Fault diagnosis-based SDG transfer for zero-sample fault symptom

Mengqin Yu [a,1], Yi Shan Lee [a,2], Junghui Chen [a,3,*]

[a] Department of Chemical Engineering, Chung Yuan Christian University, Chung-Li, Taoyuan, Taiwan, R.O.C
[1] monica159266@gmail.com; [2] bmtys1995@gmail.com; [3] jason@wavenet.cycu.edu.tw
* corresponding author

ARTICLE INFO

Article history
Received July 15, 2023
Revised September 14, 2023
Accepted November 14, 2023
Available online November 30, 2023

Selected paper from The 2023 6th International Symposium on Advanced Intelligent Informatics (SAIN'23), Yogyakarta (Virtually), September 21, 2023, http://sain.ijain.org/2023/. Peer-reviewed by SAIN'23 Scientific Committee and Editorial Team of IJAIN journal.

Keywords
Conditional variational autoencoder
Fault diagnosis
Signed directed graph
Zero-shot learning

ABSTRACT

The traditional fault diagnosis models cannot achieve good fault diagnosis accuracy when a new unseen fault class appears in the test set, but there is no training sample of this fault in the training set. Therefore, studying the unseen cause-effect problem of fault symptoms is extremely challenging. As various faults often occur in a chemical plant, it is necessary to perform fault causal-effect diagnosis to find the root cause of the fault. However, only some fault causal-effect data are always available to construct a reliable causal-effect diagnosis model. Another worst thing is that measurement noise often contaminates the collected data. The above problems are very common in industrial operations. However, past-developed data-driven approaches rarely include causal-effect relationships between variables, particularly in the zero-shot of causal-effect relationships. This would cause incorrect inference of seen faults and make it impossible to predict unseen faults. This study effectively combines zero-shot learning, conditional variational autoencoders (CVAE), and the signed directed graph (SDG) to solve the above problems. Specifically, the learning approach that determines the cause-effect of all the faults using SDG with physics knowledge to obtain the fault description. SDG is used to determine the attributes of the seen and unseen faults. Instead of the seen fault label space, attributes can easily create an unseen fault space from a seen fault space. After having the corresponding attribute spaces of the failure cause, some failure causes are learned in advance by a CVAE model from the available fault data. The advantage of the CVAE is that process variables are mapped into the latent space for dimension reduction and measurement noise deduction; the latent data can more accurately represent the actual behavior of the process. Then, with the extended space spanned by unseen attributes, the migration capabilities can predict the unseen causes of failure and infer the causes of the unseen failures. Finally, the feasibility of the proposed method is verified by the data collected from chemical reaction processes.

1. Introduction

Fault diagnosis and classification play important roles in the maintenance of chemical plants [1], [2]. To establish an effective fault diagnosis system, the faulty equipment, processes, and systems must be studied and analyzed initially. Then, based on prior information and the relationship between input and output, a mathematical model or a qualitative empirical model is established as the foundation for fault diagnosis. Finally, the system's failure is determined by evaluating measurable variables or estimated variables that are not directly measurable. If the system's output deviates from the expected range or if the system's state changes and exceeds the predetermined range, faults can be detected promptly.

They are techniques for ensuring normal operational production by quickly identifying the causes of accidents and providing real-time operational guidance. By doing so, they can prevent further deterioration to the process during accidents and improve plant reliability, safety, and economy [3]. The techniques for fault diagnosis can commonly be categorized into data-driven (DD) and model-based methods, as depicted in Fig. 1 [4].

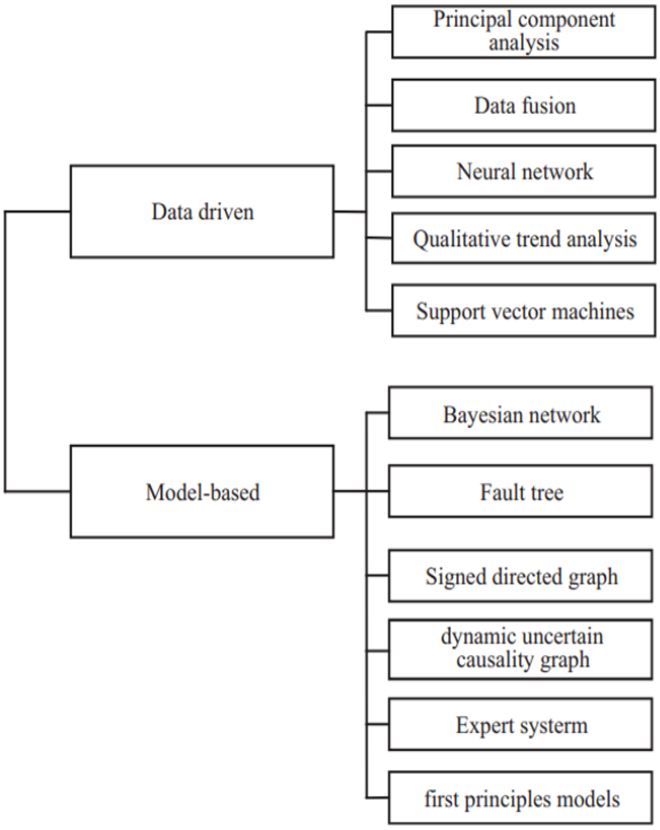

Fig. 1. Classification of common fault diagnosis methods [5]

DD methods primarily rely on a large amount of data to establish the relationship between parameters and faults. Previous developed DD methods include principal component analysis [6], [7], qualitative trend analysis [8], and other approaches [9]. On the other hand, the model-based methods utilize domain knowledge and yield qualitative rather than quantitative results. The core of the expert systems [10] is based on the application of expert knowledge. Meanwhile, Bayesian networks (BN) [11], [12] first-principle models [13], directed graphs [14], and dynamic uncertain causality graphs [15], [16], employs graph theory to visually depict the relationships between various parameters and faults. Maintaining the Integrity of the Specifications

However, chemical processes are typically characterized by nonlinearity and high dimensionality, making them highly complex. Therefore, constructing a reliable model to describe all the characteristics of such a process system is challenging. In order to enhance the ability to diagnose various faults in complex chemical production processes, researchers have proposed numerous DD methods for monitoring and diagnose faults [6], [17]. Since the emergence of neural networks [18], [19] in 2006, many deep learning models capable of solving highly nonlinear problems have been developed. These models establish latent variable representations by mapping data to low-dimensional spaces. Examples of such models include the encoder (AE) [20], stacked denoising autoencoder (SDAE) [21],[22] variational autoencoder (VAE) [23], [24] and supervised variational autoencoder (S-VAE). These models rely on a substantial amount of uniformly distributed and representative data, thereby exhibiting relatively high reliability. However, in cases where there is limited or no failure data available for the target failure during the operational process, the collected data may not be sufficiently representative to

ensure the validity of the modeling. Since many faults can lead to severe damage and substantial losses, it is rare for factories to operate under various fault states and collect samples for training fault diagnosis systems [19]. Additionally, transitioning from normal operating conditions to fault conditions is time-consuming and poses challenges in obtaining an adequate number of fault samples using DD methods [13].

