공학박사 학위논문

Data-driven Fault Detection and Diagnosis Using Machine Learning Techniques and Information Theory

머신 러닝 기법과 정보 이론을 이용한 데이터 기반 이상 감지 및 진단

2021 년 8 월

서울대학교 대학원

화학생물공학부

이 호 동

Data-driven Fault Detection and Diagnosis Using Machine Learning Techniques and Information Theory

지도교수 이 종 민

이 논문을 공학박사 학위논문으로 제출함

2021년 8월

서울대학교 대학원

화학생물공학부

이 호 동

이 호 동의 공학박사 학위논문을 인준함

2021년 8월

위 원 장 _____ 이 원 보 _____

부위원장 _____ 이 종 민 _____

위 원 _____ 남 재 욱 _____

위 원 _____ 정 동 휘 _____

위 원 _____ 나 종 걸 _____

Abstract

Data-driven Fault Detection and Diagnosis Using Machine Learning Techniques and Information Theory

Hodong Lee

School of Chemical and Biological Engineering

The Graduate School

Seoul National University

Process monitoring system is an essential component for efficient and safe operation. Process faults can affect the quality of the product or interfere with the normal operation of the process, hindering productivity. In the case of chemical processes dealing with explosive and flammable materials, process fault can act as a threat to the process safety which should be the top priority. Meanwhile, modern processes demand a more advanced monitoring system as the scope of the process expands and the process automation and intensification progress.

The framework of the process monitoring system can be classified into three stages. It is divided into process fault detection that determines the existence of process faults in a system in real time, fault diagnosis that identifies the root cause of the faults, and finally, process recovery that removes the cause of the fault and normalizes the process. In particular, various methodologies for fault detection and diagnosis have been proposed, and they can be categorized into three approaches. Data-driven methodologies are widely utilized due to the

general applicability and the conditions under which abundant process data are provided compared to analytical methods based on the detailed first principle models and knowledge-based methods on the specific domain knowledge. Furthermore, the advantage of the data-driven methods can be prominent as the scale and complexity of the process increase. In this thesis, fault detection and diagnosis methodologies to improve the performance of existing data-driven methods are proposed.

Conventional data-driven fault detection systems have been developed based on dimensionality reduction methods. The fault detection models using dimensionality reduction identify the low dimensional latent space defined by features inherent in process data, performing process monitoring based on it. As the representative methods, there are principal component analysis which is the conventional multivariate process monitoring approach, and autoencoder which is one of the machine learning techniques. Although the monitoring systems using various machine learning techniques have been widely utilized thanks to sufficient process data and good performance, a monitoring scheme that improves the performance of up-to-date methods is required due to the aforementioned factors. To improve the performance of such a data-driven monitoring system, approaches that change the structure of the model or learning procedure have been mainly discussed. Meanwhile, the nature that data-driven methods are ultimately dependent on the quality of the training dataset still remains. In other words, a methodology to enhance the completeness of the monitoring system by supplementing the insufficient information in the training dataset is required. Thus, a process fault detection method that combines data augmentation techniques is proposed in the first part of the thesis.

Data augmentation has been mostly employed to manage the deficiency of certain classes, between-class imbalance, in a classification problem. In this

case, data augmentation can be effectively applied to improve the training performance by balancing the amount of each class. Data augmentation in this study, on the other hand, is applied to alleviate the within-class imbalance. The process data in normal operation has characteristics that the data samples in the borderline of normal and abnormal state are relatively sparse. Given that the modeling of the fault detection system corresponds to defining the low-dimensional feature space and monitoring the system in it, it can be expected that the supplement of the samples on the boundary of the normal state would positively affect the training process. In this context, the proposed method is as follows.

First, variational autoencoder which is a generative model is constructed to generate the synthetic data using the original training data. The sample vector corresponding to the boundary region of the low-dimensional distribution of the normal state learned by the generative model is generated as the synthetic data and augmented to the original training data. Based on the augmented training data the fault detection system is established using autoencoder, a machine learning algorithm for feature extraction. The feature learning of autoencoder can be performed more effectively by using the augmented training data, which can lead to the improvement of the fault detection system that distinguishes between normal and abnormal states.

The dimensionality reduction methods have been also utilized as the fault isolation method known as the contribution charts. However, the approaches showed limited performance and inconsistent analysis results due to the information loss during the dimension reduction process. To resolve the limitations of the conventional method, the approaches that directly figure out the causal relationships between process variables have been developed. As one of them, transfer entropy, an information-theoretic causality measure, is generally known to have good fault isolation performance in the fault isolation of nonlinear processes because it is neither linearity assumption nor model-based method.

However, it has been limitedly applied to the small-scale process because of the drawback that the causal analysis using transfer entropy requires costly density estimation. To resolve the limitation, the method that combines graphical lasso which is a regularization method with transfer entropy is proposed.

Graphical lasso is a sparse structure learning algorithm of the undirected graph model, which can be used to sort out the most relevant subgroup in the entire graph model. As graphical lasso algorithm presents the output as a highly correlated subgroup with the rest of the variables, the iterative application of graphical lasso can substitute the entire process into several subgroups. This process can greatly reduce the subject of causal analysis by excluding relationships with little relevance in advance. Accordingly, the limitation of demanding cost of transfer entropy can be mitigated and thus the applicability of fault isolation using transfer entropy can be expanded through this process.

Combining the two methods, the following fault isolation method is proposed. First of all, the entire process variables are divided into the five most relevant subgroups based on the data when the fault has occurred. The root cause variable can be isolated from the most significant relationship by calculating the causality measure using transfer entropy only within each subgroup. It is possible to significantly reduce the computational cost due to transfer entropy by efficiently decreasing the subject of causal analysis through graphical lasso. Therefore, the proposed method is noteworthy in that it enables the application of fault isolation using transfer entropy for industrial-scale processes.

The proposed methodologies in each stage are verified by applying them to the industrial-scale benchmark process model, the Tennessee Eastman process (TEP). The benchmark process model is suitable to test the performance of the proposed methods because it is a process model with similar complexity as a real chemical process involving multiple unit operations, recycle stream, and chemical reactions in it. The performance test is performed with respect to the

28 predefined process faults scenarios in TEP model. Application results of the proposed fault detection method performed better than the case using the conventional approach in terms of the fault detection rate. In some fault cases, the fault detection delay, the time required to first detect a fault since it occurred, also showed improvement. Fault isolation results by the proposed method integrating transfer entropy with graphical lasso showed that it could effectively identify the cause of the process fault with only about 20% of the computational cost compared to the base case that directly applied the transfer entropy to the entire process for fault isolation. In addition, the demonstration results suggested that the proposed method could outperform the base case in terms of accuracy in some particular cases.

Keywords: Process monitoring; Fault detection and isolation; Autoencoder; Variational Autoencoder; Transfer Entropy; Graphical Lasso

Student Number: 2016-21049

Contents

List of Tables

List of Figures

보존용 학위논문 정오표

페이지	정정 전	정정 후
pp. 42, 48, 50, 52, 57, 84, 88, 90, 99, 101, 104, 108, 111, 114	편집상 오류 (표기한 페이지의 본문상 Table 캡션 앞 줄바꿈)	파일 변환상의 오류로 논문 내용을 파악하는데 문제가 없음을 밝히오니 참고해주십시오.

Chapter 1

Introduction

1.1. Research Motivation

Process monitoring system is an essential component for efficient and safe operation. Various types of faults such as sensor failure or process malfunction can happen in a certain module during the operation of a process, which could damage the reliability of the process operation. Due to either an increase in downtime to normalize process fault or an occurrence of a product that fails to meet necessary quality specifications, the productivity of a process can be hampered. Furthermore, the safety of the process which is of the greatest importance cannot be assured in the case of the chemical process where it deals with flammable and explosive materials. Meanwhile, modern processes demand a more advanced monitoring system due to the following factors: process automation based on the complex process control systems, intensification derived by process integration and development, and expansion of monitoring scope. Furthermore, the chemical process generally includes plenty of recycle streams to increase productivity, which makes it more difficult to appropriately monitor a process fault in a precise and timely manner. Thus, a process monitoring system should be prepared to efficiently detect and accurately diagnose process faults.

The framework of the process monitoring system can be classified into four stages as Figure 1.1 [1]. As the first step, fault detection is performed in real time to determine whether a process fault has occurred. Once any fault would be detected, fault isolation and identification are conducted to pinpoint the root cause variable and analyze the type and scale, respectively. These two processes are commonly referred to as fault diagnosis. After the fault is properly diagnosed, the process can be restored to the normal operation by eliminating the

source of a fault. In particular, the first two among these four stages in the monitoring workflow, fault detection and isolation (FDI), have been intensively studied.

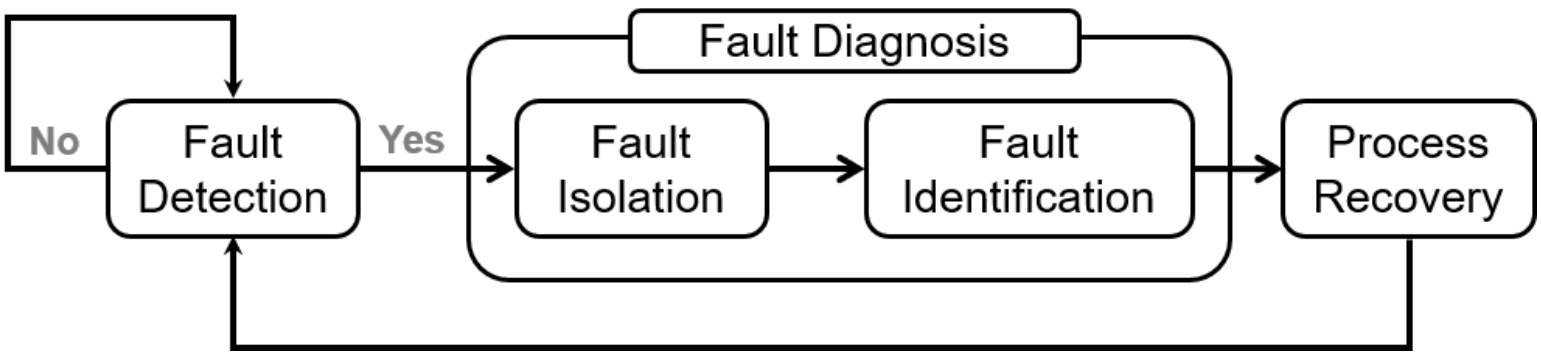

Figure 1.1. Schematic diagram for workflow of process monitoring system.

FDI system can be categorized into three approaches: analytical approaches, knowledge-based method, and historical data-driven method. The analytical approaches utilize mathematical models based on the detailed first principle modeling. However, the comprehensive first principle modeling is commonly unavailable for the large-scale industrial process. The knowledge-based monitoring systems are based on the prior domain knowledge to construct a monitoring system. Despite its interpretability about the output, the costly and time-consuming characteristics of knowledge-based methods have restricted their application and progress in the industrial-scale process. Meanwhile, the historical data-driven monitoring systems provide better applicability because they only utilize the historical data in the construction of a monitoring system. The data-driven methods have been widely adopted in the development of the monitoring system for the industrial process because a sufficient amount of data from the extensive sensor networks can be provided. Furthermore, the advantage of data-driven methods in terms of applicability compared to knowledge-based methods becomes more pronounced as the complexity and scale of the process increase.

As the conventional data-driven methodologies, various dimensionality reduction methods for multivariate process monitoring systems such as principal component analysis (PCA), independent component analysis (ICA), Fisher discriminant analysis (FDA) have been developed. These are the latent variable methods to investigate the intrinsic characteristics of the state in lower-dimensional feature space. The latent variables can be obtained by extracting feature representation that can capture the variability of the data while preserving the structural correlations between the process variables. Fault detection systems to monitor process faults can be established by defining monitoring statistics using the feature representation and its reconstruction. For example, Hotelling's T^2 and squared prediction error (SPE) correspond to the monitoring statistics of

the fault detection system using PCA. Most of them, however, have the common limitation that assumes the linearity of the process variables to characterize the representations in the feature space. It leads to the restriction of the monitoring performance with respect to the nonlinear process in general. Although kernel PCA(KPCA) [2] was suggested to tackle the nonlinear relationship, it is commonly considered impractical in the industrial-scale process as the excessive computational burden and high sensitivity of the kernel trick.

To manage the nonlinearity of the process, machine learning techniques employing the neural network framework have been recognized as an alternate approach for nonlinear feature extraction. As the representative scheme for efficient feature extraction, autoencoders (AEs) have shown better capability as a nonlinear monitoring system. AEs are neural networks with nonlinear activation functions to deal with the nonlinearity of the data. It can be also characterized by a bottleneck hidden layer which encourages the feature extraction, which can be achieved in an unsupervised fashion due to the network structure reconstructing input itself.

Despite its superiority, it is bound to reach the limit of improvement. Thus, various approaches have been proposed to further enhance the performance of the monitoring systems from different perspectives. For example, the performance of a model can be improved by adjusting the model capacity through the depth and width of the network. As a more fundamental approach to modify the training procedure, it can be also effective to adjust either the optimizer or the regularizer. Otherwise, a totally new type of layer such as a recurrent layer that takes into account the autocorrelation in the process data or a convolutional layer that extracts useful features preserving the spatial information could lead to an improvement in performance. Meanwhile, the nature that data-driven methods are ultimately dependent on the quality of the training data still remains. In other words, an insufficient amount of training data or the data-rich

but the information-poor issue can hinder the successful modeling of monitoring systems using machine learning techniques. Therefore, a methodology to enhance the completeness of the monitoring system by alleviating the insufficient information issue in the training dataset is required.

The methods for fault isolation have been extensively studied as well. The contribution chart is an extension of dimensionality reduction methods used in the modeling of the fault detection system. It can be developed by investigating the contribution of the original process variables with respect to the faulty instances detected from the fault detection system. These approaches, however, have shown inherent limitations such as information loss during the dimensionality reduction and inconsistency depending on sample points. To avoid these limitations, methodologies that directly analyze causal relationships between process variables have been suggested.

Other methods include the use of Granger causality and transfer entropy, which are the predictive causality measures. Granger causality is a causal analysis tool that examines a causal relationship by quantifying the improvement due to additional predictors of the vector autoregressive (VAR) models. Transfer entropy can be regarded as an information-theoretic interpretation of Granger causality. It was derived from mutual information, which is a common measure to estimate the correlation, by replacing the joint probability with conditional probability. Transfer entropy shows better performance than Granger causality in general due to the fact that it is free from linearity assumptions and models such as the VAR model. In spite of its superiority, the application of transfer entropy in the industrial-scale process has been limited because it includes costly density estimation during the calculation of the measure. Also, the fact that such a high-cost evaluation should be made for all of the pairwise combinations for the entire process hinders the active applications of the industrial process.

Therefore, a methodology that can be performed in a computationally efficient way while preserving reliable causal analysis performance is required.

1.2. Research Objectives

In this thesis, integrated methodologies for each component of FDI have been proposed to address the shortcomings, which can facilitate the general applicability of industrial processes. The proposed methodology for fault detection systems comprises manifold learning via autoencoder and generative modeling for data augmentation using a variant of variational autoencoder. Another proposed methodology for fault isolation includes transfer entropy for causality analysis from information theory and graphical lasso which is a regularization method to sort out the most relevant subset from the whole process.

In the first part, a fault detection system integrated with data augmentation for the purpose of supplement insufficient information in manifold learning is proposed. To do so, variational autoencoder which is a generative model is employed to supplement the informative training samples into the original training dataset, and the augmented dataset is provided as the new training dataset for a manifold learning method based on autoencoder algorithm. The artificially generated samples corresponding to the boundary region of the normal samples are informative but rarely happened. Therefore, augmentation of data in this region allows building a more effective monitoring system by providing information to promote manifold learning especially in terms of detailed borderline. The monitoring system using autoencoder is constructed based on the augmented dataset. Finally, the proposed method is verified by comparing the monitoring performance to the base case without the data augmentation.

Next, a fault isolation system that incorporates a regularization method for subset selection with a causality analysis that is transfer entropy is proposed. The proposed method is intended to leverage a reliable causality measure so that fault isolation could be performed efficiently even for industrial-scale processes. Graphical lasso, which is a regularization technique for sparse structure learning of the undirected graphical model, is employed to divide the whole

process variables into several subgroups. Afterward, causal analysis for fault isolation using transfer entropy is carried out only for the relationships in the subgroups, by which the computational cost of the redundant relations in the causal analysis stage can be saved. To verify the effectiveness of the proposed method, it is applied to a benchmark process to assess the fault isolation performance comparing to the conventional situation. The demonstrations of the proposed methods for both fault detection and isolation use the widely used chemical benchmark process, the Tennessee Eastman process.

1.3. Outline of the Thesis

The outline of the thesis is as follows. In Chapter 2, the backgrounds and required preliminaries of the elementary components for the proposed methods in the remainders are introduced. Chapter 3 proposes a methodology for fault detection systems incorporating the data augmentation strategy. In this chapter, the detailed methods for data augmentation and fault detection modeling are presented with a comprehensive case study and discussion. Chapter 4 provides an integrated methodology for fault isolation. In addition, it includes a detailed description of the procedure for applying the proposed method, and application results about two types of industrial processes. Finally, the concluding remarks with suggestions for future work are presented in Chapter 5.

Chapter 2

Backgrounds and Preliminaries

2.1. Autoencoder

Autoencoder(AE) is an unsupervised machine-learning technique for feature extraction. It encodes the input data onto low-dimensional latent features and reconstructs the data using only the encoded feature by squeezing the middle layer of the symmetric network, as shown in Figure 2.1. The compression part of the network in Figure 2.1, which produces a latent feature, is the encoder, and the opposite part, which is the decoder, applies the converse functionality. The encoder function mapping an input $\mathrm{x} \in R^n$ into a latent vector $\mathrm{h} \in R^m$ through general functions is as follows:

$$h = F(x) = f(W_1 \cdot x + b_1), \tag{2.1}$$

where W_1 is an $\mathrm{m} \times \mathrm{n}$ weight matrix, b_1 is an $\mathrm{m} \times 1$ bias vector, and $f(\cdot)$ is an activation function. The activations for hidden layers typically employ nonlinear functions such as a sigmoid, tangent hyperbolic, and rectified linear units, except for the visible layers having linear activations. By adopting a bottleneck structure in the latent space, AE is guided to extract a rich representation that is advantageous to the reconstruction of the input. The reconstruction of z is as follows:

$$x' = G(h) = g(W_2 \cdot h + b_2), \tag{2.2}$$

where W_2 is an $\mathrm{n} \times \mathrm{m}$ weight matrix, b_2 is an $\mathrm{n} \times 1$ bias vector, and $\mathrm{g}(\cdot)$ is an activation function. The weight matrices W_1 and W_2 have distinct weight values in general, but can be tied, i.e., $W_1 = W_2^T$, in some cases.

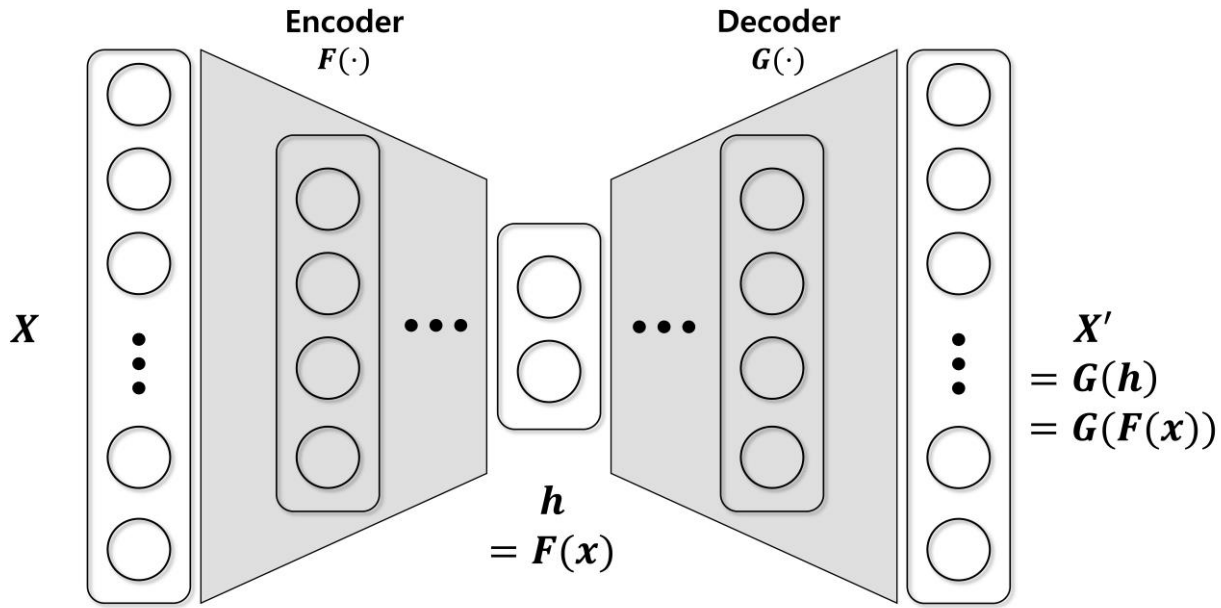

Figure 2.1. Conceptual scheme of AE.

The objective function of an AE, which is the loss function that the optimizer should minimize, has different forms depending on the data type, such as the squared error and cross-entropy. Following the typical choices in the cases of linear regression, the reconstruction loss function across a given set of training samples, D, can be represented as follows:

$$\begin{aligned} \mathrm{L} &= \min_{W,b} \frac{\sum_{x \in D} \|x - z\|^2}{\|D\|} \\ &= \min_{W,b} \sum_{x \in D} \|x - g(f(x))\|^2 / \|D\|. \end{aligned} \tag{2.3}$$

The network can be extended to an arbitrary number of hidden layers and nodes in both the encoding and decoding parts. Special attention is needed to determine the dimensions of networks depending on the applications to prevent underfitting and/or overfitting. As a typical approach in machine learning, regularization methods such as weight regularizations, where the objective function includes the norm of the weights [3], dropouts [4], batch normalization [5], and pruning [6] can implicitly help a network avoid overfitting starting from a network with sufficient model capacity.

The operations through weights and biases used to reveal the latent vector z correspond to the projection of input data from the original to feature space in PCA. If the linear activation replaces the nonlinear functions, AE is reduced to PCA, which is conceptually equivalent [7]. Thus, AE is a nonlinear generalization of PCA, which is a conventional dimensionality reduction method used for purposes such as feature extraction, visualization, and data compression. Although KPCA [2], a nonlinear extension of PCA using kernel trick, can be compared with AE using the same nonlinear dimensionality reduction method, the performance of KPCA depends entirely on the type of kernel and represents poor robustness against the kernel parameters depending on the applications. However, AE copes inherently with nonlinearity through nonlinear activation

functions in each layer. Meanwhile, there exist variants of AE to improve the limitations in terms of robustness against process noise. Denoising autoencoder (DAE) [8] can improve the robustness by intentional random corruption of the input data to promote the reconstruction ability even under noisy situations. For a similar purpose, contractive autoencoder (CAE) [9] was devised by explicitly penalizing the objective function by adding a term representing the sensitivity of hidden representations to the input perturbations, $\left\|J_f(x)\right\|_F^2 = \sum_{ij}(\frac{\partial h_j(x)}{\partial x_i})^2$.

2.2. Variational Autoencoder

Variational autoencoder(VAE) is a popular generative model that learns the data distribution to generate new samples aside from existing data in an unsupervised manner. Once the training of VAE is completed, the latent code z to be used as the input for the generation process is sampled in the feature space. The main objective of VAE is to generate new synthetic samples of the original space using the latent code z sampled from a low-dimensional feature space through a generation network that corresponds to the decoder, as shown in Figure 2.2. Meanwhile, it is insufficient to train a generation network to generate plausible samples with only randomly sampled codes drawn from a prior distribution p(z), which is typically assumed as a normal distribution that possesses little information producing meaningful samples. Thus, the encoder network is introduced to provide evidence to produce a latent code that allows the decoder to reconstruct at least the training samples well. At this point, the true posterior $p(z|x)$, which is generally intractable, is replaced by the approximated posterior $q_\phi(z|x)$ parameterized by ϕ, which is typically assumed as a multivariate Gaussian, leading to a closed-form loss term, that is the variational inference method.

As a result, the entire structure incorporating an inferential encoding network and a generative decoding network is analogous to AE in terms of compressing the input data into a low-dimensional latent space and then restoring the data. Thus, the methodology, which is an autoencoder with a variational inference method for generative modeling, is called a variational autoencoder. Given the ultimate purpose of development and the process of establishing a generative model, the basis of VAE has little to do with AE, except for the structural similarity of the final form of the objective function [10].

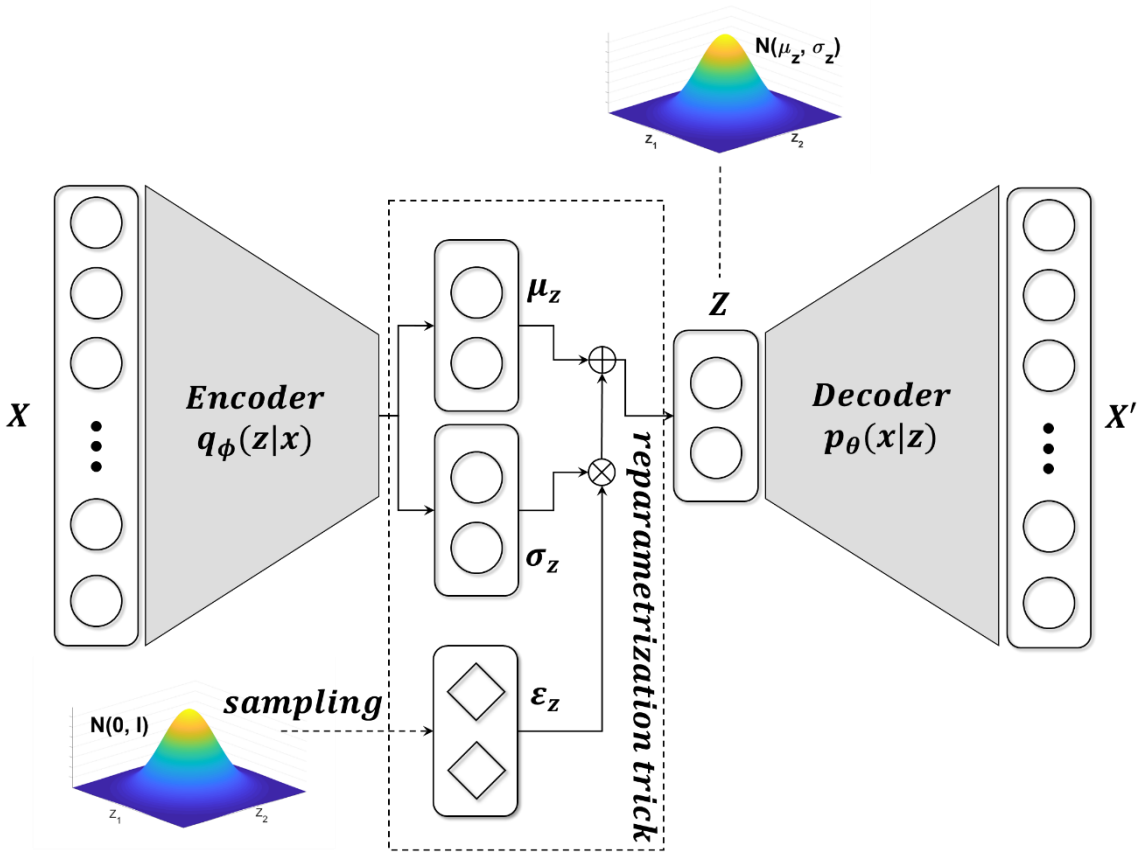

Figure 2.2. Conceptual scheme of VAE and reparameterization trick.

