Reliability Engineering and System Safety 256 (2025) 110653

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Unsupervised anomaly detection of machines operating under time-varying conditions: DCD-VAE enabled feature disentanglement of operating conditions and states

Haoxuan Zhou [a,b,d], Bingsen Wang [d,e], Enrico Zio [c,d], Zihao Lei [a,b], Guangrui Wen [a,b,*], Xuefeng Chen [a,b]

[a] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[b] State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[c] Centre de Recherche sur les Risques et les Crises (CRC), MINES ParisPSL University, Sophia Antipolis, France
[d] Energy Department, Politecnico di Milano, Via La Masa 34, Milano 20156, Italy
[e] Centre for Advances in Reliability and Safety (CAiRS), Hong Kong, SAR, China

ARTICLE INFO

Keywords:
Machines
Condition monitoring
Anomaly detection
Disentangled representation learning
Time-varying operating conditions

ABSTRACT

Anomaly detection (AD) plays a key role in condition monitoring (CM) to ensure the machine system's operating reliability and safety. When machinery operates under time-varying operating conditions (TVOCs), interference from varying operating conditions (OCs) exacerbates the difficulty of AD. To address this issue, a Disentangled Representation Learning(DRL) approach is proposed to dissociate the features linked with OCs and operating states (OSs). Expanding on the pre-existing Variational Autoencoder (VAE), Distribution Constraint Decomposition (DCD) is proposed as a regularization approach, which implements a loose-tight constraint depending on Kullback-Leibler(KL) divergence to enforce prior constraints on the latent features. As a result, DCD-VAE, which enables the selective allocation of different types of information, achieving disentanglement between OCs' information and the OSs' information, is proposed in this paper. An anomaly indicator(ANI) constructed based on the OSs features enables AD. Simulation and experiments validate the substantial advantage of the proposed approach over comparable methods, facilitating the timely and precise identification of mechanical faults.

List of Main Abbreviations

AD	Anomaly Detection
ANI	Anomaly Indicator
CM	Condition Monitoring
CVAE	Conditional Variational Autoencoder
DCD	Distribution Constraint Decomposition
DRL	Disentangled Representation Learning
DSMDA	Deep Sequence Multi-Distribution Adversarial
ELBO	Evidence Lower Bound
FDCVAE	Feature Disentanglement Conditional VAE
GI	Gini Index
HI	Health Indicator
KL	Kullback-Leibler
MI	Mutual Information
MPN	Multiple Pattern Normality
MRRAE	Memory Residual Regression Autoencoder
OCs	Operating Conditions
OSs	Operating States
PCA	Principal Component Analysis
PDF	Probability Density Function
RE	Reconstruction Error
RMS	Root Mean Square
ROC	Receiver Operating Characteristic
TVOCs	Time-Varying Operating Conditions
VAE	Variational Autoencoder

1. Introduction

Condition monitoring (CM) serves as a preliminary procedure for the

* Corresponding author.
E-mail addresses: hxzhou@stu.xjtu.edu.cn (H. Zhou), bingsen.wang@polimi.it (B. Wang), enrico.zio@polimi.it (E. Zio), zihaolei@stu.xjtu.edu.cn (Z. Lei), grwen@xjtu.edu.cn (G. Wen), chenxf@mail.xjtu.edu.cn (X. Chen).

https://doi.org/10.1016/j.ress.2024.110653
Received 22 April 2024; Received in revised form 8 October 2024; Accepted 12 November 2024
Available online 14 November 2024

health management of industrial machines, and monitoring the condition of machines with the use of monitoring information is of utmost importance in maintaining the reliability and safety of industrial processes [1–3]. CM involves regularly collecting and analyzing data from various sensors to identify changes in the equipment's behavior and detect potential issues before they escalate into major problems [4,5], and this is where anomaly detection(AD) comes into play. AD enables us to detect outliers and unusual patterns in the data by using advanced statistical techniques and machine learning algorithms [6,7]. The approaches applied in AD for machines can be categorized into two main groups, the first one can be considered the expert-knowledge-based method, based on the insight into the principles of anomalies, features related to anomalies can be extracted from monitoring data and used for anomaly identification [8–10]. The other group is the data-driven-based method, which is a technique used to automatically identify unusual patterns or outliers in data sets. Data-driven methods are unbiased as they rely solely on the data to detect anomalies, rather than on predefined rules or assumptions. This makes data-driven AD techniques adaptable and suitable for a wide range of applications in industrial processes [11,12].

The data-driven AD can also be classified into supervised and unsupervised categories. Supervised AD requires labeled data [13–15], where each data is classified as normal or anomalous. The model is trained on labeled data and learns to identify anomalies based on the patterns in the labeled data. However, obtaining labeled data can be challenging in industrial applications. On the contrary, unsupervised AD does not require labeled data and can detect anomalies based on deviations from normal behavior. It uses machine-learning techniques to identify patterns in the data and identify any deviations from the expected behavior [16]. As one of the most attractive technologies for AD, learning feature representation of normality has been realized by some exquisitely designed neural networks as deep learning flourishes. For instance, the autoencoder [17] and generative adversarial network [18] are widely utilized in industrial machines [19,20], such as bearing [21–23] and Maritime Components [24].

However, unsupervised methods can also produce miss alarms or false alarms when the bias is introduced into the feature representation due to the presence of multiple pattern normality (MPN) in the monitoring data that is utilized as training data. For instance in industry, the variation of operating conditions(OCs, which is related to characteristics such as speed, load, temperature, humidity, etc. that are either subjectively imposed by humans or objectively present in the environment) of a machine in the normal operating states (OSs, which are related to characteristics such as operational performance, state of health, and the presence of faults.) can cause the MPN to emerge in the monitoring data (which means healthy or normal machines can generate different kinds of normal monitoring data when they are operating under different OCs). In such cases, effective AD requires expert interpretation to determine the actual anomalies for each pattern normality (such as thresholds need to be determined independently for each OC). Classic AD methods that are based on the reconstruction error (RE) of the autoencoder are taken as an example to illustrate the MPN that existed in the AD, as shown in Fig. 1. The left side of the figure illustrates two scenarios of TVOCs that could lead to missed and false AD in traditional autoencoder-based AD, where OCs type A corresponds to missed AD, while OCs type B corresponds to false AD. When machines are operating under TVOCs, there are N kinds of patterns that represent the normality of machines under N different OCs(in type A, $N = 5$; in type B, $N = 2$) and are used as training data $\mathbf{x}$ to train the autoencoder, and the distribution of training samples vary due to the influence of different OCs, as illustrated by the probability density functions in Fig. 1, the RE is calculated based on the quantification of similarity of the original samples $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$ (as the reconstruction loss shown in the figure), which can force the autoencoder to fit the training data and thus the feature representation can be learned and represented by the latent feature $\mathbf{z}$. However, as the successful reconstruction of training samples relies on minimizing RE, the latent features inevitably preserve the distributional differences of each training sample caused by variations in OCs (as the feature visualization part shown in the figure). As a result, the RE for training samples belonging to different OCs varies and makes it impossible to provide a unified optimal anomaly threshold that can adapt to all various OCs(as reconstruction loss shown in the upper middle part of the figure), with the utilization of the RE as the ANI during the deployment stage [21,25]. Consequently, false positives (false alarms) and false negatives (miss alarms) are likely to occur in AD, significantly reducing the accuracy and reliability of AD. Furthermore, to the best of our knowledge, there is scant research that addresses the issues arising from TVOCs in the process of machine AD or proposes corresponding solutions.

Recent works [5,26] first proposed the conditional disentangled representation learning(DRL) [27–29] approaches to address this MPN challenge for AD of Machines, the OCs information is introduced to the DRL model by constructing a side OCs feature coding branches network, thus the feature disentanglement [30,31] is conditionally realized(i.e., the need for accurate OCs' information that is aligned with the monitoring signal). However, an obvious shortcoming of these methods is highlighted when OCs' information is not available. To address this drawback and enable the method to have a broader scope of application, it is important to propose a method that can be applied to machines AD considering the unavailability of OCs' information. Inspired by the variational autoencoder [32] (VAE) and its improvement β-VAE [33], this paper proposes an unsupervised AD method based on distribution

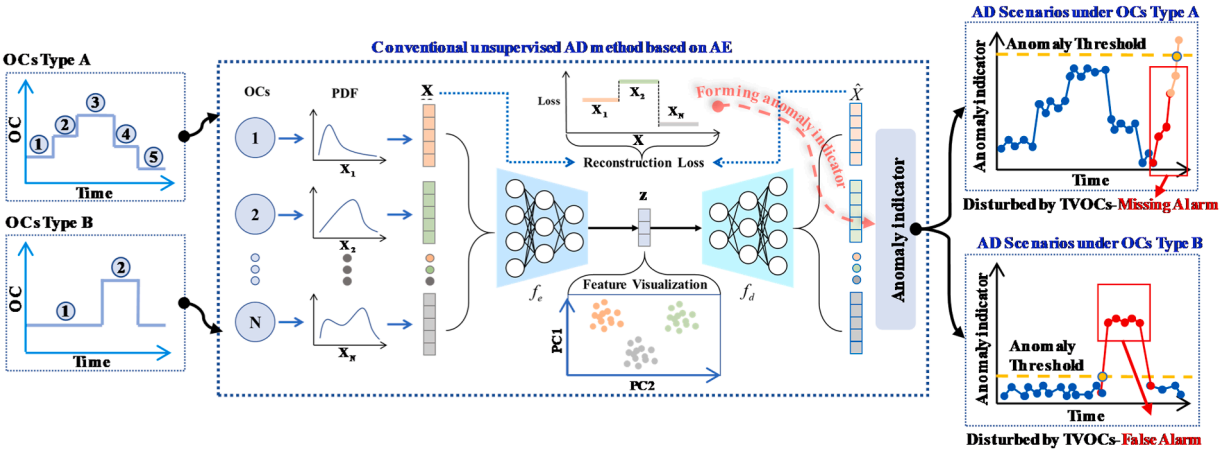

Fig. 1. The illustration of MPN encountered in the AD based on RE of the autoencoder. The vectors $\mathbf{x}$ in different colors represent model input monitoring data generated under various OCs, each with a distinct probability density function (PDF).

constraint decomposition(DCD) of the latent feature of VAE, which is the so-called DCD-VAE. The monitoring signals are as input to the proposed network without OCs signal, and two different levels of distribution constraints by two weight parameters β_o and β_s are imposed on the two separated sub-features that represent different information of the input data. The proposed method hypothesizes that there is inconsistency in the data distribution characteristics caused by discrepancies in information between OCs-related features and OSs-related features. The separation of information in latent features can be achieved by imposing different degrees of distribution constraints in the latent feature, where a low degree of constraint will make the OCs-related information flow more toward it, while a high degree of constraint will make the OSs-related information flow more toward it. Due to human intervention, the information on OCs does not conform to a Gaussian distribution, which is different from the information associated with the OSs [26]. Therefore, the DRL can be realized through the DCD based on VAE, and the AD of machines operating under TVOCs can only be implemented based on the OSs information without the interference of the varying OCs. Thus, more accurate and reliable AD of machines can be achieved. The main contributions of this paper can be concluded as follows:

(1) This study addresses the underexplored issue of data-driven AD in machine operations under TVOCs, resolving the resulting false positives and negatives is crucial for industrial applications.
(2) The unsupervised AD of machines operating under TVOCs is realized through DRL without the extra OCs' information. It has a better generalization and application scope, due to its need for less auxiliary information
(3) DCD-VAE is proposed to learn the disentanglement feature of the input data, the disentanglement is achieved through two different levels of distribution constraints based on the divisibility of the distribution of entangled features.
(4) Simulation and experiment verify the effectiveness of feature disentanglement between the OCs-related feature and OSs-related features, and further, the superiority of AD based on the disentangled OSs-related features.

This paper is organized as follows. Section 2 introduces the DRL. In Section 3, The DCD-VAE and theoretical guidelines are provided. The simulation and the accelerated degradation test on bearing in Section 4 further evaluate the proposed method. Finally, the conclusion of this paper is in Section 5.

2. Disentangled representation learning-DRL

As a feature representation learning technique, DRL provides access to disentangled representations that can reflect the data in a compositional structure perspective [34], which differs from other representation learning such as vanilla autoencoder. DRL aims to learn a representation that axis aligns with the generative factors of the data and can be achieved through an autoencoder- liked network, as shown in Fig. 2, each element of the learned disentangled feature of the letter S characterizes the independent generative factor.

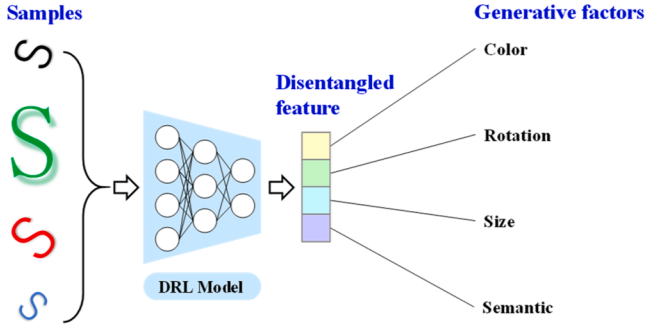

Fig. 2. The Schematic of the DRL process based on the letter S.

In general, there are *states* that describe the object in the world, and these *states* are the generative factors of the data, such as the color, and size of the letter S in Fig. 2. *Observations* are defined as the manifestations of these generative factors as perceived through various means, such as the samples of letter S in the right of Fig. 2. The generative process assumes that it can produce *observations* based on several *states*. Different observers can yield different *observations* of the same state. For instance, a machine operating in a specified *state* can be observed through vibration, sound, and other sensors.

The inference process is the counterpart of the generative process, wherein a model takes *observations* as input and produces a representation of *states*. Typically, this representation is a vector. The entire process, from the *states* of the object to the model's representation, can be described by a composite mapping that forms a bijection between the *states* and the model's representation. This mapping is considered a disentangled mapping, and the model's representation is a disentangled representation of the *observations*.

In other words, a bijection mapping indicates that the disentangled representation can be decomposed into several subgroups, each independently connected to specific *states*. Variations in each subspace of the representation affect only the corresponding *states*, leaving all other *states* unaffected, and vice versa. This property is crucial for accurately modeling system dynamics and improving AD capabilities by ensuring that the model can separately and effectively capture the variations due to different underlying factors.

Formally, let S and O be the sets of *states* and *observations*, respectively. The generative process is denoted as G: $S \to O$, and the inference process is I: $O \to Z$. Let Z be the set of the disentangled representation in vector space. Assume that the composite mapping F:$S \to Z$ is a combination of G and I, and defines the numerical disturbance Δd_i^S on the i th *state* s_i and Δd_i^Z on the i th subgroup z_i of representation. The goal of DRL is to find the mapping F that the following equation is satisfied:

$$F\left(\Delta d_i^S * s_i\right) = \Delta d_i^Z * F(s_i), s_i \in S \tag{1}$$

Where the $*$ denotes the disturbance operator, we define this disturbance operator as an algebraic operation considering the need for mathematical modeling. Eq.(1) shows that the disturbance Δd_i^S can commute with mapping F but may be in another kind of form Δd_i^Z. Thus, if the mapping F exists and Eq. (1) is satisfied, the feature $\mathbf{z}$ is a disentangled feature with respect to disturbance Δd_i^S and Δd_i^Z, and the composite mapping F is a bijection mapping.