For fault classification, Mishra et al. [25] utilize conditional variational autoencoders (CVAE) to generate samples from given attributes and employ the generated samples for classifying unknown categories. CVAE is a conditional directed graphical model where input observations modulate the prior on latent variables that generate the outputs, in order to model the distribution of high-dimensional output space as a generative model conditioned on the input observation. It aims at extracting latent representations or features in a latent space using deep neural networks [26], [27]. However, this method of data generation may introduce deviations from the real situation. Another prominent industrial scenario analysis method is fault tree analysis [28], [29] which constructs a diagnostic system based on knowledge of the fault process. It has proven successful in understanding the causes of system failures, identifying effective risk reduction strategies, and estimating failure occurrence rates [30]. However, fault tree analysis only incorporates physical knowledge and does not effectively leverage actual field operation data and conditions. Furthermore, the fault diagnosis analysis using fault tree analysis tends to be time-consuming. To address the challenge of fault classification of the unseen faults, one viable approach is to transfer knowledge acquired from easily obtained or historical faults (training faults) to those that are difficult or costly to collect (target faults) [31], [32]. Hugo et al. [33] first proposed zero-shot learning in 2008, aiming to solve the classification problem where there is not enough label data to distinguish all categories. The purpose of zero-shot learning is to solve the problem of being unable to model unseen fault data [34], [35].

In previous research on zero-shot learning, the focus was primarily on mapping between images and attributes [36]. However, when applied to fault diagnosis tasks, there are no images available to obtain various fault attributes. Visual properties are not applicable to industrial sensor signals, necessitating the need for more effective auxiliary information in zero-shot fault diagnosis tasks. In 2021, Zhao et al. [17] proposed a direct approach where the fault description is utilized as the attribute for fault detection. For example, it could be a specific feed amount change or a certain temperature change at a particular position. The method involves extracting features from the data and training the attribute learner. However, performing this process in two stages may lead to a situation where individual training performs well, but the combined result is unsatisfactory. Mou et al. [37] introduce a comprehensive zero-shot fault diagnosis model known as Distributional Semantic Embedding and Cross-Modal Reconstruction VAE (DSECMR-VAE). This model considers fault samples and fault attribute semantic vectors as distinct modalities and employs two Variational Autoencoders (VAEs) to reconstruct these inputs. Li et al.[38] delved into this area by exploring a federated zero-shot fault diagnosis framework, which introduces a novel paradigm for semantic knowledge sharing. The carefully designed network structure and aggregation strategy within the framework create a synergy between zero-shot modeling and federated aggregation processes, yielding mutual benefits. While the aforementioned methods employ fault descriptions as attributes, the primary objective of this study extends beyond fault type diagnosis. Our aim is to not only identify the fault type, but also to deduce its underlying root cause. To address this, we incorporate signed fault directed graphs (SDG) constructed using physical knowledge [39]. An SDG serves as a visual representation of the causal relationships within processes, depicting process variables as nodes on the graph and causal connections as directed arcs. These arcs, indicating the cause-effect relationship, can point in either the same or opposite directions. A solid line signifies a positive effect, while a dotted line denotes a negative effect. Within the SDG model for fault diagnostics, the nodes are quantifiable process variables. Any deviation in these variables is attributed to abnormal factors, leading to a shift in the subsequent node. It's important to note that various nodes may originate from distinct sources of faults. In this context, all nodes representing causes are considered root nodes [40]. SDGs reveal the intrinsic causal relationships among variables in complex systems. The SDG model solely focuses on the qualitative relationship of the system. In actual industrial systems, the qualitative relationship between variables happens to be the only constant property, which gives SDG an advantage

over other fault diagnosis methods. Consequently, the fault attribute value is defined by the logical relationship in the SDG, allowing the diagnosis process to encompass the causal relationship among variables in the system, facilitating further speculation on the fault's source. SDGs can be constructed with only physical knowledge or a small amount of data to directly identify the root cause of the fault. However, it does require significant time for interpretation after receiving the data.

This study utilizes a SDG that incorporates physical knowledge to depict the path of fault propagation, fault location, and fault cause. To integrate it with the data-driven approach, the CVAE is employed to mitigate the influence of measurements. This enables the training process, from data to attributes, to be completed in a single step. Additionally, zero-shot learning is incorporated, where known fault data is used to pre-learn the causes of specific faults. Subsequently, a transfer learning technique is employed to infer the causal relationship between fault variables that have not previously occurred when encountering unknown fault data. This approach allows for speculation on the root cause of the failure.

Considering the limited availability of fault samples and the challenges associated with fault propagation, this study introduces a zero-sample fault causal root analysis method. The main contributions of this study are as follows:

- Establishment of a binary attribute table using the SDG, which reveals the impact, location, and source of faults.
- Utilization of the CVAE to map process variables into a latent variable space, reducing dimensionality and eliminating measurement noise while accurately representing process behavior.
- Training the CVAE-SDG model using known fault causality attributes and measured data, and applying it to infer unknown fault causality for unknown fault paths.

2. Method

In order to effectively describe the fault attributes of various faults to provide information on each path, the attributes can be used to express the causal relationship between process variables in the SDG. SDG can show the impact of the failure, the location of the failure and the cause of the failure, etc. Take the simple process in Fig. 2 as example, x_1, x_2, x_3, x_4, x_5 in Fig. 2 represent different process variables.

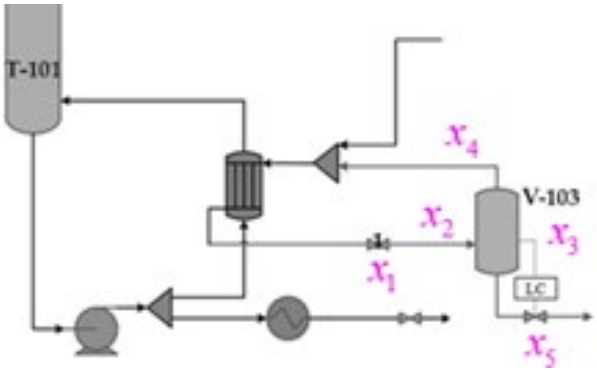

Fig. 2. Process flow chart

They include the valve opening, flow rate, liquid level, etc. Fig. 3 is the SDG of the simple process. The connecting lines represent the causal relationship between them. It shows that x_1 directly affects x_2, x_1, x_2, x_3, x_4, x_5 which in turn affects x_3. In addition, there are mutual influences between x_2 and x_4 and between x_3 and x_5 due to the return flow and the control loop. The presence or absence of the path can be represented by the attribute with the attribute value of 1 or 0. 1 represents presence and 0 represents absence. The joint attributes $a_1, a_2, \cdots, a_6$ forms the path which starts with a source. In this

way, a fault path with source x_1 can be defined and defined as a label y_1 with the joint attributes $a_1, a_2, \cdots, a_6$. Likewise, all paths are defined in the same way. Each attribute is a dimension in the vector space, expressed as $\mathbf{a} = [a_1, a_2, \cdots, a_C] \in \square^C$, where C is the number of attributes, and the measured process variables are represented by $\mathbf{x} = [x_1, x_2, \cdots, x_D]$, where D is the number of process variables. In this study, all defined paths will be divided into paths of known fault data and unknown fault data according to whether there is historically collected process variable data.

