the monitoring data, fault diagnostic models are trained by the training sets with the generated data. And, the average and standard deviation of fault-accuracy on the testing sets are adopted to evaluate the generalization performance and the model stability of different methods. The fault-accuracy is the proportion of correctly predicted fault samples in all fault samples.

The common data generation methods, SMOTE and GAN, are used as the comparison methods. At present, there are many improved algorithms based on SMOTE or GAN, which can achieve different effects. In this paper, to avoid confusion, only the classic methods are chosen, including SMOTE [9], BorderlineSMOTE [10], adaptive synthetic (ADASYN) [11] and the basic GAN [12]. The off-the-shelf classifiers, including Logistic regression (LR) [29], K-nearest neighbor (KNN) [33] and support vector machine (SVM) [5], are used as the fault diagnosis models. It should be noted that deep learning methods, such as neural networks, have high flexibility and strong learning ability, so they are not used as the basic classifiers for verifying the counterfactual data in this paper. Moreover, $\boldsymbol{Y}$ includes multiple faults, thus a group of binary classifiers is constructed to identify each fault by considering other faults as another class. SMOTE and GAN are interpolated in the same way, i.e. the identification of each fault is converted into a binary classification task.

The comparison results of different models are shown in Table 1 and Table 2. The proposed counterfactual data

generation method is noted as DCGAN in the comparison results, i.e. decoupled causal generative adversarial network.

TABLE I
THE COMPARISON RESULTS OF GENERALIZATION PERFORMANCE

Models	$Q = 10$	$Q = 15$	$Q = 20$	$Q = 30$	$Q = 50$
LR	0.13098	0.10637	0.09936	0.05474	0.00088
SMOTE-LR	0.19116	0.19346	0.19539	0.19663	0.18793
BorderlineSMOTE1-LR	0.19066	0.19322	0.19505	0.19611	0.15931
BorderlineSMOTE2-LR	0.19076	0.19344	0.19512	0.19628	0.15369
ADASYN-LR	0.19132	0.19338	0.19537	0.19666	0.18779
GAN-LR	0.20142	0.19855	0.19476	0.19174	0.18701
DCGAN-LR	**0.26477**	**0.33013**	**0.39593**	**0.41060**	**0.37301**
KNN	0.03031	0.01394	0.00852	0.00419	0.00119
SMOTE-KNN	0.12112	0.11154	0.11443	0.12523	0.11751
BorderlineSMOTE1-KNN	0.11569	0.10284	0.10237	0.10540	0.07263
BorderlineSMOTE2-KNN	0.13961	0.13410	0.13737	0.14745	0.11405
ADASYN-KNN	0.12185	0.11248	0.11554	0.12556	0.11686
GAN-KNN	0.13031	0.15118	0.13299	0.12766	0.14847
DCGAN-KNN	**0.17869**	**0.20324**	**0.21925**	**0.26638**	**0.19294**
SVM	0.15168	0.14125	0.13983	0.12389	0.00046
SMOTE-SVM	0.19172	0.19383	0.19564	0.19696	0.17908
BorderlineSMOTE1-SVM	0.19137	0.19353	0.19514	0.19619	0.15797
BorderlineSMOTE2-SVM	0.19182	0.19381	0.19545	0.19669	0.14816
ADASYN-SVM	0.19182	0.19361	0.19552	0.19688	0.17899
GAN-SVM	0.20492	0.20295	0.20103	0.19943	0.23286
DCGAN-SVM	**0.29715**	**0.28675**	**0.34695**	**0.35482**	**0.33840**