To construct the model that satisfied Eq.(1) for mapping the *states* s into a disentangled representation $\mathbf{z}$. The VAE network is introduced since it also has the inference and generative process. Same as the structure shown in Fig. 1, the VAE has the encoder and decoder structure, as shown in Fig. 3. The encoder portion of a VAE E_ϕ yields an approximate posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ that is parameterized on a neural network with weight ϕ, and the decoder portion of VAE D_θ yields a likelihood distribution $p_\theta(\mathbf{z}|\mathbf{x})$ parametrized on a neural network with weights θ. The input data $\mathbf{x} \in \mathbb{R}^{1\times n}$ is first input into the encoder E_ϕ and the latent feature $\mathbf{z}$ is obtained through the posterior distribution approximation $q_\phi(\mathbf{z}|\mathbf{x})$, this encoding process is the inference process that the posterior distribution approximation $q_\phi(\mathbf{z}|\mathbf{x})$ needs to approximate the true posterior probability $p_\theta(\mathbf{z}|\mathbf{x})$, and the Kullback-Leibler (KL) divergence is used as the measures [35]:

$$D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})\right) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z},\mathbf{x})}\right] + \log p_\theta(\mathbf{x}) \tag{2}$$

Where $D_{KL}(\cdot \parallel \cdot)$ is the calculation of KL divergence, $\log p_\theta(\mathbf{x})$ is the log-likelihood of input data $\mathbf{x}$. Then the log-likelihood function goes:

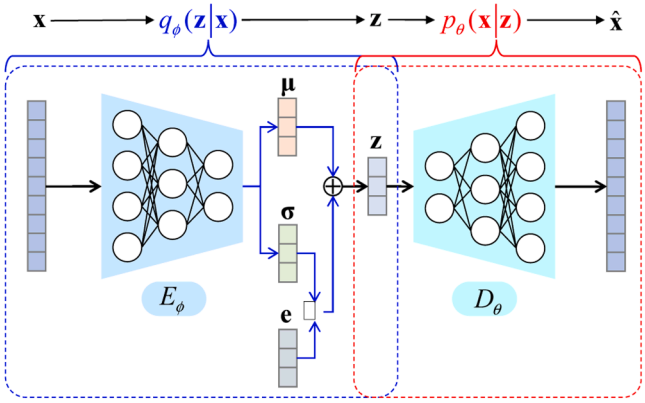

Fig. 3. The general structure of VAE. The $\hat{\mathbf{x}}$ is the reconstructed input via the generative process, which is the same as the $D_\theta(\mathbf{z})$ denoted in the main text.

$$\log p_\theta(\mathbf{x}) = D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})\right) \underbrace{-\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z},\mathbf{x})}\right]}_{\mathscr{L}(\mathbf{x},\phi,\theta)} \tag{3}$$

Since the D_{KL} is non-negative, $\mathscr{L}(\mathbf{x},\phi,\theta)$ is the Evidence lower bound (ELBO) of the log-likelihood function $\log p_\theta(\mathbf{x})$. Thus, to train the VAE model to fit the input data $\mathbf{x}$, the maximum of $\log p_\theta(\mathbf{x})$ given the model's parameters ϕ and θ can be approximated by maximizing the following form of ELBO [35]:

$$\begin{aligned}\phi^*,\theta^* &= \underset{\phi,\theta}{\operatorname{argmax}}\mathscr{L}(\mathbf{x},\phi,\theta)\\ &= \underset{\phi,\theta}{\operatorname{argmax}}\left[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})\| p_\theta(\mathbf{z})\right)\right]\end{aligned} \tag{4}$$

In this way, the VAE model can be trained based on the training data $\mathbf{x}$ by optimizing the above equation. Next, in the specific implementation of the VAE model, The real posterior distribution $p_\theta(\mathbf{z})$ in Eq.(4) is set as the standard multivariate Gaussian distribution $\mathscr{N}(\mathbf{0},\mathbf{I})$, and the latent feature $\mathbf{z}\in\mathbb{R}^{1\times n}$ in VAE can be calculated through Eq.(5):

$$\mathbf{z} = \mu_\phi(\mathbf{x}) + \sigma_\phi(\mathbf{x})\circ\mathbf{e} \tag{5}$$

Where $\mathbf{e}\in\mathbb{R}^{1\times n}$ is the auxiliary vector that follows the standard Gaussian distribution $\mathscr{N}(\mathbf{0},\ \mathbf{I})$, $\circ$ denotes Hadamard operator.$\mu_\phi(\mathbf{x})\in\mathbb{R}^{1\times n}$, $\sigma_\phi(\mathbf{x})\in\mathbb{R}^{1\times n}$ are the mean and variances vectors that are calculated by the output of the encoder E_ϕ. The maximization of $\log p_\theta(\mathbf{x}|\mathbf{z})$ is implemented as the minimization of $\| \mathbf{x} - D_\theta(\mathbf{z})\|_2^2$, which is so-called as the RE. Combining the calculation of KL divergence, the optimization of ELBO can be implemented as the following Eq.(6):

$$\begin{aligned}\underset{\phi,\theta}{\operatorname{argmax}}\mathscr{L}(\mathbf{x},\phi,\theta) = \underset{\phi,\theta}{\operatorname{argmin}}\Big[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\| \mathbf{x} - D_\theta(\mathbf{z})\ \|_2^2\right] + \frac{1}{2}\big(N\sigma_\phi(\mathbf{x})^2\\ + \mu_\phi(\mathbf{x})^2 - 2N\log\sigma_\phi(\mathbf{x})\big)\Big]\end{aligned} \tag{6}$$

Based on the VAE structure, β-VAE was proposed to realize the DRL by imposing an extra weight parameter β on the KL divergence term in Eq.(4). Thus, the ELBO of β-VAE can be rewritten as:

$$\mathscr{L}(\mathbf{x},\phi,\theta) = \underset{\phi,\theta}{\operatorname{argmax}}\left[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})\| p_\theta(\mathbf{z})\right)\right] \tag{7}$$

In general, the parameter β is larger than 1 (when β=1, β-VAE is degraded to the vanilla VAE) to push the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ more toward the prior distribution $p_\theta(\mathbf{z})$. Further, the decomposition of the KL divergence term yields [36]:

$$\begin{aligned}D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})\| p_\theta(\mathbf{z})\right) &= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z},\mathbf{x})}\left[\log\frac{q_\phi(\mathbf{z}|\mathbf{x})p_\theta(\mathbf{x})}{q_\phi(\mathbf{z})p_\theta(\mathbf{x})}\right]\\ &+ \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z})}\left[\log\frac{q_\phi(\mathbf{z})}{\prod_{j=1}^{n} q_\phi(\mathbf{z}_j)}\right] + \sum_{j=1}^{n}\log\frac{q_\phi(\mathbf{z}_j)}{p_\theta(\mathbf{z}_j)}\end{aligned} \tag{8}$$

Where $\mathbf{z}_j$ is the j-th dimension of the latent variable $\mathbf{z}$. In the above Eq. (8), The first term is the index-coding mutual information(MI), which is related to the MI between sample $\mathbf{x}$ and the latent feature $\mathbf{z}$. Higher MI indicates better disentanglement. The second term is total correlation, representing the degree of interdependence among the variables $\mathbf{z}_j$ in the latent variable space. A heavier penalty on total correlation indicates stronger statistical independence among the variables $\mathbf{z}_j$ in the posterior probability distribution. The last term is the measurement of the divergence between the marginal distribution $q_\phi(\mathbf{z}_j)$ of $\mathbf{z}_j$ and the prior distribution $p_\theta(\mathbf{z}_j)$, indicating that the dimensions of each latent variable should not deviate too much from the dimensions of the prior latent variable. It is apparent that the second term is going to ensure the disentanglement of the latent feature $\mathbf{z}$ when the KL divergence is getting smaller, but the side-effect is that the MI between $\mathbf{x}$ and $\mathbf{z}$ is eliminated. The capability of disentanglement of β-VAE can also be illustrated by the information bottom neck [37] in the information theory, in which the KL divergence forces the latent feature $\mathbf{z}$ to abandon the MI with the sample $\mathbf{x}$.

3. Methodology

3.1. Motivation

Back to the MPN problem shown in Fig. 1, and start from the information bottleneck perspective. The information flow diagram when training β-VAE is demonstrated in Fig. 4. Assume that the probability of OCs is $p(\mathbf{o})$, the relationship between OCs and monitoring data $\mathbf{x}$ is bijection mapping(It is believed that the machine working under different OCs will produce different distributions of monitoring information [38,39]), and each pattern normality in the training data follows the conditional probability $p(\mathbf{x}|\mathbf{o})$. If the randomness present in the sample $\mathbf{x}$ is ignored and the bijection mapping assumption is combined, then the conditional probability of each pattern normality can be considered the same as the OC's probability:

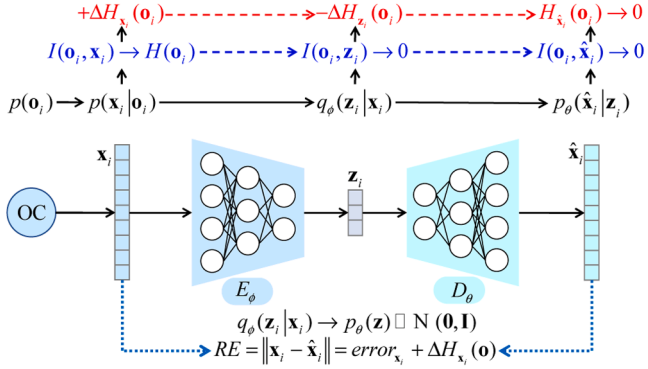

Fig. 4. Information flow diagram in β-VAE. The $H(\cdot)$ denotes the information entropy. The upper red-marked information GAP represents the state of increase or decrease in the OCs' information within each variable throughout the entire process, from the input sample $\mathbf{x}_i$, to the latent variable $\mathbf{z}_i$, and then to the reconstructed sample $\hat{\mathbf{x}}_i$ during the training process. The lower blue-marked MI similarly indicates the changes in MI between each variable and the OCs during the network transmission process.

$$p(\mathbf{x}|\mathbf{o}) = p(\mathbf{o}) \tag{9}$$

This implies that disregarding randomness, the variability information among samples in the monitoring data $\mathbf{x}$ only retains the parts associated with OCs. The distribution of the OCs corresponds to the distribution of the training samples. Consequently, the MI between the OCs $\mathbf{o}$ and the corresponding training sample $\mathbf{x}$ is as follows:

$$I(\mathbf{x},\mathbf{o}) = \mathbb{E}_{\mathbf{o}} D_{KL}(p(\mathbf{x}|\mathbf{o})\| p(\mathbf{o})) = \mathbb{E}_{o} D_{KL}(p(\mathbf{o})\| p(\mathbf{o})) \tag{10}$$

It suggests that strong correlation between the sample $\mathbf{x}$ and OCs $\mathbf{o}$. On the one hand, if the β-VAE is utilizing and training for AD, the strong penalty induced by the parameter β on the KL divergence term in Eq.(7) during the training stage will degrade the MI $I(\mathbf{x},\mathbf{z})$ between each sample $\mathbf{x}$ and its corresponding latent variable $\mathbf{z}$, and become two independent variables considering:

$$\begin{aligned} &I(\mathbf{x},\mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})p_\theta(\mathbf{x})}{q_\phi(\mathbf{z})p_\theta(\mathbf{x})}\right] \to 0 \\ &\Rightarrow q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}), q_\phi(\mathbf{z},\mathbf{x}) = q_\phi(\mathbf{z})\cdot p_\theta(\mathbf{x}) \end{aligned} \tag{11}$$

In this way, the MI $I(\mathbf{o},\mathbf{z})$ is similarly attenuated (as shown in Fig. 4), with the information of OCs lost, and also becomes two independent variables:

$$q_\phi(\mathbf{o},\mathbf{z}) = p(\mathbf{o})\cdot q_\phi(\mathbf{z}) \tag{12}$$

However, the maximization of log-likelihood driven by reconstruction loss has to ensure the distribution $q_\theta(\mathbf{x}|\mathbf{z})$ is close enough to the real distribution $p_\theta(\mathbf{x})$, but the information gap due to the missing information about $\mathbf{o}$ in $\mathbf{z}$ will make the reconstruction loss cannot be maintained at a stable level, and it will keep changing with the difference of OCs, which is harmful to AD based on RE. On the other hand, the minimization of both KL term and reconstruction loss is indeed an adversarial process same as the generative adversarial network, and if we apply the method that directly analyzes the latent feature, the information gap induced by $\Delta H_{\mathbf{x}_i}(\mathbf{o})$ (as shown in Fig. 4) in the RE will inevitably be detrimental to the optimization of KL term, even though a stronger optimization focus was placed on KL term in the β-VAE. As a consequence, the interference caused by OCs still exists in the latent feature of β-VAE, which is also an unfavorable situation for AD.

In summary, Fig. 4 illustrates the transmission process of the training sample xxx within the β-VAE framework. Initially, under OCs o_i, there exists a conditional probability $p(\mathbf{x}_i|\mathbf{o}_i)$ between the training sample $\mathbf{x}_i$ and OCs $\mathbf{o}_i$. This conditional probability causes the MI $I(\mathbf{o}_i,\mathbf{x}_i)$ to approach $H(\mathbf{o}_i)$ considering that the randomness of $\mathbf{x}_i$ is ignored, thereby increasing the entropy $+H(\mathbf{o}_i)$ in $\mathbf{x}_i$. As the encoder maps $\mathbf{x}_i$ to $\mathbf{z}_i$, the constraints imposed by the KL term guide $q_\phi(\mathbf{z}_i|\mathbf{x}_i)$ towards $\mathcal{N}(\mathbf{0},\mathbf{I})$. Consequently, the MI $I(\mathbf{o}_i,\mathbf{z}_i)$ approaches zero, resulting in a reduction of entropy $-H(\mathbf{o}_i)$ in $\mathbf{z}_i$. Finally, after the decoder reconstructs $\widehat{\mathbf{x}}_i$, the MI $I(\mathbf{o}_i,\widehat{\mathbf{x}}_i)$ between the reconstructed sample $\widehat{\mathbf{x}}_i$ and $\mathbf{o}_i$ also approaches zero, leading to a decrease in the entropy $H(\mathbf{o}_i)$ in reconstructed sample $\widehat{\mathbf{x}}_i$ approaches zero. Thus, the RE comprises not only the inherent error $error_{\mathbf{x}_i}$ within the network but also the information loss $\Delta H_{\mathbf{x}_i}(\mathbf{o})$ incurred during the transmission of the sample $\mathbf{x}_i$ through the network, as depicted in the figure.