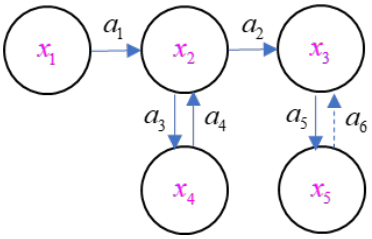

Fig. 3. The SDG in Fig. 2

The goal of this research is to use the J types fault paths of known training samples to diagnose and identify the I types fault paths of unknown training samples. The set of known fault paths is denoted as $S = \{s_1, \cdots, s_J\}$, where s_j is the fault path with samples. The set of unknown fault paths is denoted as $T = \{t_1, \cdots, t_I\}$, where t_i is the unknown fault path. T and S are disjoint to each other, i.e., $T \cap S = \varnothing$. The sample set of S is denoted by $\Im = \{\mathbf{X}^S \in \square^{N \times D}, \mathbf{y}^S \in \square^N\}$, where $\mathbf{X}^S = [\mathbf{x}_1^S, \mathbf{x}_2^S, \cdots, \mathbf{x}_N^S]$ is the data collected for training, including N training samples and D process variables. x_1, x_2, x_3, x_4, x_5 in Fig. 2 is the measured process variable. And $\mathbf{y}^S = [y_1, y_2, \cdots, y_N]$ is the label of each sample corresponding to the fault path. For all L types faults ($L = I + J$), the attribute matrix can be expressed as $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_L] = [\mathbf{A}_S, \mathbf{A}_T] \in \square^{L \times C}$. All elements in $\mathbf{A}$ are 1 or 0, which shows whether the attribute exists in a certain fault path. The task of this research is to use the training data $\mathbf{X}^S$ of known faults and the corresponding labels $\mathbf{y}^S$ to build a model f, so that the loss of known faults is as small as possible, as shown in Eq.(1).

$$y^s = f(X^s) \; and \; \min \text{CLoss}\,(\text{y}\,s, \hat{y}^s) \tag{1}$$

where $CLoss$ represents the loss of fault classification, and $\mathbf{y}^S$ is the label predicted by the training data model. When f is used on the measured variable of unknown fault, its formula is as follows Eq.(2).

$$\hat{y}^T = f(\boldsymbol{X}^T) \tag{2}$$

where $\mathbf{X}^T = [\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_M^T]$ is the measured variable of M target faults, and $\hat{\mathbf{y}}^T$ is the label predicted by the target fault model. There is a direct correspondence between the fault path label $\mathbf{y}$ and the attribute matrix $\mathbf{A}$ in SDG. Therefore, $\mathbf{A}$ can be used to replace $\mathbf{y}^S$ in Eq.(1). Then Eq.(1) and Eq.(2) can be rewritten as

$$\boldsymbol{A} = f(\boldsymbol{X}^s) \; and \; \min CLoss\,(A^S, \hat{A}^S) \tag{3}$$

$$\hat{\boldsymbol{A}}^T = f(\boldsymbol{X}^T) \tag{4}$$

Furthermore, $\mathbf{A}^S$ and $\mathbf{A}^T$ will substitute $\mathbf{y}^S$ for model training, since attribute descriptions are class-level rather than sample-level, which can be easily obtained with physical manipulation knowledge.

In order to establish the model relationship from $\mathbf{X}^S$ to $\mathbf{A}$ in Eq.(3) with the form of probability, the objective function can be defined as maximizing the conditional probability distribution of process variables $\mathbf{x}_{n,j}$ to the attributes $\mathbf{a}_{n.j}$ as:

$$max \prod_{j=1}^{J} \prod_{n=1}^{N} p(a_{n.j}|X_{n,j}) \tag{5}$$

where the subscript j refers to the corresponding fault type; J denotes the total number of seen fault types. The subscript n refers to the sample point and N denotes the total number of samples in fault type j. For the simplification in expressing the following derivations, only one fault sample is considered from now onwards, the subscripts of the variables and the attributes are tentatively ignored. As the observation variables are collected in the noise-contaminated space, the inference of latent variables ($\mathbf{z}$) at the lower noise level and lower-dimensional feature space corresponding to those observations are necessary. The latent variables can be considered as the important features between the process variables and the fault attributes. Thus, the conditional probability distribution in Eq. (5)can be expressed as

$$p(\boldsymbol{a}|\boldsymbol{x}) = \int p(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{x}) dz \tag{6}$$

The goal is to maximize the probability distribution of $p(\mathbf{a}|\mathbf{x})$, which is equivalent to the integration of $p(\mathbf{a}|\mathbf{x},\mathbf{z})\,p(\mathbf{z}|\mathbf{x})$ over the latent variable $\mathbf{z}$. To accomplish the intractable objective in Eq.(6), a variational inference posterior distribution $q(\mathbf{z}|\mathbf{x})$ is introduced to approximate the prior distribution $p(\mathbf{z}|\mathbf{x})$. By using the Bayesian theorem, the variational lower bound can be derived as:

$$\begin{aligned} &In\, p(a|x) \geq E_{z\sim q(Z|X)}[In\, p(\boldsymbol{a}|\boldsymbol{z})] - kl(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) \\ &= L(\mathbf{a}|\mathbf{x}) \end{aligned} \tag{7}$$

The variational lower bound $L(\mathbf{a}|\mathbf{x})$ consists of the conditional likelihood $E_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}\left[\ln p(\mathbf{a}|\mathbf{z})\right]$ and the KL divergence term $kl\left(q(\mathbf{z}|\mathbf{x})\,\|\,p(\mathbf{z}|\mathbf{x})\right)$ as Eq.(7). In other words, when maximizing $L(\mathbf{a}|\mathbf{x})$, it is equivalent to maximizing $p(\mathbf{a}|\mathbf{x})$. Fig. 4 shows the model structure of the CVAE. It is established based on Eq.(7). The $q(\mathbf{z}|\mathbf{x})$ represents the encoder and $p(\mathbf{a}|\mathbf{z})$ represents the decoder. The CVAE model takes the input data $\mathbf{x}_n^S$ into a multi-layer neural network structure, which is the encoder, to convert the data into a latent variable by nonlinear mapping. Then predict $\mathbf{a}_j^S$ through the decoder and optimize the model parameters in the CVAE using the variational lower bound $L(\mathbf{a}|\mathbf{x})$.