TABLE II
THE COMPARISON RESULTS OF MODEL STABILITY

Models	$Q = 10$	$Q = 15$	$Q = 20$	$Q = 30$	$Q = 50$
LR	0.00162	0.00205	0.00225	0.00145	0.00027
SMOTE-LR	0.00098	0.00063	0.00071	0.00048	0.00261
BorderlineSMOTE1-LR	0.00108	0.00059	0.00077	0.00039	0.00277
BorderlineSMOTE2-LR	0.00097	0.00070	0.00094	0.00075	0.00255
ADASYN-LR	0.00106	0.00075	0.00082	0.00063	0.00263
GAN-LR	0.00089	0.00163	0.00151	0.00165	0.01126
DCGAN-LR	**0.00009**	**0.00010**	**0.00015**	**0.00032**	**0.00047**
KNN	0.00142	0.00101	0.00090	0.00046	0.00039
SMOTE-KNN	0.00164	0.00172	0.00117	0.00169	0.00203
BorderlineSMOTE1-KNN	0.00205	0.00109	0.00098	0.00203	0.00169
BorderlineSMOTE2-KNN	0.00311	0.00062	0.00092	0.00124	0.00398
ADASYN-KNN	0.00106	0.00149	0.00097	0.00180	0.00249
GAN-KNN	0.00320	0.00296	0.00285	0.00273	0.00814
DCGAN-KNN	**0.00044**	**0.00016**	**0.00014**	**0.00018**	**0.00012**
SVM	0.00134	0.00179	0.00165	0.00227	0.00009
SMOTE-SVM	0.00108	0.00060	0.00078	0.00055	0.00261
BorderlineSMOTE1-SVM	0.00093	0.00075	0.00089	0.00055	0.00251
BorderlineSMOTE2-SVM	0.00102	0.00092	0.00082	0.00053	0.00159
ADASYN-SVM	0.00099	0.00075	0.00081	0.00070	0.00230
GAN-SVM	0.00122	0.00141	0.00176	0.00156	0.01058
DCGAN-SVM	**0.00011**	**0.00013**	**0.00015**	**0.00027**	**0.00078**

By comparing the performance of different methods, we can see that the common data generation methods, including SMOTE, BorderlineSMOTE1, BorderlineSMOTE2, ADASYN and the basic GAN, can slightly improve the model generalization performance. But, it is difficult to determine the impacts of these methods on the model stability. In fact, the stability of these methods usually depends on the distribution of the training sets. Based on the proposed counterfactual data generation method, the generalization performance of the fault diagnosis models can be improved more obviously. And, the proposed method also has higher stability.

B. Application results on real data

In this section, the effectiveness of the proposed method is verified via the real monitoring data from the electro-pneumatic (E-P) brake control system of a high-speed train. The process diagram of the E-P brake control system is shown in Fig.11. The raw monitoring data contains 73007 samples and 44 variables, including *voltage*, *current*, *traction_effort*, *vibration_frequency*, *rotation_rate*, *internal_temperature*, *external_temperature*, *humidity*, *braking mode*, *train_operation_mode*, *state*, etc. To comply with the confidentiality agreement, these variables cannot be listed one by one, and are only denoted as V_1, V_2,..., V_{44}. Among them, the combination of *braking mode* and *train_operation_mode* is the system operation mode $\boldsymbol{M}$ described before. The last variable V_{44} is the system fault status $\boldsymbol{Y}$. In the raw monitoring data, different faults are not distinguished, and the data samples are only marked as fault or normal (fault state is 1, normal state is 0). It is easy to select environmental factors $\boldsymbol{\varepsilon}$ according to the variable names, such as *external_temperature*, *humidity*, etc. The remaining 37 variables are the system monitoring parameters, which lack practical significance and prior knowledge in this case. Specifically, several sensors are placed in some positions and the same parameters are repeatedly collected. Or, several parameters of one component have been monitored, but the specific names and meanings of these parameters are unknown. Thus, more practical and comprehensive features need to be constructed.

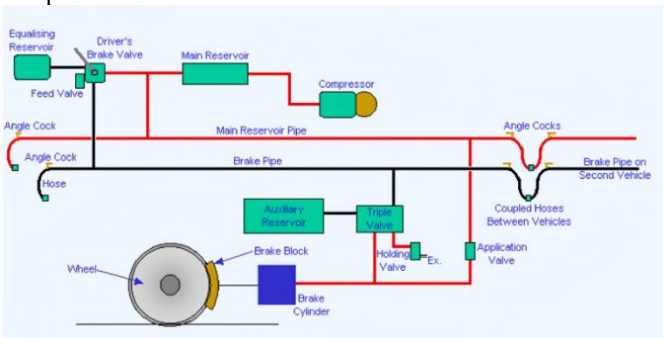


Fig. 11. The process diagram of the E-P brake control system [5].