Since both the RE-based method and the latent-feature-based method cannot avoid the interference of the OCs, and inspired by the information bottleneck mentioned in the above analysis, a constraint decomposition method is proposed to achieve feature disentanglement, i.e., decomposing KL term into information bottleneck with different degrees of constraint to guide the information to separate, so that different information in the associated $\mathbf{x}$ will be subject to differential constraint to realize the separation and disentanglement adaptively.

3.2. Distribution constraint decomposition vae

Firstly, let $\mathbf{s}$ and $\mathbf{o}$ represent the factors that are related to the OSs and OCs of the machine, respectively. $\mathbf{s}$ remains invariant until the OSs of the machine change, while $\mathbf{o}$ varies under the TVOCs. An intuitive information flow is shown in Fig. 5 to demonstrate the proposed DCD-VAE. The information about the OCs and OSs is conveyed by $H(\mathbf{o})$ and $H(\mathbf{s})$, respectively. and they are both merged into the $H(\mathbf{o},\mathbf{s})$ conveyed by monitored data $\mathbf{x}$ at the first information merge stage (CM and data acquisition). The encoder in DCD-VAE serves as two separated posterior distribution inference processes $q_\phi(\mathbf{z_o}|\mathbf{x})$ and $q_\phi(\mathbf{z_s}|\mathbf{x})$ that can produce two separated latent features $\mathbf{z_o}$ and $\mathbf{z_s}$. The main trouble solver for the disentanglement of these two factors $\mathbf{s}$ and $\mathbf{o}$ is the DCD structure for latent features $\mathbf{z_s}$ and $\mathbf{z_o}$, e.g.: the β weighted KL term in β-VAE is decomposed into two distinct regularization terms:

$$\beta D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})\| p_\theta(\mathbf{z})\right) = \beta_o D_{KL}\left(q_{\phi_o}(\mathbf{z_o}|\mathbf{x})\| p_\theta(\mathbf{z})\right) + \beta_s D_{KL}\left(q_{\phi_s}(\mathbf{z_s}|\mathbf{x})\| p_\theta(\mathbf{z})\right) \tag{13}$$

Where ϕ_o and ϕ_s are the parameters that related to the feature $\mathbf{z_o}$ and $\mathbf{z_s}$, respectively. Strong distributional penalties on latent feature $\mathbf{z}$ induced by a whole weighting parameter β will eliminate the information $H(\mathbf{o})$ as Fig. 4 illustrated, which means the soft distributional penalties on latent feature $\mathbf{z}$ are going to better maintain the information $H(\mathbf{o})$. Thus, the DCD structure can provide the opportunity to keep both information $H(\mathbf{o})$ and $H(\mathbf{s})$ in $\mathbf{z}$ while separating the two of them. Interpreted intuitively, a loose constraint will cause information about these time-varying features that vary with OCs to flow to the information bottleneck it controls, while a tight constraint will allow time-invariant information to flow to it. A noteworthy point is that the loose constraint term cannot be zero. One consequence will be that the network will degenerate into a general autoencoder if the weighting parameter that controls the loose constraint term is zero since the presence of tight constraints will lead to a direct and complete flow of information to an unconstrained information bottleneck.

The log-likelihood of the reconstructed $\widehat{\mathbf{x}}$ driven by reconstruction term ensures the information conveyed by $\mathbf{z_o}$ and $\mathbf{z_s}$ is complementary and not containing other irrelevant information. The ELBO of the proposed DCD-VAE is as follows Eq.(14):

$$\begin{aligned} \mathscr{L}(\mathbf{x},\phi_o,\phi_s,\theta) = \underset{\phi,\theta}{\operatorname{argmax}}\Big(&\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta_o D_{KL}\left(q_{\phi_o}(\mathbf{z_o}|\mathbf{x})\| p_\theta(\mathbf{z})\right) \\ &- \beta_s D_{KL}\left(q_{\phi_s}(\mathbf{z_s}|\mathbf{x})\| p_\theta(\mathbf{z})\right)\Big) \end{aligned} \tag{14}$$

By choosing and tuning the relative values of β_o and β_s, the information conveyed by $\mathbf{z_o}$ and $\mathbf{z_s}$ can be complementary while achieving the feature disentanglement. Building upon the definition of feature disentanglement outlined in Section 2, the proposed DCD-VAE can also be explained from a variational perspective to elucidate its differences from β-VAE and its suitability for the MPN problem. The feature disentanglement in β-VAE is achieved by incorporating additional prior distribution constraints into the latent features, where each element in the feature is required to follow a standard normal distribution. Once feature disentanglement is achieved through the β-VAE model, each element in the latent feature controls a specific generative factor of the samples and exhibits orthogonality and independence in the feature space. Additionally, each element follows a standard normal distribution. As depicted in Fig. 6, the distribution of the latent feature is transformed to follow a standard normal distribution.

However, this can only be achieved when the training dataset is sufficiently large and covers all possible generative factors comprehensively. For instance, the problem of feature disentanglement in image generation can be addressed by utilizing a vast training dataset that encompasses diverse generative factors. However, this approach does not apply to the MPN problem in AD of machines. In the MPN problem, the training data for the OCs' factor is relatively small compared to other generative factors that constitute the monitoring signals. As a result, the model lacks the necessary data to transform the OCs' feature into a

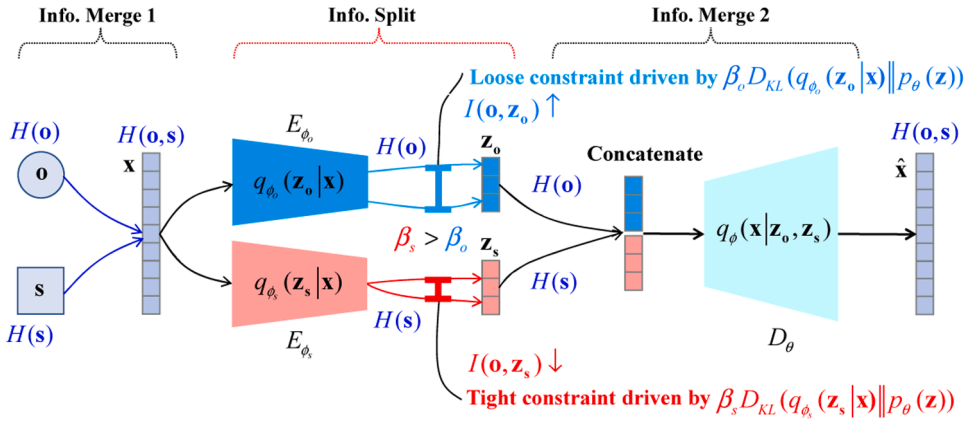

Fig. 5. Information flow diagram of the proposed DCD structure in VAE-based network.

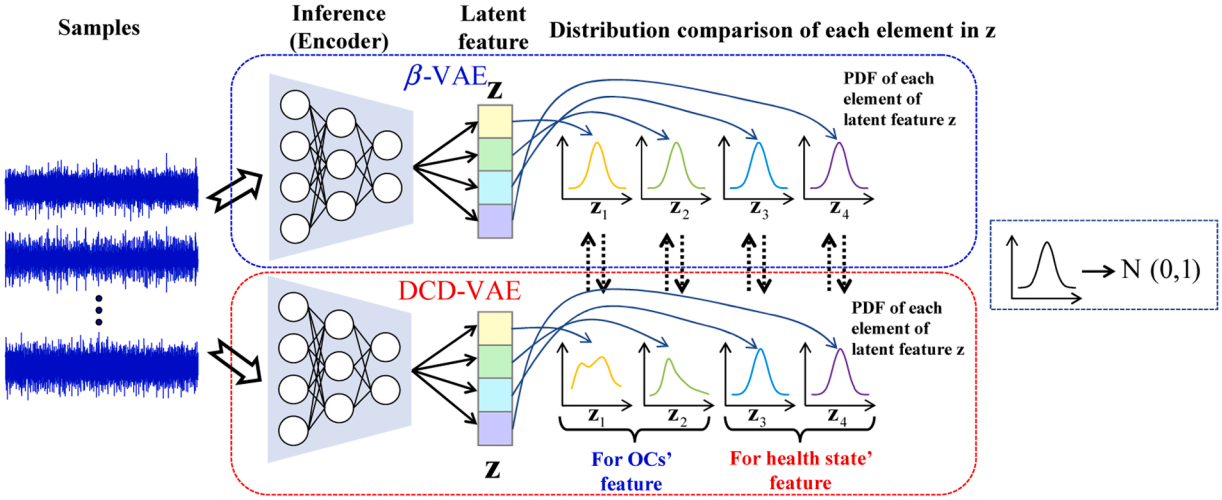

Fig. 6. The inference process of the feature disentanglement in β-VAE and DCD-VAE. The different colors in z represent elements in z.

standard normal distribution under the same constraint strength. For feature disentanglement in the context of the MPN problem, it is sufficient for each element in the latent feature to independently control different generative factors. There is no need to transform the operating conditions (OCs) into a standard normal distribution. Thus, an AD method can be constructed based on the OSs features of associated machines without the interference of OCs.

3.3. The implementation of DCD-VAE-based AD

The implementation includes the network structure and AD procedure design. The proposed DCD-VAE is based on a convolution autoencoder [40], e.g. each layer of the encoder and decoder consists of a 1-D convolution layer. Besides, the encoder is an entire encoder E_ϕ that is not separated into two single encoders E_{ϕ_o}, E_{ϕ_s} for $q_{\phi_o}(\mathbf{z_o}|\mathbf{x})$ and $q_{\phi_s}(\mathbf{z_s}|\mathbf{x})$, instead, the output feature $\mathbf{z}$ of the encoder is just split into two subparts to denote the $\mathbf{z_o}$ and $\mathbf{z_s}$ for convenience. The detailed structural information of the encoder E_ϕ and decoder D_θ are shown in Table 1.

In the encoding stage, the data $\mathbf{x}$ is inputted into the encoder, and the output is the mean and variance feature $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are acquired:

$$\boldsymbol{\mu},\boldsymbol{\sigma} = E_\phi(\mathbf{x}) \tag{15}$$

The spit operation is conducted:

$$\boldsymbol{\mu}_\mathbf{o},\boldsymbol{\mu}_\mathbf{s} = split(\boldsymbol{\mu}) \tag{16}$$

$$\boldsymbol{\sigma}_\mathbf{o},\boldsymbol{\sigma}_\mathbf{s} = split(\boldsymbol{\sigma}) \tag{17}$$

Table 1
The network structure of the encoder and decoder in the proposed DCD-VAE.

Encoder E_ϕ	Decoder D_θ
	Linear layer-1
	Linear layer-2
	Unflatten layer
Convolution layer-1 (1D)	Transpose Convolution layer-1 (1D)
MaxPool layer-1 (1D)	Upsample layer-1
BatchNormlization + Mish activation function	BatchNormlization + Mish activation function
Convolution layer-2 (1D)	Transpose Convolution layer-2 (1D)
MaxPool layer-2 1D	Upsample layer-2
BatchNormlization + Mish activation function	BatchNormlization + Mish activation function
Convolution layer-3 (1D)	Transpose Convolution layer-3 (1D)
MaxPool layer-3 (1D)	Upsample layer-3
BatchNormlization + Mish activation function	BatchNormlization + Mish activation function
Flatten layer	Flatten layer
	Linear layer-3

The KL constraints weighted by β_o,β_s are pushing the distribution of $\mathcal{N}(\boldsymbol{\mu}_\mathbf{o},\boldsymbol{\sigma}_\mathbf{o})$ and $\mathcal{N}(\boldsymbol{\mu}_\mathbf{s},\boldsymbol{\sigma}_\mathbf{s})$ close to the standard multivariate Gaussian distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$:

$$\mathscr{L}_{\mathrm{KL}}^{\mathbf{z_o}} = \frac{1}{2}\sum_{n=1}^{N}\left(1+\log\left((\boldsymbol{\sigma}_\mathbf{o}^n)^2\right)-(\boldsymbol{\mu}_\mathbf{o}^n)^2-(\boldsymbol{\sigma}_\mathbf{o}^n)^2\right) \tag{18}$$

$$\mathscr{L}_{\mathrm{KL}}^{z_s} = \frac{1}{2}\sum_{n=1}^{N}\left(1+\log\left(\left(\boldsymbol{\sigma}_{\mathbf{s}}^{\mathbf{n}}\right)^2\right)-\left(\boldsymbol{\mu}_{\mathbf{s}}^{n}\right)^2-\left(\boldsymbol{\sigma}_{\mathbf{s}}^{n}\right)^2\right) \tag{19}$$

The same resample operation is used as the vanilla VAE, and the resampled $\mathbf{z_o}$ and $\mathbf{z_s}$ are concatenated as the input of the decoder D_θ to reconstruct the input $\mathbf{x}$. The reconstruction loss $\mathscr{L}_{\mathrm{RE}}$ is calculated as Eq. (20):

$$\mathscr{L}_{\mathrm{RE}} = \| \mathbf{x} - \hat{\mathbf{x}} \|_2^2 \tag{20}$$

Therefore, the total loss function of the proposed DCD-VAE is formulated as Eq.(21):

$$\mathscr{L}_{\mathrm{DCD-VAE}} = \mathscr{L}_{\mathrm{RE}} + \beta_o \mathscr{L}_{\mathrm{KL}}^{z_o} + \beta_s \mathscr{L}_{\mathrm{KL}}^{z_s} \tag{21}$$

The most critical hyper-parameters for the proposed model are β_o and β_s, as they directly impact the disentanglement performance. Adjusting these parameters primarily involves trial and error. However, we have identified a useful approach to expedite this process. Initially, setting a relatively large β_s (e.g., 2, if using the same loss function as ours) and a small β_o (e.g., 10^(−3)) generally achieves feature disentanglement, though the fitting of input data $\mathbf{x}$ may suffer due to excessive information loss from the large β_s. By gradually reducing β_o, one can determine a value that also provides good data fit quality. Concurrently, β_o should be reduced proportionally with β_s at first, and then fine-tuned independently. The whole AD process-based DCD-VAE is shown in Fig. 7. The training data consists of the monitoring data during the initial operation of machines (the common agreement is that the initial OSs of machines can be recognized as the normal state [41,42]). And the OCs' information is not necessarily included in the monitoring data. The DCD-VAE is trained based on the training data and loss function defined in Eq.(21). In the AD stage, the encoder of the trained DCD-VAE is deployed for the feature disentanglement of the input data. The disentangled feature that is related to the OSs is utilized to form an ANI by the information fusion and dimensional reduction method (such as Principal Component Analysis, PCA). Afterward, the anomaly threshold can be successfully settled by statistical methods such as 3-sigma rules. Finally, the AD of machines can be achieved under TVOCs. To further promote the practical application of the proposed AD method, the pseudocode of the proposed DCD-VAE training and AD procedure is given in Algorithm 1.