This study presents a fault inference method that combines attribute transfer and zero-shot learning. It aims to learn knowledge from easily obtained or known fault causal data and apply it to new faults. Despite the difference in the fault propagation paths, both normal and abnormal productions typically follow the same production flow, even with the same production line. Therefore, there might be shared information among the data. Usually, fault labels are represented using one-hot encoding. For instance, Fig. 5 (left) shows two observed fault paths (star markers) in a two dimensional plane. However, the labels do not have any interpretable meaning towards the fault cause or fault propagation paths. Especially, the unknown faults (triangle markers in Fig. 5 (left)) have no way to be determined with those one-hot labels as they are unseen and physically unexplainable. Therefore, each paths can be characterized by a group of fault attribute description with possible extension in dimensional space by

incorporating physical knowledge. As in Fig. 5 (right), the two observed fault paths (star markers) are positioned on the two dimensional attribute plane with the attribute a_1 and a_2.

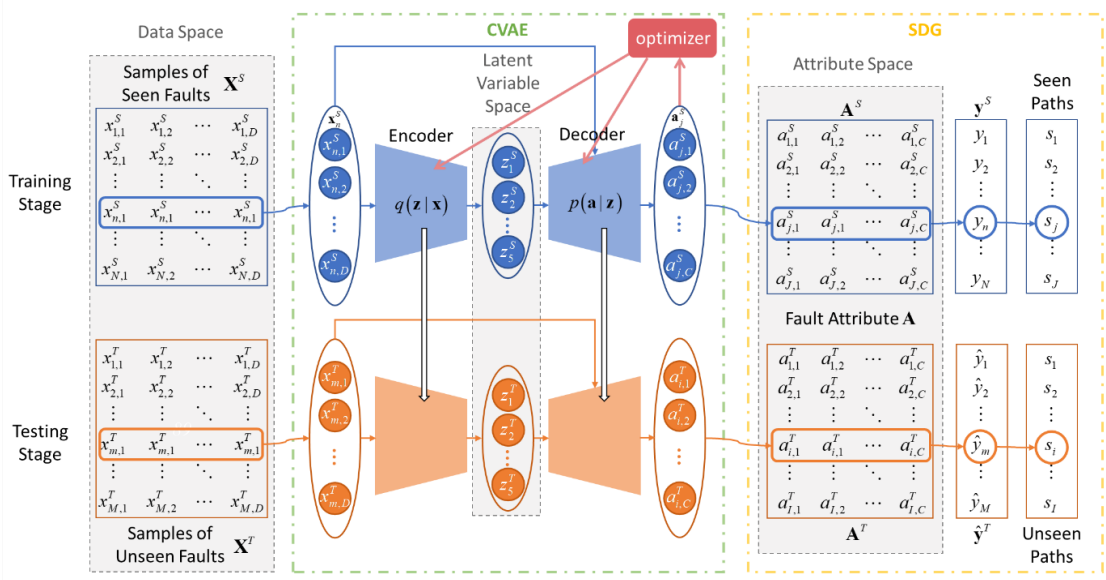

Fig. 4. Schematic diagram of fault causal diagnosis structure (top half: training phase/bottom half: testing phase)

As a solution for zero-shot learning, third attribute a_3 can be added to the attribute set as an expansion of the dimension of potential fault path (triangle markers) with physical basis. On other words, to facilitate attribute transfer and zero-shot learning, describing the fault paths with the attributes not only can assign the fault categories with physical meaning, the expansion of attribute number can prepare the model for classifying the unseen fault with meaningful explanation (zero-shot learning). Although expanding the attribute number can increase the number of possible fault path category, with the assistance of physical knowledge, categories that align with the underlying physical meaning are retained. This allows for correspondence with unseen fault data when it arises.

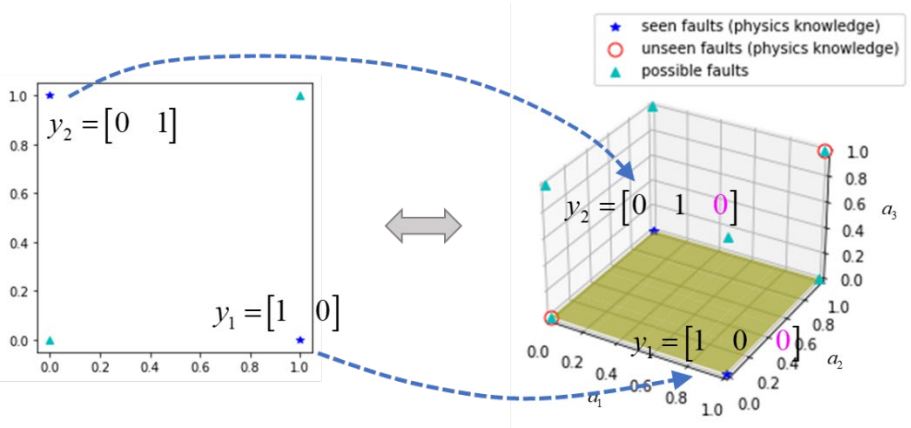

Fig. 5. Schematic diagram of entering attribute space from label space

In the framework of this zero-shot learning fault causality analysis and diagnosis method, it will be divided into training and testing stages. In the training stage in the upper half of Fig. 4, the seen fault data $\mathbf{X}^S$ is first divided into training and testing data. The training data are sent into the encoder of CVAE and their features are extracted to the latent variable space. Then the latent features are decoded for predicting fault attribute $\mathbf{A}^S$. With the variational lower bound as Eq.(7), the CVAE is trained. In the testing stage as the lower half of Fig. 4, the unseen fault samples $\mathbf{X}^T$ are used as the input of the above model to obtain the fault attribute $\mathbf{A}^T$. In the above steps, the part of extracting features and

predicting attributes is presented in the structure of CVAE. CVAE is a classification model extended from VAE, but it still has the characteristics of VAE. VAEs can infer continuous latent variables (LV) and generate reconstructed observations with complex posterior and conditional distributions. In VAE, complex nonlinearities are considered and deep neural networks are used to approximate the corresponding posterior distribution. LV in industrial systems includes those variables that contribute to the process system, usually including features of the uncertainties such as unmeasured disturbance changes, measured disturbance changes, etc. [18]. Therefore, by extending CVAE from VAE, the model can be expressed in a probabilistic manner, and the training process of extracting features from samples corresponding to attributes can be completed in one step. Finally, attribute $\mathbf{A}$ corresponds to each fault path and finds the root cause of the fault, that is, the right half of Fig. 4, which is achieved by looking into the SDG of the fault attribute.