For a vanilla VAE, first introduced by Kingma et al. [11], the objective function is

$$\log p_\theta\left(x^{(i)}\right) = D_{KL}[q_\phi(z|x^{(i)})||p_\theta\left(z|x^{(i)}\right)] + L(\theta, \phi; x^{(i)}), \tag{2.4}$$

where the first term on the right-hand side, which represents the variational inference process, forces the approximate posterior $q_\phi(z|x^{(i)})$ to match the true posterior $p_\theta(z|x^{(i)})$ by using Kullback–Leibler (KL) divergence, and the second term is the evidence lower bound (ELBO) on the marginal likelihood of data point i. Because the KL divergence is non-negative, the marginal likelihood is greater than the ELBO. The ELBO can be further decomposed as follows:

$$\begin{aligned} \log p_\theta\left(x^{(i)}\right) &\geq L\left(\theta, \phi; x^{(i)}\right) \\ &= \mathbb{E}_{q_\phi(z|x)}\left[log p_\theta\left(x^{(i)}|z\right)\right] - D_{KL}(q_\phi(z|x^{(i)})||p(z)). \end{aligned} \tag{2.5}$$

Instead of directly maximizing the marginal likelihood, the ELBO is maximized with respect to both the variational parameters ϕ and the generative parameters θ. It is noteworthy that the reparameterization trick suggested by Kingma et al. [11] makes it possible for the VAE formulation to jointly optimize both parameters in the encoder and decoder by utilizing the stochastic gradient descent method, even though it includes a non-differentiable sampling process, as shown in Figure 2.2. In addition, by assuming both the prior $p_\theta(z)$ and the inferential posterior $q_\phi(z|x^{(i)})$ as having a multivariate Gaussian distribution, and the generative posterior $p_\theta\left(x^{(i)}|z\right)$ as a multivariate Gaussian or Bernoulli distribution depending on the application, the ELBO can be represented as a closed form using the parameters of the encoder and decoder network. The detailed proof and formula can be found in the study by Kingma et al. and the appendix thereof [11]. In conclusion, the two terms in Eq. (2.5) can be respectively interpreted as a reconstruction error and a regularization encouraging the

approximated posterior to fit into the prior, which will eventually be used as a sampling distribution.

2.3. Transfer Entropy

Previous studies suggested various information-theoretic measures to quantify informational content as part of the identification of a causal structure in the system. One of the basic and most important concepts of information theory is the Shannon entropy. It is defined as the measure of the average number of bits required to optimally encode the independent variable *I* of the process following a probability distribution *p(i)* [12]. The mathematical definition of the Shannon entropy is

$$H_I = -\sum_i p(i) \log p(i) \tag{2.6}$$

In information theory, the measure of information, the Shannon entropy, corresponds to the uncertainty of an event, which is inversely related to the probability of a state for a certain variable. In other words, the higher the probability of occurrence, the smaller the amount of information that observation can provide. Through the semantic relationship in the subsequent metrics, quantification of the amount of information was used to identify the dependency structure between variables.

For the system consisting of more than one variable, there may exist information transfer if the subsystems interact with each other. Therefore, by measuring the extent of information transference, one can figure out the inherent relationship that does not manifest itself. As the second fundamental concept of information theory, the Kullback entropy(K_I) [13], also known as Kullback-Leibler divergence, signifies an excess bit required by using a hypothetical distribution $q(i)$ instead of the original distribution $p(i)$ and is defined as

$$K_I = \sum_i p(i) \log \frac{p(i)}{q(i)} \tag{2.7}$$

As a special case of the Kullback entropy, mutual information (MI), the excess amount of code induced by erroneously assuming that the two subsystems

are independent, is defined as

$$M_{I,J} = \sum p(i,j) \log \frac{p(i,j)}{q(i,j)} = \sum p(i,j) \log \frac{p(i,j)}{p(i)p(j)} \tag{2.8}$$

The mutual information indicates the degrees of deviation from the assumption of independence for each subsystem. With manipulations based on the definition of the Shannon entropy, mutual information can be decomposed into separate terms of Shannon entropies.

$$M_{IJ} = H_I + H_J - H_{IJ} \tag{2.9}$$

Although mutual information can discern whether the information can be mutually exchanged through the deviations from independence assumptions, its directionality cannot be confirmed due to its symmetric property under the exchange of two systems. This point mainly motivated the suggestion of the advanced information-theoretic measure, the transfer entropy, which is one of the two key components of the methodology proposed in this study. Another deficiency of mutual information is that the static probabilities constructing mutual information are insufficient to reflect the structure of dynamical information transfer.

Consequently, the transition probability can substitute the static probability in mutual information to incorporate the dynamical structure and clarify the direction of information flow. The probability of the occurrence of a particular subsequent state conditioned upon the prior state can represent the transition probability. By introducing conditional probability, we can calculate the entropy rate, which measures the average number of bits required to express the successor state given all prior states [12]. The larger the entropy rate, the more bits are needed to express the relationship between the preceding and subsequent state, suggesting that there is a significant association between process data. The entropy rate can be interpreted as the difference between two Shannon entropies using Bayes' rule, $\mathrm{p}\left(i_{n+1}\middle|i_n^{(k)}\right) = p\left(i_{n+1}^{(k+1)}\right)/p(i_n^{(k)})$.

$$h_I = -\sum p\left(i_{n+1}, i_n^{(k)}\right) \log p(i_{n+1}|i_n^{(k)})$$
$$= H_{I^{(k+1)}} - H_{I^{(k)}} \tag{2.10}$$
$$\text{(n: time index,} \quad i_n^{(k)} = (i_n, \dots, i_{n-k+1}))$$

To extend this discussion, the dynamics of shared information between subsystems in the target process need to be determined. In this case, contrary to the case of the mutual information rate, which measures the deviation from the independence assumption, transfer entropy evaluates the deviation from the generalized Markov property:

$$\mathrm{p}\left(i_{n+1}\middle|i_n^{(k)}\right) = \mathrm{p}\left(i_{n+1}\middle|i_n^{(k)}, j_n^{(l)}\right)$$
$$(i_n^{(k)} = (i_n, \dots, i_{n-k+1}), \quad j_n^{(l)} = (j_n, \dots, j_{n-l+1}), \tag{2.11}$$
$$\text{k: embedding dimension of variable } i,$$
$$\text{l: embedding dimension of variable } j\text{)}$$

Based on the generalized Markov property, if there is no information transfer from subsystem *J* to *I*, the historical data of subsystem *J* are insignificant for predicting future states of subsystem *I*. Otherwise, the degrees of deviation from the generalized Markov property would not be negligible. The extent of deviation can also be quantified by the Kullback entropy, which is the definition of transfer entropy. Analogous to mutual information, Schreiber[12] formulated the transfer entropy, which utilizes two sets of augmented data of variables under investigation. The two variables regarded as the cause and effect contain their previous k and l samples, respectively. Using conditional probability, transfer entropy incorporates the time dependency and thus predicts the next value of the dependent variable as shown in the definition (2.12).

$$T_{J\to I} = \sum p(i_{n+1}, i_n^{(k)}, j_n^{(l)}) \log \frac{\mathrm{p}\left(i_{n+1}\middle|i_n^{(k)}, j_n^{(l)}\right)}{\mathrm{p}\left(i_{n+1}\middle|i_n^{(k)}\right)} \tag{2.12}$$

Transfer entropy, contrary to mutual information, has an asymmetric property,

and one can detect the direction of information transfer, i.e., causality, by comparing the magnitudes of information transfer in each direction.

$$T_{J\to I} = (H_{I^{(k+1)}} - H_{I^{(k)}}) - (H_{I^{(k+1)},J^{(l)}} - H_{I^{(k)},j^{(l)}}) \tag{2.13}$$

Many diagnosis methodologies were mainly developed based on Wiener's causality: "*Process X could be termed as to cause process Y if the predictability of Y is improved by incorporating information about X"* [14]. While Granger causality, which formulates predictability as a variance of an autoregressive model, is a mathematical interpretation of Wiener's principle, transfer entropy interprets predictability through uncertainty from the viewpoint of information-theoretic understanding. In other words, the transfer entropy defines the causal relationship as a reduction of uncertainty. If the predictability of i_{n+1} with the additional knowledge of another process, $j_n^{(l)}$, is improved, the numerator in Eq. (2.12) would be greater than the denominator. Then, the transfer entropy, $T_{J\to I}$, becomes positive and one can deduce a causal relationship between the two processes according to its relative magnitude. As a new measure of causality, the net amount of transfer entropy was devised by Bauer et al. [15].

$$t_{I\Rightarrow J} = T_{J\to I} - T_{I\to I} \tag{2.14}$$

A positive value of $t_{I\Rightarrow J}$ means that process *I* serves as the cause of process *J* and vice versa. If the values of $t_{I\Rightarrow J}$ are close to zero, the process is likely to have no causal relationship.

2.4. Graphical Lasso

Least absolute shrinkage and selection operator (lasso) is known as a form of regression problem penalized by the l_1-norm of the coefficient of predictors. Based on a previous study by Frank and Friedman [16], lasso originated from the bridge regression (l_q-norm $(\mathrm{q} > 0)$) by Tibshirani [17] as a special case of the bridge with $\mathrm{q} = 1$. By means of regularization, lasso derives the coefficients of irrelevant predictors with respect to the response to be set to zero, which can increase the accuracy of regression models. Recently, with the improvement of computational power and advances in the algorithm, many variants of the lasso were suggested, including the implementation for graphical models.

Graphical lasso, which was first introduced by Friedman et al. [18], is a methodology to find a sparse inverse covariance of the design matrix of variables based on the Gaussian log-likelihood with lasso penalty. According to the fact that the zeros in the inverse covariance matrix of Gaussian distributed variables correspond to the conditional independence, a convex optimization problem can be formulated to obtain a sparse form of the maximum likelihood estimation [19]. Assuming that $y_i, i = 1, \dots, N$ are independently sampled from $N(\mu, \Sigma)$, the log-likelihood of the observations $L(\mu, \Sigma) = log \prod_i f(y_i)$ can be formulated as

$$\begin{aligned} L(\mu, \Sigma) &= -\frac{N}{2} log det\Sigma - \frac{1}{2}\sum_{i=1}^{N}(y_i - \mu)^T\Sigma^{-1}(y_i - \mu) \\ &= \frac{N}{2}\left(log det\Sigma^{-1} - tr(S\Sigma^{-1}) - (\mu - \bar{\mu})^T\Sigma^{-1}(\mu - \bar{\mu})\right) \end{aligned} \tag{2.15}$$

where $\bar{\mu}$ and S are the sample mean and covariance

$$\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i,\ S = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{\mu})(y_i - \bar{\mu})^T.$$

The maximum likelihood estimation of the log-likelihood can be represented as

$$\max_{\Sigma^{-1}} \; logdet\Sigma^{-1} - tr(S\Sigma^{-1}) - (\mu - \bar{\mu})^T\Sigma^{-1}(\mu - \bar{\mu}). \qquad (2.16)$$

As the optimal value for mean is the sample mean $\bar{\mu}$, the last term in Eq. (2.16) can be dropped. By changing the inverse covariance Σ^{-1} to precision matrix Θ and introducing lasso penalty term $\|\boldsymbol{\Theta}\|_1$ the graphical lasso problem can be completely formulated as

$$\max_{\boldsymbol{\Theta}} \log\det(\boldsymbol{\Theta}) - \mathrm{tr}(\mathbf{S}\boldsymbol{\Theta}) - \rho\|\boldsymbol{\Theta}\|_1 \qquad (2.17)$$

where tr denotes the trace operator and $\|\Theta\|_1$ is the l_1-norm (regularization) term with the regularization parameter ρ.

As a result, the graphical lasso problem in Eq. (2.17) is the Gaussian log-likelihood expression of the precision matrix including the l_1-norm regularization. The estimated inverse covariance matrix using graphical lasso corresponds to a subset of the original graph, where irrelevant relationships are eliminated since a zero in the resulting precision matrix corresponds to a missing edge in the original graph. Accordingly, it allows for increasing the sparsity of the original graph of the target process.

To solve the optimization problem of the graphical lasso in Eq. (2.17), Friedman [18] adopted a block coordinate descent approach to estimate $\boldsymbol{W}$ (estimate of $\boldsymbol{\Sigma}$) instead of $\boldsymbol{\Theta}$ (estimate of $\boldsymbol{\Sigma^{-1}}$).

$$\text{W= } W11w12w12Tw22\,, \qquad \text{S= } S11s12s12Ts22 \qquad (2.18)$$

By partitioning W and S, they exploited the fact that $\boldsymbol{w_{12}}$ satisfies Eq. (2.19).

$$w_{12} = \underset{y}{\mathrm{argmin}}\{y^T W_{11}^{-1} y : \; \|y - s_{12}\|_\infty \leq \rho\} \qquad (2.19)$$

Using convex duality, Eq. (2.19) can be written as the equivalent dual problem of

$$\min_{\beta}\{\frac{1}{2}\left\|W_{11}^{1/2}\boldsymbol{\beta} - b\right\|^2 + \rho\|\beta\|_1\} \qquad (2.20)$$

where $\mathrm{b} = \boldsymbol{W_{11}^{-1/2}s_{12}}$. Substituting the solution to Eq. (2.20), $\boldsymbol{\beta}$, into Eq.

(2.19) yields $\boldsymbol{w_{12}} = \boldsymbol{W_{11}\beta}$. The subgradient equation of Eq. (2.17) can be derived as Eq. (2.21) using the relation $\mathbf{W\Theta} = \mathbf{I}$ and the fact that the derivative of $\log\det\boldsymbol{\Theta}$ equals to $\boldsymbol{\Theta^{-1}}$,

$$\mathrm{W} - \mathrm{S} - \rho \cdot \Gamma = 0 \tag{2.21}$$

where $\boldsymbol{\Gamma}_{ij} \in sign(\boldsymbol{\Theta}_{ij})$; that is $\boldsymbol{\Gamma}_{ij} = sign(\boldsymbol{\Theta}_{ij})$ if $\boldsymbol{\Theta}_{ij} \neq 0$; otherwise, $\boldsymbol{\Gamma}_{ij} \in [-1,1]$. Similarly, the sub-gradient equation of the dual problem becomes

$$W_{11}\beta - s_{12} + \rho \cdot \nu = 0 \tag{2.22}$$

where $\nu \in sign(\beta)$.

In conclusion, the irrelevant relationships are eliminated from the original graph, and the subgraph that contains only strongly associated relationships is formed by a single implementation of the graphical lasso. In addition, the uniqueness of the graphical lasso can be guaranteed under the same condition because it consistently estimates the strongly associated relationships as suggested by Meinshausen [20] and Banerjee [21]. By iteratively applying graphical lasso to the remaining variables that are not included in the subgraph in the previous implementation, the entire graph can be divided into subgraphs, which reduces the computational complexity by excluding the redundant relationships from the costly calculation of the causality measures.

Chapter 3

Process Fault Detection Using Autoencoder with Data Augmentation via Variational Autoencoder[1]

3.1. Introduction

Multivariate statistical process monitoring (MSPM) is an indispensable part of the successful operation of chemical processes used to guarantee the safety and quality of the products. There are various MSPM methods, which can be classified into two approaches: prior knowledge-based methods, such as first principle equations or empirical equations, and historical data-driven methods [22]. Historical data-driven methods have the advantage of generality, and thus there is no need for process-specific domain knowledge. These have the advantage of general applicability owing to the fast and straightforward model construction in general. Data-driven process monitoring models make use of the data under normal operation in developing the monitoring statistics and defining a boundary of normal states that detect the process faults by checking whether the online monitoring statistics violate the boundary. Conventional multivariate statistical models using latent variables for process monitoring, such as PCA and partial least squares (PLS), have been widely used as dimensionality reduction methods. PCA, which defines orthogonal latent variables that maximize the variance of the original data, is used as a dimensionality reduction method for monitoring in a reduced dimensional feature space. PLS is an extension of PCA and incorporates quality variables under inspection. ICA

[1] This chapter is an adapted version of H. Lee., C. Kim., D. H. Jeong., and J. M. Lee, "Data-driven Process Fault Detection for Chemical Processes Using Autoencoder with Data Augmentation", *Korean Journal of Chemical Engineering*, Accepted.

utilizing higher-order statistics, unlike PCA, which only employs second-order statistics such as the mean and variance, performs better on data following a non-Gaussian distribution. However, it still has certain limitations with respect to the nonlinearity of the data owing to a linearity assumption. To deal with nonlinearity, KPCA has been suggested [2]. KPCA exploits the kernel trick to map the nonlinear data into a higher dimensional linear space such that it can perform feature extraction better than a directly applied PCA on nonlinear data. However, it has a limitation in that the computational complexity in the kernel method increases exponentially as the number of dimensions and samples increases. In addition, it is well known that the kernel method has limitations in that it exhibits an inconsistent performance that is significantly dependent on the kernel type and hyperparameters. Autoencoders (AEs), which are a type of neural network for unsupervised dimensional reduction, have recently been suggested as a notable alternative to overcome these limitations with the help of recent advances in machine learning techniques. Various studies have demonstrated that AEs show better performance compared to a conventional dimensionality reduction method in process monitoring [23].

Hinton [11] compared the performances of the AEs and PCA as a conventional dimensionality reduction technique for various types of data. Notable results also suggest that AEs achieve a better performance in reducing the dimensionality of the data than conventional methods such as PCA, ICA, and KPCA when sufficient computational resources, sufficient numbers of training data, and a plausible initialization of the weight parameters are secured [11]. Since it was first reported by Hinton, many studies on process monitoring using AEs have been actively conducted and have proved the competitiveness of AEs when combined with nonlinear activations in terms of the effectiveness of nonlinear feature extraction in process monitoring [24,25]. Advancing from the classical form of AEs, various AEs used to cope with noisy process data have

been proposed, such as denoising autoencoder (DAE) [8], contractive autoencoder (CAE) [9], and robust autoencoder [26]. DAE and CAE were used to demonstrate the improvement in monitoring performance over the basic AE and PCA for the Tennessee Eastman process, a benchmark chemical process selected as the target process in the present study [23]. The AEs were integrated with another averaging approximator such as k-nearest neighbor (kNN) to suggest a newly refined monitoring metric [27] or combined with a regularization method such as elastic net to enhance the robustness of the monitoring model under abundant training data [28]. Even if sufficient amounts of training data can be provided, the data-rich but information-poor problem still remains, resulting in a typical overfitting issue of the models [29]. For this reason, there have been diverse attempts to supplement information through data augmentation, which makes manifold learning robust for both overfitting and underfitting.

Data augmentation techniques can be classified into two approaches: conventional methods and generative models. Two representative methods for conventional data augmentation have been developed in the fields of image processing and computer vision applications: data warping [30] and the synthetic minority over-sampling technique (SMOTE) [31]. Data warping involves the synthesis of data by applying a deformation from intuitive features in the original data space, such as translation, rotation, and skewing from existing samples. Whereas SMOTE can be applied in both the data space and the feature space to produce artificial samples, it was primarily proposed to alleviate class imbalance problems during classification. Being implemented through an affine transformation in the feature space as well, SMOTE has the advantage of being applied independently of the applications owing to the fact that a feature space can represent the salient structure of the data. Wong [30] reported that both warping and SMOTE can improve the performance of a classification model.

The generative model that belongs to the neural network-based method can

be categorized into two groups: variational autoencoder (VAE) and generative adversarial network (GAN). Unlike a conventional method, generative models generally estimate the underlying distribution in the feature space. Based on the feature space, new vectors, which are latent codes, are sampled and then fed into a generator unit corresponding to each generative model to create artificial data. By leveraging the inherent manifold knowledge rather than directly manipulating the sample data in the original space, the generative modeling approach has proven its superiority in terms of the quality and effectiveness of augmentation in previous studies conducted in diverse fields [32,33]. This property becomes more significant as the number of dimensions increases because the Euclidean distance commonly used as the distance metric weakens the meaning as a similarity measure in the original space. To make use of the indispensable merits of stable convergence of VAE in modeling compared to those of GAN, which possesses an adversarial training process between two networks, VAE has been employed to augment the supplementary training data. Because most of the chemical process data might not violate the assumption of VAE in which the class for encoding and the prior distribution are restricted as multivariate Gaussians, it makes use of the VAE characteristic in which the latent vectors can be sampled from the explicit distribution in the feature space. This enables the manipulation of the latent vectors to reflect the intentions of the data augmentation, such as selective sampling within the boundary regions of a normal distribution, which correspond to rare samples. The capability of a selective production of artificial data to convey the intention for augmentation can contribute to the improvement of the process monitoring modeling by providing insufficient information.

Various studies have used generative models as tools for data augmentation, particularly in the field of computer vision. Several studies have improved the performance of image classifiers through data augmentation using generative

models, such as VAEs, GANs, and their variants [33,34,35]. The models for speech recognition [36] or translation [32] can be supported by data augmentation techniques to alleviate the class imbalance problem or allow the reuse in another domain, as applied in transfer learning. Although they applied a variant of GAN to construct a generative model, and not VAE, Gao et al. [37] suggested that augmentation in the case of process data can also contribute to improvements as a classifier for process monitoring.

This chapter was motivated by previous studies promoting the quality of manifold learning, which is an essential part of modeling for fault detection, through data augmentation. Integrated with the idea of an exclusive augmentation of data that rarely appears but should be classified as a normal state similar to the training dataset, the proposed method suggests a framework to boost the monitoring performance for fault detection by supplementing the insufficient information of the training dataset. Based on a specific strategy to reflect the intention of the augmentation, an edge-based oversampling scheme [38] is utilized with a general transformation to explicitly aim the boundary region of the normal state within the feature space [10]. The synthetic samples generated from the latent codes of the border of a normal region are augmented in the training dataset to promote manifold learning by imposing more weight.

The remainder of chapter 3 is organized as follows. In Section 2, the preliminaries of the proposed method are introduced. A description of the Tennessee Eastman process which is the target process used in the case study, the implementation of the proposed methodology, and its discussion are presented in Section 3.

3.2. Process Fault Detection Model Integrated with Data Augmentation

In this section, we propose a method that makes use of the advantage of data augmentation, particularly of the boundary of the normal operation data, such that it can help the classifier generalize better in terms of the manifold learning of the normal state. The method for edge-based sampling and data generation, proposed under the term DOPING technique[38], showed an improvement of image classification for the well-known MNIST dataset. Although the study was tested in different domains and utilized a different type of generative model and classifier from this study, it revealed the effectiveness of data augmentation based on the edge of a certain class. The borderline-SMOTE[31], which is a modified minority over-sampling method used only to generate samples near the borderline of the minority class, also presents evidence of further improvements with the help of borderline samples.

In this study, we propose an approach to supplementing rare samples in the same class to mitigate the in-class data imbalance, unlike previous studies that augment the minority class to resolve the between-class imbalance. From the viewpoint of process fault detection, the proposed method was designed to augment relatively large amounts of rare samples that occur with low probability distributed within the boundary region of a normal state. As a result, by deliberately adding rare normal instances to the training data, the monitoring system can better perform in terms of increasing the fault detection rate (FDR) while keeping the false alarm rate (FAR) below the acceptable level.

The workflow of the proposed method is summarized in Figure 3.1. To prepare the data at similar scales and variabilities, a preprocessing step is first required. The generative model using Info-VAE was trained on the original data to generate artificial data for augmentation. Once the generative model is prepared, various sets of data are generated by sampling in the latent space and

retrieving data of the original space through the decoder network. The generated data are merged with the original data as the augmented training data for modeling the fault-detection model using AE.

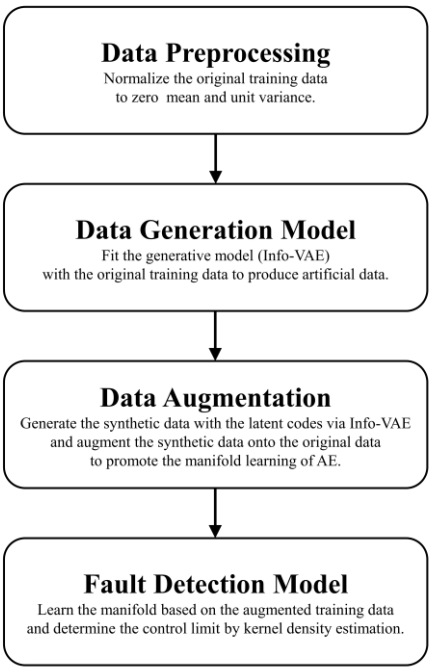

Figure 3.1. Flowchart of the proposed method for process fault detection model.

3.2.1. Info-Variational Autoencoder for Data Augmentation

InfoVAE [39] has recently been proposed to improve the problem of vanilla VAE ignoring the latent vectors, i.e., so-called uninformative latent codes because it has been shown that a decoding network with sufficient capacity can take over the role of reconstructing inputs even with meaningless random codes. In other words, the latent codes, which must potentially retain significant information needed to restore the data, are forced to fit the prior distribution by minimizing the second term in Eq. (2.5). This is a fatal limitation in that a latent vector cannot contain any data features, particularly when there is a significant manipulation to impose any intention in the latent space. Zhao et al. [39] introduced an additional regularization term in the objective that allows the encoded distribution in the latent space to preserve the data features. Organizing the ELBO objective of the original VAE as an equation,

$$\begin{aligned} & L_{ELBO}(\theta, \phi; x^{(i)}) \\ = \; & \mathbb{E}_{q_\phi(z|x)}\left[log p_\theta\left(x^{(i)}|z\right)\right] \\ - \; & D_{KL}(q_\phi(z|x^{(i)})||p(z)) \\ = \; & - D_{KL}(q_\phi(z)||p_\theta(z)) \\ & - \mathbb{E}_{q(z)}\left[D_{KL}(q_\phi\left(x^{(i)}|z\right)||p_\theta\left(x^{(i)}|z\right)\right] \end{aligned} \tag{3.1}$$

The objective function of InfoVAE, including a mutual information maximization term that leads to meaningful features, is defined as follows:

$$\begin{aligned} & L_{InfoVAE}(\theta, \phi; x^{(i)}) \\ \equiv \; & \mathbb{E}_{q_\phi(z|x)}\left[log p_\theta\left(x^{(i)}|z\right)\right] \\ - \; & D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)) + \alpha I_q \\ = \; & \mathbb{E}_{p_\mathcal{D}(x)}\mathbb{E}_{q_\phi(z|x)}[log p_\theta(x|z)] - (1 \\ & - \alpha)\mathbb{E}_{p_\mathcal{D}(x)} D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)) \end{aligned} \tag{3.2}$$

$$-(\alpha - 1)D_{KL}(q_\phi(z)||p(z)),$$

where the scaling parameter λ in the original study [39] was assumed to be one for simplicity. According to the proof and derivation in InfoVAE, the final form of the objective function can be computed by replacing the last term in Eq. (3.2) with an equivalent divergence family, i.e., the maximum mean discrepancy (MMD). Hence, an operation in the latent space can convey implications to be reflected on the generated data. It provides conditions under which operations in the learned latent space may have meaningful implications for the generated data, which cannot be utilized by an uninformative latent code. In other words, it is noteworthy that the characteristic of Info-VAE that promotes the extraction of the informative features enables the exclusive sampling and generation of the boundary samples which encourage high-fidelity manifold learning.