4. Case study

In this paper, two cases are introduced to verify the effectiveness of the proposed method. Case 1 is a simulation that simulates the CM process of machines operating under TVOCs. Case 2 includes the experiment that is conducted on an accelerative bearing life test bench to test the bearing OSs degradation under time-varying speed conditions.

4.1. Case 1

4.1.1. Dataset details

The benchmark case study is considered in this paper, the dataset is from the Aramis challenge [43] and it is redesigned to simulate the run-to-failure trajectory of the multi-sensor monitoring under TVOCs in this paper. The sensor number is $M = 10$ and D_t represents the degradation information at time t, which can be also recognized as the health parameters of the monitoring machines. The machines are working in normal healthy conditions for a long period in most cases until the change point τ emerges as an anomaly and indicates the beginning of the degradation. Thus, D_t is modeled as a two-phase degradation process as shown in Eq.(22):

$$D_t = \begin{cases} 0, t \le \tau \\ D_{t-1} + \alpha(\alpha_E E_{t-1} + w_{t-1} + 1), t > \tau \end{cases} \tag{22}$$

Where $D_t = 0$ denotes the machine is operating under normal (healthy) conditions. And D_t starts to vary based on the combination of the degradation rate α, α_E, environment E_{t-1}, and random parameters

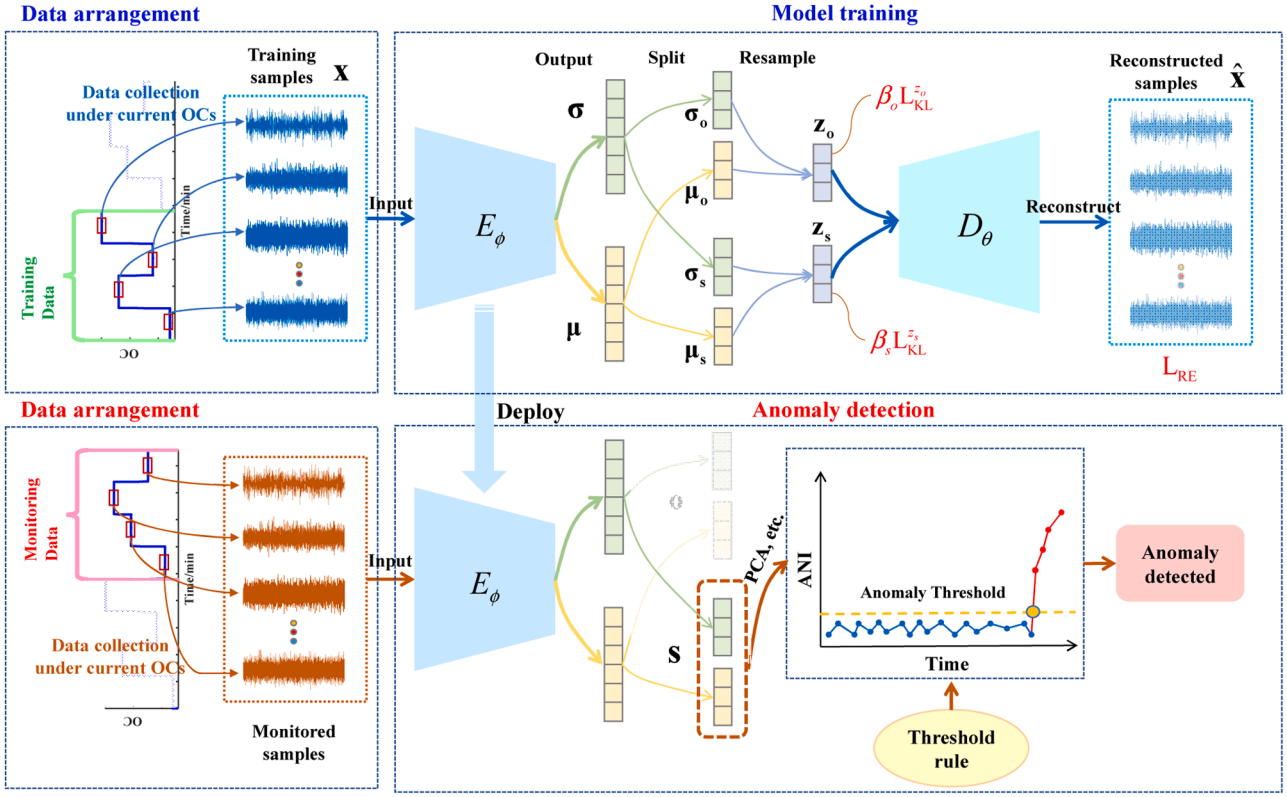

Fig. 7. The AD process based on the proposed DCD-VAE.

Algorithm 1
DCD-VAE-based AD.

Input: The training data $\mathbf{x}_{Train}$, and testing data $\mathbf{x}_{Test}$. DCD-VAE network consists of an encoder and decoder E_ϕ,D_θ with their network parameters θ_E,$\theta_D$respectively, learning rate lr, trade-off parameter β_o and β_s.
Output: the AD result of the testing data.
1. **While** *not done* **do:**
Model training:
2. initialize the parameters of the E_ϕ,D_θ.
3. **Repeat:**
4. Using the dataset $\mathbf{x}_{Train}$ to update E_ϕ,D_θ.
5. $\theta_E \leftarrow \theta_E - lr\left(\frac{\partial \mathscr{L}_{KL}}{\partial \theta_E} + \frac{\partial \mathscr{L}_{\mathrm{RE}}}{\partial \theta_E}\right)$
6. $\theta_D \leftarrow \theta_D - lr\left(\frac{\partial \mathscr{L}_{\mathrm{RE}}}{\partial \theta_o}\right)$
8 **Until** θ_E,θ_D have converged.
Deploy:
9. **For** each data in the dataset $\mathbf{x}_{Train}$,$\mathbf{x}_{Test}$:
10. Input the data to the encoder E_ϕ to acquire the disentangled feature μ_s
11. Using PCA to reduce the dimension of the feature μ_s into one-dimension ANI
12. **End**
13. Calculate the anomaly threshold with ANI in the dataset $\mathbf{x}_{Test}$ by the 3-sigma rule.
16. Detect the anomaly in dataset $\mathbf{x}_{Test}$
17. **For** each ANI of dataset $\mathbf{x}_{Test}$:
18. **If** the ANI>threshold
19. ANI is detected as an anomaly
20. **Else** ANI is normal
21. *done*

w_{t-1} once time passes the change point. The monitored signals are assumed to be influenced by three factors: (1) the bias induced by environmental conditions, which are so-called OCs in this paper. (2) the degradation of the machine's health. (3) Uncertainty exists both in the degradation of the machine and the measurement process of the monitoring system. Therefore, 10 monitoring signals can be simulated using the following Equations:

$$
\begin{aligned}
&s_t^1 = \sin(1/2D_t + 2E_t), s_t^2 = \sin(D_t + \tan(E_t)), s_t^3 = \tan(1/2E_t - 2/3w_t), s_t^4 = \cos^3(w_t)\sin(E_t)\\
&s_t^5 = \sin(3/10D_t + 3w_t) - E_t, s_t^6 = \sin\left(D_t + w_t - \tan^2(E_t + 1)\right), s_t^7 = \sin(1/2D_tE_t) + 1/2w_t,\\
&s_t^8 = \cos^2(D_t - \tan(E_t) + 2\tan(w_t)), s_t^9 = \cos(D_t - \tan(E_t) + 2\tan(w_t)),\\
&s_t^{10} = \tan(1/4w_t) - 1/4\cos(1/3D_t\sin(E_t))
\end{aligned}
\tag{23}
$$

The length of the trajectory is set as 1000, and the degradation rate α is set as a random value that follows the Gaussian distribution $\mathscr{N}(0.015, 0.007)$ truncated in the range [0,0.02], α_E=0.3, τ=600, w_t are random value sampled from the Gaussian distribution $\mathscr{N}(0, 0.1)$ truncated in the range [−0.5,0.5]. The simulated evolution of the degradation D_t is shown in Fig. 8(a) where the true label of the data samples are defined and the E_t is generated according to ten cities' temperatures and humidity same as the original dataset, and it is shown in Fig. 8(b). The generated environment parameter E_t demonstrates the strong TVOCs scenario due to the random and periodic characteristics induced by the city's temperature and humidity, which is exactly the scenario that this work focuses on. According to Eq.(23), 10 monitoring signals are simulated and shown in Fig. 9. The monitoring signals are strongly disturbed by the TVOCs, and emerge the same quasi-periodic pattern shown in E_t. Besides, a rough qualitative observation of the change point can be made, but setting a quantitative threshold to determine the change point remains a challenge in terms of most of the signals. It is a typical MPN situation once the data-driven method is utilized for AD due to the varying environmental parameter E_t, and there are different normalities in the training data with different environmental values if a portion of earlier monitored data is used as a training sample. However,

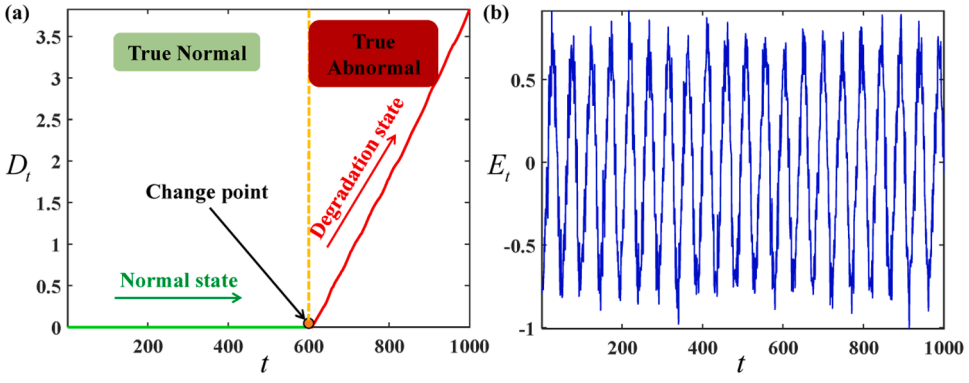

Fig. 8. The simulated degradation indicator and environment E_t served as TVOCs.

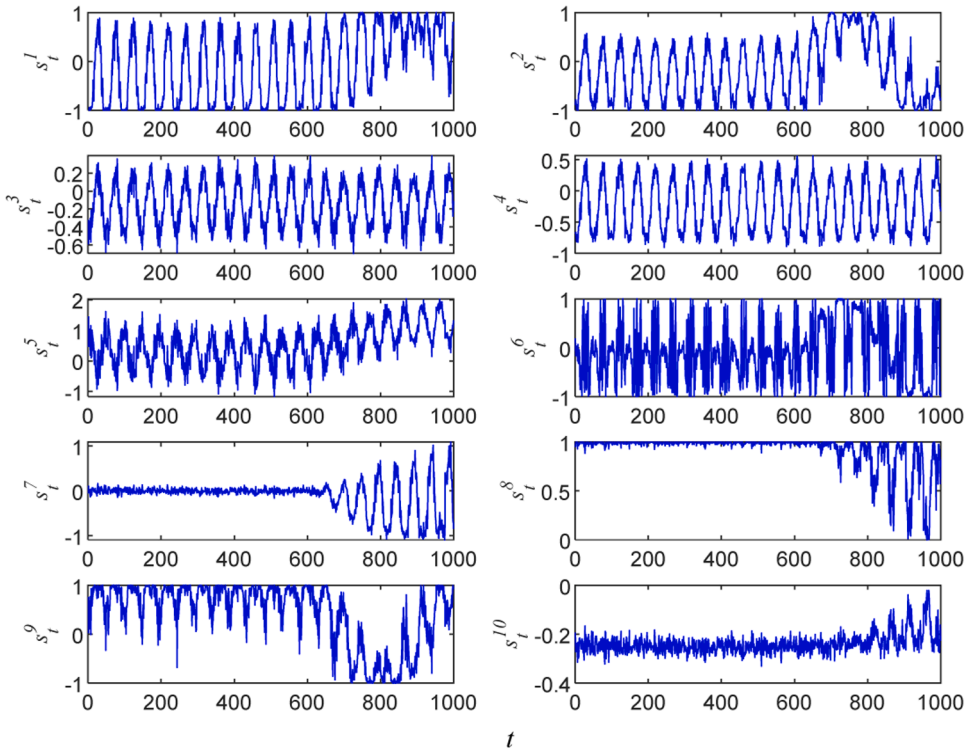

Fig. 9. 10 Simulated signals according to the degradation information under the time-varying environment.

some signals such as s_t^7,s_t^8,s_t^{10}, are free from disturbance of E_t before the change point while suffering the same interference after the change point during the degradation stage. Therefore, these three signals are not considered as one of the inputs to the proposed DCD-VAE for better validation rigor.

4.1.2. The result of the proposed method

The proposed DCD-VAE is used to verify its effectiveness in this case, the whole trajectories which consist of 7 monitored signals are manually divided into training data and testing data. The training data comprises the signals of the first 400 time points in the trajectories, while the subsequent 600 time points are used as testing data. Therefore, the training data is arrayed with the size of 400 × 7 and the testing data is 600 × 7. The weighted parameters β_o and β_s are experimentally turned according to the strategy in Section 3.3, which is β_o=0.0001,β_s=1.9, respectively. Each convolution kernel size is set as 3 in both the encoder and decoder. The latent dimension of the z is set as 128(Here since the dimension of the potential features is higher than the dimension of the input data, the encoder becomes an embedder). The optimizer uses Adam with a learning rate of 1e-3. The number of training epochs is set as 100.

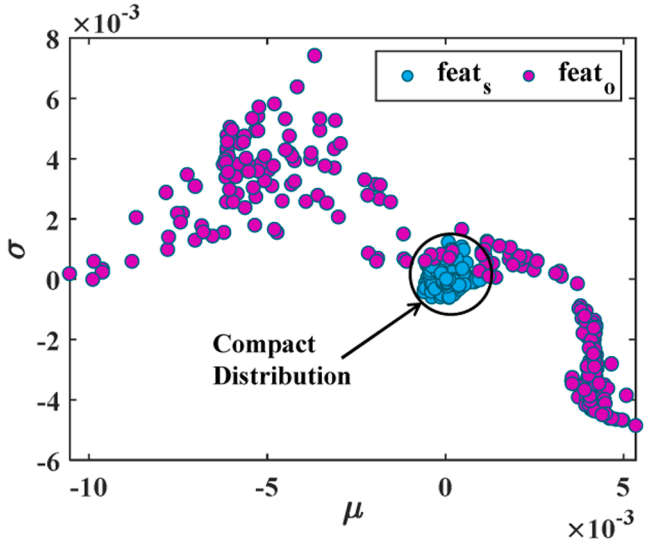

Fig. 10. The training result of the proposed DCD-VAE. (a) is the loss curve of the three reweighted losses during the whole training process; (b) is the scatter plot of latent feature z distribution for training samples characterized by mean μ and variance σ.