Summarizing the entire design process above, the research steps can be organized and presented as follows :

- Based on the background physics knowledge related to chemical industry, establish the logical symbol relationship in SDG and the corresponding fault path. Then draw an attribute description table from the fault path including the seen fault and the unseen fault.
- Collect seen fault data in the past, and maximize the objective function (Eq.(7)) to train the CVAE model with the training samples. Get the model parameters after the training is complete and predict the attributes (top half of Fig. 4).
- Substitute the unseen fault data into CVAE to obtain the output attributes (the lower half of Fig. 4)
- Infer the causal relationship of the fault according to the SDG, and explain the source of the fault (right half of Fig. 4)

3. Results and Discussion

The following analysis will utilize simulation data from two jacketed continuous stirred tank reactors in series. The purpose is to compare and evaluate the effectiveness of the method proposed in this study with previous zero-sample research articles. Fig. 6 illustrates the schematic diagram of the simulation used in this study.

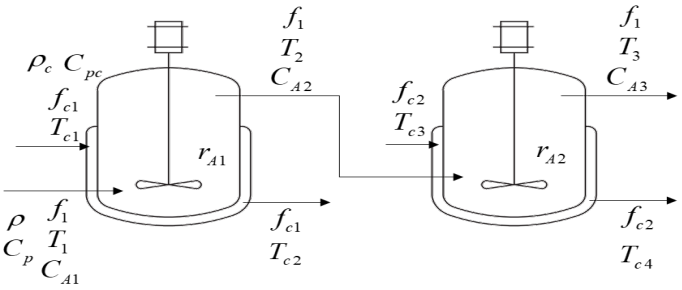

Fig. 6. Schematic diagram of simulation for continuous stirred tank reactor

Eq.(8) represents the equation for generating the simulated data, comprising a total of 13 variables (as listed in Table 1).

$$f_1 C_{A1} - f_1 C_{A2} - V k_0 e - \frac{E}{RT_2} C_{A2}^2 = 0$$

$$f_1 \rho C_\rho T_1 - UA(T_2 - T_{C2}) - f_1 \rho C_\rho T_2 - V k_0 e - \frac{E}{RT_2} C_{A2}^2 \Delta H_r = 0$$

$$f_{c1} \rho_c C_{\rho c} T_{c1} + UA(T_2 - T_{C2}) - f_{c1} \rho_c C_{\rho c} T_{c2} = 0$$

$$f_1 C_{A2} - f_1 C_{A3} - V k_0 e - \frac{E}{RT_2} C_{A3}^2 = 0$$

$$f_1\rho C_\rho T_2 - UA(T_3 - T_{C4}) - f_1\rho C_\rho T_3 - Vk_0 e - \frac{E}{RT_2} C_{A3}^2 \Delta H_r = 0$$

$$f_{c2}\rho_c C_{\rho c} T_{c3} + UA(T_3 - T_{C4}) - f_{c2}\rho_c C_{\rho c} T_{c4} = 0 \quad (8)$$

Table 1. Variable description of the simulation

Variable	Description
f_{c1}	Reaction tank 1 jacket cold water flow rate
T_{c1}	Reaction tank 1 jacket cold water inlet temperature
T_1	Reaction tank 1 inlet temperature
C_{A1}	Concentration in reaction tank 1
f_1	Reaction tank 1 inlet flow rate
T_{c3}	Reaction tank 2 jacket cold water inlet temperature
f_{c2}	Reaction tank 2 jacket cold water flow rate
T_{c2}	Reaction tank 1 jacket cold water outlet temperature
T_2	Reaction tank 1 outlet temperature/reaction tank 2 inlet temperature
C_{A2}	Reaction tank 1 outlet concentration/reaction tank 2 inlet concentration
T_3	Reaction tank 2 outlet temperature
C_{A3}	Reaction tank 2 outlet concentration
T_{c4}	Reaction tank 2 jacket cold water outlet temperature

Among these variables, the first 7 are defined as the fault source, and their relationships are represented by the SDG based on physical knowledge (Fig. 7).

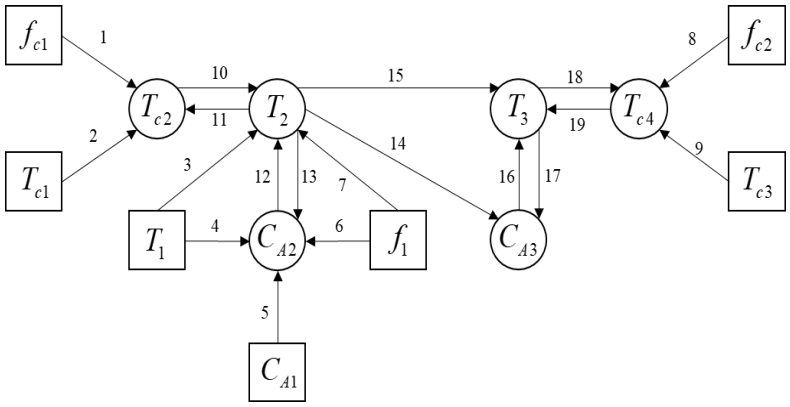

Fig. 7. SDG of the simulation program

In this case, a total of 28 fault paths are generated, depending on the size of the fault source. The attributes are established based on the presence or absence of line segments connecting variables within the SDG. Fig. 7 displays the SDG with 19 distinct attribute categories, and their corresponding descriptions are provided in Table 2.

Table 2. SDG attribute description table

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}	a_{19}
y_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
y_2	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
y_3	1	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	0	0	0
⋮											⋮								
y_{27}	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1
y_{28}	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	1

A total of 4,500 data records were generated for this case study, with 500 records assigned to three unseen fault paths (types 7, 8, and 9). Fig. 8 displays 100 data points for the 9th, 10th, and 11th types of fault paths, (including variables in Table 1) including the values of the 13 variables. From the data points, one can never determine the fault path category by visualization. Additionally, it is not possible to distinguish between known and unknown fault paths from Fig. 8 alone.