3.2.2. Autoencoder for Process Monitoring

The construction of statistics for process monitoring using AE is carried out using the same procedure as that used in PCA. After completing the network training, test statistics defined in the two spaces are used to monitor the abnormality of a process. One is H^2, which is the squared scalar value of the latent vector corresponding to T^2 in PCA calculated in the feature space by utilizing the output in the feature space calculated from Eq. (2.1) as follows:

$$H^2 = h^T \cdot h = F^T(x) \cdot F(x) \tag{3.3}$$

The other is the squared prediction error (SPE) defined in the residual space [23] based on the reconstruction error as follow:

$$SPE = e^T \cdot e = (x - G(F(x)))^T \cdot (x - G(F(x))) \tag{3.4}$$

Subsequently, the control limit used to characterize the normal operating region is defined by a non-parametric density estimator called kernel density estimation (KDE). Based on the KDE results for the normal operation training samples, the 95 percentile values for each space are typically determined criteria for process monitoring. After finishing the offline training procedure based on a set of training samples, the test samples are mapped into the low-dimensional manifold and reconstructed into the original dimensional space online. Process monitoring is conducted by comparing the statistics of the test data to the control limit in each space.

3.3. Case study and Discussion

In this section, we demonstrate the proposed method on an industrial-scale benchmark process, Tennessee Eastman process to verify its effectiveness. The benchmark process to be used in a case study is first introduced. A detailed description of the fault detection modeling with the proposed method can be found in the subsequent section and the result of the implementation is also thoroughly discussed.

3.3.1. Tennessee Eastman Process

TEP is a widely used benchmark chemical process for performance comparison in process monitoring algorithms or control structures. It consists of five modules: a reactor, condenser, product separator, stripper, and compressor for the recycle stream, as shown in Figure 3.2. The irreversible and exothermic gas-phase catalytic reactions of reactants A, C, D, and E occur to produce two liquid products, G and H. The following steps, including condensation, separation, and compression, recycle the unconverted reactants in the product streams and make up the fresh reactants rectified through the stripper. Some reactions involve inert gas B and byproduct F, which are primarily removed by the purge stream. A detailed description of the TEP can be found in the original suggestion of the Fortran [40] and revised MATLAB [41] models. The model used in this study contains the control strategy proposed by Ricker [42] based on the revised MATLAB model.

There are 41 measurements, i.e., 22 continuous variables and 19 composition variables from the installed analyzers. The model also includes 12 manipulated variables used in the process control. In this study, 50 variables, excluding three manipulated variables remaining as fixed values (compressor recycle valve, stripper steam valve, and agitator speed), were investigated. The target variables for the analysis are listed in Table 3.1. The MATLAB model modified based on the original Fortran model includes 28 pre-defined fault cases, and 8 (21–28) more fault cases were added to the 20 fault cases in the original model [40]. A total of 28 faults in the TEP cover various types, such as step change, random variation, slow drift, and sticking of a certain variable. The fault scenarios in the TEP are summarized in Table 3.2. The proposed methodology for process fault detection, which is described in detail in the following sections, is validated and analyzed using the TEP in the following sections.

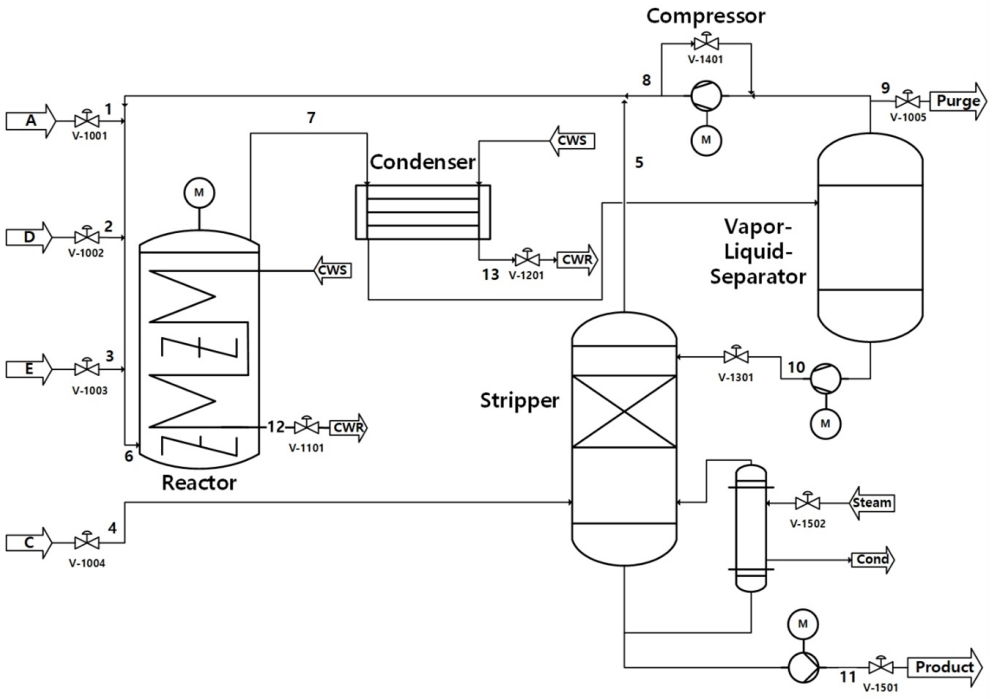

Figure 3.2. Process flow diagram of Tennessee Eastman process [43].

Table 3.1. Process variables of TEP subject to process monitoring.

Variable No.	Variable Name	Variable No.	Variable Name
1	A feed flowrate (stream 1)	**18**	Stripper temperature
2	D feed flowrate (stream 2)	**19**	Stripper steam flowrate
3	E feed flowrate (stream 3)	**20**	Compressor work
4	A & C feed flowrate (stream 4)	**21**	Reactor c/w outlet temperature
5	Recycle flowrate (stream 8)	**22**	Condenser c/w outlet temperature
6	Reactor feed rate (stream 6)	**23–28**	Reactor feed analysis (A–F mol%) (stream 6)
7	Reactor pressure	**29–36**	Purge gas analysis (A–H mol%) (stream 9)
8	Reactor level	**37–41**	Product analysis (D–H mol%) (stream 11)
9	Reactor temperature	**42**	D feed flow valve (stream 2)
10	Purge rate (stream 9)	**43**	E feed flow valve (stream 3)
11	Product separator temperature	**44**	A feed flow valve (stream 1)
12	Product separator level	**45**	A & C feed flow valve (stream 4)
13	Product separator pressure	**46**	Purge valve (stream 9)
14	Product separator under flowrate (stream 10)	**47**	Separator pot liquid flow valve (stream 10)
15	Stripper level	**48**	Stripper liquid product flow valve (stream 11)
16	Stripper pressure	**49**	Reactor c/w flow valve
17	Stripper under flowrate (stream 11)	**50**	Condenser c/w flow valve

Table 3.2. Process faults in TEP.

No.	Description	Type
IDV(1)	A/C feed ratio, B composition constant (stream 4)	Step
IDV(2)	B composition, A/C ratio constant (stream 4)	Step
IDV(3)	D feed temperature (stream 2)	Step
IDV(4)	Reactor cooling water inlet temperature	Step
IDV(5)	Condenser cooling water inlet temperature	Step
IDV(6)	A feed loss (stream 1)	Step
IDV(7)	C header pressure loss–reduced availability (stream 4)	Step
IDV(8)	A, B, C feed composition (stream 4)	Random variation
IDV(9)	D feed temperature (stream 2)	Random variation
IDV(10)	C feed temperature (stream 4)	Random variation
IDV(11)	Reactor cooling water inlet temperature	Random variation
IDV(12)	Condenser cooling water inlet temperature	Random variation
IDV(13)	Reaction kinetics	Slow drift
IDV(14)	Reactor cooling water valve	Sticking
IDV(15)	Condenser cooling water valve	Sticking
IDV(16)	*Unknown (Deviation of heat transfer within stripper heat exchanger)	*Unknown (Random variation)
IDV(17)	*Unknown (Deviation of heat transfer within reactor)	*Unknown (Random variation)
IDV(18)	*Unknown (Deviation of heat transfer within condenser)	*Unknown (Random variation)
IDV(19)	*Unknown (recycle valve, stripper steam valve, underflow separator (stream 10), underflow stripper (stream 11))	*Unknown (Sticking)
IDV(20)	*Unknown	*Unknown
IDV(21)	A feed temperature (stream 1)	Random variation
IDV(22)	E feed temperature (stream 3)	Random variation
IDV(23)	A feed pressure (stream 1)	Random variation
IDV(24)	D feed pressure (stream 2)	Random variation
IDV(25)	E feed pressure (stream 3)	Random variation
IDV(26)	A & C feed pressure (stream 4)	Random variation
IDV(27)	Reactor cooling water pressure	Random variation
IDV(28)	Condenser cooling water pressure	Random variation

* Unknown: Uncovered by A. Bathelt in the revised version of MATLAB model.

3.3.2. Implementation of the Proposed Methodology

As suggested in Figure 3.1 in section 3.2, the data scaling process comes first as the data preprocessing before employing the modeling of the generative and monitoring models. The data scaling process, which imposes equal importance against all variables in the model, works as a critical role, as in other machine learning algorithms. This is also important in terms of the stable convergence of the model, which is valid for all methods employed in this study. The standardization to scale each variable is as follows:

$$X_{scaled} = \frac{X-\mu}{\sigma}, \tag{3.5}$$

where X is the original variable, and μ and σ are the mean and standard deviation of each variable based on the training data, respectively. To establish the stopping criteria of the training process, the original dataset was divided into training and validation sets, each having 6,000 and 1,200 samples out of a total of 7,200 samples.

The structure of the Info-VAE used in this study is summarized in Table 3.3. First, the distributions of the inferential posterior, $q_\phi(z|x^{(i)})$, and the generative posterior, $p_\theta\left(x^{(i)}\middle|z\right)$, are assumed to be multivariate Gaussians because process variables with continuous values are considered. The input layer dimension is matched with the dimensions of the benchmark process system, TEP, which has 50 variables, as suggested in Table 3.1. As one of the most critical parameters in the application of the autoencoders, the reduced dimensions of the latent space should be determined. There exist several heuristics to determine the dimensionality of the latent space such as the elbow of the scree plot, the cutoff of the eigenvalues greater than 1, or the cumulative percentage of explained variance (CPV). As shown in Figure 3.3, the first criteria, the elbow of the scree plot, is insufficient to account for the variance of the TEP data in

the latent space because it results in excessive loss of information during dimension reduction. Therefore, a case study was performed to decide the dimension of the latent space that is the most efficient in terms of the monitoring performance. Based on the settings of the base case which would be introduced later in this chapter, the case study was conducted by varying the dimensionality of the latent space from 60% to 90% CPV. According to each case, the averages of the FDR over the 28 fault cases in TEP were derived and compared as Figure 3.4. According to the result of the case study, it can be concluded that the 80% CPV is enough to efficiently reduce the dimension of the latent space based on the monitoring performance. Thus, the number of nodes in the bottleneck layer was set to 30. The result is reflected in the output dimension of the second layer in the encoder. In this study, the number of hidden layers between the input and output, which is another essential hyperparameter determining the performance of the model, was determined to be one in both the encoding and decoding networks to prevent overfitting. The nonlinear activation functions for the hidden layers are set to the leaky rectified linear unit (ReLU), except for the output levels, such as the feature and reconstruction layers. In general, no activation function is adopted for the output layers in the regression models, which corresponds to the linear activations to fit the means of a multivariate Gaussian for the feature code and reconstruction of the input. In the case of layers used to fit the standard deviations of the feature codes or the reconstructions, another type of activation function, i.e., a softplus activation, is employed to explicitly impose a positive definite constraint for the standard deviations. The relevant hyperparameters for configuring the generative model using Info-VAE are listed in Table 3.4. The weight parameter adjusting the relative importance between the reconstruction loss term and the replaced divergence term from the KL divergence, i.e., the MMD, should be selected to balance the relative scale of each element.

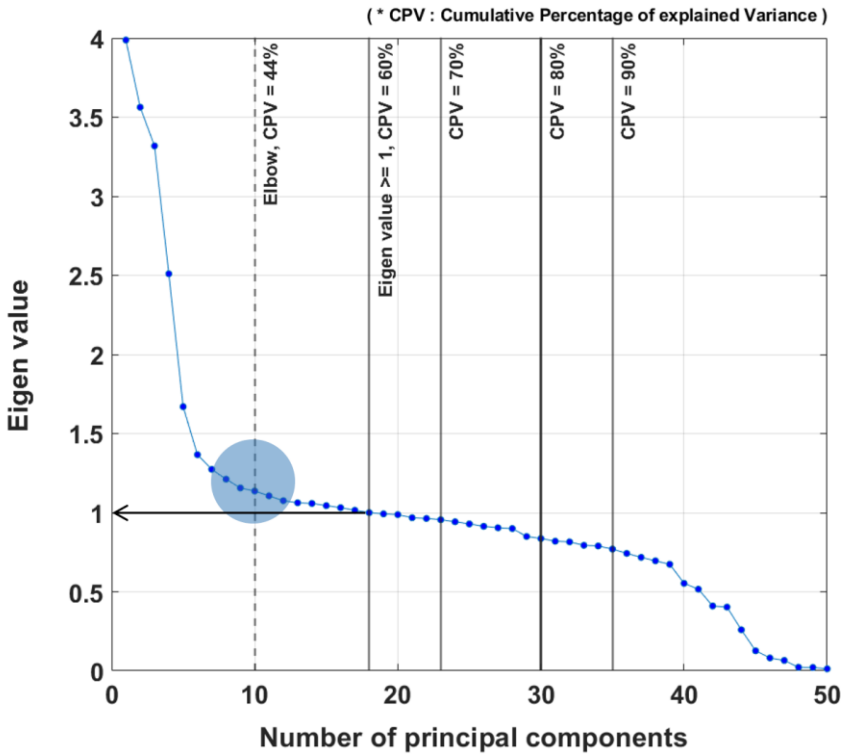

Figure 3.3. Scree plot for selection of the reduced dimension of the latent space.

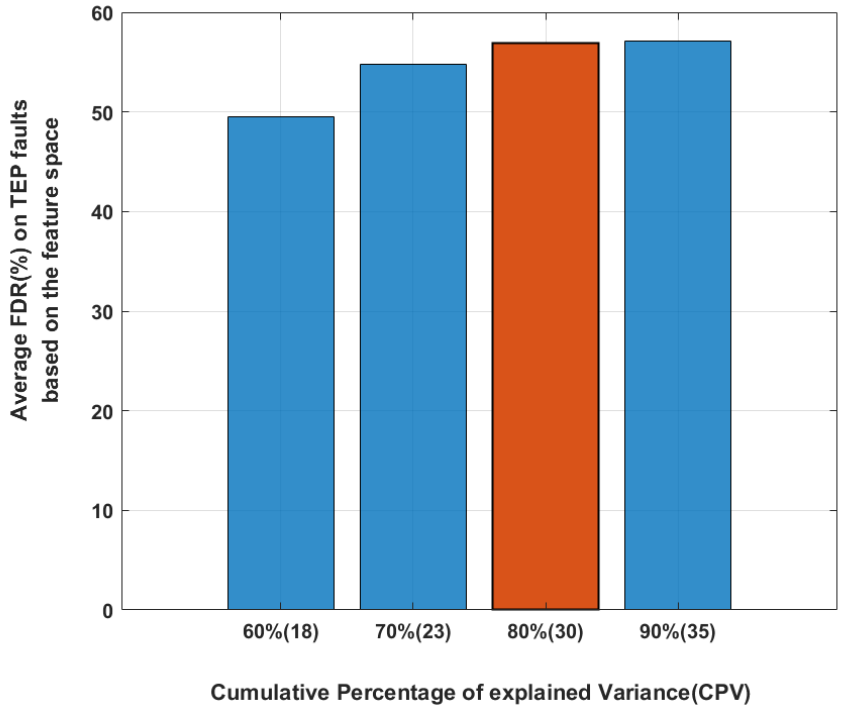

Figure 3.4. Case study to decide the reduced dimension of the latent space. (The dimensionalities of the latent space in each case are represented in the parenthesis.)

Table 3.3. Structure of the generative model using Info-VAE.

Layer	Dimension	Activation	Remarks
Input	50	-	
Encoder 1	40	Leaky ReLU	Alpha: 0.2
Encoder 2 for Mean	30	Linear	
Encoder 2 for STD	30	Softplus	
Feature	30	-	
Decoder 1	40	Leaky ReLU	Alpha: 0.2
Decoder 2 for Mean	50	Linear	
Decoder 2 for STD	50	Softplus	
Output	50	-	

Table 3.4. Hyperparameters of Info-VAE.

Variable Name	Value
MMD weight	50
Mini batch size	256
Optimizer	RMSProp
Learning rate	0.001

After finishing the training process, the Info-VAE model was used to augment the original training dataset with artificial samples. The model generates samples that can help the manifold learning of AE by emphasizing the boundary of the training data based on the latent distribution. To selectively specify the boundary of the normal samples distributed by a multivariate Gaussian, a ring-shape transformation is applied that can be easily extended to a shell of a sphere or a hypersphere within a higher space. First, random samples are extracted from the prior distribution, which has a multivariate normal distribution in a typical VAE having the same dimensionality as the feature space. A specific mapping of the samples from the Gaussian to ring-shape distribution is then applied to rearrange the sample codes based on the latent space, such that the sample codes suggest the meaning of the boundary region based on the original dimensional space. The mapping to the boundary is defined as follows:

$$R(z) = \frac{z}{a} + b \cdot \frac{z}{\|z\|} \ , \tag{3.6}$$

where a and b are responsible for the scatteredness and radius of the resulting ring, respectively.

The results of the case study for various sets of parameters a and b based on two-dimensional Gaussian data are shown in Figure 3.5. By using this mapping from the randomly sampled points from a normal distribution, as shown in Figure 3.5 (a), we can exclusively select the input codes representing boundaries in the latent space, which is based on the notion that the latent space contains the inherent features of the original data. By adjusting parameters a and b in Eq. (3.6) to point to the objective region corresponding to the boundaries of the normal state based on the two-dimensional feature space, the desired area in the feature space can be specified as shown in Figure 3.5 (b), (c), and (d) depending on the purpose. Although the case study is only demonstrated for two-dimensional data, it can also be expanded into higher-dimensional data without a loss of generality.

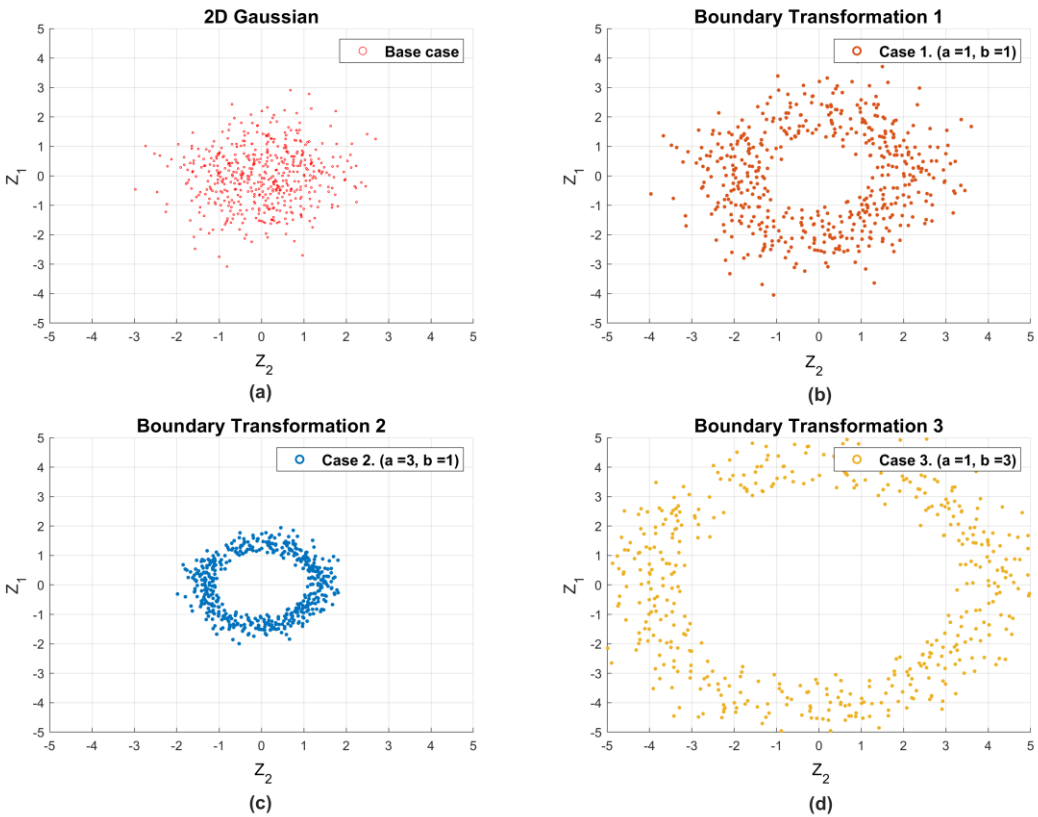

Figure 3.5. Case study for various parameters of boundary transformation in 2D Gaussian data.

The generated samples can be classified as distinct groups representing different regions of the normal samples to adjust the number of the different augmentation groups by manipulating the scatteredness and radius through a and b, respectively. To flexibly control the number of augmented datasets with different characteristics, a strategy that divides boundaries into several specific groups and then merges a different number of samples for each group for augmenting into the original dataset was used in this study.

The detailed methodology for augmenting the synthetic data is explained based on a case study of TEP. Although the dimension of the feature space is beyond the visualizable limit, the main idea of the proposed method for data augmentation can be conceptually explained in a two-dimensional space. The candidate groups for augmentation were divided into five groups, as shown in Figure 3.6. The groups were chosen to be able to thoroughly cover the areas that were originally described by the prior distribution while not overlapping each other. Each group can be distinguished based on its distance from the mean.

The groups of infrequent samples that exist far from the mean have a higher weight among the augmented data to supplement the deficient information in the original data. The sample codes near the center, such as G1, G2, and G3, as well as the outer groups such as G4 and G5 representing the boundary, are also included in the samples to generate artificial data for augmentation to avoid a data imbalance problem owing to an excessive supplementation of the boundary data indiscriminately. Instead, relatively high weights are assigned to the outer groups to emphasize the meaning of the augmentation of the boundary samples that correspond to rare normal samples. The parameters of the boundary transformation in each group are presented in Table 3.5.

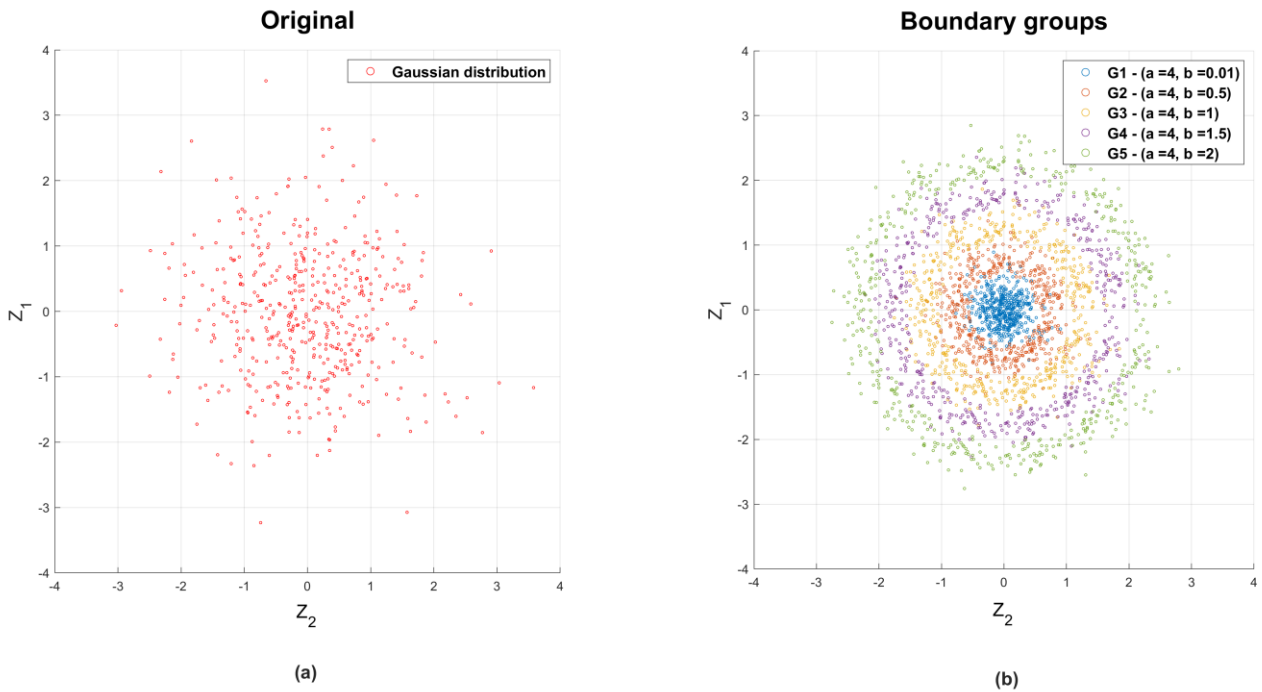

Figure 3.6. (a) Sampling from 2D Gaussian distribution
(b) Candidate groups of samples transformed by the boundary mapping.

Table 3.5. Parameters of boundary groups for the TEP case study.

Boundary groups	Parameters	
	a	**b**
G1	4	0.01
G2	4	0.5
G3	4	1.0
G4	4	1.5
G5	4	2.0

The total amount of the augmented training dataset was determined by a case study to analyze the sensitivity against the relative amount of the augmented samples. The augmentation of the training dataset for four relative sizes compared to the original training dataset was tested as suggested in

Table 3.6. To examine the effectiveness of the augmentation of the boundary samples, the average FDR on the 28 fault cases of the TEP based on the latent space was compared. As shown in Figure 3.7, the improvement of the monitoring performance by the data augmentation can be maximized in the case of the half size of the original training dataset. Thus, the total amount of augmented samples was designed to be half of the original training data. The relative amounts of each group in the augmented samples were set to be linearly proportional from the center to the outside. Hyperparameters such as the relative scale of the augmentation compared to the original data, the importance among the various groups, and the number of different groups suggested in Table 3.6 are adjustable depending on the applications.