After the training process, the visualization of feature distribution parameters of the training data in Fig. 10 demonstrates a compact distribution in features related to the OSs of the machine, while the feature related to OCs exhibits a contrasting pattern with a scattered and non-compact distribution. These two different distributional properties indicate that the separation of feature distributions is achieved in the training set. Due to the higher variability of OCs-related features, they exhibit a stronger divergence in their distribution. Conversely, the features associated with the OSs achieve a more compact distribution under the stronger constraint of KL loss after disentangling. Furthermore, the average reconstruction loss after training convergence for vanilla VAE and β-VAE is shown in Table 2 for comparison. The trade-off relationship between $\mathscr{L}_{\mathrm{RE}}$ and $\mathscr{L}_{\mathrm{KL}}$ indicates the adversarial process of each other during the training stage. The optimal result is that both $\mathscr{L}_{\mathrm{RE}}$ and $\mathscr{L}_{\mathrm{KL}}$ can be optimized into the lowest level as much as possible for the promotion of better information reservation(lower $\mathscr{L}_{\mathrm{RE}}$) and model generation(lower $\mathscr{L}_{\mathrm{KL}}$). Compared with the mean converged value of reconstruction loss within these three networks, the proposed DCD-VAE can still maintain a well-converged value of 0.3064 which is a benefit

Table 2
The average reconstruction loss of the proposed model, VAE and β-VAE after the training convergence.

Model	DCD-VAE	VAE	β-VAE
Average $\mathscr{L}_{\mathrm{RE}}$	0.3064	0.3151	0.3235

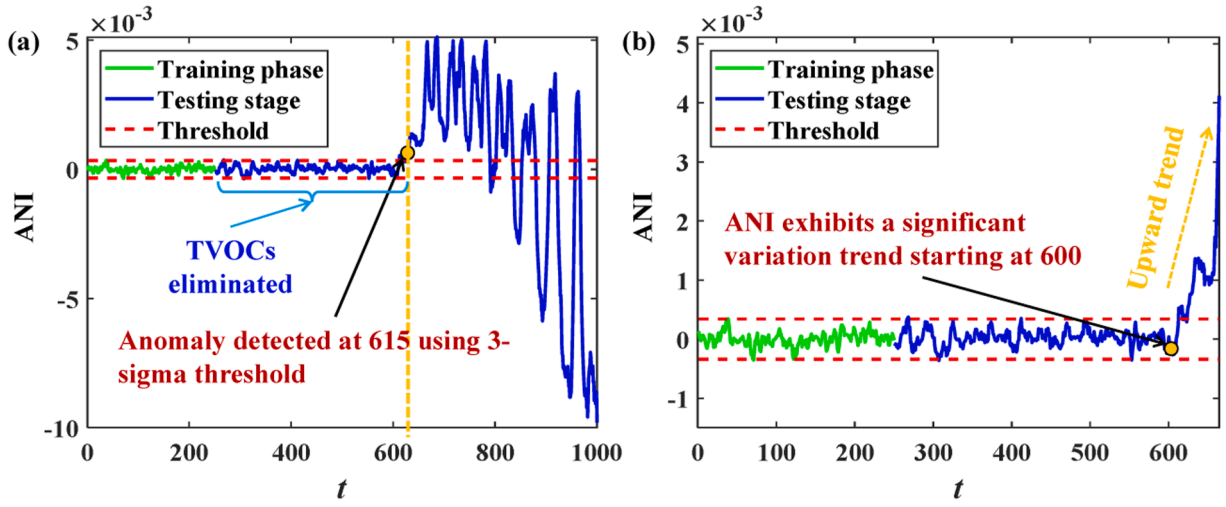

Fig. 11. The result of AD based on ANI derived from the proposed DCD-VAE. (a) is the detection result that the anomaly happened at 615; (b) is the enlarged view that suggests an earlier change point at 600. The yellow dashed line denotes the actual occurrence point of degradation, namely the anomaly change point τ.

from the special design DCD structure in the network loss term.

According to the AD process indicated in Fig. 7, the ANI needs to be constructed for better intuitive observation. As suggested above, the ANI can be constructed by the dimensional reduction method based on the latent feature. During the actual implementation, the first principle component of the mean of the latent feature $\mathbf{z}_s$ after PCA can be recognized as the ANI. The reason why the mean of the latent feature is adopted for ANI construction is that the resample trick used in a network with a VAE-like structure introduces additional noise into the feature $\boldsymbol{z}$. Therefore, the ANI based on the whole trajectory is shown in Fig. 11(a). The threshold is acquired by the 3-sigma rule, which is formulated as Eq. (24):

$$\left|ANI_i - \mu_{ANI_{train}}\right| > 3\sigma_{ANI_{train}} \tag{24}$$

Where $\mu_{ANI_{train}}$ and $\sigma_{ANI_{train}}$ denote the means and standard deviations of ANI in training data, the result indicates that the anomaly is first detected at 615, which is a little bit late compared with the true change point at 600. However, the enlarged view in Fig. 11(b) suggests that the upward trend emerges at 600. Despite the inability to automatically detect this change using thresholding methods, it is undeniable that the ANI is responsive to anomalies at that specific point, and this phenomenon is attributed to the persistent interference of noise in the ANI. Besides, the time before the change point τ as marked in curly brackets shows the disturbance introduced by TVOCs is successfully eliminated.

4.1.3. Comparison study

Five advanced and classical AD models are introduced for comparison. Specifically, these include the recently proposed FDCVAE [26], especially for AD under TVOCs, the MRRAE [41] and DSMDA [42] models for stable OCs, as well as the classical VAE and β-VAE models. The network structures of the encoder-decoder in all these comparative models are consistent with the proposed DCD-VAE for a fair comparison. Other hyperparameters of the networks' training are also kept the same as DCD-VAE. The calculation of anomaly thresholds in all comparative methods is based on the 3-sigma rule. Firstly, a direct comparison of the results is provided by presenting the constructed ANIs and their corresponding detection outcomes. Fig. 12 presents the AD results of three comparative methods recently published in the literature. In general, due to the interference of the TVOCs, the anomalies detected by these three methods are delayed compared to the proposed method. It is evident that FDCVAE, a model specifically designed for AD under TVOCs, generates ANI that is less influenced by varying OCs compared to those produced by the MRRAE and DSMDA models, with better distinguishability between normal and anomalous states. However, FDCVAE is prone to a higher rate of false alarms. Conversely, the ANIs generated by MRRAE and DSMDA fluctuate with OCs, exhibiting poorer timeliness in AD compared with FDCVAE, and the first correct anomalies were detected only at points 721 and 768, respectively.

The AD results of VAE which consists of three different ANIs are shown in Fig. 13, These constructed ANIs are more or less affected by TVOCs, particularly the RE-based ANI, as shown in Fig. 13(b), which is most sensitive to TVOCs. All curves exhibit a delayed detection of anomalies compared to the proposed method, leading to an increased false negative rate. Additionally, it is observed that the results depicted in the figure also exhibit some instances of false positives. Furthermore, AD results based on loss functions as shown in Fig. 13(b) and Fig. 13(c)

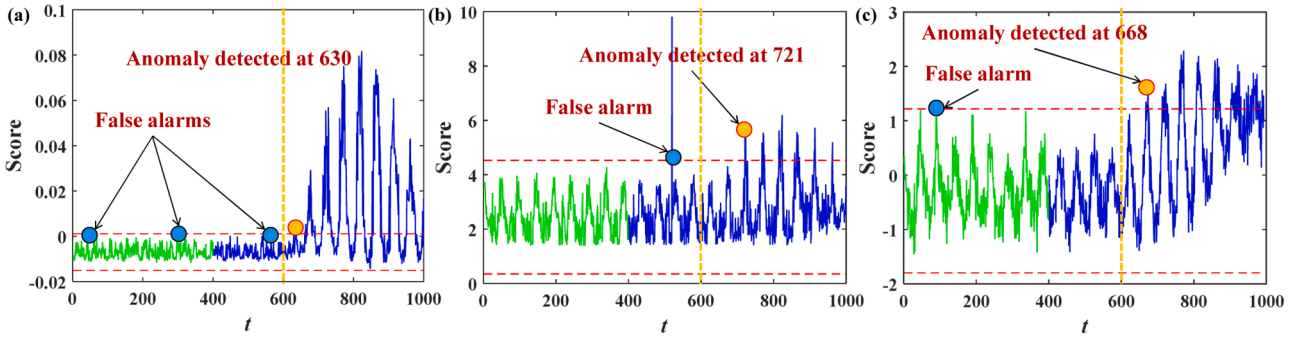

Fig. 12. The AD result of the comparative models in Case 1. (a) is the result of FDCVAE; (b) is the result of MRRAE; (c) is the result of DSMDA. The green line and blue line denote the training and testing stage, the red dashed line is the threshold, yellow dashed line is the dividing line between truly normal and abnormal.

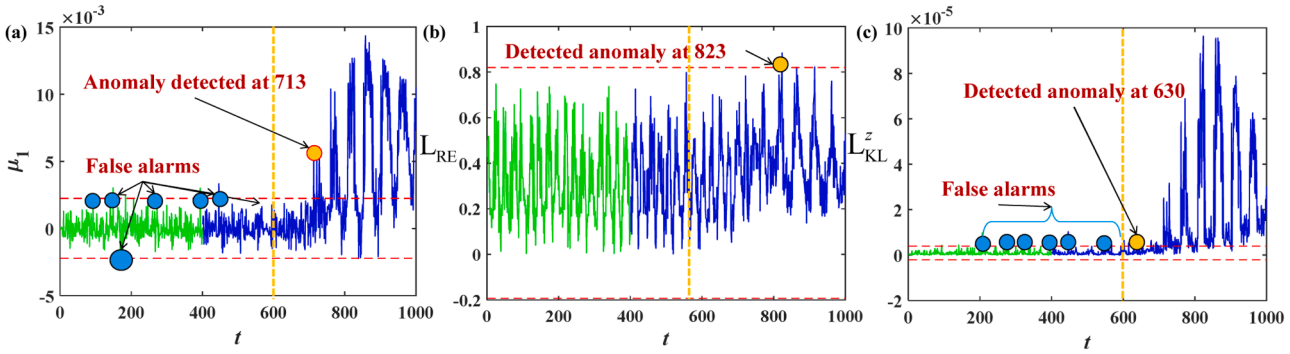

Fig. 13. The AD result of the comparative model VAE. (a) is mean feature of latent feature z; (b) and (c) are the reconstruction loss and KL loss. The green line and blue line are the training and testing stage, red dashed line is the threshold. The yellow dashed line is the dividing line between truly normal and abnormal.

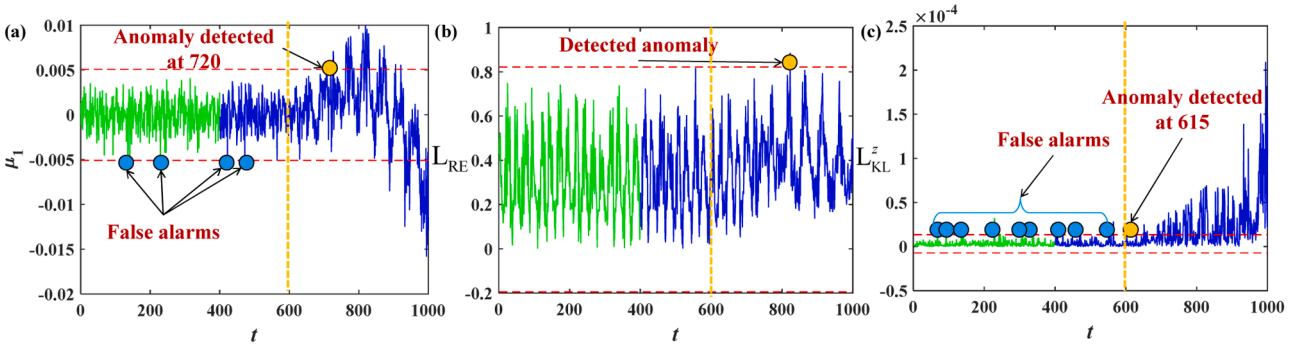

Fig. 14. The AD result of the comparative model β-VAE. (a) is mean feature of latent feature z; (b) and (c) are the reconstruction loss and KL loss. The green line and blue line are the training and testing stage, red dashed line is the threshold. The yellow dashed line is the dividing line between truly normal and abnormal.

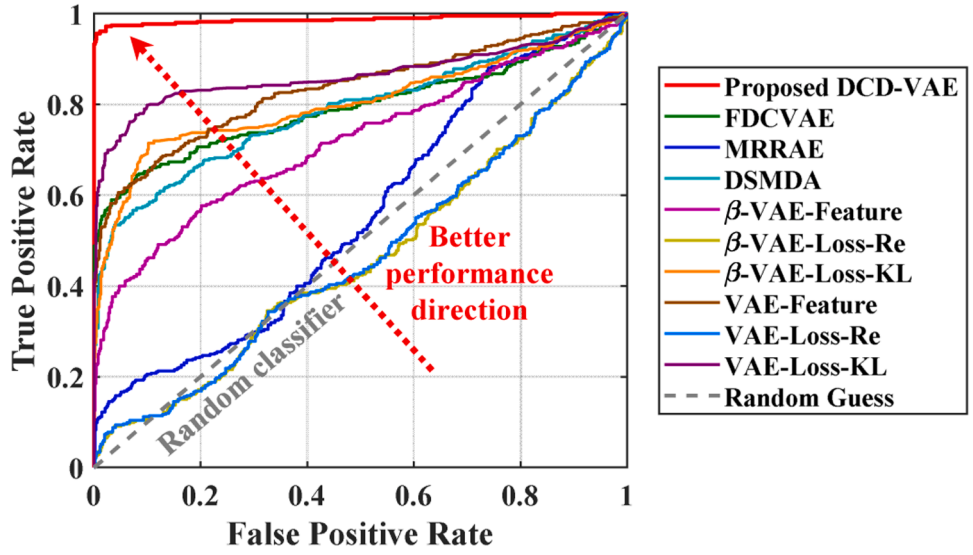

Fig. 15. ROC curves of the proposed method and comparative methods in Case 1.

demonstrate that the commonly used method based on reconstruction loss is ineffective in these AD circumstances. Additionally, the method based on the KL loss function exhibits a high sensitivity to noise, resulting in a significant presence of false alarms.

The AD results based on the mean feature and loss functions of comparative model β-VAE are shown in Fig. 14. Similar to Fig. 13, these constructed ANIs are also affected by TVOCs. Particularly, the RE-based ANI, as shown in Fig. 13(b), remains the most sensitive to TVOCs. Issues such as false alarms and missing detection also arise in the detection results based on β-VAE. As a result, the proposed approach utilizing the disentangled feature construction for AD achieves superior performance. Specifically, the constructed ANI based on the disentangled feature facilitates early AD and results in fewer false alarms.

To better compare and analyze the AD performance of the proposed method against the comparative methods in this case study, Fig. 15 presents the receiver operating characteristic (ROC) curves of these methods. The horizontal and vertical axes represent the false positive rate (FPR) and true positive rate (TPR), respectively. The curves

Table 3
Evaluation metrics of AD. *TN* and *FN* denote true negative and false negative samples, *TP* and *FP* denote true positive and false positive.