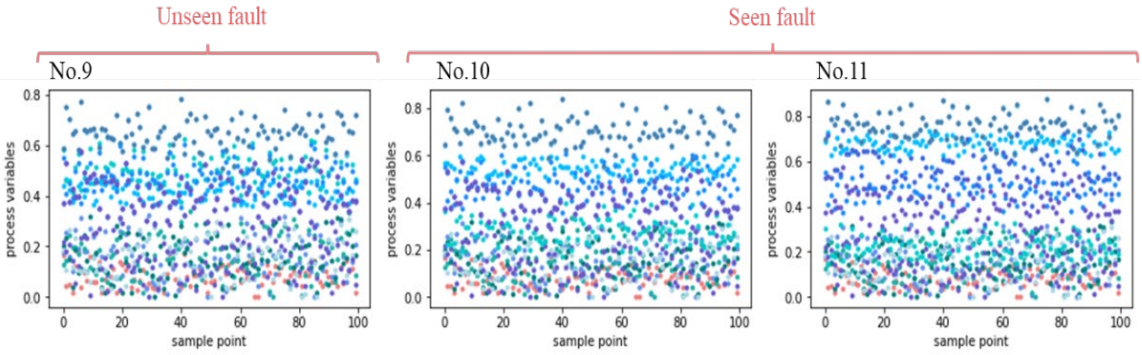

Fig. 8. Type 9, 10, and 11 fault path data (type 9 is unseen fault)

To explain the expansion of unseen fault attributes, Fig. 8 is used to map the changes in three specific attributes (a12, a17, and a18) that occur within the fault attributes (as depicted in Fig. 9). This observation aligns with the pattern observed in Fig. 5. Initially, with only two attributes (a17 and a18), the attribute points in the dimensional space lie on a plane. However, upon introducing a12, they move outside the plane, thereby increasing the potential fault paths. With the aid of physical knowledge, the position of another unseen fault path (Category 9) can be identified. This illustrates that utilizing attributes instead of labels effectively expands the space for unseen faults, facilitating the attribute transfer and enabling inference of unseen fault paths.

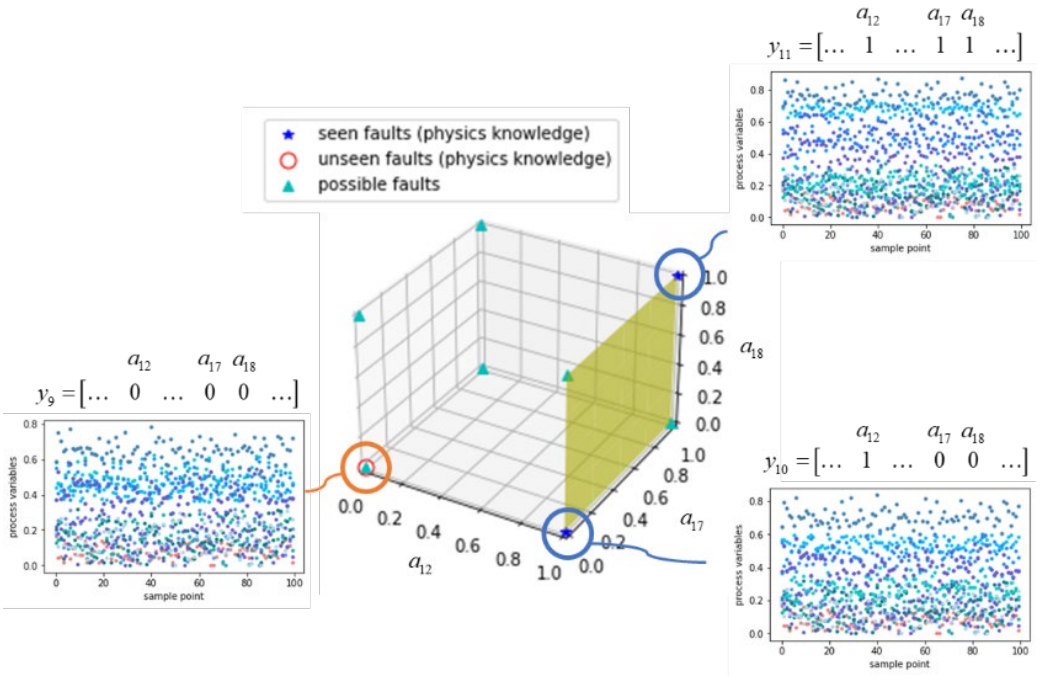

Fig. 9. Attribute spaces of the 9th, 10th, and 11th types of fault paths

Utilize Eq. (7) to train the CVAE model with both the encoder and decoder comprising three hidden layers. The activation function is set to tanh, and each layer consists of 180 neurons. The training process will run for 800 epochs. Subsequently, the trained model will be tested using both the training data and unseen fault data for evaluation. The accuracy of the training and test data is calculated separately. In the

testing phase, the trained model parameters are applied to the unseen fault data to calculate the accuracy of the unknown paths. In the method proposed by Zhao et al. [13] in 2021, the supervised principal component analysis method is initially employed to extract attribute-related features. Then, the attribute learner is trained using three different machine learning algorithms: linear support vector machine (LSVM), nonlinear random forest (RF), and probabilistic naive Bayesian (NB). Among these algorithms, RF and NB demonstrate superior test results. Consequently, the simulated data is trained using RF and NB in combination with the SDG for comparison studies. The accuracy of RF and NB, as well as the results obtained using the CVAE in this study, are presented in Table 3.

Table 3. Comparison of accuracy test results

Method	Training path	Test path	Unseen path
NB	0.51	0.515	0.784
RF	1.0	0.984	0.49
CVAE	0.839	0.84	0.966

Based on the resuls in Table 3, it is obvious that the results obtained using the NB algorithm are not as favorable as those obtained using the CVAE model structure. When employing the RF algorithm, accurate prediction of fault paths with known fault data is achieved, but it struggles to effectively predict fault paths without any fault data. This indicates that the RF algorithm lacks good generalization capability. This limitation may arise from the two-step process of feature extraction and attribute training, which could result in suboptimal cooperation between the extracted features and attributes, leading to less-than-ideal prediction outcomes. In contrast, the CVAE exhibits not only favorable results in training and testing but also excellent performance in predicting unknown fault paths. This outcome directly up-raised the effectiveness of the proposed method.

By obtaining the attribute description table from the SDG path relationship, it becomes possible to predict the attributes and thereby determine the corresponding fault path and the source of the error. This enables the diagnosis of unknown fault roots, thereby achieving the intended objective. In essence, the method proposed in this study enables the diagnosis of previously unobserved faults. This capability is invaluable in promptly identifying and addressing unforeseen faults, thereby averting potential dangers or substantial financial losses that may arise from the malfunction of a chemical plant.

However, it's worth noting that the current model and simulation data utilized in this research are static, whereas real-world data collected in industrial settings is dynamic. As a result, future enhancements to the existing model structure should focus on incorporating dynamic elements, ensuring it is better suited for real-world factory scenarios. This refinement would enable early fault detection, ultimately leading to a reduction in plant losses.