The process of generating counterfactual data is as follows:

1) The system monitoring parameters are grouped according to the identifiable subsystems (components). A variable that cannot be grouped by subsystems is considered as a separate group. In this work, the 37 monitoring parameters $\boldsymbol{V}$ are divided into 22 groups.

2) Setting the cumulative variance contribution of PCA upper 85%, the principal components of each variable group

can be obtained. These principal components are regarded as the system monitoring parameter data $\boldsymbol{C}$. Here, $\boldsymbol{C}$ contains 29 features. The PCA models of all variable groups are combined and recorded as:

$$\boldsymbol{C} = (C_1, C_2, ..., C_{29}) = (V_1, V_2, ..., V_{37})\boldsymbol{\Omega} = \boldsymbol{V\Omega} \quad (22)$$

where $V_1, V_2, ..., V_{37}$ arranged by variable group; $\boldsymbol{\Omega} \in R^{37\times29}$ is the matrix with the component score coefficient matrix of each PCA model arranged in the corresponding position and the other elements are 0. Here, the corresponding position means that, for the group containing variable V_i, V_{i+1}, ..., V_{i+m} and corresponding to components C_j, C_{j+1}, ..., C_{j+n}, the score coefficient matrix of $m\times n$ dimension obtained by its PCA model is placed in the $i\sim i+m$ rows and $j\sim j+n$ columns of $\boldsymbol{\Omega}$.

3) Based on the dataset ($\boldsymbol{C}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{M}$, $\boldsymbol{Y}$), the proposed priori constrained FCI algorithm is used to mine causality related to fault diagnosis in the E-P brake control system. The used priori constraints are as shown in Fig.6. The output causal network of ($\boldsymbol{C}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{M}$, $\boldsymbol{Y}$) is shown in Fig.12. In Fig.12, E1, E2 and E3 are used to represent the three environmental factors.

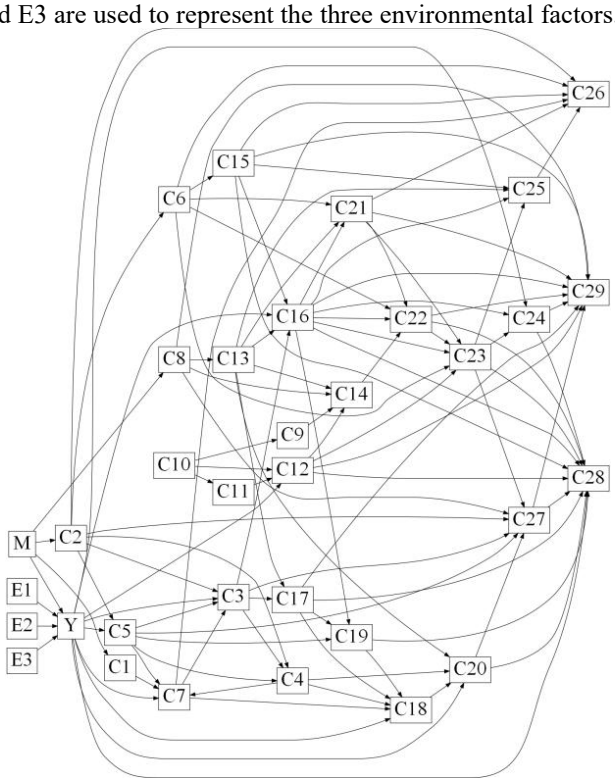


Fig. 12. The causal network of ($\boldsymbol{C}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{M}$, $\boldsymbol{Y}$).

4) The proposed graph decoupling network is used to decouple the causal mechanism in $\boldsymbol{C}$. For $\boldsymbol{C}$ in this case, by minimize the loss function in (10), the decoupling network is finally optimized to be as a two-layer structure. And, the component-level degradation status representation $\boldsymbol{U}$ with interpretability for $\boldsymbol{Y}$ and $\boldsymbol{M}$ is obtained. For ease of observation, the causal network of ($\boldsymbol{U}$, $\boldsymbol{Y}$) is presented in two parts as shown in Fig.13. It verifies that there is no causal relationships between the features $\boldsymbol{U}$ and each feature serves as a cause of $\boldsymbol{Y}$.