After the augmented training dataset for the construction of the fault detection system using AE is prepared, the training of the normal state for the process fault detection system is performed by defining the normal manifold to be used as a monitoring model. The 6,000 samples for the training data out of the total 7,200 samples of the original data from the TEP simulation model were set apart from the validation data after a random shuffling process, which is in accord with the structure of the AE assuming each sample as being independent. The validation data were used to determine the termination point of the training to prevent overfitting using the early stopping criteria. Both the training and validation data in the original dataset were the same as those used in the modeling

of Info-VAE, thus standardization was applied as a scaling process. Because the synthetic data obtained from the generative model are scaled, the data for augmentation are attached to the original training and validation dataset resulting from the generation by Info-VAE. Finally, the detailed configuration of the training and validation datasets after the augmentation of the synthetic data are summarized in Table 3.6. The relative size of the total training dataset was set to five times that of the validation data, which was determined through a case study for various amounts of augmentation.

		Base case		Case 1 (25%)		Case 2 (50%)		Case 3 (75%)		Case 4 (100%)	
		Training dataset	Validation dataset	Training dataset	Validation dataset	Training dataset	Validation dataset	Training dataset	Validation dataset	Training dataset	Validation dataset
Original dataset		6,000	1,200	6,000	1,200	6,000	1,200	6,000	1,200	6,000	1,200
Augmented datasets	**G1**	-	-	100	20	200	40	300	60	400	80
	G2	-	-	200	40	400	80	600	120	800	160
	G3	-	-	300	60	600	120	900	180	1,200	240
	G4	-	-	400	80	800	160	1,200	240	1,600	320
	G5	-	-	500	100	1,000	200	1,500	300	2,000	400
Total		6,000	1,200	7,500	1,500	9,000	1,800	10,500	2,100	12,000	2,400

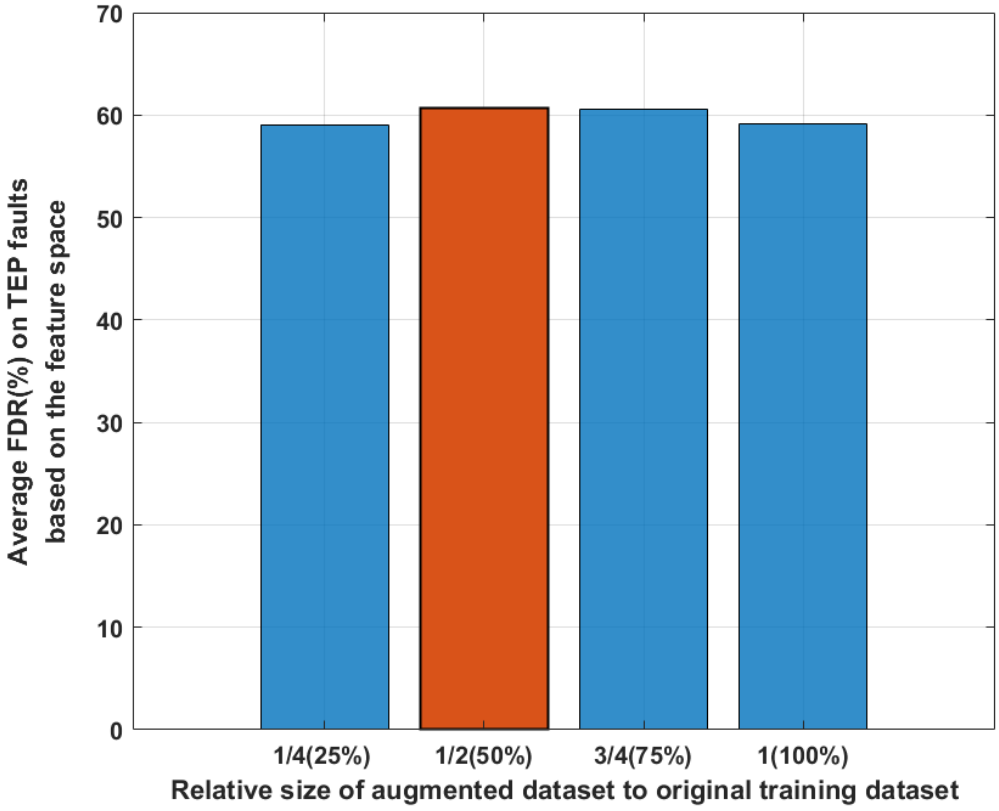

Figure 3.7. Case study to decide the relative amount of the augmentation to the original training dataset.
(Ratio of augmented data to the size of the original training dataset.)

Table 3.6. Configurations for the sensitivity analysis of the relative size of the augmented datasets to the original dataset.

		Base case		**Case 1 (25%)**		**Case 2 (50%)**		**Case 3 (75%)**		**Case 4 (100%)**	
		Training dataset	Validation dataset	Training dataset	Validation dataset	Training dataset	Validation dataset	Training dataset	Validation dataset	Training dataset	Validation dataset
Original dataset		6,000	1,200	6,000	1,200	6,000	1,200	6,000	1,200	6,000	1,200
Augmented datasets	**G1**	-	-	100	20	200	40	300	60	400	80
	G2	-	-	200	40	400	80	600	120	800	160
	G3	-	-	300	60	600	120	900	180	1,200	240
	G4	-	-	400	80	800	160	1,200	240	1,600	320
	G5	-	-	500	100	1,000	200	1,500	300	2,000	400
Total		6,000	1,200	7,500	1,500	9,000	1,800	10,500	2,100	12,000	2,400

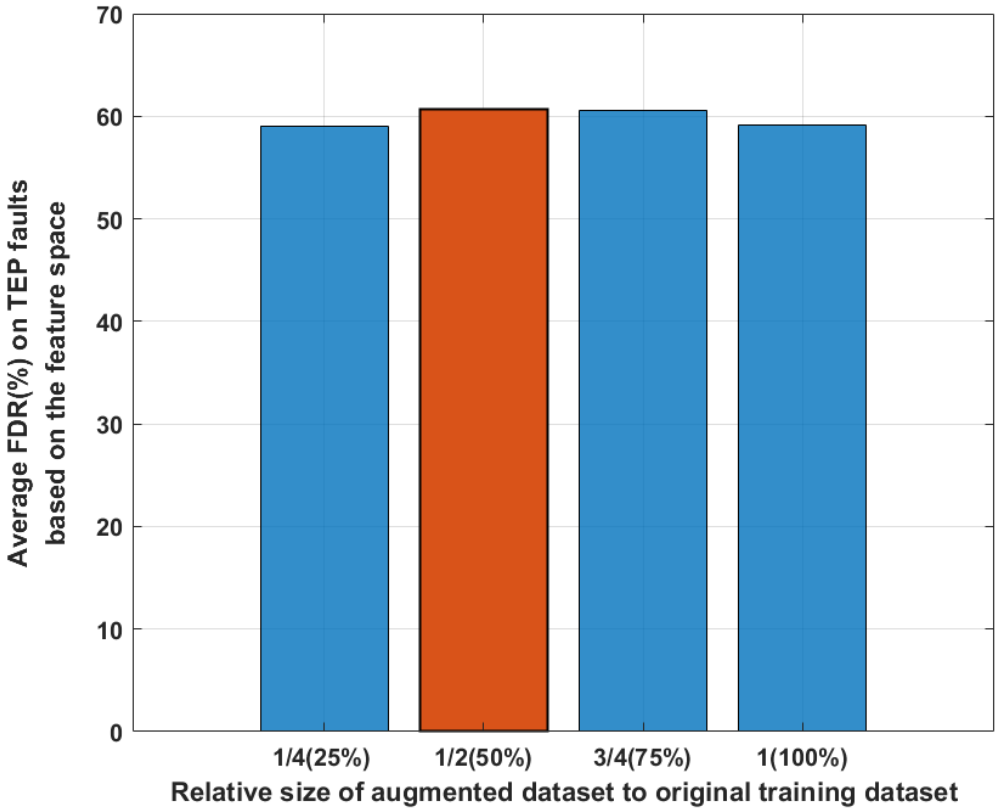

Figure 3.7. Case study to decide the relative amount of the augmentation to the original training dataset.
(Ratio of augmented data to the size of the original training dataset.)

As the dimensions of the inner part, the number of nodes of the feature layer was set to 30, which is the same as in the case of generative modeling using Info-VAE. However, the intermediate structure of the monitoring system using AE can be set differently from that of the generative model, where the capacity of the model is limited owing to the lack of the original training data. To make full use of AE for the monitoring system, the number of hidden layers and the size of each layer can be adjusted according to the application. A case study to tune the hyperparameters, such as the number of hidden layers and nodes of the AE monitoring system, determined the final structure, as shown in

Table **3.7**. All layers employed a fully connected layer, and the weights in all cases were initialized using a truncated normal distribution. The nonlinear activation functions of the AE monitoring model used to cope with the nonlinearity of the chemical process data were set to a rectified linear unit (ReLU) with the same hyperparameters. The nonlinear activations for the output layers for the encoder and decoder of the AE monitoring model, which correspond to the feature and reconstruction layers, respectively, were not applied to leave them as linear units following the convention used in regression problems. Kernel regularizations were adopted in the first layers of the encoding and decoding networks to control the weight parameters from being excessively large by penalizing them.

Table 3.7. Structure of the monitoring system using AE.

Layer	Dimension	Activation	Remarks
Input	50	-	
Encoder 1	46	ReLU	Alpha: 0.2; Kernel_regularizer: L2(0.2)
Encoder 2	42	ReLU	Alpha: 0.2
Encoder 3	38	ReLU	Alpha: 0.2
Encoder 4	34	ReLU	Alpha: 0.2
Feature	30	Linear	
Decoder 1	34	ReLU	Alpha: 0.2; Kernel_regularizer: L2(0.2)
Decoder 2	38	ReLU	Alpha: 0.2
Decoder 3	42	ReLU	Alpha: 0.2
Decoder 4	46	ReLU	Alpha: 0.2
Output	50	Linear	

The additional hyperparameters used to set up the training conditions are listed in

Table 3.8. The loss to be minimized during the training process is set by the mean squared error (MSE) between the input and its reconstruction at the end of the network. Adam with a default learning rate of 0.001 was applied as the optimizer. For the reproducibility of the monitoring system under the same conditions, early stopping criteria were introduced during the training process. The early stopping criteria are a methodology suggesting the termination of the training process if no improvements more than the minimum changes are made, that is, min_delta in

Table 3.8, during a predefined patience epoch by monitoring the validation loss. To compare the proposed method under the same conditions as the base case, which establishes the monitoring system using only the original training data, the same specifications for the training process are applied to the proposed case, as shown in

Table 3.8.

Table 3.8. Hyperparameters for the training of AE.

Variable Name		Value
Mini batch size		256
Loss		MSE
Optimizer		Adam
Learning rate		0.001
Early stopping	min_delta	$5*10^{-4}$
	patience	320
	mode	min

The configurations of the KDE, which are used to determine the control limit for the monitoring system, are presented in

Table 3.9. Although the original process data follow a Gaussian distribution, the hidden representations and reconstructions used to obtain the monitoring statistics might not follow the same distribution after passing through AE. Hence, KDE is utilized as the general approach to estimate the probability density function of the monitoring statistics, which is the basis of the decision of the control limit in each space. The Gaussian kernel, the most common type of kernel, was used to estimate the densities of each monitoring statistic. The bandwidths, which are the most significant parameters of KDE influencing the results of the estimation, were selected based on 20-fold cross-validation to cover all data samples in determining the hyperparameter. Since the control limits are determined based on the models of the base case and the proposed case respectively, they have different values in each case, as shown in

Table 3.9.

Table 3.9. Settings for KDE and the control limits for each case.

		Base case	**Proposed case**
Kernel type		Gaussian	Gaussian
Bandwidth (20-fold cross-validation)	H^2	6.158	2.335
	SPE	1.438	1.128
Control Limits	H_α^2	57.85	63.75
	SPE_α	31.40	32.25

To monitor the process fault, two monitoring statistics are defined in the feature space and the residual space, similar to that of PCA [44]. Instead of T^2 in the case of PCA, H^2 can be analogously defined based on the hidden representations in the feature space as follows:

$$\begin{aligned} H^2 &= h^T \cdot h, \\ h &= f_{En^M}(f_{En^{M-1}} \cdots (f_{En^1}(x))), \end{aligned} \tag{3.7}$$

where f_{En^i} represents the i^{th} hidden layer in the encoder network, and M is the number of intermediate layers between the input and feature layers. Similar to the other statistics in PCA, the SPE can be calculated from the reconstruction error between the input and its reconstruction as

$$\begin{aligned} SPE &= e^T \cdot e, \\ e &= x - g_{De^M}(g_{De^{M-1}} \cdots (g_{De^1}(h))), \end{aligned} \tag{3.8}$$

where g_{De^i} denotes the i^{th} hidden layer in the decoder network. With the proposed method, the two statistics are observed in real time against the process data for fault detection.

Once the training process is completed, the original training data under normal operating conditions are fed into the network. Based on the two statistics, H^2 and SPE, calculated based on the original training data, KDE was applied to predefine the control limits for each monitoring chart [45]. The typical choice for a significance level of $\alpha = 0.05$ is adopted such that the confidence limits in detecting the faulty conditions when the monitoring statistics of the new samples exceed the limits are set to 95%.

3.3.3. Discussion of the Results

In this section, the monitoring system based on the proposed method is tested on the TEP fault cases, and the monitoring results are analyzed. To demonstrate the advantage of data augmentation in building a fault detection system, the performance of the proposed method is compared to that of the base case, which only utilizes the original training data in constructing a monitoring system. The simulation was run for a total of 7200 samples with a sampling frequency of 0.01 hr/sample in the Simulink model, which corresponds to 72 hr of plant operation. The simulation data of the faulty condition have the same size as the training data under normal operations, although the process faults are introduced at 1000 simulation times for all cases of faulty conditions.

To compare the performance of the monitoring systems quantitatively, two performance metrics were set up: FDR and FAR [46]. These two metrics were defined based on the results of the binary classification test. The monitoring results of the data points can be classified into four groups, as shown in Figure 3.8 [47]. FDR and FAR can be calculated based on the number of instances in each group as follows:

$$FDR = \frac{TP}{TP + FN}, \qquad FAR = \frac{FP}{TN + FP} \tag{3.9}$$

FDR is the ratio of the samples exceeding the control limit to the entire sample time since the fault has been introduced. Conversely, FAR is the number of samples falsely going beyond the control limit per total number of normal operation samples. It needs to maximize FDR on the abnormal data while keeping FAR for the normal data as low as possible, which is generally determined as 5%. These two metrics should be compared simultaneously because a monitoring system with a high FDR and high FAR under a normal state is undesirable.

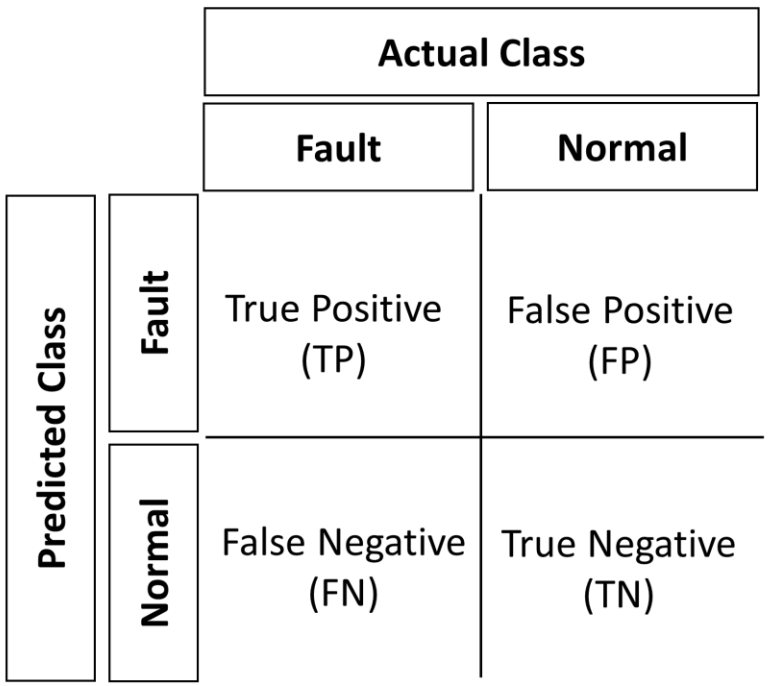

Figure 3.8. Binary classification criteria based on the monitoring results of data points.

For the first case, the monitoring chart of fault 1 of the TEP is shown in Figure 3.9. The blue and orange lines represent the monitoring statistics of the test samples in the feature space and residual space, H^2 and SPE, respectively. The red horizontal lines in each monitoring chart are the respective control limits, H_α^2 and SPE_α, as determined by KDE in

Table 3.9. For fault 1 in the TEP, both statistics can detect the process fault immediately after the occurrence of the process anomaly, similar to other methodologies used in previous research [23]. In the investigation during the first 1000 sample times before the fault was introduced, it was confirmed that more than 95% of the samples were classified as being in a normal state, distributed within the control limits. Considering the scenario of fault 1, which incurs a step deviation of the feed ratio of streams A and C, it is obvious that the majority of the process variables deviate from their nominal values during normal operation. These results verify that the monitoring statistics in both spaces can properly define a normal manifold and differentiate the faulty process condition from it.

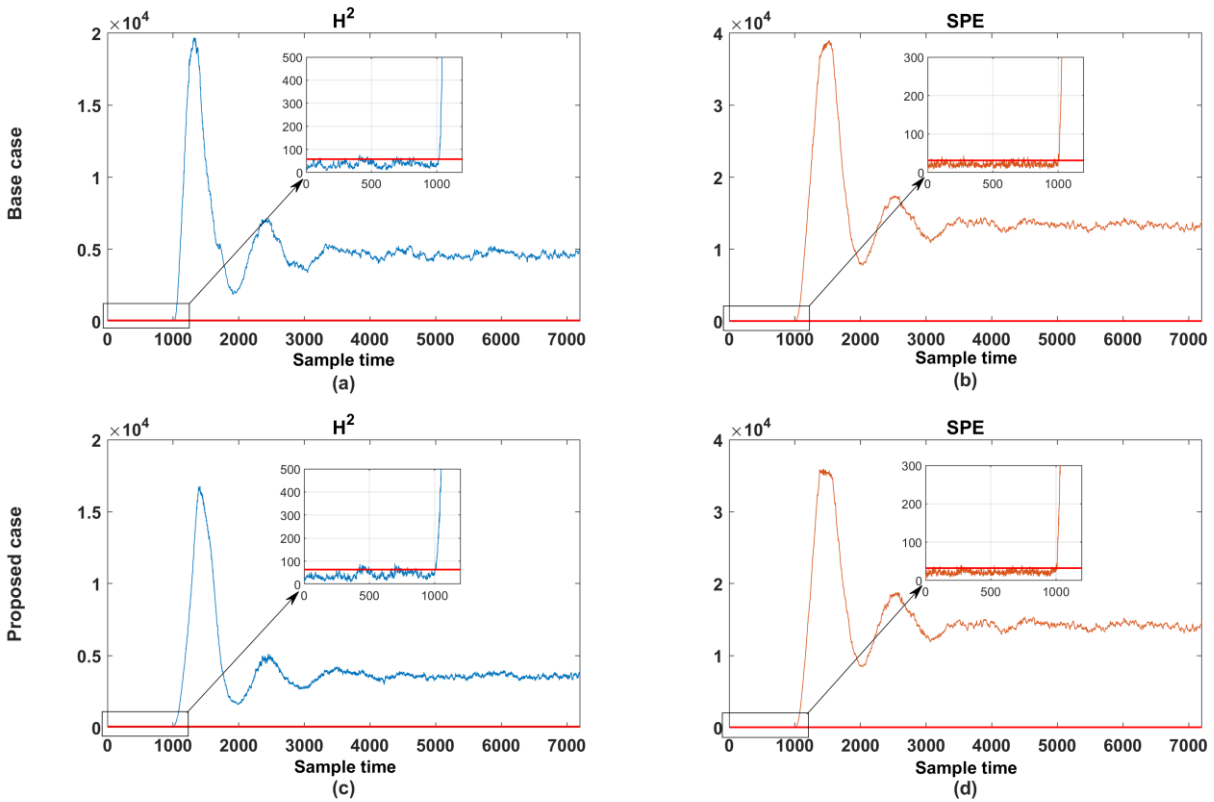

Figure 3.9. Monitoring charts of fault 1 for the base case ((a) and (b)) and the proposed case((c) and (d)).

For fault 11, which is the random variation of the reactor cooling water inlet temperature, the monitoring performance of the proposed method does not show a significant improvement in terms of the FDR or FAR compared to the base case. However, based on the results of the monitoring charts in the feature spaces shown in Figure 3.10 (a) and (c), the proposed method showed more pronounced isolation with a larger magnitude in the monitoring statistics for the faulty samples compared to the normal operation samples. The false-negative rate, similar to the type II error in the statistical analysis, was reduced from 10.9% to 5.5%. Therefore, data augmentation can improve monitoring systems. The improvement in the feature space is also noteworthy because it is in the feature space where data augmentation is designed to emphasize the boundary region of the normal space.

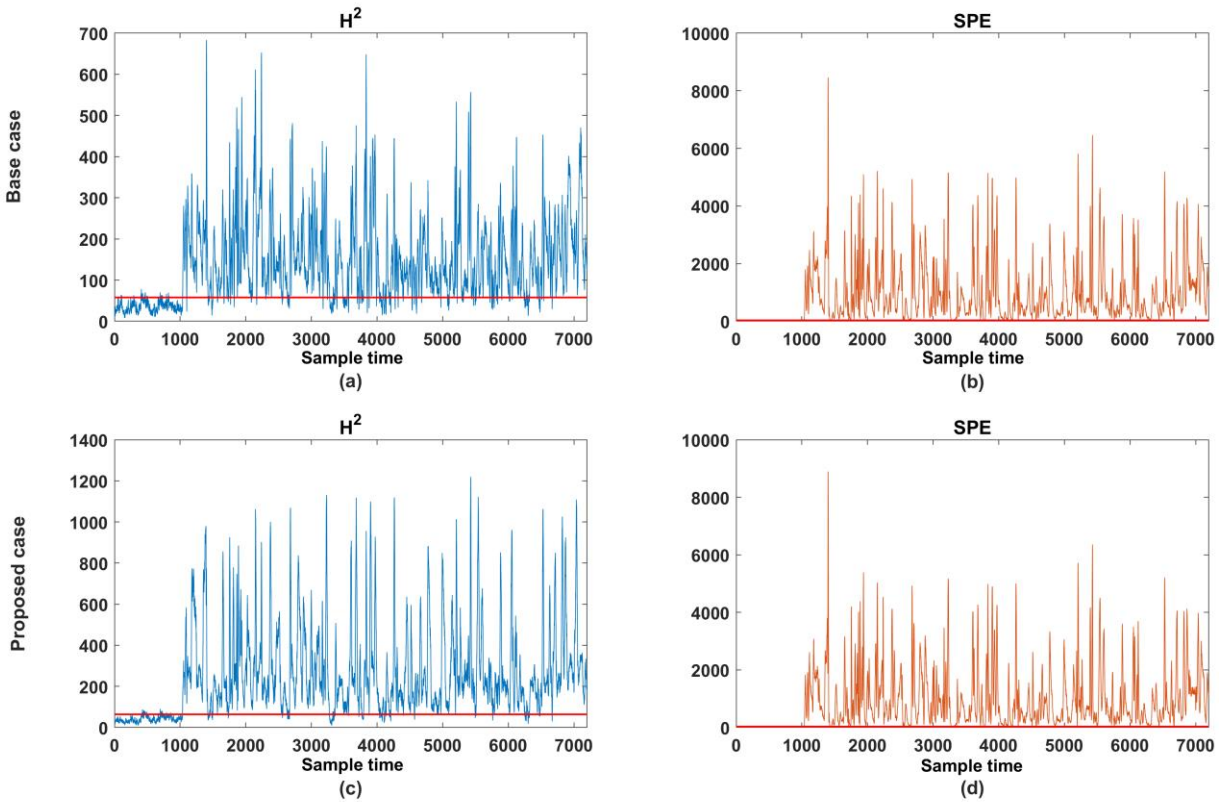

Figure 3.10. Monitoring charts of fault 11 for the base case ((a) and (b)) and the proposed case((c) and (d)).

The monitoring result of fault 14, which incurs the sticking of the reactor water cooling valve, is shown in Figure 3.11. Although more than 7.5% of the monitoring statistics of the base case in Figure 3.11 (a) improperly stay below the monitoring limit since a fault occurs in the 1000 sample time, the results of the proposed method in Figure 3.11 (c) neatly exceed the limit for all but 0.15% of the faulty samples. In terms of the fault detection rate, the fault was detected with high accuracy by the proposed method 99.85% of the time, with 92.37% being the base case. This demonstrates the effectiveness of the proposed method, particularly in the feature space. In addition, it can be confirmed that the proposed method is effective with other types of faults, such as the sticking of a valve as fault 14.

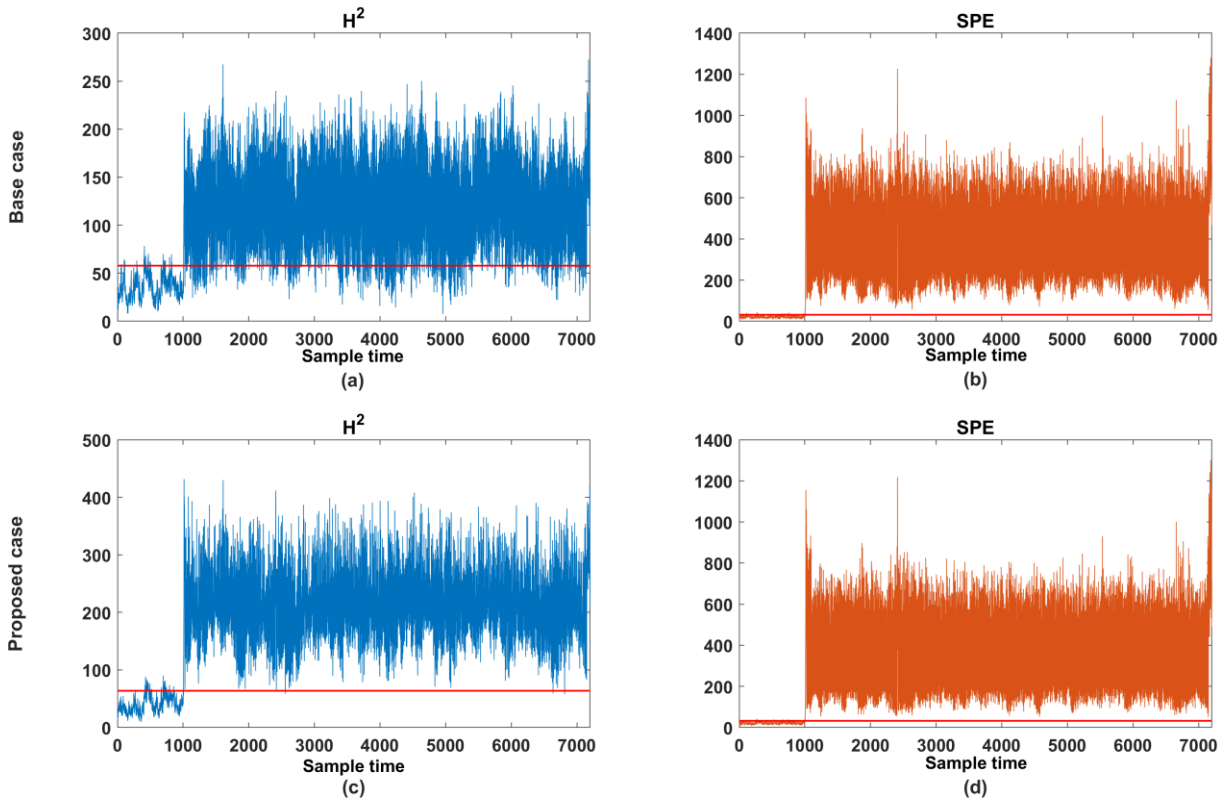

Figure 3.11. Monitoring charts of fault 14 for the base case ((a) and (b)) and the proposed case((c) and (d)).