Metrics	Formula	Definition
Accuracy (Acc)	$\frac{TP+TN}{TP+FP+FN+TN}$	Percentage of correctly classified samples out of the total number of samples.
Precision (Pr)	$\frac{TP}{TP+FP}$	Percentage of true positive samples among all samples predicted as positive.
Recall (*Re*)	$\frac{TP}{TP+FN}$	Percentage of true positive samples among all actual positive samples.
F1 Score	$2\times\frac{\mathrm{Pr}\times\mathrm{Re}}{\mathrm{Pr}+\mathrm{Re}}$	The harmonic mean of Precision and Recall, balancing both metrics
False Positive Rate (FPR)	$\frac{FP}{TP+FN}$	Percentage of negative samples incorrectly predicted as positive out of all actual negative samples.
False Negative Rate (FNR)	$\frac{FN}{TP+FN}$	Percentage of positive samples incorrectly predicted as negative out of all actual positive samples.
AUROC	$\int_0^1 TPR(FPR)d(FPR)$	Area Under the Receiver Operating Characteristic curve, measuring the overall performance of a classifier across all threshold levels. $TPR=\frac{TP}{TP+FN}$ is true positive rate.

Table 4
Evaluation results of the proposed method and other rivals in Case 1. The up and down arrows indicate the optimal corresponding metric value. The bolded numbers in the table indicate the best results in each column of metrics.

Metrics / Models	Acc (%)↑	Pr (%)↑	*Re* (%)↑	F1 (%)↑	FPR (%)↓	FNR (%)↓	AUROC (%)↑
FDCVAE	78.80	98.45	47.45	64.30	0.52	52.25	77.98
MRRAE	62.40	96.15	6.25	11.73	0.16	93.75	55.64
DSMDA	69.48	98.98	24.50	39.27	0.16	75.50	77.25
β-VAE-Feature	76.20	96.55	42.00	58.53	1.00	58.00	80.58
β-VAE-Loss-*Re*	60.20	**100**	0.50	0.99	**0.00**	99.50	46.19
β-VAE-Loss-KL	77.80	93.51	50.50	65.58	2.33	49.50	83.84
VAE-Feature	75.20	96.34	39.50	56.02	1.00	60.50	82.03
VAE-Loss-*Re*	60.20	**100**	0.50	0.99	**0.00**	99.50	46.28
VAE-Loss-KL	77.80	94.05	47.50	63.12	2.00	52.50	86.67
DCD-VAE	**92.40**	98.21	**82.50**	**89.67**	0.50	**17.50**	**97.87**

illustrate the trends of FPR and TPR values under different anomaly thresholds. The red dashed arrow in the figure indicates the direction of the optimal ROC curve, where the Area Under the ROC curve (AUROC) is 1. From the results shown in the figure, it is evident that the proposed DCD-VAE method exhibits superior AD performance compared to the benchmark methods. Additionally, the figure reveals that the performance of MRRAE, *β*-VAE-Loss_*Re*, and VAE-Loss_*Re* can be considered akin to that of random classifiers in this case, as they fail to effectively achieve AD. Moreover, to more accurately compare the performance of various methods in this AD case, a series of quantification metrics for binary classification is further introduced, as detailed in Table 3.

The evaluation results of the comparison are shown in Table 4, the proposed DCD-VAE model demonstrates superior performance across multiple dimensions, highlighting its efficacy in identifying anomalies. Specifically, advantages were achieved over other methods in terms of accuracy, recall, F1 score, false negative rate, and AUROC. In addition, it can be found that *β*-VAE-Loss-*Re* and VAE-Loss-*Re* achieve better accuracy and false positive rate. However, combined with the false negative rate and the results in Fig. 13 and Fig. 14, it can be found that the results are almost unable to detect the targeted anomalies, and there are a large number of fault negatives so it is at the bottom of the F1 score in the comprehensive evaluation metric. Besides, the proposed DCD-VAE model does not optimize the accuracy and false alarm rate, but it still maintains a good level.

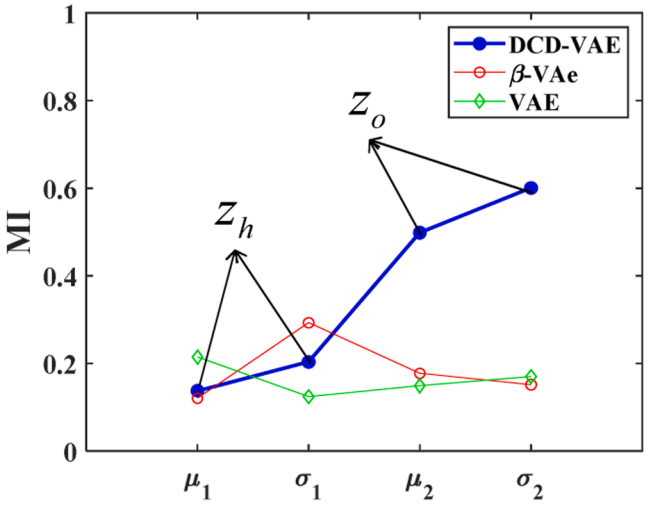

Fig. 17. Comparison of MI between features and E_t in different networks.

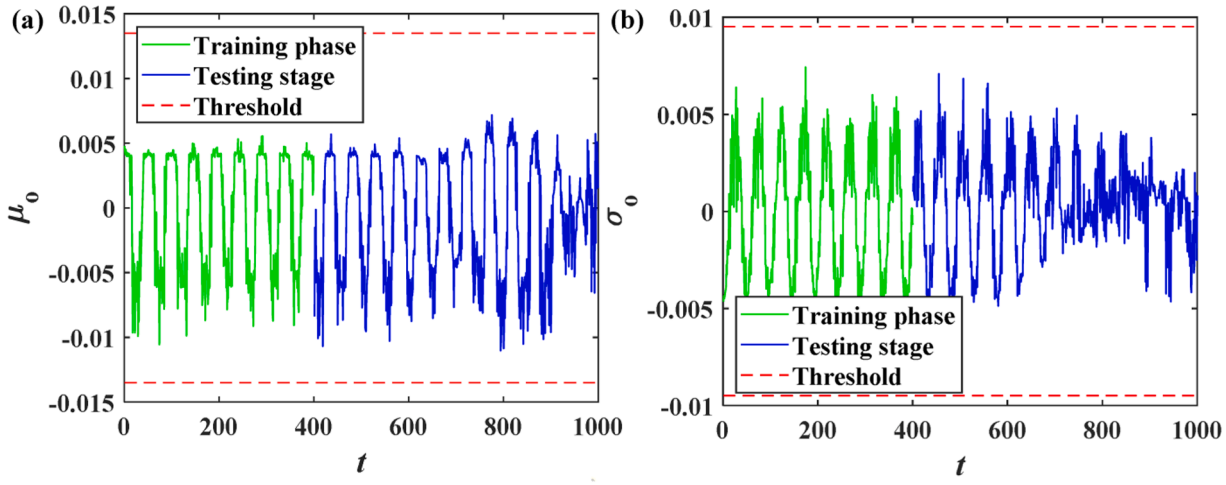

Fig. 16. The variation of values associated with $\mathbf{z_o}$ throughout the trajectory. (a) is the mean feature; (b) is the variance feature.

The mean and variance features related to $\mathbf{z_o}$ are further visualized in Fig. 16. It can be observed that they exhibit a striking resemblance to the trends of environmental parameters E_t, displaying similar fluctuations. Hence, it can be inferred that feature $\mathbf{z_o}$ carries information about the environment. To quantitatively analyze this phenomenon, the MI between the two decoupled features (four distributional statistics) and the environmental parameters E_t are presented in Fig. 17. It can be observed that the features belonging to $\mathbf{z_o}$ exhibit higher MI with E_t compared to all features in the comparative methods. Meanwhile, the MI related to $\mathbf{z_s}$ remains at a relatively low level, further demonstrating the effective disentanglement of features achieved by DCD-VAE.

4.1.4. Parameters sensitivity analysis

As discussed in Section 3.2, the weighting parameters β_o and β_s not only restrict the total amount of information in the features $\mathbf{z_o}$ and $\mathbf{z_s}$, but also control the feature disentanglement performance of the model. Given that the proposed method involves tuning these two parameters, it is essential to analyze their sensitivity to AD results for a deeper understanding of the model's characteristics and to refine the proposed method. This section analyzes the AD performance of the proposed method by setting different combinations of the parameters (β_o, β_s), where the performance is quantified using the comprehensive evaluation metric F1 score introduced in the previous section. Specifically, the range of β_o is (1e-6: 4), and the range of β_s is (1:4). The resulting parameter combinations are illustrated in Fig. 18.

In Fig. 18(b), the red dashed line represents the case where $\frac{\beta_s}{\beta_o} = 1$, corresponding to the parameter settings in VAE and β-VAE. In both Fig. 18(a) and Fig. 18(b), the yellow regions indicate $\beta_s > \beta_o$ while the blue regions indicate $\beta_s < \beta_o$. The other hyperparameter settings for this experiment are consistent with those in Section 4.1.2. The heatmap illustrating parameter sensitivity, shown in Fig. 19, reveals that when the parameter combination (β_s,β_o) is in the yellow region depicted in Fig. 18(a) and (b), the F1 score in Fig. 19(a) achieves better results. Conversely, when the parameter combination (β_s,β_o) falls within the blue region shown in Fig. 18(b), a clear drop in the F1 score can be observed at the red dashed line ($\frac{\beta_s}{\beta_o} = 1$) in Fig. 19(b). This indicates that when $\beta_s < \beta_o$, the $\mathbf{z}_s$ controlled by the smaller parameter β_s fails to achieve accurate AD, this further validates that the proposed DCD architecture can indeed guide the flow of different information through varying constraints, thereby disentangling relevant features and ultimately achieving accurate AD under TVOCs.

Additionally, Fig. 20 presents the two-dimensional projection of Fig. 19. It can be observed that under different parameter combinations of the proposed method from Fig. 20(a), the F1 score remains at a relatively high level, with the minimum value of 0.8317 still significantly outperforming the results of the comparison methods listed in Table 4. The range remains low (Range=0.0888), indicating that the sensitivity of the two most critical parameters to the final AD result is relatively low. Fig. 20(b) presents the two-dimensional projection results under the linear scale of β_o, further illustrating the abrupt change in the F1 score values along the parameter combination (β_s,β_o) distribution, with a clear boundary indicated by the red dashed line. The deep blue section of the F1 score in the lower right of Fig. 20(b) represents the AD performance evaluation results for parameters satisfying $1 < \beta_s < \beta_o$. This section is symmetrically opposed to the yellow F1 score values in the upper left part of the figure. This symmetry indicates that in these cases, $\mathbf{z_s}$ is essentially the same as $\mathbf{z_o}$ in the $\beta_s > \beta_o$ region of the upper left part. Consequently, it can be observed that the disentangled feature $\mathbf{z_o}$ derived from DCD-VAE fails to achieve accurate AD.

4.2. Case 2

4.2.1. Experiment details

The experimental data of bearings was employed as the validation for the proposed method in this case study. The experimental setup consists of the test rig, data acquisition devices (NI-9232), and a host computer. The host computer controlled the OCs of the test rig and collected the monitoring data. As depicted in Fig. 21(a), the test rig comprises an alternating current motor (labeled as 1), a key phase sensor for measuring the rotating speed (labeled as 2), two support bearings (labeled as 3), two hydraulic loading systems for axial and radial loading (labeled as 4,6), and the testing deep groove ball bearings (ER-16 K). An accelerometer (PCB068A11) was mounted in the radial direction of the bearing housing, as shown in Fig. 21(b). TVOCs were set up for testing, as demonstrated in Fig. 22(a), the simulated OCs variations over the entire life cycle of the bearings. The rotating speed remained constant for 10 h at each speed. The test bearings were subjected to a constant axial load of 0.5 kN and a radial load of 4 kN since the focus of this study was on the time-varying rotating speed.

The vibration signals were sampled at a frequency of 20.48 kHz, with a sampling interval of 10 min. Each sample contained 3 s of vibration data. The vibration data generated by the testing bearing during its respective full-service lives, as shown in Fig. 22(b). A total of 894 sets of vibration data were collected, with each set containing 3 s of data. Therefore, the overall shape of the entire dataset is 894 × 3 × 20,480. Different types of faults occurred on the testing bearing, as shown in Fig. 23. Compared with preset files for OCs in Fig. 22(a), the monitored

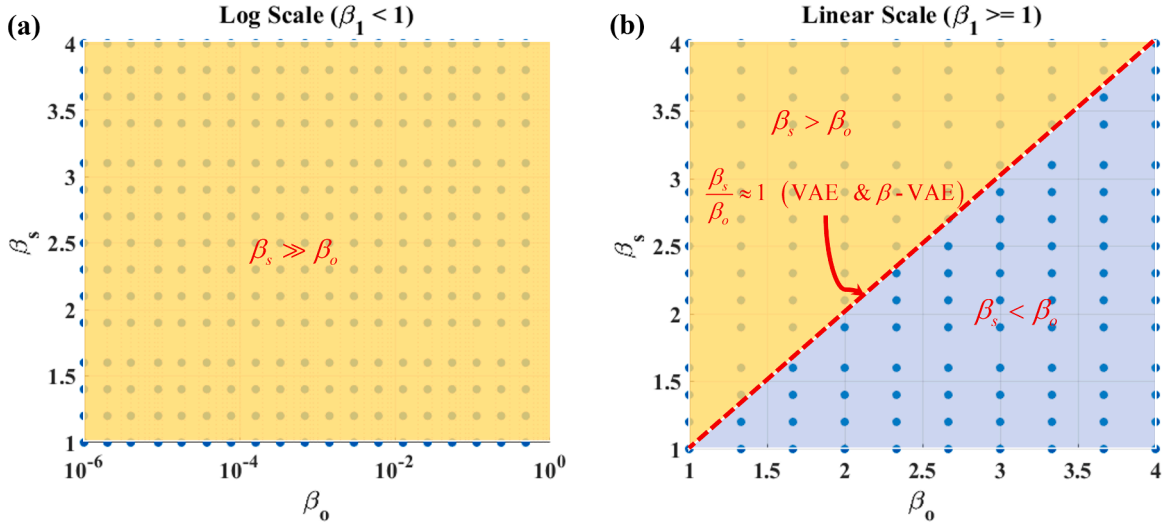

Fig. 18. The combination of β_s and β_o utilized in parameters sensitivity analysis for Case 1.