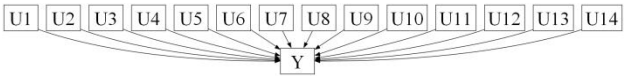


Fig. 13. (a) The causal network of (U_1, U_2, ..., U_{14}, $\boldsymbol{Y}$)

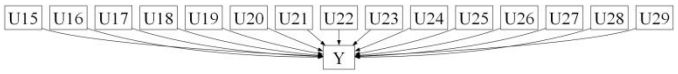


Fig. 13. (b) The causal network of (U_{15}, U_{16}, ..., U_{29}, $\boldsymbol{Y}$).

5) The decoder, which is used to restore the degradation status representation to the system monitoring parameters, is trained based on ($\boldsymbol{U}$, $\boldsymbol{C}$).

6) Based on the decoupled data ($\boldsymbol{U}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{M}$, $\boldsymbol{Y}$), the CMM of the E-P brake control system is established by using a group of MLP models. Each operation mode corresponds to a MLP model. MLP is considered here because the estimated CMM requires relatively high fitting accuracy.

7) By changing each feature in $\boldsymbol{U}$ into white noise in turn, while making the generated data $(\boldsymbol{C}', \boldsymbol{M}, \boldsymbol{\varepsilon}, \boldsymbol{Y}')$ conform to the causal relationship shown in Fig.12, multiple counterfactual datasets can be obtained.

8) The generated data $(\boldsymbol{C}', \boldsymbol{M}, \boldsymbol{\varepsilon}, \boldsymbol{Y}')$ is converted into $(\boldsymbol{V}', \boldsymbol{M}, \boldsymbol{\varepsilon}, \boldsymbol{Y}')$ by:

$$\boldsymbol{V} = (V_1, V_2, ..., V_{37}) = (C_1, C_2, ..., C_{29})\boldsymbol{\Theta} = \boldsymbol{C\Theta} \quad (23)$$

where $\boldsymbol{\Theta} \in R^{29\times37}$ is the pseudo inverse matrix of $\boldsymbol{\Omega}$, which can also be obtained by arranging the load matrix of each PCA model in corresponding position. Here, the corresponding position means that, for the group containing variable V_i, V_{i+1}, ..., V_{i+m} and corresponding to components C_j, C_{j+1}, ..., C_{j+n}, the load matrix of $n\times m$ dimension obtained by its PCA model is placed in the $j\sim j+n$ rows and $i\sim i+m$ columns of $\boldsymbol{\Theta}$.

Also, the monitoring data is randomly divided into a training set and a test set with a sample ratio of 4:1. And, due to serious imbalance of the monitoring data (fault samples only account for 0.42%), *precision* and *recall* of the fault class on the testing set are applied as the evaluation measures to comprehensively compare different methods. Whereas *TP*, *TN*, *FP*, *FN* represent the number of true positive, true negative, fault positive and fault negative samples, respectively, *precision* and *recall* can be calculated by:

$$precision = \frac{TP}{TP + FP} \quad (24)$$

$$recall = \frac{TP}{TP + FN} \quad (25)$$

By adding the generated data to the training set, under the off-the-shelf classifiers (LR, KNN and SVM), the generalization performances of different data generation methods are shown in Table 3. The comparison results are the average and standard deviation of the the evaluation measures under 5-fold cross-validation (recorded as *precision_mean*, *precision_std, recall_mean* and *recall_std*).