For fault 18, in which the deviation of the heat transfer within the condenser occurs as a random variation type, a similar result can be observed in the monitoring result as shown in Figure 3.12. The monitoring charts in both cases have common trends where the fault pushes the state far from the normal condition, followed by the control actions compensating it iteratively. Given the control scheme applied to the TEP model used in this study [42], the trends of the monitoring charts in Figure 3.12 are the result of controlling the separator temperature by utilizing the condenser coolant valve. Meanwhile, the proposed method showed a distinct result, minimizing the restoration of the normal state and emphasizing the magnitude of the fault compared to the base case result, as shown in Figure 3.12 (a) and (c). Considering the monitoring results in the residual space, Figure 3.12 (b) and (d) shows a better performance than that of the feature space in both cases, and the improvement of the monitoring performance in the feature space from the base case, as shown in Figure 3.12 (c), can be interpreted as evidence that the data augmentation encourages manifold learning. As another advantage of the proposed method, the monitoring indices exhibit a larger magnitude of deviation in the monitoring statistics, which means that the proposed method can isolate the fault condition better.

The fault detection rates for all 28 faults in the TEP are summarized in Table 3.10. The results of PCA as the linear dimensionality reduction method are also included to compare the monitoring performance with the base case and the proposed case. While PCA based on the feature space represented slightly better monitoring performance than the base case which employed AE as the dimensionality reduction method, the base case showed a much higher FDR than PCA based on the residual space, especially in the hard-to-detect cases. This demonstrates the benefits of the nonlinear dimensionality reduction method of AE.

The detection rate in the residual space, SPE, is slightly higher in the base case than in the proposed case, but the difference is negligible considering that

the base case maintains a relatively higher FAR of 8.58% on the normal operation data than that of the proposed case (6.92%). It is also noteworthy that the proposed method in the feature space outperforms the base case for most situations while maintaining a lower FAR than the base case, which means that it can distinguish between normal and abnormal states more accurately.

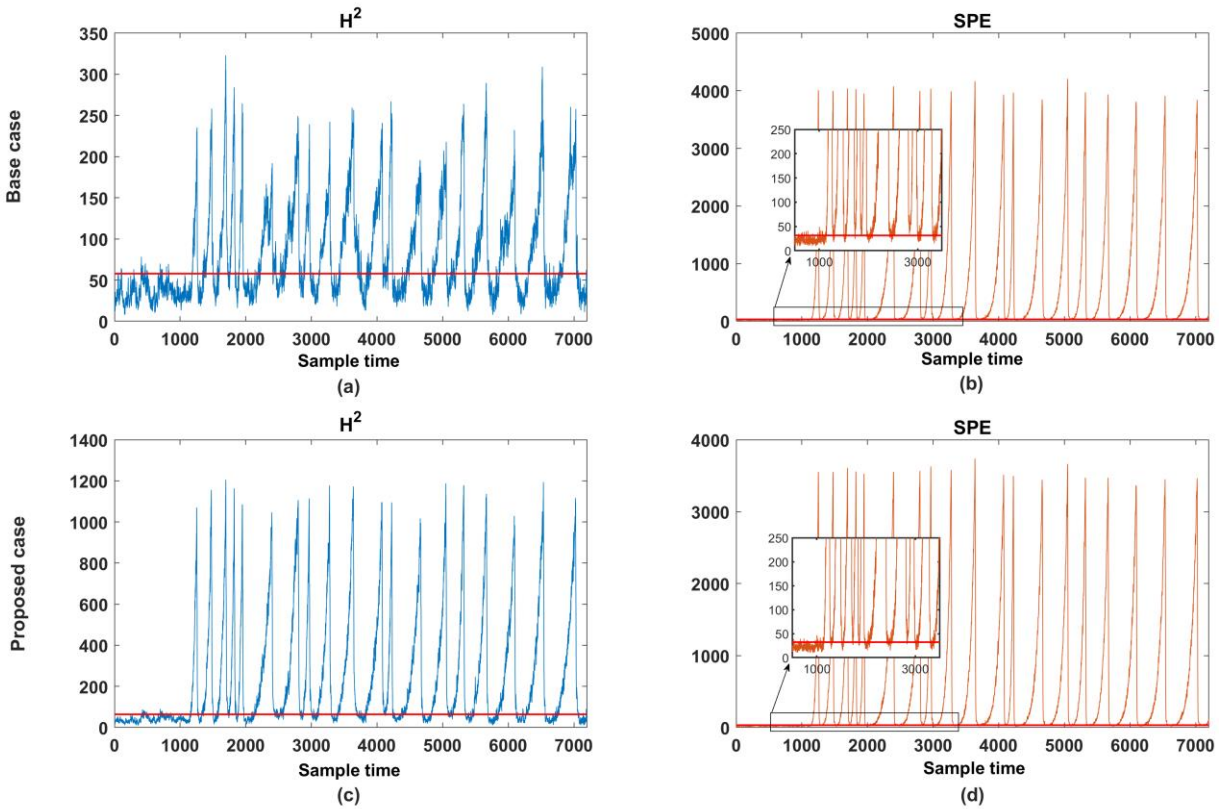

Figure 3.12. Monitoring charts of fault 18 for the base case ((a) and (b)) and the proposed case((c) and (d)).

Table 3.10. FDR (%) of PCA, the base case, and the proposed case for all 28 faults in the TEP model. (The value in the parenthesis corresponds to FAR (%) in each space.)

Fault No.	PCA		Base case		Proposed case	
	T^2 (1.00)	SPE (2.80)	H^2 (4.45)	SPE (8.58)	H^2 (3.50)	SPE (6.92)
1	99.82	99.85	99.69	99.95	99.90	99.90
2	99.39	99.48	99.37	99.81	99.37	99.55
3	0.53	12.33	1.63	25.98	1.53	15.26
4	99.97	99.97	99.40	99.97	99.97	99.97
5	0.84	9.69	2.90	26.95	2.79	17.38
6	99.72	99.72	99.72	99.72	99.72	99.72
7	99.97	99.97	99.97	99.97	99.97	99.97
8	98.11	98.0	97.69	98.87	98.00	98.50
9	1.55	13.53	2.21	32.24	7.08	20.79
10	60.95	92.7	62.34	94.44	75.78	93.61
11	96.37	98.36	89.15	98.87	94.48	98.68
12	28.72	43.03	20.35	65.34	39.64	56.51
13	99.18	99.27	98.15	99.47	99.29	99.40
14	99.13	99.92	92.37	99.97	99.85	99.95
15	0.61	7.51	2.06	22.21	1.32	14.47
16	0.37	7.01	1.63	19.29	0.84	12.56
17	95.50	98.47	91.69	98.69	97.5	98.61
18	62.69	84.88	57.15	87.60	70.54	85.02
19	97.23	99.39	92.45	99.45	97.94	99.39
20	96.97	97.6	92.21	97.87	97.18	97.73
21	1.05	7.13	4.55	22.98	2.87	14.92
22	0.98	18.09	4.45	34.19	3.43	21.53
23	0.74	8.58	3.21	24.77	2.64	15.93
24	88.39	97.21	75.89	98.21	92.74	98.06
25	45.84	89.73	36.09	92.90	67.39	89.92
26	63.03	91.36	64.89	93.78	77.23	92.08
27	71.19	90.50	50.43	94.05	63.55	92.84
28	1.32	9.67	3.47	26.93	5.61	18.26

As another performance index for the monitoring system, we can investigate the detection delay, which is the time required to detect a fault for the first time since it occurred. Except for a few hard-to-detect fault cases, such as faults 3, 9, and 15, for which the monitoring system could not effectively identify the process, the detection delay was significantly reduced by the proposed method in some fault cases. In terms of minimizing the loss of profitability due to process faults and securing process safety, the proposed method can inform engineers of faults more rapidly, allowing them to handle such faults as quickly as possible. Figure 3.13 shows that the delay was significantly reduced by the proposed method. Fault 10, where a random variation in the temperature of the C feed occurs, is the case with the greatest reduction in the fault detection delay while improving the detection accuracy by more than 10%. The delay in the base case was 351 samples, which corresponds to 210 minutes considering the sampling frequency of the TEP, whereas the proposed method can cut down on it by 168 samples, thereby reducing the fault detection delay by approximately 100 minutes. Even if the time when a large fault appears is equivalently assumed in terms of the monitoring statistics in the feature space, the detection delay can be reduced by 51 samples, corresponding to 30 minutes.

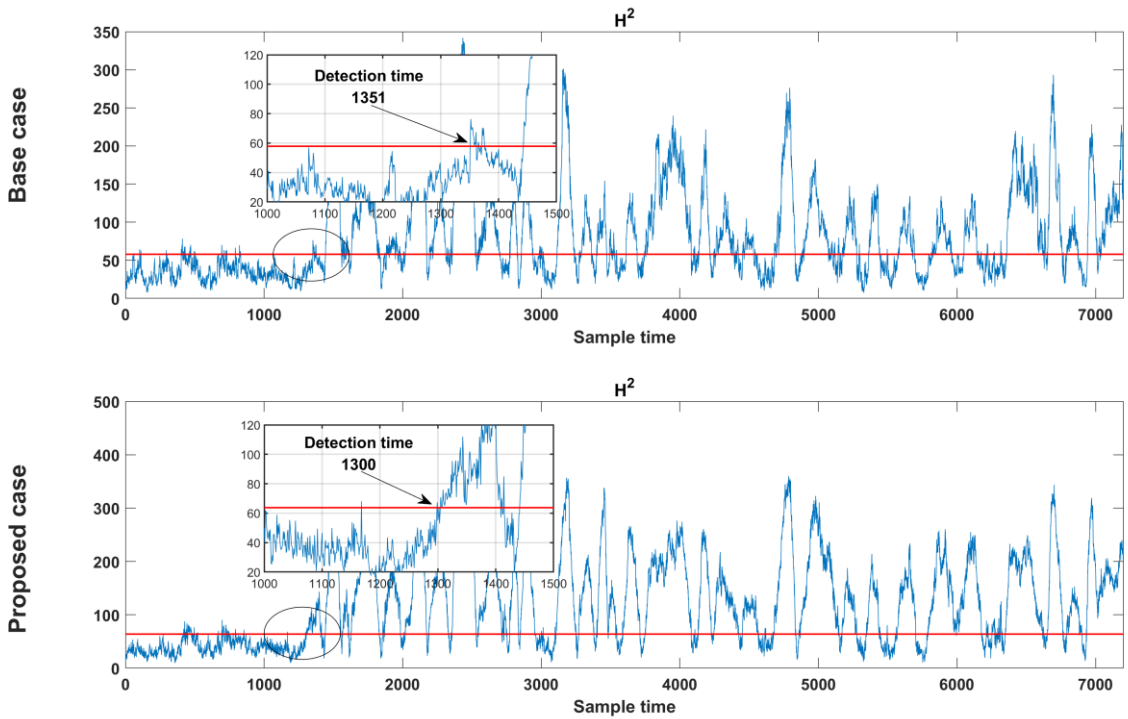

Figure 3.13. Comparison of the fault detection delay to first alarm for fault 10.

base case: 351 samples (210 min), proposed case: 300 samples (180 min). (A fault is introduced at 1000 samples.)

For fault 17, where the heat transfer within the reactor deviates from the nominal condition, the proposed method also shows an improvement in the detection accuracy and a delay reduction. The monitoring charts in the feature space for both cases are shown in Figure 3.14. As shown in the enlarged view of the plots, the detection delay in the proposed case was decreased by 21 sample times, which corresponds to approximately 12 minutes; thus, the monitoring accuracy is also improved by the proposed method.

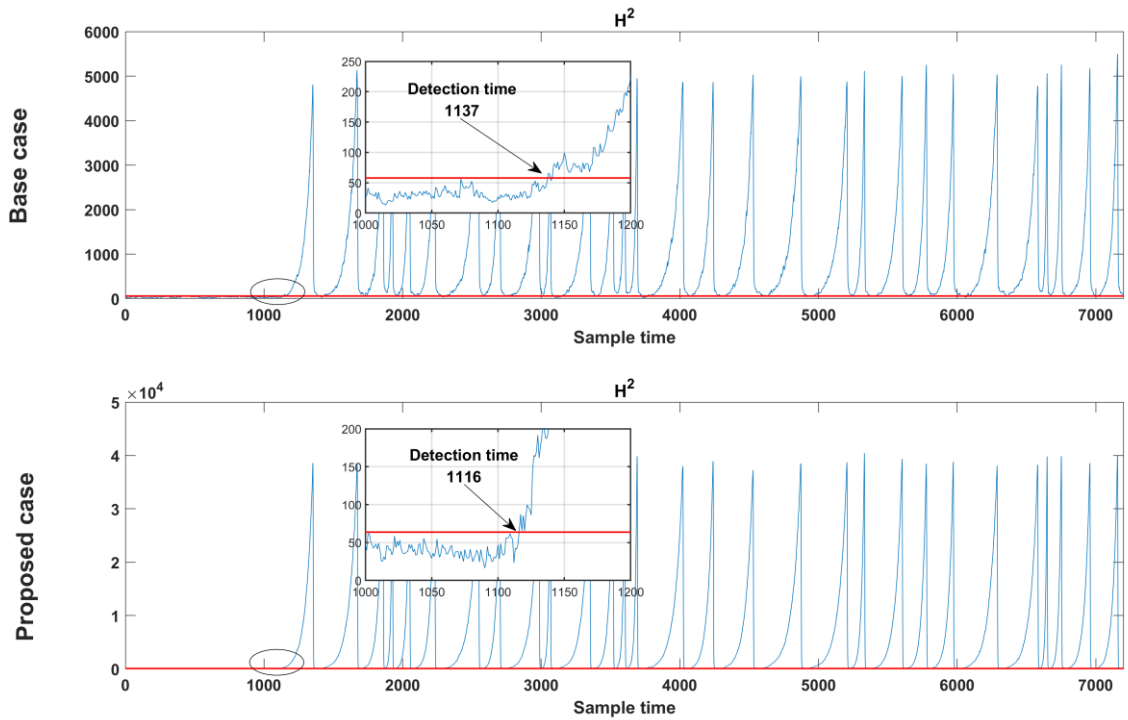

Figure 3.14. Comparison of the fault detection delay to first alarm for fault 17.

base case: 137 samples (82 min), proposed case: 116 samples (70 min). (A fault is introduced at 1000 samples.)

Chapter 4

Process Fault Isolation using Transfer Entropy and Graphical Lasso[2]

4.1. Introduction

Statistical process monitoring (SPM) usually comprises three steps. The first step is fault detection that observes process faults in real-time. Once a certain process fault is detected, the root cause of the fault should be identified as quickly and accurately as possible in the second step, the fault diagnosis. At the final step of the process monitoring, process recovery, the root cause that was identified in the second step should be eliminated so that the process can recover its original normal state. Even though the first and last steps of SPM corresponding to the process fault detection and process recovery are the crucial ones, the fault diagnosis step is notably challenging and difficult. To tackle fault diagnosis, various methodologies have been suggested, which can be categorized into knowledge-based and historical data-based methods.

Knowledge-based methods have the characteristics of computational tractability and human readability, and the collective knowledge of all ages based on experience is the most trustworthy resource for the knowledge-based method. The notable examples are adjacent matrix or reachability matrix which can be built by using process flow diagrams (PFD) or piping & instrumentation diagrams (P&ID), and such matrices are straightforward to construct and interpret-

[2] This chapter is an adapted version of H. Lee., C. Kim., S. Lim., and J. M. Lee, "Data-driven fault diagnosis for chemical processes using transfer entropy and graphical lasso.", *Computers & Chemical Engineering*, 142, 107064.

able. In addition, structural models composed of a set of fundamental first principles can be used to capture the causal relations in the process together with process knowledge, which are commonly called grey-box models.

However, knowledge-based methods that depend entirely on domain knowledge of experts are not only a waste of human resources, but they are becoming a more impractical and exhausting task as the complexity and scale of a process consistently increases. Although knowledge-based methods provide promising results in mechanical and aeronautical applications in which the mature linear theory works, they are unsuitable for the inherently nonlinear chemical process [22]. In addition, knowledge-based methods accompany the validation by process historical data [48].

Process historical data-based methods show their advantage over knowledge-based methods as the dimension and complexity of a process increases. When developing a diagnosis system for a nonlinear and large-scale process, the historical data-based methods are more plausible than the knowledge-based methods since the diagnosis system requires relevant domain knowledge to build an explicit system model in different areas. In addition, doubtful or erroneous information may disturb a modeling result, which degrades the performance of the knowledge-based methods. As historical data-based methods do not require any prior knowledge of the process for diagnosis, they can be applied to general forms of nonlinear processes. Knowledge-based methods generally go through a qualitative procedure to infer the results of the diagnosis, while the historical data-based methods can carry out quantitative analysis to obtain more detailed analysis results without the exact knowledge of an expert.

Various studies on historical data-based diagnosis methodologies were conducted since the early development stage of SPM. PCA, proposed by Pearson and further developed by Hotelling, is the first multivariate statistical technique for a data-based diagnosis system. However, as a dimensionality reduction method, it is inevitable that the loss of information for capturing the root cause

of the fault would occur in the reconstruction process from the reduced to the original dimensional data space. Other representative studies using dimensionality reduction methods such as PLS, ICA, KPCA, multiway principal component analysis (MPCA) [49], dynamic principal component analysis (DPCA), and vertices principal component analysis (VPCA) [50] are also confronted with the same limitations of information loss in the reconstruction. To alleviate these limitations, various studies exploited the use of neural networks to develop diagnostic classifiers. The representative studies include incorporation of prior domain knowledge or expert system, data preprocessing, filtering, and integration of another type of neural network [22]. The fault diagnosis system, as the neural network is based on the supervised learning framework using an identified training database, is bound to have limitations exhibiting unreliable performance in unknown situations when it goes out of training data. To rectify these shortcomings the following methods that identify the causal relation of the abnormality in an unsupervised manner are proposed using the characteristics of the data. Causal analysis was investigated from various viewpoints. Cross-correlation analysis [48], which calculates the correlation between a pair of time series, estimates the time delay on the basis of the corresponding lag. While cross-correlation analysis is practical and easy to implement, it is vulnerable to non-linear relationships and it cannot reflect the trend in time series in the estimation of time delay. At the other end of the spectrum, Granger causality [51], a methodology based on regression models, can be used to reveal a causal relationship taking into account the dynamics. It concludes the causal relations between two variables by comparing the two regression models. One is the regression model of variables based only on lagged values of itself and the other is the augmented regression model with the addition of lagged values of another variable. If a significant improvement in the regression performance compared to that of the original model exists, Granger causality would capture the causal relationship between those variables [52]. Although Granger causality makes

up for the weakness of cross-correlation, which is the inability to give consideration to dynamics, it also assumes a linear relationship among the variables and the accuracy of the regression model considerably affects the result of Granger causality analysis. Apart from the time domain methodologies, frequency domain methods to find causality were also attempted, which are represented by the directed transfer function (DTF) and partial directed coherence (PDC) [48]. However, the frequency domain methods also possess drawbacks similar to those of the time domain methods. Recently, transfer entropy [12], which is an information-theoretic approach of causality analysis, was suggested as an up-to-date analysis measure. Transfer entropy was first applied in the neuroscience area, in which an influential network between genes is difficult to infer due to internal stimuli and complexity of the system. As a new time domain data-based methodology, transfer entropy is a more effective way of dealing with causal relationships in the nonlinear process than the previous ones based on the linearity assumption between subsystems. The application of transfer entropy to chemical processes started in the early 2000s and turned out to be a promising analysis measure [12].

Various studies utilized the concept of transfer entropy for fault diagnosis problems in chemical processes. At the early stage of using transfer entropy diagnosis methodology, Bauer et al. [53] conducted a preliminary study that focused on the structural transition of information transfer from normal to abnormal state in transfer entropy. The results revealed the inconsistency issue of transfer entropy in a different condition of the process, which means that the direction of the causal relationship may change depending on the state. In addition, in the earlier studies transfer entropy was only applied to small-scale processes for demonstration due to the critical drawback that the calculations required immense computational cost when extended to industrial scale. The process demonstrated in the study of Bauer et al. [53] consisted of a total of 12 variables, of which only 5 variables were used in causal analysis using transfer

entropy.

Bauer et al. [15] suggested in a subsequent study that the fault propagation path can be resolved by proposing a causality measure utilizing transfer entropy without any observable time delay in a chemical process. Nevertheless, the causal map derived from the new measure was limited to presenting just the result of a reactor with 10 variables.

By expanding the dimension of the process, Bauer and Thornhill [54] adopted transfer entropy to inspect the cause-effect relations of the data from TEP, which is a benchmark chemical process. However, the study analyzed causal relationships only for a subset of the entire set of variables. Due to the long sampling intervals, they excluded 19 concentration measurement variables ('XMEAS #23~41' in the 41 TEP) and exploited only 22 measurement variables, which is only half of the total measurement variable set. Recently, Lindner et al. [55] compared transfer entropy against Granger causality for a chemical process with seven process variables which was merely a simple water tank simulation, and observed the same limitation in terms of the scale of the target process. It is not a constructive way to exclude available measurable variables from the analyzed process, because it not only limits the scope of the process monitoring but also generalizes the methodology. Furthermore, the TEP has 12 additional variables that represent manipulated variables used in the control loops. Even though these variables were also considered and put on record with the set of measured variables, the manipulated variables were not evaluated in this study. However, according to the guidance of Isermann [56], the input variables of the control scheme corresponding to the manipulated variables should be included in the object of investigation because some types of fault may be compensated for by the control scheme.

This study was motivated by the main shortcoming of previous studies that transfer entropy is only applicable to limited small-scale processes primarily due to the high computational cost. The proposed method for the root cause

analysis suggests an integrated use of transfer entropy with a regularization method, graphical lasso, to alleviate the drawback in the costs by eliminating redundant relationships among the entire variable set prior to costly calculation of causality measures. Therefore, in this study, a new methodology using transfer entropy was designed to improve the viability of fault diagnosis that can be effectively applied to the industrial-scale process in terms of cost-effectiveness while embracing all of the available variables.

The proposed method was tested for two processes to compare its performance with that of the conventional transfer entropy method. The proposed method was first applied to a selective catalytic reduction (SCR) system, to test its performance and verify the operating mechanism during the fault diagnosis process on an open-loop small-scale process. Afterward, the method was applied to the widely used TEP benchmark problem to test its performance on a closed-loop industrial-scale process.

This chapter is outlined as follows. In Section 2, the preliminaries essential for describing the proposed methodology are introduced. Then, in Section 3 the proposed methodology is applied to SCR system and the result of it is discussed. Finally, the case study and discussion about the TEP system are presented in Section 4.

4.2. Fault Isolation using Transfer Entropy Integrated with Graphical Lasso

In a previous study by C. Kim et al.[47], a new fault detection algorithm that incorporates the Markov random field (MRF) and graphical lasso is proposed. With the help of the iterative graphical lasso, the proposed fault detection method improves monitoring performance while reducing the computational cost of constructing the MRF by extracting only the significant relationships between the highly related variables before the monitoring process. Meanwhile, transfer entropy, statistics for causal analysis, has an analogous limitation to estimate the probability density functions (PDFs). The inherent limitation of transfer entropy is that it incurs excessive computational cost in the pairwise calculation in terms of each process variable. This needs to be managed to make it practical for the industrial-scale process. Therefore, along with the basis of previous work, we propose a fault diagnosis methodology that takes advantage of the superior performance of the causal analysis of transfer entropy while resolving the limit of high cost by incorporating the regularization method. In addition, another issue with transfer entropy is that performance strongly depends on the structural configuration which corresponds to the embedding parameters of the cause and effect variables and prediction horizon. Thus, the regularization method also has a positive effect in terms of making parameter tuning feasible.

The flowchart of the proposed methodology is shown in Figure 4.1. When a fault is detected by the real-time fault detection algorithm, the fault diagnosis proceeds according to the proposed method in the order shown in the gray shaded box. According to the study of Bauer et al. [53], the directionality of the causal relationship between process variables can be changed in abnormal process conditions. The notable difference of the proposed fault diagnosis method

from the previous fault detection algorithm is that subgroups modeling via iterative graphical lasso is performed based on the fault data rather than normal data. As the next step, transfer entropy is calculated with respect to each subgroup. The causality measure proposed by Bauer et al. [15] is derived from the transfer entropy. Lastly, the root cause analysis is carried out based on the relative magnitude of the causality measures. After the fault diagnosis is done, the entire procedure of process monitoring should be completed through the process recovery step. The detailed procedure of fault diagnosis in the gray box is described in the following sections for each case study.

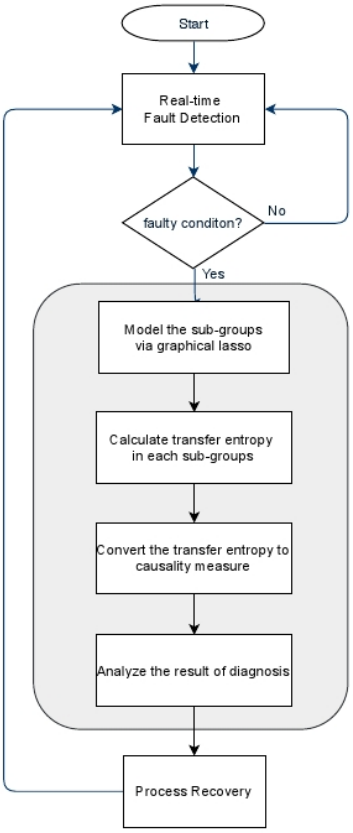

Figure 4.1. Flowchart of process monitoring with proposed fault diagnosis method. (gray-shaded)

4.2.1. Graphical Lasso for Sub-group Modeling

The elements of the inverse covariance matrix based on standardized data represent the structure of the undirected graphical model. The relationship between variables with high relevance can be obtained using the L_1-penalized regularization, $LASSO$, based on the undirected graph [18]. As the preprocessing of the proposed root cause methodology, subgroup modeling, which extracts relevant relationships among process variables in each fault condition, is performed via the iterative graphical lasso. The variables that are excluded from the subgroup of the preceding iteration are reconstructed into subgroups by iteratively setting up the graphical lasso problem. The lasso penalty parameters (ρ, in Eq. (2.20) of section 2. 4.) for each iteration, an important hyperparameter that determines the sparsity of the resulting lasso problem, are determined by the number of pre-determined subgroups (G) such that a similar number of variables is distributed to each subgroup. The number of subgroups (G) is determined by case studies based on process monitoring performance [47].