4. Conclusion

To address causal diagnosis of unseen faults in factories, this study proposes utilizing SDG to establish fault attributes, thereby creating a framework for unseen faults. Subsequently, a CVAE model is constructed to capture the relationship between known fault data and attributes, and attribute migration is employed for application to unseen faults. Through testing on a case study involving two CSTRs connected in series, the effectiveness of this model in enhancing the accuracy of zero-sample fault diagnosis tasks is demonstrated. It effectively mitigates deviations in zero-sample fault diagnosis results, confirming the model's superiority. In terms of process safety and risk engineering, the proposed zero-sample fault diagnosis model stands out for its capacity to perform fault diagnosis without relying on fault samples, rendering it highly practical for industrial processes, particularly within the context of Industry 4.0. Our upcoming research will pivot towards transitioning the model from a static to a dynamic framework. Additionally, we plan to broaden our test examples by incorporating the Tennessee Eastman process for larger-scale verification.

Acknowledgment

The authors would like to gratefully acknowledge Ministry of Science and Techonology, Taiwan, R.O.C. (MOST 109-2221-E-033-013-MY3). Also, they would like to thank China General Plastic Corporation, Taiwan, which provides the industrial data for the case study.

Declarations

Author contribution.
Mengqin Yu: Conceptualization, Methodology,Validation, Writing – original draft, Writing – review & editing, Visualization.
Yi Shan Lee: Writing – review & editing, Validation, Supervision
Junghui Chen: Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Funding acquisition.
Funding statement. This work was supported by the Ministry of Science and Technology, Taiwan, R.O.C. under Grant MOST 109-2221-E-033-013-MY3.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] Y. Ma, B. Song, H. Shi, and Y. Yang., "Fault detection via local and nonlocal embedding," *Chem. Eng. Res. Des.*, vol. 94, pp. 538–548, Feb. 2015, doi: 10.1016/j.cherd.2014.09.015.

[2] Z. Zhu *et al.*, "A review of the application of deep learning in intelligent fault diagnosis of rotating machinery," *Measurement*, vol. 206, p. 112346, Jan. 2023, doi: 10.1016/j.measurement.2022.112346.

[3] G. Wu, J. Tong, L. Zhang, Y. Zhao, and Z. Duan, "Framework for fault diagnosis with multi-source sensor nodes in nuclear power plants based on a Bayesian network," *Ann. Nucl. Energy*, vol. 122, pp. 297–308, Dec. 2018, doi: 10.1016/j.anucene.2018.08.050.

[4] J. Xu, S. Liang, X. Ding, and R. Yan, "A zero-shot fault semantics learning model for compound fault diagnosis," *Expert Syst. Appl.*, vol. 221, p. 119642, Jul. 2023, doi: 10.1016/j.eswa.2023.119642.

[5] J. Ma and J. Jiang, "Applications of fault detection and diagnosis methods in nuclear power plants: A review," *Prog. Nucl. Energy*, vol. 53, no. 3, pp. 255–266, Apr. 2011, doi: 10.1016/j.pnucene.2010.12.001.

[6] S. Gajjar, M. Kulahci, and A. Palazoglu, "Real-time fault detection and diagnosis using sparse principal component analysis," *J. Process Control*, vol. 67, pp. 112–128, Jul. 2018, doi: 10.1016/j.jprocont.2017.03.005.

[7] Y. Han, G. Song, F. Liu, Z. Geng, B. Ma, and W. Xu, "Fault monitoring using novel adaptive kernel principal component analysis integrating grey relational analysis," *Process Saf. Environ. Prot.*, vol. 157, pp. 397–410, Jan. 2022, doi: 10.1016/j.psep.2021.11.029.

[8] M. R. Maurya, R. Rengaswamy, and V. Venkatasubramanian, "Fault Diagnosis by Qualitative Trend Analysis of the Principal Components," *Chem. Eng. Res. Des.*, vol. 83, no. 9, pp. 1122–1132, Sep. 2005, doi: 10.1205/cherd.04280.

[9] M. Žarković and Z. Stojković, "Analysis of artificial intelligence expert systems for power transformer condition monitoring and diagnostics," *Electr. Power Syst. Res.*, vol. 149, pp. 125–136, Aug. 2017, doi: 10.1016/j.epsr.2017.04.025.

[10] M. A. Kramer and B. L. Palowitch, "A rule-based approach to fault diagnosis using the signed directed graph," *AIChE J.*, vol. 33, no. 7, pp. 1067–1078, Jul. 1987, doi: 10.1002/aic.690330703.

[11] C. Kang, "A Bayesian belief network-based advisory system for operational availability focused diagnosis of complex nuclear power systems," *Expert Syst. Appl.*, vol. 17, no. 1, pp. 21–32, Jul. 1999, doi: 10.1016/S0957-4174(99)00018-4.

[12] N. Liu, M. Hu, J. Wang, Y. Ren, and W. Tian, "Fault detection and diagnosis using Bayesian network model combining mechanism correlation analysis and process data: Application to unmonitored root cause variables type faults," *Process Saf. Environ. Prot.*, vol. 164, pp. 15–29, Aug. 2022, doi: 10.1016/J.PSEP.2022.05.073.

[13] C. C. Pantelides and J. G. Renfro, "The online use of first-principles models in process operations: Review, current status and future needs," *Comput. Chem. Eng.*, vol. 51, pp. 136–148, Apr. 2013, doi: 10.1016/j.compchemeng.2012.07.008.

[14] Y.-K. Liu, G.-H. Wu, C.-L. Xie, Z.-Y. Duan, M.-J. Peng, and M.-K. Li, "A fault diagnosis method based on signed directed graph and matrix for nuclear power plants," *Nucl. Eng. Des.*, vol. 297, pp. 166–174, Feb. 2016, doi: 10.1016/j.nucengdes.2015.11.016.

[15] Z. Zhou and Q. Zhang, "Model Event/Fault Trees With Dynamic Uncertain Causality Graph for Better Probabilistic Safety Assessment," *IEEE Trans. Reliab.*, vol. 66, no. 1, pp. 178–188, Mar. 2017, doi: 10.1109/TR.2017.2647845.

[16] X. Bu, H. Nie, Z. Zhang, and Q. Zhang, "An Industrial Fault Diagnostic System Based on a Cubic Dynamic Uncertain Causality Graph," *Sensors*, vol. 22, no. 11, p. 4118, May 2022, doi: 10.3390/s22114118.

[17] L. Feng and C. Zhao, "Fault Description Based Attribute Transfer for Zero-Sample Industrial Fault Diagnosis," *IEEE Trans. Ind. Informatics*, vol. 17, no. 3, pp. 1852–1862, Mar. 2021, doi: 10.1109/TII.2020.2988208.