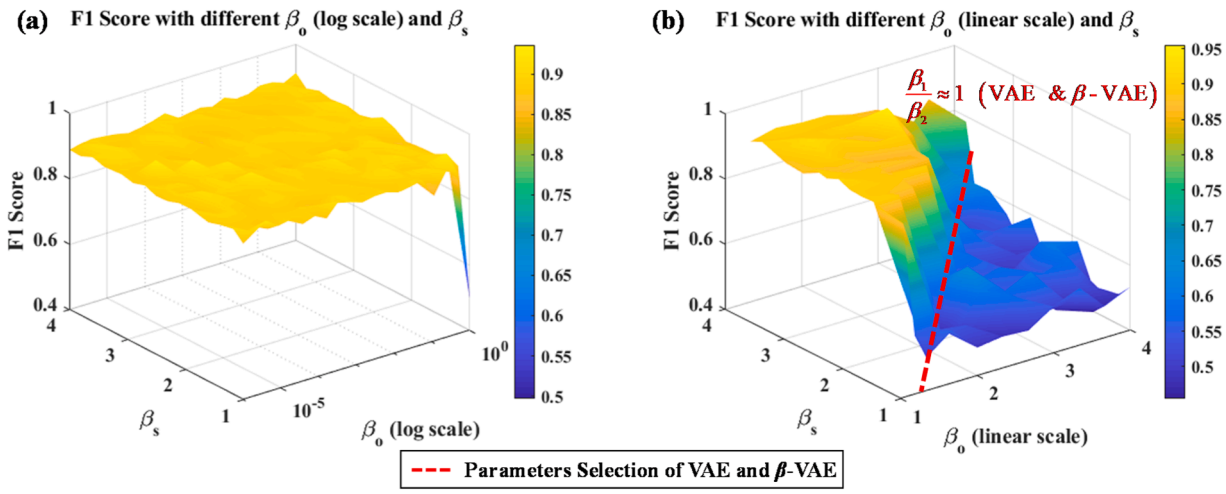

Fig. 19. CE performance (F1 Scores) for different combinations of parameters β_s and β_o in Case 1.

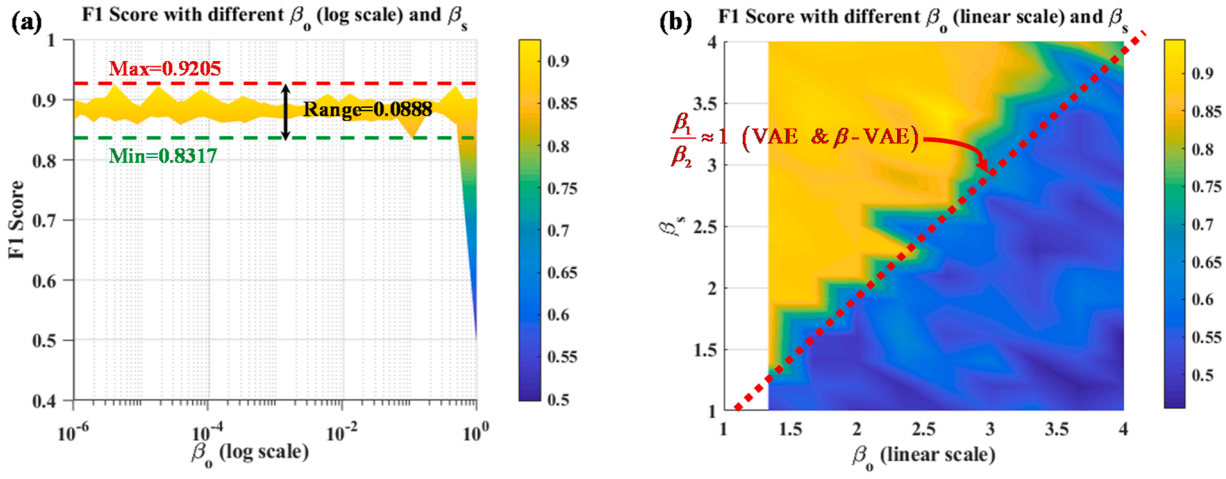

Fig. 20. CE Performance (F1 Scores) for Different Combinations of Parameters β_s and β_o(2D version).

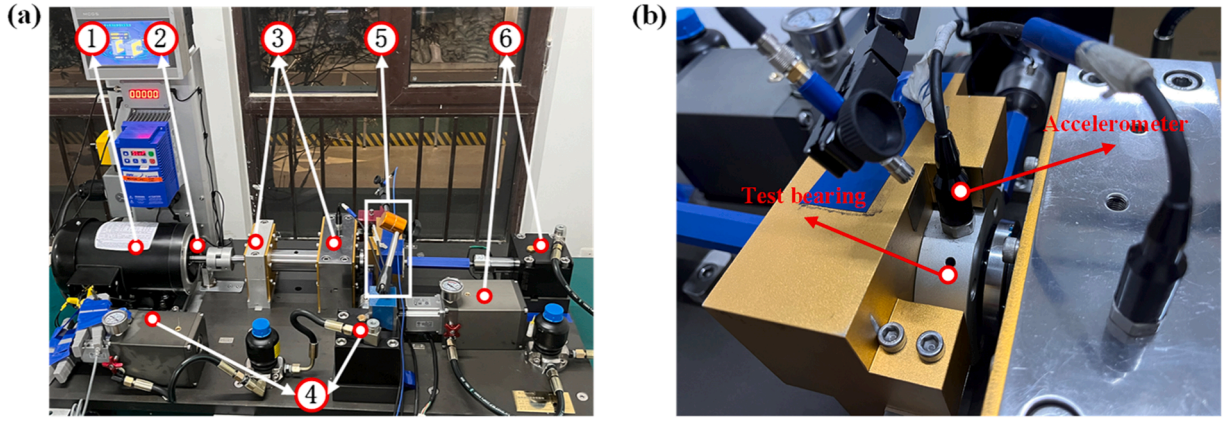

Fig. 21. The specific test rig setup of the experiment. (a) is the test rig; (b) is the accelerometer and the testing bearing setup.

signal reveals the presence of components similar to the variations in OCs, indicating the vibration signals are influenced by the OCs and manifested in the vibration time-domain signal.

4.2.2. *The result of the proposed method*

The network structure and hyperparameters of DCD-VAE are maintained the same as in Case 1. The main difference is that the trade-off parameters of the loss function need to be redetermined. Following the guiding strategy mentioned in DCD, specific values 0.5 and 2 were obtained through experimental tuning. As shown in Fig. 24, we divide the whole monitored vibration signal into training and testing datasets at time point 250, the former is used as training data, and the rest is used

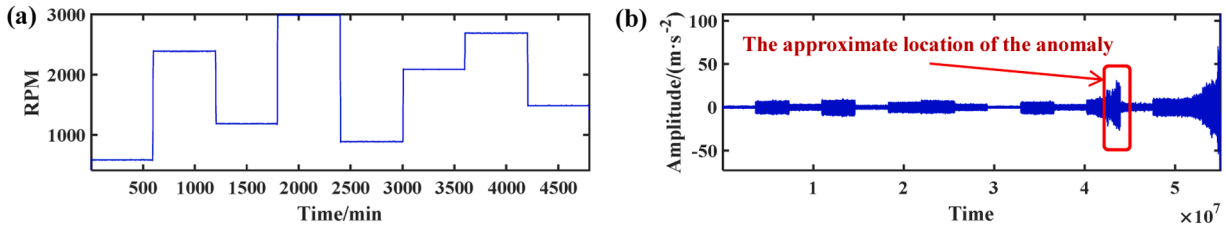

Fig. 22. Schematic diagram of the TVOCs setting and monitored vibration signal in the experiment. (a) represents the OCs setting of B1, and (b) is the vibration signal during the whole life cycle of the tested bearing.

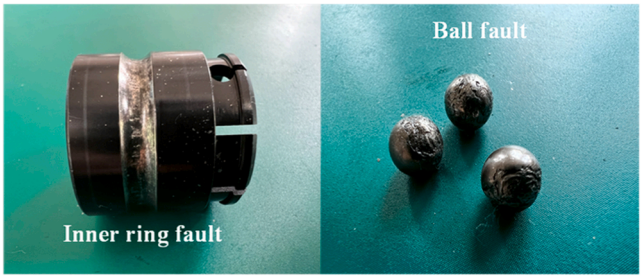

Fig. 23. Inner ring fault and ball fault found on the testing bearings after the test.

as testing data. The training data does not include data recorded after the occurrence of faults, thus satisfying the conditions for unsupervised AD. Besides, as indicated by the speed values in Fig. 24, the ranges of OCs included in the training and test sets differ. This variation is important for evaluating the model's generalization performance to different unseen OCs.

The training data are input into DCD-VAE for training and the training loss is shown in Fig. 25(a). It can be observed that the reconstruction loss and the KL loss associated with the correlated states feature $\mathbf{z_s}$ gradually converge as the number of epochs increases. However, due to the loose constraint resulting from smaller weights, the KL loss related to the correlated OCs feature $\mathbf{z_o}$ exhibits continuous fluctuations throughout the entire training process. Moreover, the mean and variance features of the latent feature z are visualized, and the whole training data and part of the testing data before time point 650 consist of the visualized data, as shown in Fig. 25(b). The visualized features in the figure demonstrate a clear separation between the distributions of mean and variance, indicating an effective disentanglement of the related OSs and OCs. Furthermore, the test data points within the range of 400–650 are distributed within the existing clusters of the training set, providing

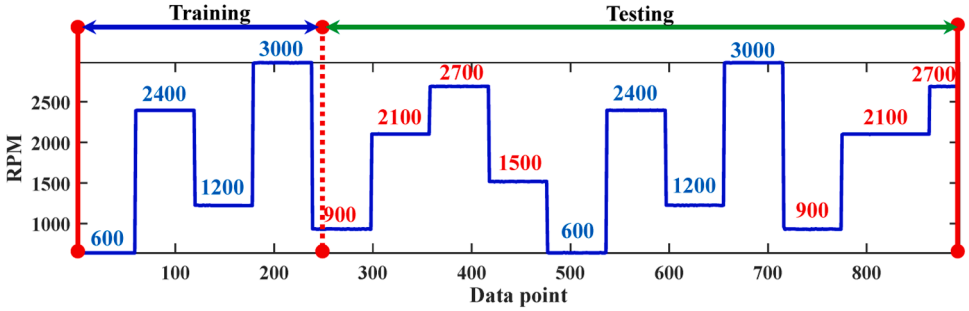

Fig. 24. Schematic diagram of the division of the training test dataset used in the experiment. The red numbers labeled therein indicate the test data generated under the OCs that are hardly covered in the training set.

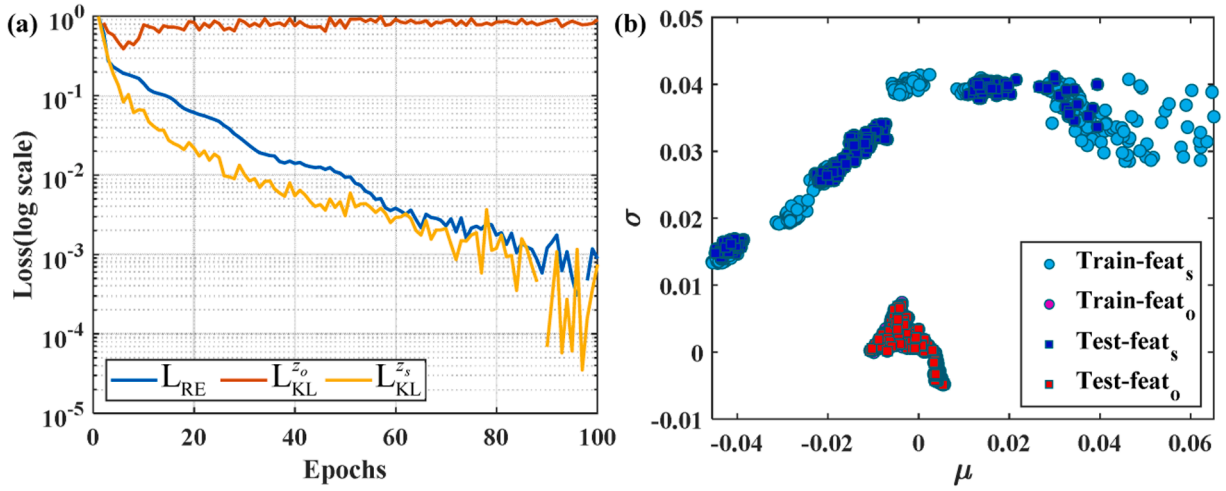

Fig. 25. The curve depicting the variation of the loss function during the training process, along with the two-dimensional visualization of the mean and variance features of latent feature z. (a) is the training loss during the 100 epochs training; (b) is the feature visualization of the mean and variance features.

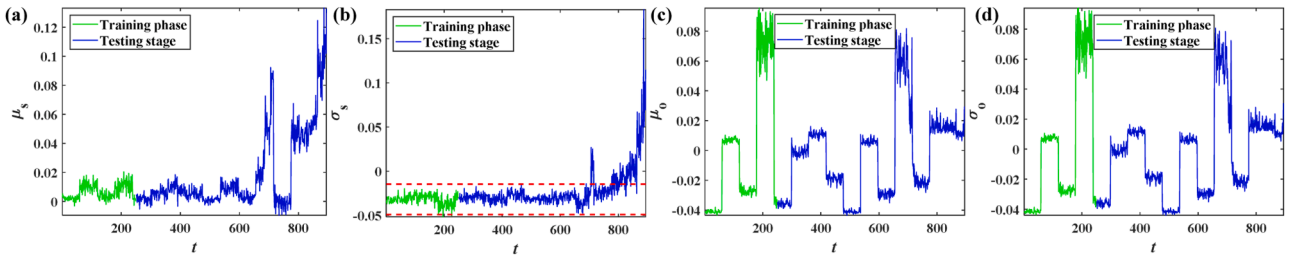

Fig. 26. Feature visualization of the mean and variance feature of latent feature $\mathbf{z}_s$ and $\mathbf{z}_o$. (a) and (b) are the first principle component of mean and variance feature of $\mathbf{z}_s$; (c) (d) are the same visualization while belonging to $\mathbf{z}_o$.

further validation of the successful feature disentanglement achieved by the proposed model.

To achieve the AD of bearing, the first principle component of each feature after the dimension reduction through PCA is shown in Fig. 26. The parameters μ_s and σ_s are associated with the latent feature $\mathbf{z_s}$, while μ_o and σ_o are associated with the latent feature $\mathbf{z_o}$. It is evident from Fig. 26(a) and Fig. 26(b) that the interference caused by TVOCs is almost eliminated. Moreover, Fig. 26(c) and Fig. 26(d) demonstrate a strong correlation between the OCs and the features. This indicates that the OCs' information and the OSs' information in the signal are effectively separated within the information bottleneck $\mathbf{z_s}$ and $\mathbf{z_o}$, while ensuring a decrease and convergence of reconstruction loss. The ANI for bearing AD is constructed as shown in Fig. 27, it can be observed that during the training and initial testing stages, as indicated by the braces in the figure, the constructed ANI through DCD-VAE overcomes the interference caused by TVOCs' information. By applying a 3- 3-sigma rule, the initial anomalies related to bearing faults can be effectively detected. The zoomed-in view in Fig. 27(b) reveals that with the automatically set threshold, an initial anomaly occurring at data point 675 is detected.