4.2.2. Transfer Entropy for Fault Isolation

After the preprocessing step, transfer entropies for each subgroup are calculated to obtain a metric for the causality analysis as in Eq. (2.13) and (2.14), respectively. In this study, the transfer entropy is computed using the algorithm suggested by Lindner et al [43,45]. The transfer entropy is calculated as the difference between the Shannon entropies with and without considering the effect of the cause variables as in Eq. (2.13). The probability density estimation in the Shannon entropy was performed by KDE, and the number of bins of KDE was set to 10. As mentioned before, the analytical performance of transfer entropy depends heavily on hyperparameters, which include L and K, each corresponding to the embedding parameters of the cause and effect variables. There are also prediction horizon, H, and the estimated time delay between the process variables X and Y, τ. The embedding parameter of the effect variable, K, is generally set to one by the 'Self Prediction Optimality' requirement to identify the causal relationship that excludes the effect of its information storage [14]. When the system dynamics can be identified, the optimal parameters can be determined accordingly. However, if this is not the case, it is recommended to use the same value for the prediction horizon (H) and time delay (τ). In addition, the embedding parameter of the effect variable (L) is generally set as small a value as possible considering the computational cost [15]. The prediction horizon and embedding parameter for the cause variable are determined depending on the system under consideration.

Once the calculation of transfer entropy is completed within each subgroup, the causality measure is derived to perform the causal analysis. The causality measure between X and Y (Eq. (2.14)) is defined as the difference between the information transfer from X to Y and that of the opposite direction. The fact that the causality measure has a large positive value means that the causal influence of X on Y is significant, and if it is a large negative value, vice versa.

No causal relationship is captured if the causality measure is less than the threshold value. As the metric of transfer entropy has its meaning in relative magnitude rather than absolute value, the measures need to be scaled for comparison. Based on the largest value from the causality measure, all the values in subgroups are normalized and then significant causal relationships above the threshold value are classified.

4.3. Case study and Discussion 1

In this section, we apply the proposed methodology to fault data of an industrial process to verify its performance. The proposed root cause methodology is first applied to a relatively small-scale open-loop system, the SCR process. The description of the process and the detailed procedure for applying the proposed methodology can be found in the subsequent sections.

4.3.1. Selective Catalytic Reduction Process

The proposed methodology for root cause analysis of process fault is validated using a process model based on a real system. As the first case study for validation, the SCR process is utilized. The SCR process is the after-treatment system to reduce NO_x in diesel engine exhaust, which is an essential part to satisfy the environmental regulations of diesel vehicles. The exhaust of diesel engines is injected into the catalytic reactor and the target material NO_x is reduced by the redox reaction with the reductant NH_3. The one-dimensional dynamic model for the SCR reactor is divided into four sections along with the flow direction to improve the accuracy of the simulation according to the Method of Lines as Figure 4.2. The governing equations of the model are derived based on the previous work of Depcik et al. [58] and Kim et al. [59]. They include mass and energy balances in the bulk gas and catalytic surface phase and interfacial phenomena between the two phases. The main difference between the studies of Depcik et al. [58] and Kim et al. [59] is that the process itself considered in the former study is 'Urea SCR', which exploits the urea as the precursor of ammonia, rather than the Urealess SCR combined with a lean NO_X trap (LNT) system. Therefore, the relevant reaction kinetics describing the whole SCR system are presented in

Table **4.1**. The reactions include adsorption, desorption (R1–R4), and oxidation (R5–R7) of NH_3, oxidation of NO (R8), and reduction of NO_x (R9–R12). The kinetic parameters of the model are estimated by the particle swarm optimization algorithm using the chassis dynamometer test data. There are 18 variables in the SCR model, and the details are provided in Table 4.2. The entire set of target variables subject to diagnosis is composed of 8 inlet variables and 10 outlet variables. Various types of faults such as step deviation, random variation, and sticking can be applied in the inlet variables. To test the proposed method on the SCR system, a random variation fault in the mole fraction of NH_3 of the SCR inlet gas was introduced and diagnosed using the proposed methodology.

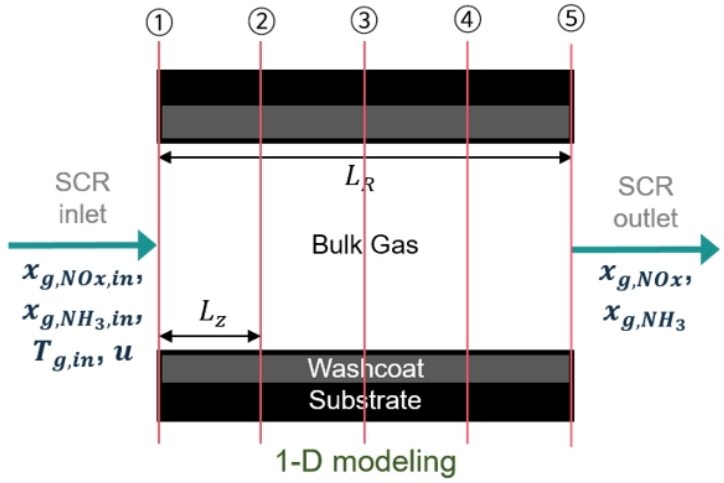

Figure 4.2. Schematic diagram of the 1-dimensional dynamic SCR model.
(L_R: $Total\ length$, L_z: $Length\ of\ discretized\ section$.)

Table 4.1. Reaction kinetics and kinetic parameters of selective catalytic reduction (SCR).

(R1) $NH_3 + S1 \rightarrow NH_3 - S1$	$r_1 = k_1 x_{NH_3}(1 - \theta_{m,1})\psi_1$
(R2) $NH_3 - S1 \rightarrow NH_3 + S1$	$r_2 = k_2 \theta_{m,1} \psi_1$
(R3) $NH_3 + S2 \rightarrow NH_3 - S2$	$r_3 = k_3 x_{NH_3}(1 - \theta_{m,2})\psi_2$
(R4) $NH_3 - S2 \rightarrow NH_3 + S2$	$r_4 = k_4 \theta_{m,2} \psi_2$
(R5) $NH_3 + 5/4\, O_2 \rightarrow \mathrm{NO} + 3/2\, H_2O$	$r_5 = k_5\, x_{NH_3} x_{O_2}/(T_{wc}G)$
(R6) $NH_3 + 3/4\, O_2 \rightarrow 1/2\, N_2 + 3/2\, H_2O$	$r_6 = k_6\, x_{NH_3} x_{O_2}/(T_{wc}G)$
(R7) $NH_3 + O_2 \rightarrow 1/2\, N_2\mathrm{O} + 3/2\, H_2O$	$r_7 = k_7\, x_{NH_3} x_{O_2}/(T_{wc}G)$
(R8) $\mathrm{NO} + 1/2\, O_2 \leftrightarrow NO_2$	$r_8 = k_8\,(x_{NO} x_{O_2}^{0.5} - x_{NO_2}/K_p)/(T_{wc}G)$ *
(R9) $NH_3 - S2 + NO + 1/4\, O_2 \rightarrow \mathrm{S2} + N_2 + 3/2\, H_2O$	$r_9 = k_9\, x_{NO} x_{O_2} \theta_{m,2} \psi_2 / G$
(R10) $NH_3 - S2 + 1/2\, NO + 1/2\, NO_2 \rightarrow \mathrm{S2} + N_2 + 3/2\, H_2O$	$r_{10} = k_{10} x_{NO} x_{O_2} \theta_{m,2} \psi_2$
(R11) $NH_3 - S2 + 3/4\, NO_2 \rightarrow \mathrm{S2} + 7/8\, N_2 + 3/2\, H_2O$	$r_{11} = k_{11} x_{NO_2} \theta_{m,2} \psi_2$
(R12) $NH_3 - S2 + 5/4\, NO_2 \rightarrow \mathrm{S2} + 1/8\, N_2 + NO_2 + 3/2\, H_2O$	$r_{12} = k_{12} x_{NO_2} \theta_{m,2} \psi_2$

$\theta_{m,k}$: Coverage fraction of 'k' site (0~1), ψ_k: $Storage\ capacity\ of\ k\ site$ $[mol/m^3]$,
G: Inhibition factor, k_i : Kinetic parameters.
(k : intermediate index (S1, S2), i : reaction index (1~12),
* $K_p = \exp[-\Delta G/RT_{wc}]$, $\Delta G = \Delta H - T\Delta S$)

Table 4.2. Process variables of the SCR model.

Variable No.	Variable Name	Variable No.	Variable Name
1	Temperature of SCR inlet gas	**10**	Mole fraction of NO of SCR outlet gas
2	Volumetric flow rate of SCR inlet gas	**11**	Mole fraction of NO_2 of SCR outlet gas
3	Mole fraction of NH_3 of SCR inlet gas	**12**	Mole fraction of N_2O of SCR outlet gas
4	Mole fraction of NO of SCR inlet gas	**13**	Mole fraction of N_2 of SCR outlet gas
5	Mole fraction of NO_2 of SCR inlet gas	**14**	Mole fraction of O_2 of SCR outlet gas
6	Mole fraction of N_2 of SCR inlet gas	**15**	Mole fraction of H_2O of SCR outlet gas
7	Mole fraction of O_2 of SCR inlet gas	**16**	Coverage fraction of S1 site
8	Mole fraction of H_2O of SCR inlet gas	**17**	Coverage fraction of S2 site
9	Mole fraction of NH_3 of SCR outlet gas	**18**	Temperature of SCR outlet gas

4.3.2. Implementation of the Proposed Methodology

To apply the proposed methodology combining the graphical lasso and transfer entropy, several parameters of each step must be determined. In this example, a total of 18 SCR variables were analyzed in two groups of nine each by adjusting the hyperparameter ρ of the graphical lasso, thus reducing the unnecessary cost of transfer entropy calculations. With the help of the first step, the next step for the transfer entropy analysis requiring pairwise analysis about the whole variables can be reduced by more than 50%. The embedding parameters of transfer entropy such as $L, H,$ and τ were determined through a case study of the various faults in terms of each parameter, which has a critical influence on the performance of the transfer entropy. The hyperparameters are provided in

Table 4.3. The simulation was run for 2000 sample times, and a process fault for the SCR example started at the beginning of the simulation. The threshold of the causality measure, which is the cutoff value to filter out only meaningful causal relationships exceeding it, was set to 0.85. This value should be determined depending on the target process and can be adjusted to a smaller value to perform a more conservative diagnosis.

Table 4.3. Hyperparameters of transfer entropy in SCR.

Hyperparameter	Value
Embedding dimension of cause variable (L)	2
Embedding dimension of effect variable (K)*	1
Prediction horizon (H)	2
Estimated delay (τ = H)	2

* Self Prediction Optimality: K = 1.

4.3.3. Discussion of the Results

The fault scenario of random variation of the inlet mole fraction of NH_3, variable 3 in

Table 4.2, was applied to the SCR system. For each subgroup divided in this condition, the result of causality measure from the proposed method is provided in Figure 4.3. Since it is the relative magnitude of the causality measures that have a significant meaning in the interpretation of the diagnosis result, the measures were shown as a bar graph after normalization based on the largest absolute value. The relationships exceeding the threshold value of the causality measure, indicated by the transparent yellow plane, are summarized in

Table **4.4** in the order of magnitude. From the relationships detected in subgroup 2, variable 3, the root cause of the process fault, is well identified as the cause variable. As the random variation occurs in the NH_3 mole fraction of the inlet gas, the mole fraction of the other inlet gas components and that of NO_2, the main component of the NO_x, are mainly affected. It is also noteworthy that no causal relationship is identified in subgroup 1, given that the cause is distinctly isolated by the proposed method.

Table **4.4** also includes the analysis result on all 18 variables, which indicates the proposed method requires only about half of the computational cost compared to the conventional diagnostic method. Thus, the proposed method has the advantage of reducing the cost of diagnosis. In conclusion, the proposed methodology could capture the root cause of the process fault in the open-loop

system.

Table 4.4. Application results and causal analysis of SCR example.

		Process variables in each group	**Calculation time (sec)**	**Causal relationship (causality measure*)**
Proposed Method	**Subgroup 1**	1,2,9,10,12,14,15,16,17	55	-
	Subgroup 2	3,4,5,6,7,8,11,13,18	54	**3** → 8 (1) **3** → 11(0.9347) **3** → 7(0.9213) **3** → 5(0.8675)
		Total: 109		
Conventional Transfer Entropy Method	**Whole process**	1 ~ 18	236	**3** → 14 (1) **3** → 8 (0.9191) **3** → 11 (0.8591)

* Relative values of causality measure are represented in parenthesis.

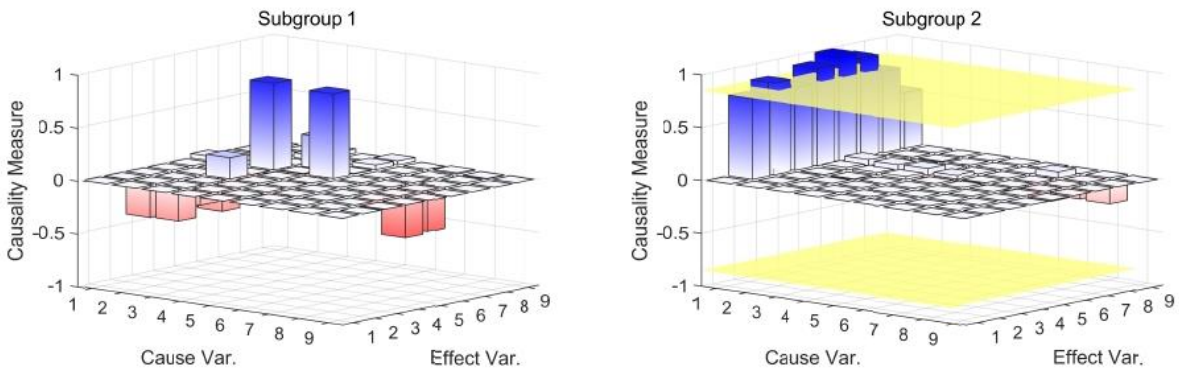

Figure 4.3. Causality measure plot of fault scenario of SCR example.

4.4. Case study and Discussion 2

In this section, we apply the proposed method to the fault data of TEP model, to validate the performance of the proposed method in the industrial-scale process. It is also noteworthy that TEP is a closed-loop benchmark chemical process that includes various control structures. The diagnosis of TEP, which includes recycle stream and complex control logic, was conducted to examine the applicability of the proposed methodology on the actual processes.

4.4.1. Tennessee Eastman Process

TEP is a benchmark process widely used for testing process monitoring and control algorithms. It consists of five process units, which are a reactor, condenser, separator, stripper, and recycle compressor, as shown in Figure 4.4. Four gaseous feeds, A, C, D, and E are used to produce two liquid products G and H through gas-phase catalytic reactions, followed by the separation process to obtain the liquid products from the unconverted reactants. Some of the reactions involve inert gas 'B' and are accompanied by two side reactions producing by-product 'F'. In this study, among three modes in the study of Downs and Vogel [40], the base case mode which produces two products at the ratio of 50:50 is adopted. The analysis was based on the revised MATLAB version [41], which contains a control structure proposed by Ricker [41]. There are 41 measured output variables with noise: 22 continuous measurements and 19 sampled composition measurements from analyzers. There are also 12 manipulated variables for process control. Of the total 53 parameters, 50 variables excluding three manipulated variables (compressor recycle valve, stripper steam valve, and agitator speed) with a fixed value in the base case were analyzed. The 50 variables are listed in Table 4.5. The revised MATLAB model has a total of 28 fault cases, including 20 pre-defined faults equipped in the original TEP model. A total of

28 faults cover various fault types such as step change, random variation, and sticking, as well as multiple fault cases where more than one fault occurs simultaneously. All the types of disturbances defined in TEP are summarized in Table 4.6.

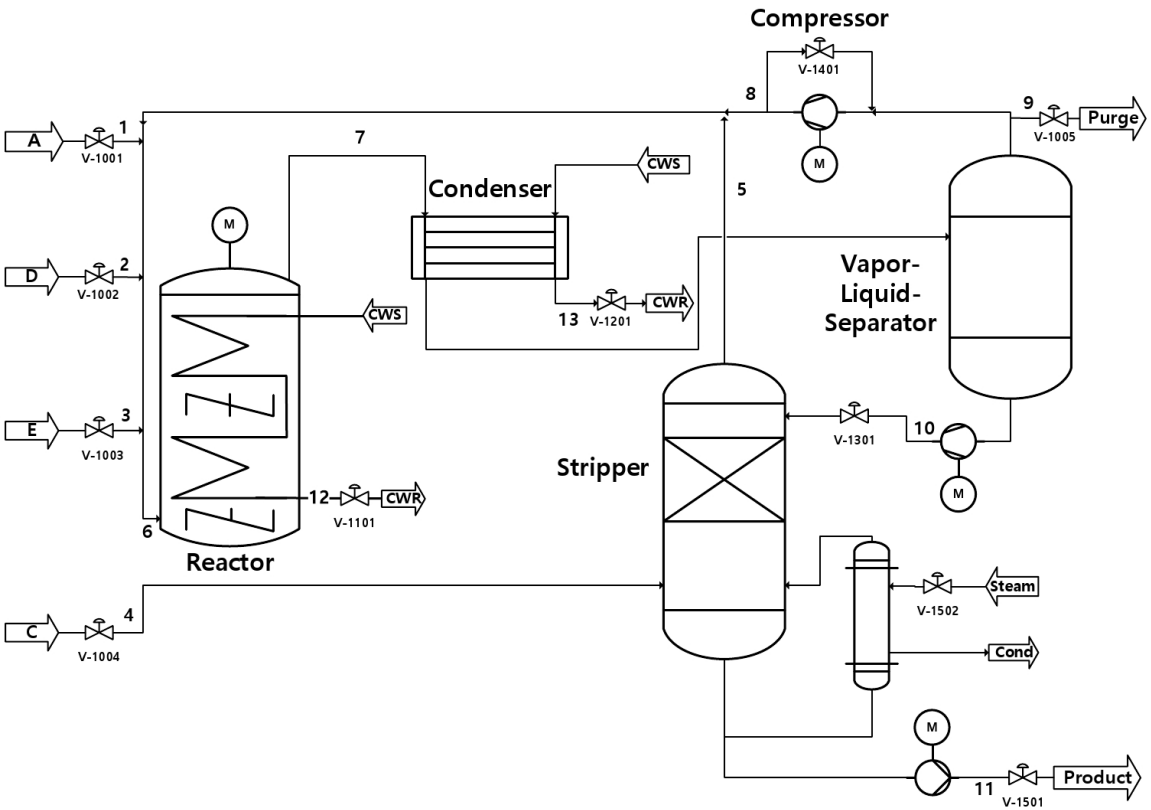

Figure 4.4. Process Flow Diagram of Tennessee Eastman process. Revised MATLAB version.

Table 4.5. Process variables of TEP subject to causal analysis.

Variable No.	Variable Name	Variable No.	Variable Name
1	A feed flowrate (stream 1)	**18**	Stripper temperature
2	D feed flowrate (stream 2)	**19**	Stripper steam flowrate
3	E feed flowrate (stream 3)	**20**	Compressor work
4	A & C feed flowrate (stream 4)	**21**	Reactor c/w outlet temperature
5	Recycle flowrate (stream 8)	**22**	Condenser c/w outlet temperature
6	Reactor feed rate (stream 6)	**23~28**	Reactor feed analysis (A~F mol%) (stream 6)
7	Reactor pressure	**29~36**	Purge gas analysis (A~H mol%) (stream 9)
8	Reactor level	**37~41**	Product analysis (D~H mol%) (stream 11)
9	Reactor temperature	**42**	D feed flow valve (stream 2)
10	Purge rate (stream 9)	**43**	E feed flow valve (stream 3)
11	Product separator temperature	**44**	A feed flow valve (stream 1)
12	Product separator level	**45**	A & C feed flow valve (stream 4)
13	Product separator pressure	**46**	Purge valve (stream 9)
14	Product separator under flowrate (stream 10)	**47**	Separator pot liquid flow valve (stream 10)
15	Stripper level	**48**	Stripper liquid product flow valve (stream 11)
16	Stripper pressure	**49**	Reactor c/w flow valve
17	Stripper under flowrate (stream 11)	**50**	Condenser c/w flow valve

Table 4.6. Process faults in Tennessee Eastman process (TEP).

No.	Description (Root Cause variable)	Type
IDV(1)	A/C feed ratio, B composition constant (stream 4)	Step
IDV(2)	B composition, A/C ratio constant (stream 4)	Step
IDV(3)	D feed temperature (stream 2)	Step
IDV(4)	Reactor cooling water inlet temperature	Step
IDV(5)	Condenser cooling water inlet temperature	Step
IDV(6)	A feed loss (stream 1)	Step
IDV(7)	C header pressure loss – reduced availability (stream 4)	Step
IDV(8)	A, B, C feed composition (stream 4)	Random variation
IDV(9)	D feed temperature (stream 2)	Random variation
IDV(10)	C feed temperature (stream 4)	Random variation
IDV(11)	Reactor cooling water inlet temperature	Random variation
IDV(12)	Condenser cooling water inlet temperature	Random variation
IDV(13)	Reaction kinetics	Slow drift
IDV(14)	Reactor cooling water valve	Sticking
IDV(15)	Condenser cooling water valve	Sticking
IDV(16)	*Unknown (Deviation of heat transfer within stripper heat exchanger)	*Unknown (Random variation)
IDV(17)	*Unknown (Deviation of heat transfer within reactor)	*Unknown (Random variation)
IDV(18)	*Unknown (Deviation of heat transfer within condenser)	*Unknown (Random variation)
IDV(19)	*Unknown (recycle valve, stripper steam valve, underflow separator (stream 10), underflow stripper (stream 11))	*Unknown (Sticking)
IDV(20)	*Unknown	*Unknown
IDV(21)	A feed temperature (stream 1)	Random variation
IDV(22)	E feed temperature (stream 3)	Random variation
IDV(23)	A feed pressure (stream 1)	Random variation
IDV(24)	D feed pressure (stream 2)	Random variation
IDV(25)	E feed pressure (stream 3)	Random variation
IDV(26)	A & C feed pressure (stream 4)	Random variation
IDV(27)	Reactor cooling water pressure	Random variation
IDV(28)	Condenser cooling water pressure	Random variation

* Unknown: Uncovered by A. Bathelt in revised MATLAB version model [41].

4.4.2. Implementation of the Proposed Methodology

To implement the proposed methodology, 50 variables of TEP were divided into five subgroups using the graphical lasso, and fault diagnosis was carried out for each of the groups. Like the previous case study, the hyperparameter ρ of the graphical lasso was adjusted to include ten process variables in each subgroup on average. The embedding parameter L had the best diagnostic performance through a case study of the various faults, as shown in

Table **4.7**. The remaining hyperparameters were specified by the values used in previous work by Bauer et al. [15]. and Wibral et al. [14]. The hyperparameters used in this study are summarized in

Table **4.7**. The simulation was run during a total of 7200 sample points by setting the disturbance to occur with the start of the simulation. The cutoff value to find causal relationships was set to the same as the previous example of the chemical process, 0.85.

Table 4.7. Hyperparameters of transfer entropy in TEP.

Hyperparameter	Value
Embedding dimension of cause variable (L)	4
Embedding dimension of effect variable (K)*	1
Prediction horizon (H)	4
Estimated delay ($\tau = H$)	4

* Self Prediction Optimality : $K = 1$

4.4.3. Discussion of the Results

Upon implementation of the proposed methodology, the fault diagnosis results of IDV (4) are shown in Figure 4.5. The 11th variable of the second subgroup, variable 49 in the original numbering of Table 4.5, was detected as the root cause variable that exceeded the threshold value of the transparent yellow plane, as summarized in

Table 4.8. In addition, variables 21 and 9, which correspond to the 4th variable in the second subgroup and the 4th variable in the 5th subgroup, respectively, were indicated as the next most likely root cause variables, although they did not exceed the threshold value. IDV (4) is a fault case where a step change deviation occurs at the inlet water temperature of the reactor cooling water. Meanwhile, although the fault in the inlet water temperature is not represented by a specific variable, the proposed method successfully points out variable 49, 'reactor cooling water flow valve', as the most relevant process variable of the actual root cause. The diagnosis result is reasonable considering the reactor temperature controller, the 16th control loop in the control strategy of the TEP model [41], in which the temperature of the reactor is controlled by the reactor cooling water flow used as a manipulated variable. Although the causality measure has a value less than the threshold and is not presented as a result of the diagnosis algorithm, considering that the 21st and 9th variables captured by the relatively high causality measure are 'reactor cooling water outlet temperature' and 'reactor temperature', respectively, the proposed methodology could pinpoint the root cause variable of IDV (4).

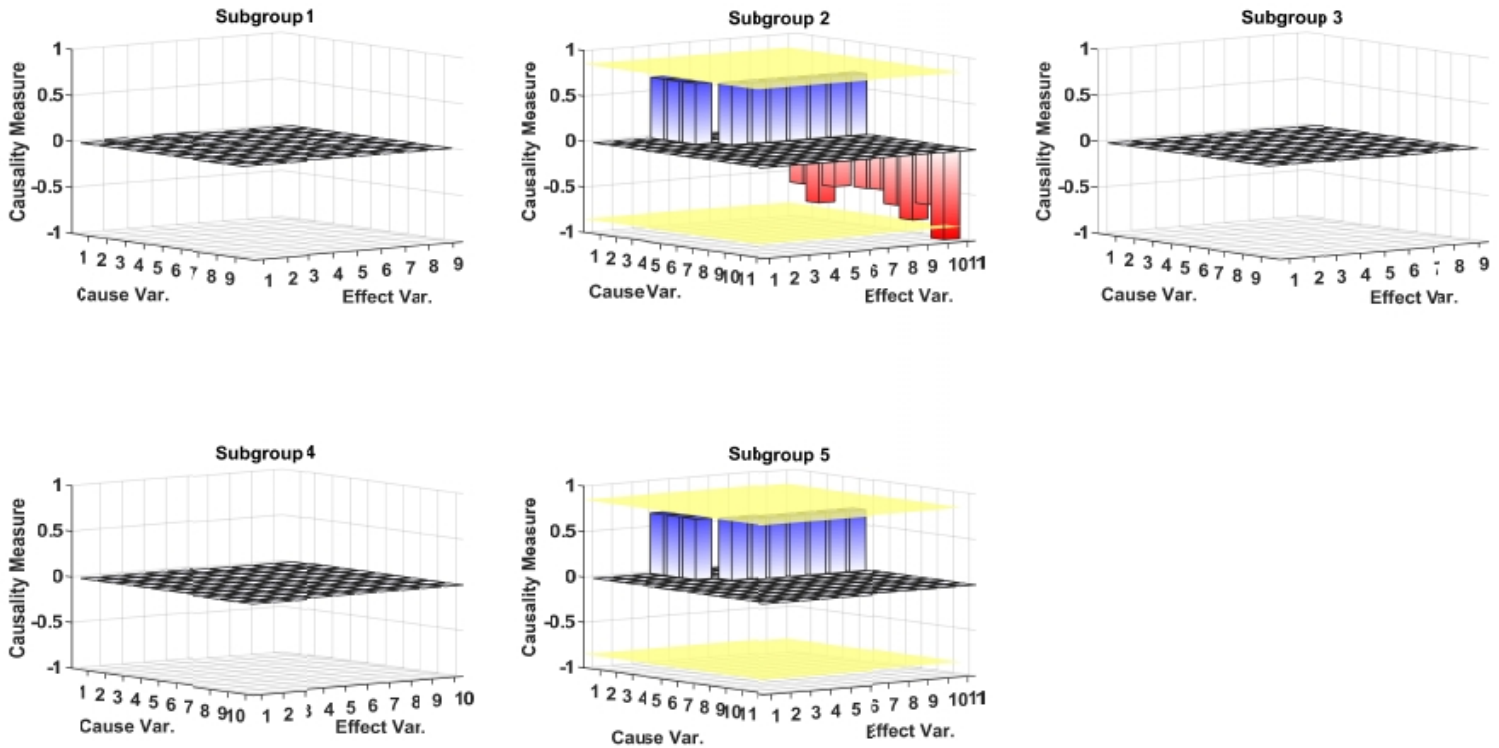

Figure 4.5. Causality measure plot of IDV (4) in each subgroup.