[18] K. Hadad, M. Pourahmadi, and H. Majidi-Maraghi, "Fault diagnosis and classification based on wavelet transform and neural network," *Prog. Nucl. Energy*, vol. 53, no. 1, pp. 41–47, Jan. 2011, doi: 10.1016/j.pnucene.2010.09.006.

[19] K. Mo, S. J. Lee, and P. H. Seong, "A dynamic neural network aggregation model for transient diagnosis in nuclear power plants," *Prog. Nucl. Energy*, vol. 49, no. 3, pp. 262–272, Apr. 2007, doi: 10.1016/j.pnucene.2007.01.002.

[20] J. Qian, Z. Song, Y. Yao, Z. Zhu, and X. Zhang, "A review on autoencoder based representation learning for fault detection and diagnosis in industrial processes," *Chemom. Intell. Lab. Syst.*, vol. 231, p. 104711, Dec. 2022, doi: 10.1016/j.chemolab.2022.104711.

[21] M. Sun, H. Wang, P. Liu, S. Huang, and P. Fan, "A sparse stacked denoising autoencoder with optimized transfer learning applied to the fault diagnosis of rolling bearings," *Measurement*, vol. 146, pp. 305–314, Nov. 2019, doi: 10.1016/j.measurement.2019.06.029.

[22] C. Zhang, D. Hu, and T. Yang, "Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost," *Reliab. Eng. Syst. Saf.*, vol. 222, p. 108445, Jun. 2022, doi: 10.1016/J.RESS.2022.108445.

[23] K. Wang, M. G. Forbes, B. Gopaluni, J. Chen, and Z. Song, "Systematic Development of a New Variational Autoencoder Model Based on Uncertain Data for Monitoring Nonlinear Processes," *IEEE Access*, vol. 7, pp. 22554–22565, 2019, doi: 10.1109/ACCESS.2019.2894764.

[24] X. Yan, D. She, and Y. Xu, "Deep order-wavelet convolutional variational autoencoder for fault identification of rolling bearing under fluctuating speed conditions," *Expert Syst. Appl.*, vol. 216, p. 119479, Apr. 2023, doi: 10.1016/j.eswa.2022.119479.

[25] A. Mishra, S. K. Reddy, A. Mittal, and H. A. Murthy, "A Generative Model for Zero Shot Learning Using Conditional Variational Autoencoders," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2018, vol. 2018-June, pp. 2269–22698, doi: 10.1109/CVPRW.2018.00294.

[26] A. A. Pol, V. Berger, C. Germain, G. Cerminara, and M. Pierini, "Anomaly Detection with Conditional Variational Autoencoders," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Dec. 2019, pp. 1651–1657, doi: 10.1109/ICMLA.2019.00270.

[27] Y. Wei, D. Wu, and J. Terpenny, "Learning the health index of complex systems using dynamic conditional variational autoencoders," *Reliab. Eng. Syst. Saf.*, vol. 216, p. 108004, Dec. 2021, doi: 10.1016/j.ress.2021.108004.

[28] K. A. Reay and J. D. Andrews, "A fault tree analysis strategy using binary decision diagrams," *Reliab. Eng. Syst. Saf.*, vol. 78, no. 1, pp. 45–56, Oct. 2002, doi: 10.1016/S0951-8320(02)00107-2.

[29] Z. Masalegooyan, F. Piadeh, and K. Behzadian, "A comprehensive framework for risk probability assessment of landfill fire incidents using fuzzy fault tree analysis," *Process Saf. Environ. Prot.*, vol. 163, pp. 679–693, Jul. 2022, doi: 10.1016/j.psep.2022.05.064.

[30] R. M. Sinnamon and J. D. Andrews, "Improved efficiency in qualitative fault tree analysis," *Qual. Reliab. Eng. Int.*, vol. 13, no. 5, pp. 293–298, Sep. 1997, doi: 10.1002/(SICI)1099-1638(199709/10)13:5<293::AID-QRE110>3.0.CO;2-Y.

[31] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly Accurate Machine Fault Diagnosis Using Deep Transfer Learning," *IEEE Trans. Ind. Informatics*, vol. 15, no. 4, pp. 2446–2455, Apr. 2019, doi: 10.1109/TII.2018.2864759.

[32] Y. Pan, F. Mei, H. Miao, J. Zheng, K. Zhu, and H. Sha, "An Approach for HVCB Mechanical Fault Diagnosis Based on a Deep Belief Network and a Transfer Learning Strategy," *J. Electr. Eng. Technol.*, vol. 14, no. 1, pp. 407–419, Jan. 2019, doi: 10.1007/s42835-018-00048-y.

[33] "Zero-data learning of new tasks," in *Proceedings of the 23rd national conference on Artificial intelligence*, 2008, pp. 646–651. [Online]. Available at: https://dl.acm.org/doi/10.5555/1620163.1620172.

[34] S. Rahman, S. Khan, and F. Porikli, "A Unified Approach for Conventional Zero-Shot, Generalized Zero-Shot, and Few-Shot Learning," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5652–5667, Nov. 2018, doi: 10.1109/TIP.2018.2861573.

[35] J. Yang, C. Wang, and C. Wei, "A novel Brownian correlation metric prototypical network for rotating machinery fault diagnosis with few and zero shot learners," *Adv. Eng. Informatics*, vol. 54, p. 101815, Oct. 2022, doi: 10.1016/j.aei.2022.101815.

[36] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 951–958, doi: 10.1109/CVPR.2009.5206594.

[37] M. Mou, X. Zhao, K. Liu, and Y. Hui, "Variational autoencoder based on distributional semantic embedding and cross-modal reconstruction for generalized zero-shot fault diagnosis of industrial processes," *Process Saf. Environ. Prot.*, vol. 177, pp. 1154–1167, Sep. 2023, doi: 10.1016/j.psep.2023.07.080.

[38] B. Li and C. Zhao, "Federated Zero-Shot Industrial Fault Diagnosis With Cloud-Shared Semantic Knowledge Base," *IEEE Internet Things J.*, vol. 10, no. 13, pp. 11619–11630, Jul. 2023, doi: 10.1109/JIOT.2023.3243401.

[39] Y. Zhang, S. Zhang, X. Jia, X. Zhang, and W. Tian, "A novel integrated fault diagnosis method of chemical processes based on deep learning and information propagation hysteresis analysis," *J. Taiwan Inst. Chem. Eng.*, vol. 142, p. 104676, Jan. 2023, doi: 10.1016/j.jtice.2023.104676.

[40] H. Ali, A. S. Maulud, H. Zabiri, M. Nawaz, H. Suleman, and S. A. A. Taqvi, "Multiscale Principal Component Analysis-Signed Directed Graph Based Process Monitoring and Fault Diagnosis," *ACS Omega*, vol. 7, no. 11, pp. 9496–9512, Mar. 2022, doi: 10.1021/acsomega.1c06839.