TABLE III
THE COMPARISON RESULTS UNDER THE OFF-THE-SHELF CLASSIFIERS

Model	*precision_mean*	*precision_std*	*recall_mean*	*recall_std*
LR	0.42137	0.10914	0.05198	0.01911
SMOTE-LR	0.03826	0.00145	0.90883	0.03854
BorderlineSMOTE1-LR	0.04412	0.00226	0.90894	0.05608
BorderlineSMOTE2-LR	0.03348	0.00069	0.93506	0.02289
ADASYN-LR	0.03215	0.00135	0.92533	0.03630
GAN-LR	0.08366	0.04356	0.93548	0.05947

DCGAN-LR	**0.43467**	**0.00020**	**0.97985**	**0.00017**
KNN	0.51684	0.03043	0.73014	0.06051
SMOTE-KNN	0.38326	0.01897	0.85759	0.06021
BorderlineSMOTE1-KNN	0.43177	0.03315	0.84738	0.03170
BorderlineSMOTE2-KNN	0.41100	0.03954	0.86049	0.04893
ADASYN-KNN	0.39180	0.01356	0.88334	0.05105
GAN-KNN	0.51812	0.23182	0.93443	0.16338
DCGAN-KNN	**0.58333**	**0.00036**	**0.99270**	**0.00119**
SVM	0.36548	0.19305	0.02597	0.00793
SMOTE-SVM	0.08225	0.00400	0.93184	0.04004
BorderlineSMOTE1-SVM	0.09997	0.00556	0.91887	0.03943
BorderlineSMOTE2-SVM	0.07993	0.00254	0.92871	0.03138
ADASYN-SVM	0.07085	0.00423	0.93517	0.02326
GAN-SVM	0.07264	0.02258	0.80645	0.03161
DCGAN-SVM	**0.41638**	**0.00030**	**0.94506**	**0.00094**

These methods of generating virtual data supplementary training sets can effectively improve the generalization *recall* of faults. Simultaneously, it also makes the classification hyperplane of the classifier be biased to the fault class, causing the low generalization *Precision* of faults. Moreover, due to the serious imbalance of data and the application of simple models, the comparison results are generally poor. KNN is based on nearest neighbor samples and has no classification hyperplane, so the results are relatively good. Nevertheless, it can also be seen that the proposed counterfactual data generation method achieves better performance in both generalization performance and model stability. This shows that the proposed method is effective on real data with complex nonlinear relations.

IV. Conclusion

When machine learning models are utilized for fault diagnosis in complex electromechanical systems, inaccurate or faulty diagnostic logic may be discovered due to the suboptimal quality and quantity of available monitoring data. For addressing this challenge, a prevalent approach involves synthetically generating fault samples. Nevertheless, most data generation methods rely solely on data distribution and Euclidean distance, overlooking the fundamental causal information among the variables. Recognizing this, the present paper proposes a counterfactual data generation approach grounded in causal relationship and fault mechanism discovery. Initially, a priori-constrained FCI algorithm is devised to uncover causalities within complex electromechanical systems. Subsequently, a graph decoupling network is engineered to disentangle the degradation status representations. Finally, counterfactual data is generated by a causal-based generative adversarial network. By integrating prior knowledge into both the causal network and data generation process, the model's reliance on historical monitoring data is mitigated. By incorporating counterfactual data into the training set, the generalization and stability of fault diagnosis models can be improved. The efficacy of this proposed methodology is demonstrated through both simulation and real-world data.

The proposed counterfactual data generation method assumes that the system causality and failure mechanism are invariant. The system causality is likely to be changed in new environments, and the causal mechanisms of novel failures cannot be obtained. Therefore, it is impossible to generate counterfactual data for new environments or faults based on the proposed method. Moreover, the construction of the causal network relies on prior knowledge. Additionally, the approach of first mining causal relationships between subsystems at the subsystem level and then using the FCI algorithm to mine inter-system causal relationships has yet to be fully verified, as it may face challenges in accurately capturing complex interactions between subsystems. Future studies will delve deeper into analyzing the quality and reliability of counterfactual data, as well as generating counterfactual data for unprecedented environments and faults. Furthermore, we will investigate common causal discovery approaches to simplify complex systems, and explore other ways to change component health status. Additionally, exploring the application of the proposed method to other types of electromechanical systems and investigating the impact of various parameter settings on method performance will also be key areas of focus.

References

[1] Dinh T Q, Senatore A, Birrell S,et al. Editorial electromechanical as an Enabler for Intelligent Transportation Systems[J]. IEEE transactions on intelligent transportation systems, 2021(9):22.