Table 4.8. Application results and causal analysis of IDV(4).

	Process variables in each group	**Calculation time (sec)**	**Causal relationship (causality measure*)**
Subgroup 1	1,7,10,13,16,42,43,44,46	5,903	-
Subgroup 2	11,18,20,21,22,29,35,38,47,48,49	9,996	48 → **49** (-1)
Subgroup 3	2,3,4,24,28,31,34,40,45	6,579	-
Subgroup 4	23,25,26,30,32,33,36,37,39,41	8,062	-
Subgroup 5	5,6,8,9,12,14,15,17,19,27,50	10,027	-

* Negative value means a causal relationship in the opposite direction.

* Relative values of causality measure are represented in parenthesis.

Not only in the case of IDV (4) but also for most disturbance cases of TEP, process variables do not include the direct root cause variables of faults. Since not all the variables in the process are measured, it is not possible to present the variables that are precisely the cause, particularly the fault cases. Moreover, unlike the SCR system, which is an open-loop system, TEP is a closed-loop simulation that reflects the control strategy to satisfy the process specifications such as production rate, inventory control, and hard constraints of the process. It changes the process to alleviate the effect of disturbances, which makes it more difficult to identify the exact root cause variable. It is also a limitation that the faults can be propagated back to the upstream as well as the downstream of the process due to the presence of recycle streams. In contrast to the previous studies, which reported the diagnosis results by analyzing only some intentionally selected variables, it is worth noting that the proposed approach can be applied to an entire process without specifying candidate variables in general.

Next, the diagnosis results of IDV(11) obtained by the proposed methodology to verify the performance of random variation disturbances rather than step changes are presented in Figure 4.6 and

Table 4.9. The location of the root cause in IDV (11) is the same as that in IDV (4), but there is a difference in the type of the disturbance which is a random variation not a step change as IDV (4). Although there are some causal relationships monitored from the proposed method in the first subgroup, the prominent root cause variable with the largest relative causality measure was presented in the second subgroup, variable 49. In the first subgroup, the variables detected as a relationship exceeding the threshold value are 'D feed flowrate valve' and 'stripper pressure', indicated as variables 42 and 16, respectively. As a result of a disturbance in the reactor cooling water, the selectivity

of the four reactions is upset, which affects the product quality. Consequently, according to the quality control strategy using feed rates, the D feed flow valve, which has the most availability among the feed streams, is primarily concerned. These changes caused by the shift of reaction rate result in composition changes of the unconverted reactants and products, affecting the pressure downstream of the reactor. It can be presumed that the stripper pressure is sensed as a consequence considering the fact that the stripper has no mechanism controlling the pressure, unlike a separator that can indirectly control the pressure with a temperature controller. For this reason, it can be interpreted that the stripper pressure is presented as an additional candidate for the indirect root cause.

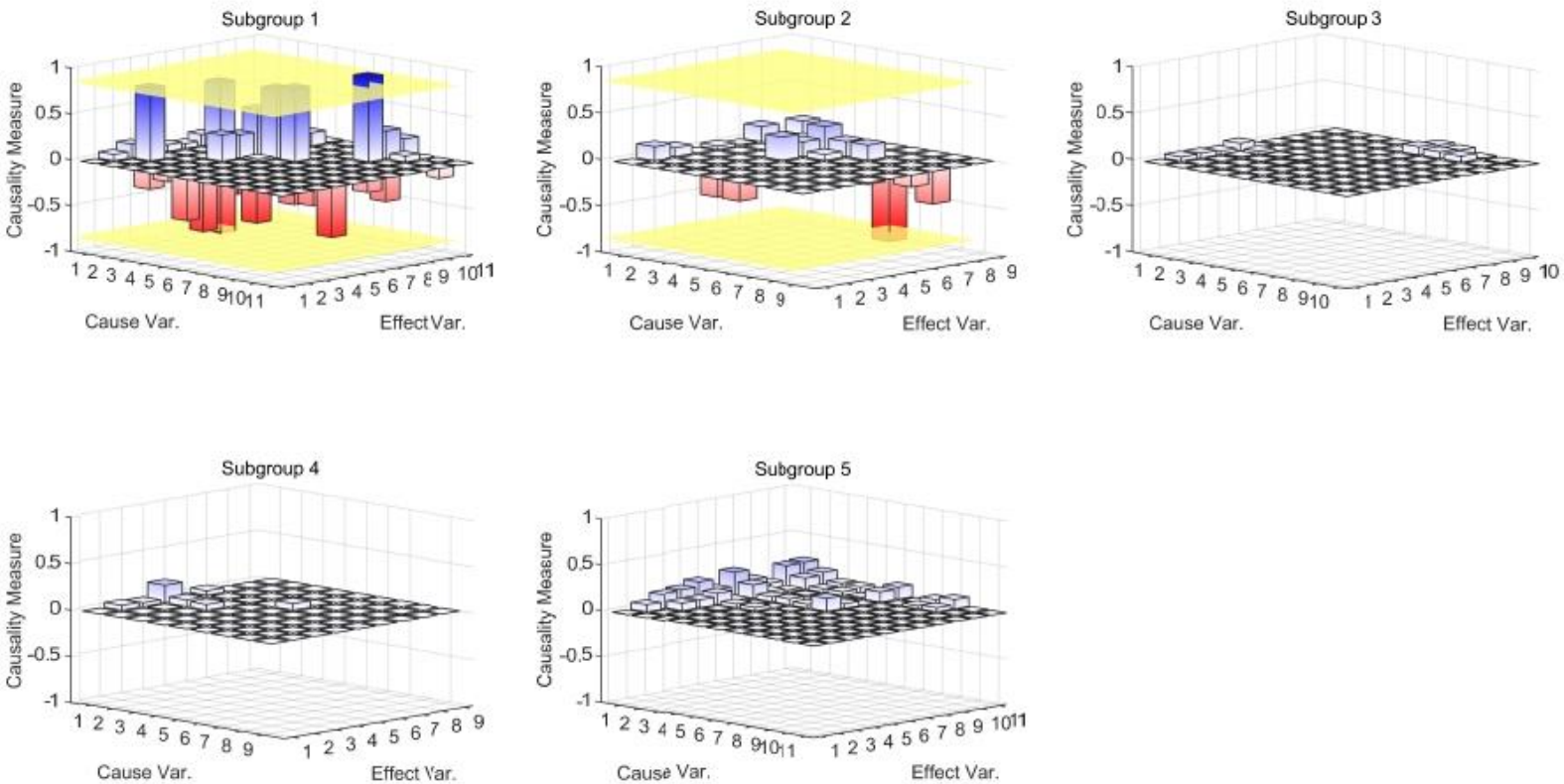

Figure 4.6. Causality measure plot of IDV (11) in each subgroup.

Table 4.9. Application results and causal analysis of IDV(11).

	Process variables in each group	**Calculation time (sec)**	**Causal relationship (causality measure*)**
Subgroup 1	1,7,9,10,13,16,21,42,43,44,46	5,391	**42** → 43 (0.923) 9 → **16** (-0.855)
Subgroup 2	11,12,14,18,20,22,35,47,49	3,514	20 → **49** (-1)
Subgroup 3	3,4,28,30,34,39,40,41,45,48	4,397	-
Subgroup 4	15,23,25,29,31,33,36,37,38	3,516	-
Subgroup 5	2,5,6,8,17,19,24,26,27,32,50	5,373	-

* Negative value means a causal relationship in the opposite direction.
* Relative values of causality measure are represented in parenthesis.

It is also worth noting that in some cases, the proposed method improved the diagnosis performance compared to the conventional method which directly applies the transfer entropy analysis to the entire process. The transfer entropy analysis method for the entire TEP process with 50 variables in the case of IDV (11) is shown in Figure 4.7. A comparison of the results of the proposed diagnosis method and the conventional transfer entropy based on the same threshold value is displayed in

Table **4.10**. In the conventional method, the most significant causality measure suggested the D feed flow valve, variable 42, as the most likely indirect root cause, followed by variable 46, although it was only slightly below the threshold. Hence, the diagnostic results derived from the proposed methodology outperformed the results from the analysis of the conventional transfer entropy because it represented a more directly related variable to the root cause of the fault. The proposed approach required only 19% of the computational cost compared to the conventional transfer entropy method. This advocates that the proposed method is more competitive than the conventional method in terms of the performance of the fault diagnosis and the computational cost.

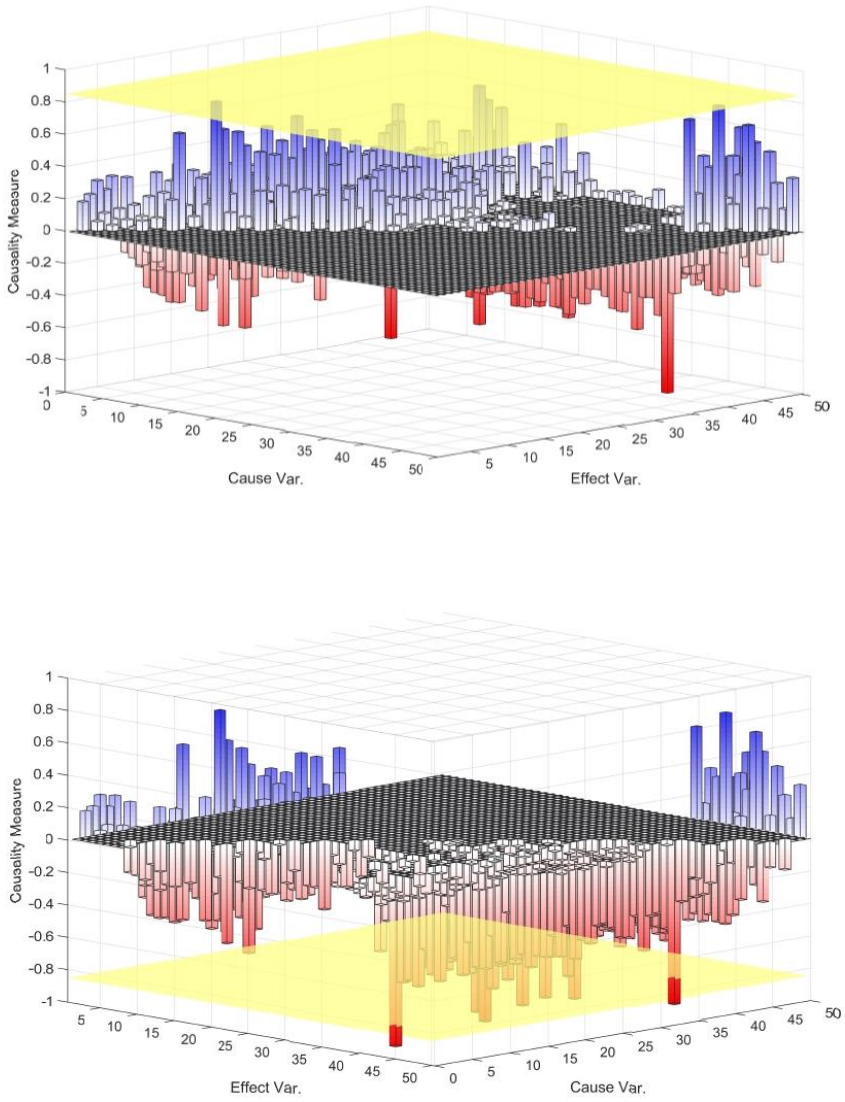

Figure 4.7. Causality measure plot of IDV(11) in the overall process.

(Transparent yellow planes indicate the threshold value of causality measure.)

Table 4.10. Causal analysis of IDV(11) in the proposed method and conventional transfer entropy.

	Calculation time (sec)		Causal relationship (causality measure*)
Proposed Method	**Subgroup 1**	5,391	20 → **49** (-1) **42** → 43 (0.923) 9 → **16** (-0.855)
	Subgroup 2	3,514	
	Subgroup 3	4,397	
	Subgroup 4	3,516	
	Subgroup 5	5,373	
	Total : 22,191		
Conventional Transfer Entropy Method	**119,727**		40 → **42** (-1) 3 → **42** (-0.966) 11 → **46** (-0.847)

* Negative value means a causal relationship in the opposite direction.
* Relative values of causality measure are represented in parenthesis.
* Bold numbers in a causal relationship indicate the diagnosed root cause variables.

To consider the case of multiple faults, the case of IDV (18) that was invoked simultaneously with IDV (5) was analyzed following the instruction of the TEP model. In this multiple fault case, because the random variation in the heat transfer of the condenser and the step deviation in the temperature of the condenser cooling water inlet occurs, the 'condenser cooling water outlet temperature', variable 22, is deemed as the variable most related to the root cause variable. In Figure 4.8, variable 22 in the second subgroup was detected as the highest priority for the root cause variable. Considering the comprehensive relationships of the last four causal relationships of

Table **4.11**, variable 9, 'reactor temperature', is another indirect root cause. This result implies that the root cause of the fault mentioned above is attributed to the temperature change of the reactor located at the front end. Likewise, since there is no process variable corresponding to the root cause, it is reasonable to attribute the root cause to the process variable of the device directly upstream. The fault in the root cause variable can, of course, affect the downstream of the process and also be propagated upstream by the recycle stream in TEP. Even though it does not thoroughly match the process flow diagram, the diagnosis result can elucidate the propagation of the faults. Given the relationship between variable 18, ‘stripper temperature’, and variable 20, ‘compressor work’ in

Table **4.11**, it can be inferred that the root cause variable, the upstream variables of variable 9, ‘reactor temperature’, and variable 22, ‘condenser cooling water outlet temperature’, can cause another downstream process fault.

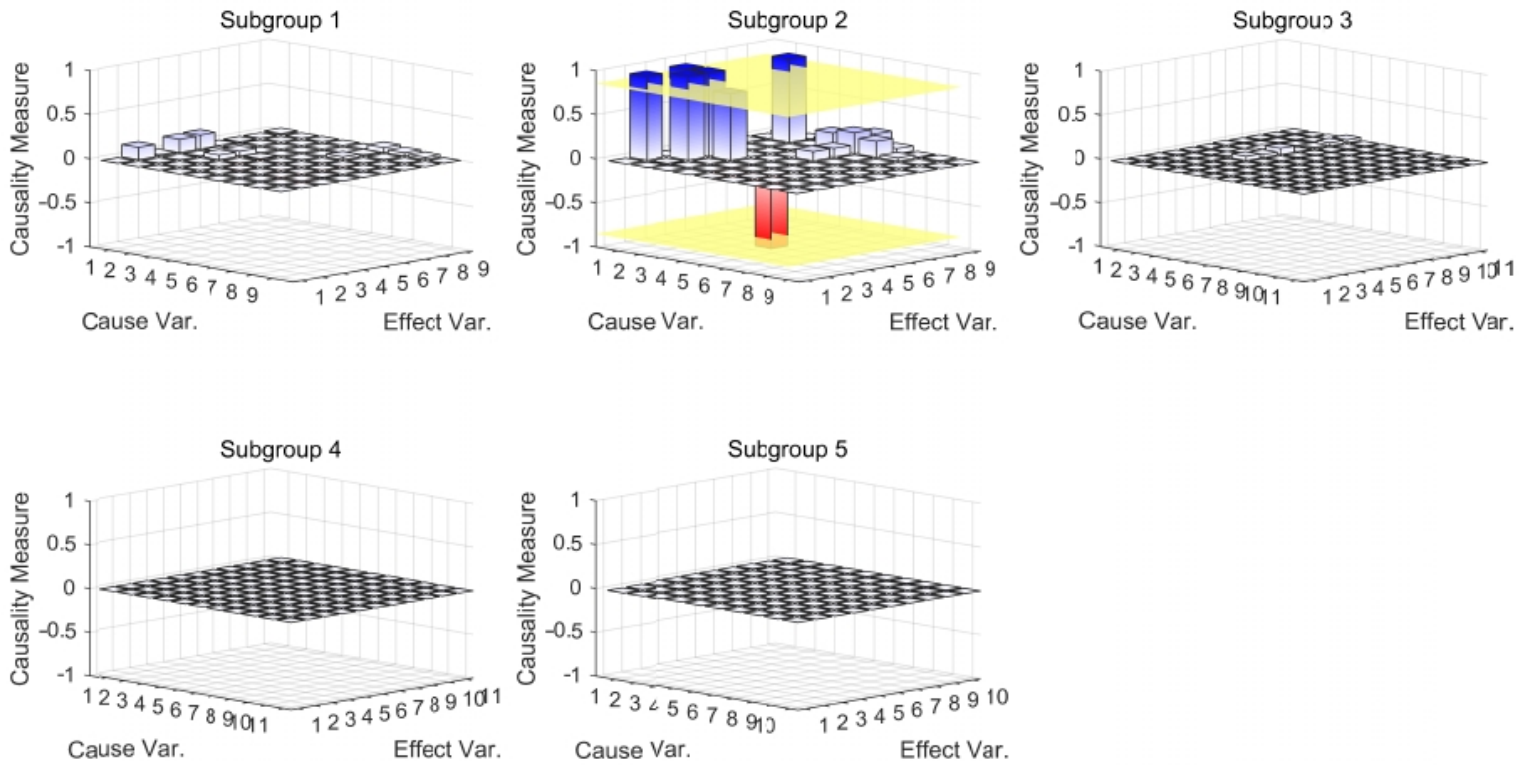

Figure 4.8. Causality measure plot of IDV (18) in each subgroup.

Table 4.11. Application results and causal analysis of IDV(18).

	Process variables in each group	**Calculation time (sec)**	**Causal relationship (causality measure*)**
Subgroup 1	1,7,10,11,13,16,42,43,44,46	7,794	-
Subgroup 2	9,18,20,21,22,47,48,49,50	6,181	21 → **22** (-1) **9** → 18 (0.947) **18** → 20 (0.947) **18** → 21 (0.947) **9** → 21 (0.947)
Subgroup 3	3,4,24,28,30,31,34,35,36,40,45	9,458	-
Subgroup 4	12,14,25,27,29,32,33,37,38,39,41	9,459	-
Subgroup 5	2,5,6,8,15,17,19,23,26	6,175	-

* Negative value means a causal relationship in the opposite direction.

* Relative values of causality measure are represented in parenthesis.

As a representative case where the actual root cause variable is included within the monitored process variables, IDV (23) was tested. The result of the analysis is shown in Figure 4.9, and detailed information of each subgroup is provided in

Table 4.12. IDV (23) is a random variation type fault in the pressure of the feed A streamline, and the effect appears in the form of a random variation of the feed A flowrate. To control the composition of the reactor feed stream, a ratio controller is installed to manipulate the flow valves of the A feed stream (stream 1) and the A & C feed stream (stream 4) to compensate for the influence of the fault. The A feed flow valve, variable 44 in the second subgroup, was correctly diagnosed having the highest causality measure value, presented in Figure 4.9. As the next priority, variable 46, the 'purge flow valve position', was captured as another cause variable candidate in the first subgroup. It can be understood that the manipulated variable of the pressure controller controlling the change in reactor pressure is captured as the feed composition changes.

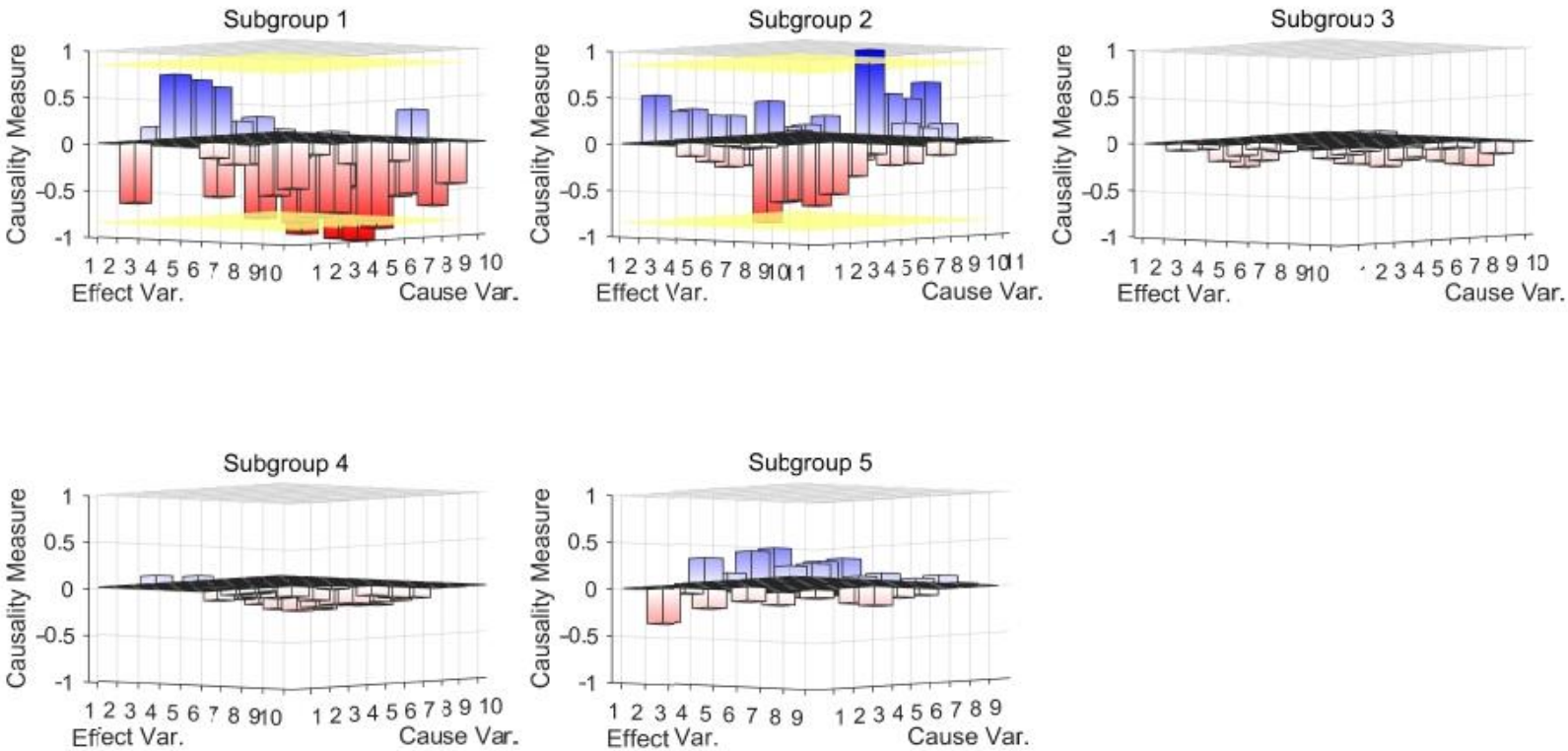

Figure 4.9. Causality measure plot of IDV (23) in each subgroup.

Table 4.12. Application results and causal analysis of IDV (23).

	Process variables in each group	**Calculation time (sec)**	**Causal relationship (causality measure*)**
Subgroup 1	7,10,11,13,16,18,40,42,43,46	7,349	13 → **46** (-0.993) 11 → **46** (-0.962) 7 → **46** (-0.887) 16 → **46** (-0.877)
Subgroup 2	1,4,20,21,22,30,44,45,47,48,49	8,975	**44** → 45 (1)
Subgroup 3	2,3,23,24,27,28,29,31,34,37	7,343	-
Subgroup 4	14,25,26,32,33,35,36,38,39,41	7,342	-
Subgroup 5	5,6,8,9,12,15,17,19,50	5,873	-

* Negative value means a causal relationship in the opposite direction.
* Relative values of causality measure are represented in parenthesis.

Similar to IDV (4), (11), (18), and (23), the rest of the fault cases were diagnosed using the proposed method. The entire diagnostic results for 28 faults are presented in Table 4.13. The second column is the diagnosis result obtained from the proposed method, and the third column shows the analysis result of each fault. In some cases, the root cause is not captured by the transfer entropy analysis. One of the reasons for the difficulty in using the transfer entropy for particular examples is that the entire process is affected by a complex control structure to maximize the production rate while guaranteeing the inventory management of the process units. In this case, the knowledge of historical data of root cause variables cannot significantly decrease the uncertainty of the prediction of the effect variables, which means a negligible amount of information is transferred. So, the root cause of the faults cannot be precisely diagnosed. Therefore, these fault cases need to be diagnosed by reflecting the control logic included in the TEP.

In the case of IDV (3), (9), (21), and (22), although those are all of the disturbances in feed temperature, the feed flow valves were consistently suggested as the result of the analysis. This corresponded to the product quality controller and the change in reaction rate induced by the temperature change as described above. In IDV (13), as the reactor pressure changes due to the drift in the reaction kinetics, the result proposed the cancellation of the purge valve and flow rate, which are manipulated variables of the reactor pressure controller. The root cause of the IDV (20) was not revealed even in the revised TEP model, but the results of the fault diagnosis using the proposed methodology are included. Among all cases, it was difficult to find the root causes of the fault cases associated with the condenser. The analysis is expected to be challenging because the condenser variables are associated with the secondary controller of the cascade control loop for the reactor level. Besides, even when a disturbance occurs in the feed composition, the root cause cannot be deduced by the diagnostic

method because the analyzer was inappropriately positioned. Nonetheless, a thorough analysis of the results of the proposed methodology would provide insight to the process engineer to investigate the source of the fault.

Table 4.13. Fault diagnosis result and analysis for the whole cases in the TEP.