[2] Peng Y, Liu D, Peng X. A review: Prognostics and health management[J]. Journal of Electronic Measurement and Instrument, 2010, 24(1):1-9.

[3] Qin S J. Survey on data-driven industrial process monitoring and diagnosis[J]. ANNUAL REVIEWS IN CONTROL, 2012, 36(2):220-234.

[4] Hu G, Zhou T and Liu Q. Data-Driven Machine Learning for Fault Detection and Diagnosis in Nuclear Power Plants: A Review[J]. Frontiers in Energy Research, 2021, 9:663296. doi: 10.3389/fenrg.2021.663296

[5] Liu J, Hu Y, Yang S. A SVM-Based Framework for Fault Detection in High-Speed Trains[J]. Measurement, 2020, 172:108779.

[6] Zhai Y, Yang K, Zhao Z,et al.Geometric characteristic learning R-CNN for shockproof hammer defect detection[J]. Engineering Applications of Artificial Intelligence: The International Journal of Intelligent Real-Time Automation, 2022, 116:105429-105441.

[7] Sun J, Liu Z, Wen J,et al. Multiple hierarchical compression for deep neural network toward intelligent bearing fault diagnosis[J]. Engineering Applications of Artificial Intelligence: The International Journal of Intelligent Real-Time Automation, 2022, 116:105498-105508.

[8] Shaheen B, Kocsis A, Nemeth I. Data-driven failure prediction and RUL estimation of mechanical components using accumulative artificial neural networks[J]. Engineering Applications of Artificial Intelligence, 2023, 119:105749-105765.

[9] Bernardo A, Valle E D. An extensive study of C-SMOTE, a Continuous Synthetic Minority Oversampling Technique for Evolving Data Streams[J]. Expert Systems with Applications, 2022, 196:116630.1-116630.21.

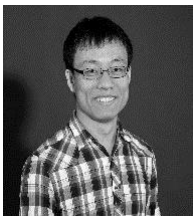
[10]Han H, Wang WY, Mao BH. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning[C]. In: Huang, DS., Zhang, XP., Huang, GB. (eds) Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg. 2005(3644): 878-887.

[11]He H, Yang B, Garcia E A , et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]// Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE, 2008.

[12]Li Z, Zheng T, Wang Y, et al. A Novel Method for Imbalanced Fault Diagnosis of Rotating Machinery based on Generative Adversarial Networks[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70:3500417.1-3500417.17.