No.	Diagnosis result of the proposed methodology	Analysis of the result	Remarks
IDV(1)	Subgroup 3 (38, 47)	Feed composition disturbance	
IDV(2)	Subgroup 1 (10, 46)	Purge rate due to the change of inert gas (‘B’ component)	
IDV(3)	Subgroup 1 (42, 44, 46)	Feed valves due to the feed temperature disturbance	
IDV(4)	Subgroup 2 (49)	Reactor cooling water valve due to the change of the inlet temperature disturbance	
IDV(5)	Subgroup 2 (18)	Temperature disturbance due to the disturbance in condenser	
IDV(6)	Subgroup 5 (1)	‘A’ feed flow rate due to the loss	
IDV(7)	Subgroup 1 (42)	Feed valve due to reduction of availability	
IDV(8)	Subgroup 1 (1, 44)	‘A’ feed flow rate due to feed composition disturbance	
IDV(9)	Subgroup 1 (46) Subgroup 2 (18)	Feed temperature disturbance	
IDV(10)	Subgroup 1 (18)	Stripper temperature due to the feed temperature disturbance	
IDV(11)	Subgroup 2 (49) Subgroup 1 (16, 42)	Reactor cooling water flow valve due to the disturbance in the inlet temperature	
IDV(12)	Subgroup 1 (1, 44)	Temperature disturbance	
IDV(13)	Subgroup 1 (10, 46)	Purge flow and valve due to the reaction kinetics drift	
IDV(14)	Subgroup 1 (9, 49)	Reactor temperature and cooling water valve due to the valve sticking	IDV(4)
IDV(15)	Subgroup 1 (46) Subgroup 2 (18)	Temperature disturbance	IDV(5)
IDV(16)	Subgroup 1 (18)	Stripper temperature due to the deviation of heat transfer and temperature disturbance	IDV(10)
IDV(17)	Subgroup 2 (21)	Reactor cooling water outlet temperature due to the temperature disturbances	IDV(11)
IDV(18)	Subgroup 2 (18, 22)	Temperature disturbance in condenser	IDV(5)

IDV(19)	Subgroup 3 (47)	Underflow of separator due to the valve stiction	
IDV(20)	Subgroup 2 (20) Subgroup 1 (13, 16)	Compressor work due to the pressure disturbance in separator	
IDV(21)	Subgroup 1 (42, 44)	Feed temperature disturbance	
IDV(22)	Subgroup 1 (42, 46)	Feed temperature disturbance	
IDV(23)	Subgroup 2 (44) Subgroup 1 (46)	'A' feed flow valve due to the pressure disturbance in feed stream	
IDV(24)	Subgroup 2 (42, 18)	'D' feed flow valve due to the pressure disturbance in feed stream	
IDV(25)	Subgroup 1 (43)	'E' feed flow valve due to the pressure disturbance in feed stream	
IDV(26)	Subgroup 2 (45)	'A&C' feed flow valve due to the pressure disturbance in feed stream	
IDV(27)	Subgroup 1 (42) Subgroup 2 (49)	Reactor cooling water flow valve due to the pressure disturbance in cooling water stream	
IDV(28)	Subgroup 1 (46) Subgroup 2 (18)	Temperature disturbance	

*Remarks indicate common disturbance used with the existing one.

Chapter 5

Concluding Remarks

5.1. Summary of the Contributions

This thesis focused on the improvement of the existing FDI systems. In particular, different approaches that can improve performance from viewpoints that have not been noted in traditional studies for process monitoring were suggested in each part of the process monitoring system. In addition, it can relieve the computational burden significantly, leading to the extension of applicability on large-scale processes.

The previous studies for fault detection systems mainly focused on the innovation in the model structure to better characterize the given data while keeping the given training data intact. Meanwhile, in the first part of the thesis, the given training data was targeted as the object to be improved through the data augmentation scheme. Data augmentation has been extensively studied to alleviate the class imbalance problem, especially for classification problems where there exists an imbalance between classes in the training data. Under the situation, data augmentation to balance the amount of data in each class allows utilizing all of the training data in a balanced manner, which can provide a favorable condition for classifier modeling. On the other hand, data augmentation has been employed in the proposed method to alleviate the within-class imbalance where the normal samples presenting the borderline of normal and abnormal state are relatively sparse compared to the typical normal sample distributed in the center of the normal manifold. Given that the modeling of the fault detection system corresponds to the manifold learning of the normal state, it can be expected that the supplement of the samples on the boundary of the normal state would positively affect the training process.

In this context, a monitoring framework was proposed that integrates manifold learning with data augmentation to supplement insufficient information for training. The main idea is to augment the synthetic data into the original training data using a generative model, Info-VAE, to supplement the training data for the construction of the fault detection system using AE. The synthetic data for augmentation is aimed at the region of the boundary of the normal training data, which contains infrequent but informative samples. The sample vectors for the augmentation were designed into several groups with different characteristics from center to boundary based on the latent space. At this moment, it should be noted that the total amount of the augmented samples on the boundary section needs to be carefully adjusted to avoid skewness of the original characteristics of the normal state. Moreover, the relative weights of the augmentation groups need to be considered as one of the critical factors that affect the performance of the proposed method. The effectiveness of the proposed method was verified by the improvement of the monitoring performance through the demonstration of the benchmark process. The analysis results showed that the fault detection accuracy was improved for most fault cases in the feature space in accord with the intention of the data augmentation, and the fault detection delay was also reduced.

In another work for fault isolation, a methodology was proposed in which the transfer entropy, the information-theoretic measure for causal analysis, was integrated with the graphical lasso to mitigate the downside of the costly causal evaluation. By sorting out the most relevant relationship to be analyzed in fault diagnosis among the whole relations, the graphical lasso reduces the load from the unnecessary causal analysis. The proposed method first divides the entire set of process variables into several related subgroups by iteratively applying graphical lasso to the remaining parts in each step, and then performs the root

cause analysis based on the subsets using transfer entropy. The root cause variable can be isolated if the causality measure exceeds the predefined threshold based on the relative magnitude. It should be noted that the proposed method inevitably possesses a tradeoff between the benefit from the sparsity in the correlations and the fidelity of the diagnostic performance. Therefore, the hyperparameters related to the graphical lasso are needed to be carefully determined depending on the applications. Nevertheless, the applicability of transfer entropy analysis for industrial-scale processes could be extended by significantly reducing the computational cost of transfer entropy which has been the most critical restriction. Furthermore, it is noteworthy that the diagnostic performance of the proposed method was superior to that of the conventional method in some fault cases.

5.2. Future Work

Although the proposed methodologies were able to mitigate the practical issues and improve the performance, there remain various future works for further improvement. Several issues can be suggested with respect to each part as follows.

First of all, the generative model can be further investigated to improve the fidelity of the synthetic sample for data augmentation. As the hybrid of VAE and GAN, adversarial autoencoder(AAE) [60] was proposed, which replaces the KL divergence penalizing the encoding distribution to fit the prior distribution with the discriminative network. By the modification, the assumption that the encoding posterior $q_\phi(z|x)$ should be multivariate Gaussian is no longer constrained, thus allowing the arbitrary distribution for the latent vector z. As AAE retains the structure of VAE that can fit the data distribution in the latent space to certain distribution, selective sampling and generation such as the boundary region of the data distribution still can be achieved.

Regarding the study of the fault detection systems, the model structure could be altered to a more sophisticated type. Recurrent neural networks (RNN), for example, can be effectively applied to improve the performance of the fault detection system considering the dynamics of the process data. This can be also a valuable approach for generative modeling as well as fault detection system modeling in terms of being able to obtain more reliable samples. Even though the hyperparameters such as the ratio of the augmented samples and the total amount of them were tuned by case studies, they could be determined optimally by formulating an optimization problem, where the objective function is the performance index and the decision variables are set to those parameters in data augmentation process.

There are also some issues that can be further investigated to boost the performance of the proposed fault isolation method. First of all, the most critical parts of the transfer entropy measure, the hyperparameters such as the prediction horizon and the embedding dimensions of each variable, could be optimized. Considering that the isolation process based on the conventional way for the entire 50 variables with the current hyperparameters takes more than 24 hours, it is practically impossible to perform comprehensive sensitivity analysis about other various settings of hyperparameters for more demanding conditions. Thanks to the advantage of saving the computational cost by the regularization process in the proposed method, there is room for an iterative investigation to find the optimal settings depending on the applications. In addition, another concern in the fault diagnosis, the fault propagation path, could be analyzed based on the subgroups in the context of the proposed method. Because the variables within a subgroup are closely related in terms of the physical principle, process control scheme, and layout of the process, the analysis between the subgroups could suggest meaningful information about the fault propagation path which is valuable evidence for the following process recovery.

Bibliography

[1] L. H. Chiang, R. D. Braatz, and E. L. Russell, *Fault Detection and Diagnosis in Industrial Systems*, Springer Science & Business Media. 2005.

[2] J. Lee, C. Yoo, S. Wook, P. A. Vanrolleghem, and I. Lee, "Nonlinear process monitoring using kernel principal component analysis," vol. 59, pp. 223–234, 2004.

[3] J. A. Krogh, A., & Hertz, "A Simple Weight Decay Can Improve Generalization," *Adv. Neural Inf. Process. Syst.*, pp. 950–957, 1992.

[4] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[5] C. Ioffe, S., & Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *Int. Conf. Mach. Learn.*, pp. 448–456, 2015.

[6] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 2015-January, pp. 1135–1143, 2015.

[7] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural Networks*, vol. 2, no. 1, pp. 53–58, 1989.

[8] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, 2010.

[9] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," *Proc. 28th Int. Conf. Mach. Learn. ICML 2011*, no. 1, pp. 833–840, 2011.

[10] C. Mellon and U. C. Berkeley, "Tutorial on Variational Autoencoders," *arXiv Prepr. arXiv1606.05908.*, pp. 1–23, 2016.

[11] A. J. Holden *et al.*, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. July, pp. 504–507, 2006.

[12] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, no. 2, pp. 461–464, 2000.

[13] S. Kullback, *Information theory and statistics*, Courier Corporation, 1997.

[14] M. Wibral *et al.*, “Measuring Information-Transfer Delays,” *PLoS One*, vol. 7, no. 4, pp. 1–10, 2012.

[15] M. Bauer, J. W. Cox, M. H. Caveness, J. J. Downs, and N. F. Thornhill, “Finding the direction of disturbance propagation in a chemical process using transfer entropy,” *IEEE Trans. Control Syst. Technol.*, vol. 15, no. 1, pp. 12–21, 2007.

[16] LLdiko E. Frank and J. H. Friedman, “A Statistical View of Some Chemometrics Regression Tools,” *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.

[17] R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso,” *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[18] J. Friedman, T. Hastie, and R. Tibshirani, “Sparse inverse covariance estimation with the graphical lasso,” *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[19] J. Dahl, V. Roychowdhury, and L. Vandenberghe, “Maximum likelihood estimation of Gaussian graphical models : Numerical implementation and topology selection,” *Preprint*, pp. 1–29, 2005.

[20] N. Meinshausen and P. Bühlmann, “High-dimensional graphs and variable selection with the Lasso,” *Ann. Stat.*, vol. 34, no. 3, pp. 1436–1462, 2006.

[21] O. Banerjee, L. El Ghaoui, and A. D’Aspremont, “Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data,” *J. Mach. Learn. Res.*, vol. 9, pp. 485–516, 2008.

[22] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, “A review of process fault detection and diagnosis part I: Quantitative model-based methods,” *Comput. Chem. Eng.*, vol. 27, no. 3, pp. 293–311, 2003.

[23] W. Yan, P. Guo, and Z. Li, “Chemometrics and Intelligent Laboratory Systems Nonlinear and robust statistical process monitoring based on variant autoencoders,” *Chemom. Intell. Lab. Syst.*, vol. 158, pp. 31–40, 2016.

[24] F. Lv, C. Wen, Z. Bao, and M. Liu, “Fault Diagnosis Based on Deep Learning,” *2016 Am. Control Conf.*, no. 2, pp. 6851–6856, 2016.

[25] J. Fan and W. Wang, “AutoEncoder based High-Dimensional Data Fault Detection System,” *IEEE*, pp. 1001–1006, 2017.

[26] L. Jiang, Z. Song, Z. Ge, and J. Chen, “Robust Self-Supervised Model and Its Application for Fault Detection,” *Ind. Eng. Chem. Res.*, 2017.

[27] Z. Zhang, T. Jiang, S. Li, and Y. Yang, “Automated feature learning

for nonlinear process monitoring – An approach using stacked denoising autoencoder and k-nearest neighbor rule," *J. Process Control*, vol. 64, pp. 49–61, 2018.

[28] W. Yu and C. Zhao, "Robust Monitoring and Fault Isolation of Nonlinear Industrial Processes Using Denoising Autoencoder and Elastic Net," *IEEE Trans. Control Syst. Technol.*, vol. PP, pp. 1–9, 2019.

[29] H. Zhao, "Neural component analysis for fault detection," *Chemom. Intell. Lab. Syst.*, vol. 176, pp. 11–21, 2018.

[30] S. C. Wong and M. D. Mcdonnell, "Understanding data augmentation for classification : when to warp ?," *2016 Int. Conf. Digit. Image Comput. Tech. Appl.*, pp. 1–6, 2016.

[31] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Lecture Notes in Computer Science*, 2005, vol. 3644, pp. 878–887.

[32] T. Devries and G. W. Taylor, "Dataset Augmentation in Feature Space," *arXiv Prepr. arXiv1702.05538*, pp. 1–12, 2017.

[33] Z. Wan, Y. Zhang, and H. He, "Variational Autoencoder Based Synthetic Data Generation for Imbalanced Learning," *IEEE*, pp. 1–7, 2017.

[34] N. Etworks, A. Storkey, and H. Edwards, "Data Augmentation Generative Adversarial Networks," *arXiv Prepr. arXiv1711.04340*, pp. 1–14, 2017.

[35] J. Jorge, R. Paredes, J. A. Sanchez, and M. Bened, "Empirical Evaluation of Variational Autoencoders for Data Augmentation," *VISIGRAPP (5 VISAPP)*, vol. 5, pp. 96–104, 2018.

[36] W.-N. Hsu, Yu Zhang, and J. Glass, "Unsupervised Domain Adaptation for Robust Speech Recognition via Variational Autoencoder-based Data Augmentation," *IEEE Autom. Speech Recognit. Underst. Work.*, no. 1, pp. 16–23, 2017.

[37] X. Gao, F. Deng, and X. Yue, "Data augmentation in fault diagnosis based on the Wasserstein," *Neurocomputing*, vol. 396, pp. 487–494, 2019.

[38] S. K. Lim, Y. Loo, N. Tran, N. Cheung, G. Roig, and Y. Elovici, "DOPING : Generative Data Augmentation for Unsupervised Anomaly Detection with GAN," *2018 IEEE Int. Conf. Data Min.*, pp. 1122–1127, 2018.

[39] S. Zhao, J. Song, and S. Ermon, "InfoVAE : Balancing Learning and Inference in Variational Autoencoders Two Problems of Variational Autoencoders," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp.

5885–5892, 2019.

[40] J. J. Downs and E. C. Company, “A plant-wide industrial process control problem,” *Comput. Chem. Eng.*, vol. 17, no. 3, pp. 245–255, 1993.

[41] A. Bathelt, N. L. Ricker, and M. Jelali, “Revision of the Tennessee Eastman Process,” *IFAC-PapersOnLine*, vol. 48, no. 8, pp. 309–314, 2014.

[42] N. L. Ricker, “Decentralized control of the Tennessee Eastman Challenge Process,” *J. Process Control*, vol. 6, no. 4, pp. 205–221, 1996.

[43] H. Lee, C. Kim, S. Lim, and J. Min, “Data-driven fault diagnosis for chemical processes using transfer entropy and graphical lasso,” *Comput. Chem. Eng.*, vol. 142, p. 107064, 2020.

[44] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, “A review of process fault detection and diagnosis part III: Process history based methods,” *Comput. Chem. Eng.*, vol. 27, no. 3, pp. 327–346, 2003.

[45] R. T. Samuel and Y. Cao, “Nonlinear process fault detection and identification using kernel PCA and kernel density estimation,” *Syst. Sci. Control Eng.*, vol. 4, no. 1, pp. 165–174, 2016.

[46] D. L. Olson and D. Delen, *Advanced data mining techniques*, no. December 2013. Springer Science & Business, 2008.

[47] C. Kim, H. Lee, K. Kim, Y. Lee, and W. B. Lee, “Efficient Process Monitoring via the Integrated Use of Markov Random Fields Learning and the Graphical Lasso,” *Ind. Eng. Chem. Res.*, vol. 57, no. 39, pp. 13144–13155, 2018.

[48] T. Yang, Fan and Duan, Ping and Shah, Sirish L and Chen, *Capturing connectivity and causality in complex industrial processes*. Springer Science & Business Media, 2014.

[49] H. Chen, B. Jiang, and N. Lu, “A Multi-mode Incipient Sensor Fault Detection and Diagnosis Method for Electrical Traction Systems,” *Int. J. Control. Autom. Syst.*, vol. 16, no. 4, pp. 1783–1793, 2018.

[50] I. Gueddi, O. Nasri, K. Benothman, and P. Dague, “Fault Detection and Isolation of spacecraft thrusters using an extended principal component analysis to interval data,” *Int. J. Control. Autom. Syst.*, vol. 15, no. 2, pp. 776–789, 2017.

[51] C. J. W. Granger, “Investigating Causal Relations by Econometric Models and Cross-spectral Methods,” *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[52] G. Li, S. J. Qin, and T. Yuan, “Data-driven root cause diagnosis of faults in process industries,” *Chemom. Intell. Lab. Syst.*, vol. 159, no. July, pp. 1–11, 2016.

[53] M. Bauer, N. F. Thornhill, and A. Meaburn, “Specifying the directionality of fault propagation paths using transfer entropy,” *IFAC Proc. Vol.*, vol. 37, no. 9, pp. 203–208, 2004.

[54] M. Bauer and N. F. Thornhill, “A practical method for identifying the propagation path of plant-wide disturbances,” *J. Process Control*, vol. 18, no. 7–8, pp. 707–719, 2008.

[55] B. Lindner, L. Auret, and M. Bauer, “Investigating the Impact of Perturbations in Chemical Processes on Data-Based Causality Analysis. Part 1: Defining Desired Performance of Causality Analysis Techniques,” *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 3269–3274, 2017.

[56] R. Isermann, *Fault-diagnosis systems: an introduction from fault detection to fault tolerance*. Springer Science & Business Media, 2005.

[57] B. Lindner, L. Auret, and M. Bauer, “Investigating the Impact of Perturbations in Chemical Processes on Data-Based Causality Analysis. Part 2: Testing Granger Causality and Transfer Entropy,” *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 3275–3280, 2017.

[58] C. Depcik, D. Assanis, and K. Bevan, “A one-dimensional lean NOx trap model with a global kinetic mechanism that includes NH3 and N2O,” *Int. J. Engine Res.*, vol. 9, no. 1, pp. 57–77, 2008.

[59] Y. Kim, T. Park, C. Jung, C. H. Kim, Y. W. Kim, and J. M. Lee, “Hybrid Nonlinear Model Predictive Control of LNT and Urealess SCR Aftertreatment System,” *IEEE Trans. Control Syst. Technol.*, vol. 27, no. 5, pp. 2305–2313, 2018.

[60] A. Makhzani, B. Frey, and I. Goodfellow, “Adversarial Autoencoders,” *arXiv Prepr. arXiv1511.05644*, 2014.

초 록

공정 모니터링 시스템은 효과적이고 안전한 공정 운전을 위한 필수적인 요소이다. 공정 이상은 목표 생성물의 품질에 영향을 주거나 공정의 정상 가동을 방해하여 생산성을 저해할 수 있다. 폭발성 및 인화성 물질을 주로 다루는 화학공정의 경우 공정 이상은 가장 중요한 요소인 공정의 안전을 위협하는 요소로 작용할 수 있다. 한편, 현대의 공정의 범위가 확장되고 자동화와 고도화가 진행됨에 따라 점점 더 신뢰도 높은 모니터링 시스템이 요구되고 있다.

공정 모니터링은 크게 세 단계로 구분될 수 있다. 실시간으로 공정의 이상 여부를 판단하는 공정 이상 감지, 다음으로 감지된 이상의 원인을 파악하는 이상 진단, 마지막으로 공정 이상의 원인을 제거하고 정상 상태로 회복시키는 복원으로 나뉘어진다. 특히 공정 이상 감지와 진단 시스템을 위해 다양한 방법론들이 제안되어왔으며, 그 방법론들은 크게 세 가지로 구분할 수 있다. 물리 이론을 기반으로 한 모델 분석 방법과 특정 분야의 경험 지식을 바탕으로 한 지식 기반 방법론에 비해 범용적인 적용 가능성과 현대 공정의 풍부한 공정 데이터가 제공되는 조건의 충족으로 인해 데이터 기반 방법론이 널리 활용되어지고 있다. 또한, 데이터 기반 공정 모니터링 방법론들은 공정의 규모와 복잡도가 증가함에 따라 그 장점이 더욱 극대화되는 특징을 갖는다. 본 연구에서는 기존의 데이터 기반 공정 모니터링 방법론들의 성능을 개선하기 위한 공정 이상 감지 방법론과 이상 진단 방법론을 제안한다.

전통적인 공정 이상 감지 시스템은 차원 축소방법들을 기반으로 개발되었다. 차원 축소를 기반으로 한 공정 이상 감지 모델은 공정 데이터에 내재되어 있는 특징으로 정의되는 저차원의 잠재 공간을 정의하고, 이를 기준으로 모니터링을 수행한다. 대표적인 방법으로는 전통적인 다변량 공정 모니터링 방법인 주 성분 분석과 머신 러닝 기법인

오토인코더가 있다. 최근 풍부한 학습 데이터와 우수한 성능 덕분에 다양한 머신 러닝 기법을 사용한 이상 감지 시스템이 널리 활용되고 있지만, 앞서 소개한 현대 공정의 다양한 특징으로 인해 더욱 향상된 성능의 모니터링 기법의 개발이 요구되어지고 있다. 이러한 데이터 기반 모니터링 시스템의 성능 향상을 위해서 모델의 구조를 변경하거나 모델의 학습 절차를 변형하는 접근법들이 주로 제안되었다. 하지만, 데이터 기반 방법론들은 궁극적으로 학습 데이터의 품질에 의존적이라는 특성은 여전히 남아있다. 즉, 학습 데이터의 부족한 정보를 보완함으로써 모니터링 시스템의 완성도를 높일 수 있는 방법론이 요구된다. 따라서, 본 연구는 첫 번째 주제로 데이터 증강 기법을 결합한 공정 이상 감지 방법론을 제안한다.

데이터 증강 기법은 여러 집합을 구분하는 분류기 모델링시에 특정 집합의 학습 데이터가 부족한 경우에 주로 활용되었다. 이러한 경우 데이터 증강을 통해 학습 데이터의 균형을 맞춤으로써 모델의 학습 효율을 증진시킬 수 있다. 반면에, 본 연구에서의 데이터 증강은 한 집합 내에서의 불균형을 완화하기 위한 목적으로 사용되었다. 정상 조건의 공정 데이터는 정상과 이상의 경계에 분포하는 데이터가 희박하게 존재하는 특징을 갖는다. 이상 감지 시스템이 정상 상태의 저차원 특징 공간을 학습하고, 이를 통해 정상과 이상을 구분하는 모델이라는 점을 고려하면 경계 영역의 데이터의 증강이 특징 공간 학습에 긍정적으로 작용할 것을 기대해 볼 수 있다. 이와 같은 맥락에서 제안된 방법론은 다음과 같다.

먼저, 기존의 학습 데이터를 이용하여 인공 데이터를 생성하기위한 생성모델인 변분 오토인코더를 학습한다. 생성 모델로 학습한 정상 운전 데이터의 저차원 분포의 경계영역에 해당하는 데이터들을 인공 데이터로 생성하여 학습데이터에 증강시킨다. 이렇게 증강된 학습 데이터를 기반으로 이상 감지 모델을 위한 머신 러닝 기반 차원 축소 방법인 오토인코더를 학습하여 이상 감지 시스템을 구축한다. 증강된 학습 데이터를 사용함으로써 오토인코더의 잠재 공간 학습이 더 효과적

으로 수행될 수 있고, 이는 곧 정상과 이상 상태를 구분하는 이상 감지 시스템의 성능 개선으로 이어질 수 있다.

차원 축소 기법은 전통적인 이상 진단 방법으로도 활용되었다. 하지만, 이는 차원 축소시의 정보의 손실로 인해 저조하고 일관성이 부족한 성능을 보였다. 전통적인 방법의 한계점을 개선하기 위해 공정 변수 간의 인과 관계를 직접적으로 분석하는 기법들이 개발되었다. 그 중 하나인 정보 이론 기반의 전달 엔트로피는 특정 모델이나 선형 가정을 기반으로 하지 않기 때문에 비선형 공정의 이상 진단에 대해 일반적으로 우수한 성능을 보인다고 알려져 있다. 하지만, 전달 엔트로피를 이용한 인과관계 분석 방법은 고비용의 밀도 추정을 필요로 한다는 단점으로 인해 소규모 공정에 대해서만 제한적으로 적용되어 왔다. 이러한 한계점을 개선하기 위한 방안으로 그래프 라쏘라는 조정 방법을 전달 엔트로피와 결합한 방법론을 제안하였다.

그래프 라쏘는 비 방향성 그래프 모델에서 성긴 구조를 학습하기 위한 방법론으로 전체 공정 그래프로부터 상관 관계가 높은 부분 그래프를 추출해낼 수 있다. 가장 높은 상관 관계를 갖는 부분 그래프와 독립된 나머지 변수들이 그래프 라쏘의 출력으로 제시되기 때문에, 나머지 변수들에 대한 반복적인 적용을 통해 전체 공정 변수들을 연관성이 높은 몇몇의 부분 그래프로 변환할 수 있다. 연관성이 낮은 관계를 사전에 배제함으로써 인과 관계 분석의 대상을 크게 축소할 수 있다. 즉, 이 단계를 통해 고비용의 전달 엔트로피의 한계점을 완화하고, 그 적용 가능성을 확장할 수 있도록 한다.

두 방법을 결합하여 다음과 같은 이상 진단 방법론을 제안하였다. 먼저, 공정 이상이 발생한 데이터를 대상으로 반복적 그래프 라쏘를 적용하여 전체 공정 변수들을 연관성이 높은 5개의 부분 집합으로 구분한다. 구분된 각각의 부분 집합을 대상으로 전달 엔트로피를 이용한 인과관계 척도를 계산하고, 가장 유력한 원인 변수를 판별해낸다. 즉, 그래프 라쏘를 통해 효과적으로 인과관계 분석의 대상을 축소함으로써 불필요한 전달 엔트로피 계산으로 발생하는 비용을 크게 절감할

수 있다. 따라서, 제안된 방법론은 대규모 산업 공정에 대해서도 전달 엔트로피를 이용한 이상 진단 기법의 적용을 가능하게 했다는 점에서 의의가 있다.

본 연구에서 제안된 방법론의 성능을 검증하기 위하여 산업 규모의 벤치마크 공정 모델인 테네시 이스트만 공정에 이를 적용하고 결과를 분석하였다. 벤치마크 공정 모델은 다수의 단위 공정을 포함하고, 재순환 흐름과 화학 반응을 포함하고 있어 실제 공정과 같은 복잡도를 갖는 공정 모델로서 제안한 방법론들의 성능을 시험해보기에 적합했다. 성능 테스트는 테네시 이스트만 공정 모델에 포함되어 있는 사전에 정의된 28개 종류의 공정 이상에 대하여 수행하였다. 제안한 데이터 증강을 접목한 공정 이상 감지 방법론은 기존 방법론 대비 높은 이상 감지율을 보였다. 일부의 경우 이상 감지 지연측면에서도 개선을 확인할 수 있었다. 또한, 이상 진단을 위해 전달 엔트로피와 그래프 라쏘를 결합한 제안한 방법론은 전체 공정에 전달 엔트로피를 직접 적용한 기존의 방법론 대비 약 20%의 계산 비용만으로도 효과적으로 이상의 원인을 파악해내는 것을 확인할 수 있었다. 또한, 성능 테스트 결과는 일부 공정 이상의 경우 제안한 방법론이 기존의 방법보다 더 정확한 이상 진단 결과를 제시할 수 있음을 보였다.

주요어: 공정 모니터링; 공정 이상 감지 및 진단; 오토인코더; 전달 엔트로피; 그래프 라쏘

학번: 2016-21049