[13]Cui Z, Lu Y, Yan X,et al. Compound fault diagnosis of diesel engines by combining generative adversarial networks and transfer learning[J].Expert Systems With Applications, 2024, 251(123969).
[14]Wang C, Liu J, and Zio E. A Modified Generative Adversarial Network for Fault Diagnosis in High-Speed Train Components with Imbalanced and Heterogeneous Monitoring Data[J]. Journal of Dynamics, Monitoring and Diagnostics, 2022, 1(2):84-92.
[15]Liu H, Zhao H, Wang J, et al. LSTM-GAN-AE: A Promising Approach for Fault Diagnosis in Machine Health Monitoring[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71:1-13.
[16]Lyu P, Cheng Y, Zhang M, et al. GPSC-GAN: A Data Enhanced Model for Intelligent Fault Diagnosis[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73:3532116.
[17]Sameer A, Matteo T, Ankit A, et al. SKDCGN: Source-free Knowledge Distillation of Counterfactual Generative Networks using cGANs[J], In book: Computer Vision - ECCV 2022 Workshops, 2023, vol 13804, PP: 679–693, Springer, Cham.
[18]Pearl J. Causality: Models, Reasoning, and Inference[M], second edition. Cambridge University Press, New York, United States, 2000.
[19]Glymour C, Zhang K, Spirtes P. Review of Causal Discovery Methods Based on Graphical Models[J]. Frontiers in Genetics, 2019, 1(10):524-539.
[20]Shen X, Ma S, Vemuri P, et al. Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology[J]. Scientific Reports, 2020, 10(1):2975.
[21]Tsamardinos I. Constraint-based Causal Discovery from Multiple Interventions over Overlapping Variable Sets[J]. Journal of Machine Learning Research, 2015, 16:2147-2205.
[22]Peter S A and Clark G. An Algorithm for Fast Recovery of Sparse Causal Graphs[J]. Social Science Computer Review, 1991, 9(1):62-72.
[23]Spirtes P, Meek C, Richardson T, An algorithm for causal inference in the presence of latent variables and selection bias, in: C. Glymour, G. Cooper (Eds.), Computation, Causation, and Discovery, MIT Press, Cambridge, MA, 1999, pp. 211–252.
[24]Zhang J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias[J]. Artificial Intelligence, 2008, 172(16-17):1873-1896.
[25]Abdi H, Williams L J. Principal component analysis[J]. Wiley Interdisciplinary Reviews Computational Statistics, 2010, 2(4):433-459.
[26]Smith J, Johnson P. The role of environmental factors in fault diagnosis of complex systems[J]. Journal of Reliability Engineering, 2022, 18(3), 234-246.
[27]Thomas N. K, Welling M. Semi-Supervised Classification with Graph Convolutional Networks[C] In the 5th International Conference on Learning Representations (ICLR-17), Toulon, France, 2017, pp:1-14.
[28]Peter R J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational & Applied Mathematics, 1987, 20:53-65.
[29]Menard S. Logistic Regression[J]. American Statistician, 2004, 58(4):364.
[30]Dombrowski S C, Watkins M W, Brogan M J. An Exploratory Investigation of the Factor Structure of the Reynolds Intellectual Assessment Scales (RIAS)[J]. Journal of Psychoeducational Assessment, 2009, 27(6):494-507.
[31]Nahler, Gerhard. Pearson correlation coefficient[J]. Springer Vienna, 2009, 10.1007/978-3-211-89836-9(Chapter 1025):132-132.
[32]Wang J. Partial Correlation Coefficient[J]. Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensis, 2013:1634-1635.
[33]Dudani S A. The Distance-Weighted k-Nearest-Neighbor Rule[J]. IEEE Trans Systems Man & Cybernetics, 1976, 6(4):325-327.

Chong Wang received the B.S. degree in Applied Mathematics, the M.S. degree in Statistics, and the Ph.D. degree in Systems Engineering from Beihang University, Beijing, China, in 2014, 2017, and 2023, respectively. She is currently a postdoctoral researcher in the Department of Automation, Tsinghua University, Beijing, China. Her research interests include fault diagnosis and imbalanced data.



Jie Liu received the B.S. degree in Mechanical Engineering and the M.S. degree in Physics from Beihang University, Beijing, China, in 2009 and 2012, respectively. He received the Ph.D. degree in CentraleSupélec France. He is currently an associate professor in School of Reliability and Systems Engineering, Beihang University, Beijing, China. His research focuses on fault detection, diagnostics, and prognostics.

Junwei Cao received the B.E. and M.S.E. degrees from Tsinghua University, Beijing, China, in 1995 and 1998, respectively. He received the Ph.D. degree from the University of Warwick, UK, in 2001. He is currently a Professor in Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China. His research focuses on advanced computing technologies and their applications ever since.

Xi Chen received the B.E. and M.S. degrees from Guangdong University of Technology, Guangzhou, Guangdong Province, China, in 2003 and 2015, respectively. He is now a Ph.D. in the Department of Information Science and Technology, Southwest Jiaotong University, Chengdu, Sichuan Province, China. His research focuses on Intelligentization within the Transportation Engineering.

Lihua Chen received the B.E. degree in Automatic Control from Lanzhou Jiaotong University, Lanzhou, Gansu, China, in 2001. He is now a Ph.D. in the Department of Information Science and Technology, Southwest Jiaotong University, Chengdu, Sichuan Province, China. His research primarily centers on Railway Transportation.

Yindong Ji received the B.E. and M.S. degrees from Tsinghua University, Beijing, China, in 1985 and 1989, respectively. He is currently a professor at the Department of Automation, Tsinghua University, Beijing, China. His research interests include fault detection, intelligent control and integrated optimization.