To provide better clarification, the mechanism of the rotating machinery fault is introduced. Fig. 28(a) shows the vibration signal in the time domain, and Fig. 28(b) presents its corresponding spectrum. For comparison, the monitored data at data point 674, just before the detected anomaly at data point 675, are shown in Fig. 28(c) and 28(d). It can be observed that there are no significant differences between these two data points in both the time and frequency domains.

To further analyze the specific manifestation of the anomalies detected by the proposed method, we introduce the classical spectral kurtosis filtering method [41], which is widely used for analyzing weak fault features in rotating machinery. The filtering results of these two signals are shown in Fig. 29.

In Fig. 29(a) and 29(b), the components related to the rotational frequency are prominently highlighted in the filtered signals, and there is a significant presence of impulsive components in the time domain. In contrast, Fig. 29(c) and Fig. 29(d) show that the impulsive components related to the rotational frequency are not as abundant. This indicates that the proposed AD method identifies a subtle rotational frequency impact anomaly at data point 675, which can be considered a rubbing anomaly caused by increased friction due to internal bearing wear.

Furthermore, attention should be paid to the content highlighted in red in Fig. 29(c) and 29(d). Some impulsive components begin to appear in the time domain, accompanied by frequency clusters centered around the rotational frequency and its harmonics in the envelope spectrum. This suggests that the fault has already begun to show initial signs, although these signs are not as pronounced as the significant rotational frequency impact observed at data point 675.

4.2.3. Comparison study

The comparison research is conducted in this case. In addition to the various data-driven comparative models and methods used in Case 1, this section introduces some health indicators(HIs) based on expert knowledge as an extra comparative method for AD, considering the specificity of bearing objects. The details are as follows:

Data-driven based: FDCVAE, MRRAE, DSMDA, and conventional loss-error-based one and latent-feature-based one(VAE-Feature, VAE-Loss-*Re*, VAE-Loss-KL, VAE-Feature, β-VAE-Loss-*Re*, β-VAE-Loss-KL, β-VAE-Feature).

Expert-knowledge-based: the HI used in bearing fault detection based on expert knowledge is involved, which consists of Root-Mean-Square(RMS), Kurtosis [1], Gini index [1], SESGI [44], Hoyer [45], entropy [1].

The comparison results of the data-driven method are shown in Fig. 30, Fig. 31, and Fig. 32. The results of FDCVAE in Fig. 30(a) demonstrate the enhanced resistance to disturbances from TVOCs

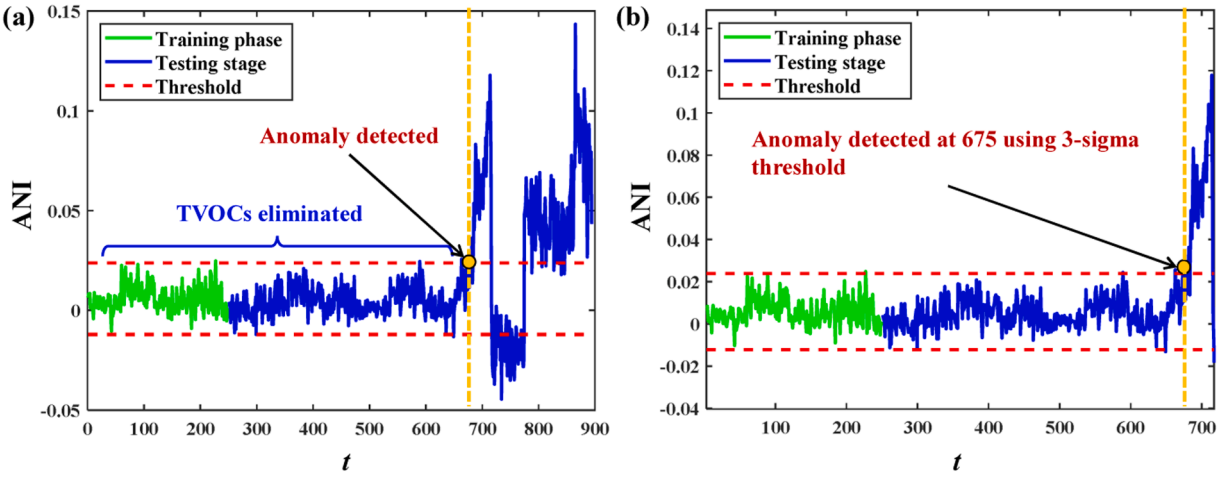

Fig. 27. The constructed ANI and the AD results of the proposed method in Case 2. (a) is the result of the AD of bearing; (b) is an enlarged version of the figure, and clearly indicates the anomaly.

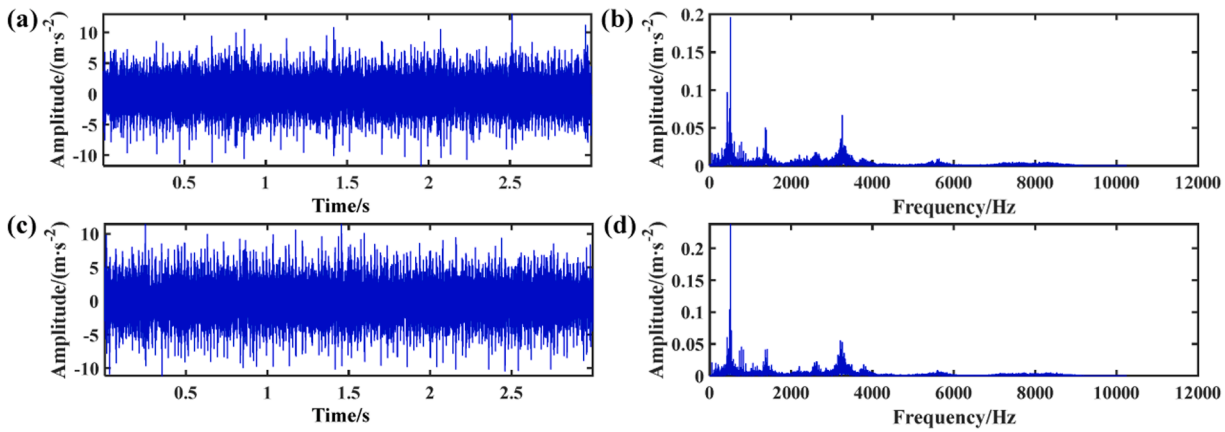

Fig. 28. The monitored vibration signal and its spectrum at detected anomaly data point 675 and data point 674. (a) and (b) are the time-domain signal and corresponding spectrum of 675; (c) and (d) are the corresponding 674.

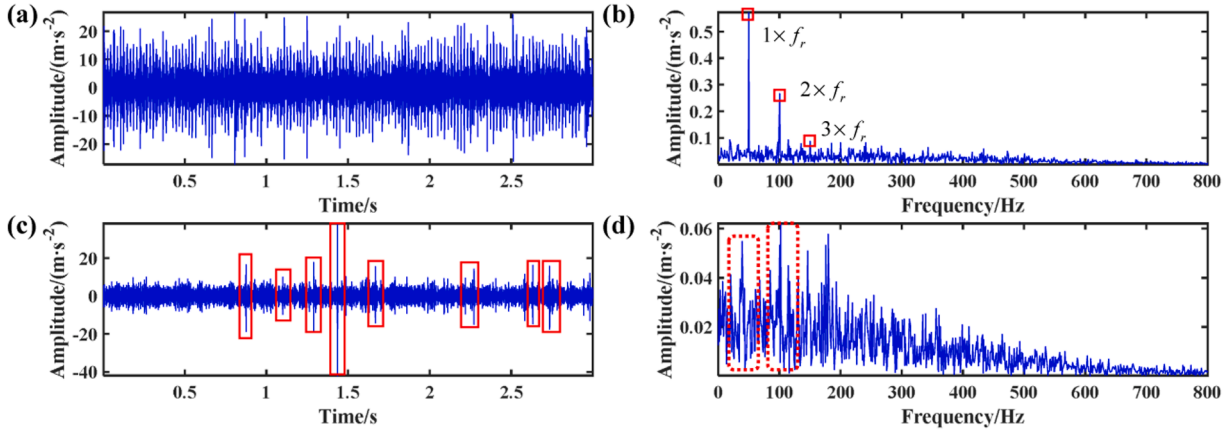

Fig. 29. The filtering result of the detected anomaly data point 675 and data point 674. (a) and (b) are the filtered signal and envelop spectrum of 675; (c) and (d) are the corresponding data point 674.

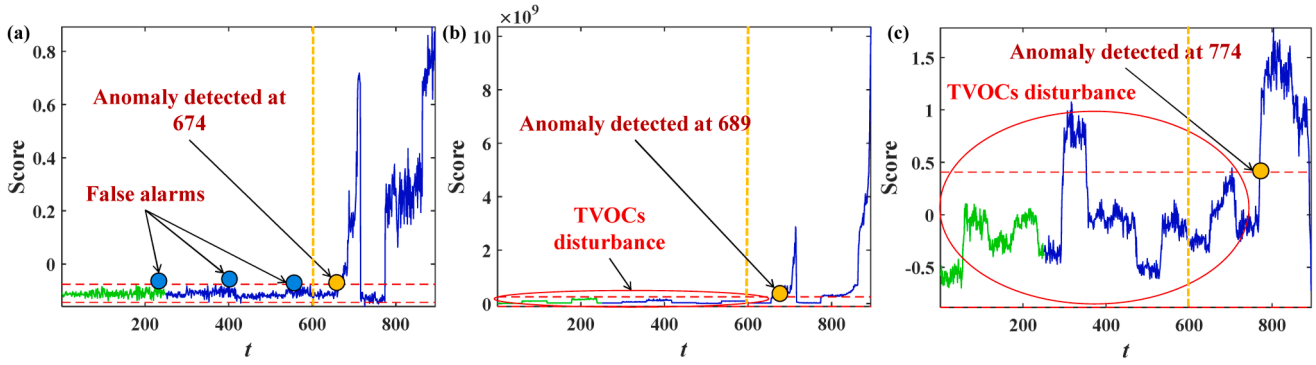

Fig. 30. The AD result of the comparative models in Case 2. (a) is the result of FDCVAE; (b) is the result of MRRAE; (c) is the result of DSMDA. The green line and blue line denote the training and testing stage, the red dashed line is the threshold, yellow dashed line is the dividing line between truly normal and abnormal.

compared to the other two models, facilitating earlier detection of anomalies and thus reducing missed detections. However, it also exhibits false positives, unlike the method proposed in this paper. Both the MRRAE and DSMDA models are more sensitive to TVOCs, and show delayed AD, which results in a series of missed detections. In Fig. 31 and Fig. 32, the AD results of the comparative models VAE and β-VAE exhibit certain similarities. Both the latent-feature-based and KL-loss-based ANI fail to detect the correct anomalies, with ANIs being extremely susceptible to OCs and becoming completely ineffective under the disturbances of TVOCs. Although the RE-based ANI in both models can initially detect anomalies at data point 689, it still exhibits latency in AD compared to the method proposed in this paper, leading to missed alarms. Furthermore, as illustrated in the figures, there is a significant presence of disturbance components from TVOCs.

Fig. 33 illustrates the AD results of the expert-knowledge-based methods. It can be observed that these methods are significantly affected by TVOCs, exhibiting a chaotic and time-varying trend. Consequently, it is challenging to accurately determine the occurrence

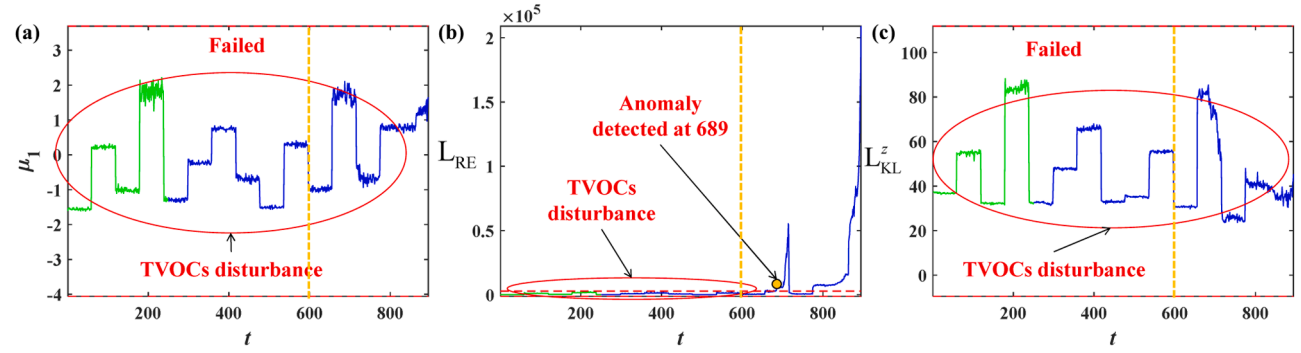

Fig. 31. Comparison results of different ANI based on VAE for AD of Tested bearing. (a) is the latent mean feature statistics;(b) is the reconstruction loss; (c) is the KL divergence loss. The green line and blue line are the training and testing stage, the red dashed line is the threshold, yellow dashed line is the dividing line between truly normal and abnormal.

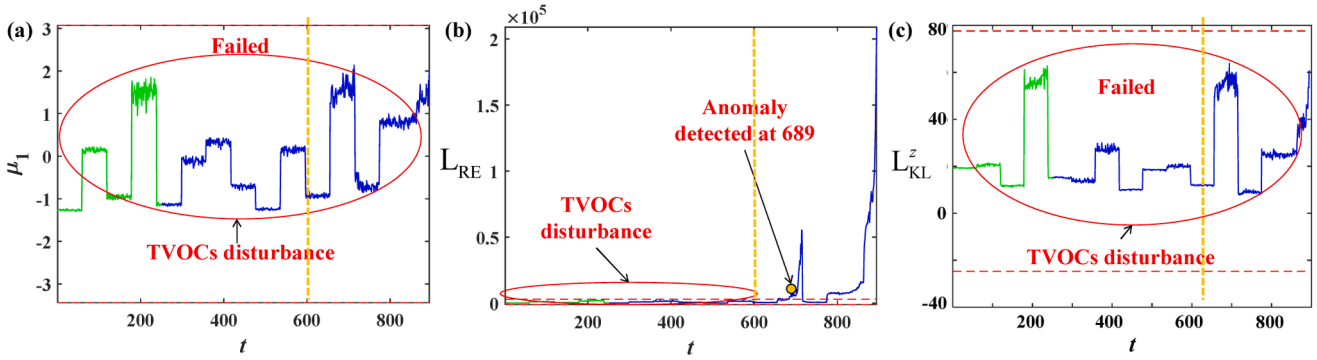

Fig. 32. Comparison results of different ANI based on β-VAE for AD of Tested bearing. (a) is the latent mean feature statistics;(b) is the reconstruction loss; (c) is the KL divergence loss. The green line and blue line are the training and testing stage, the red dashed line is the threshold, yellow dashed line is the dividing line between truly normal and abnormal.

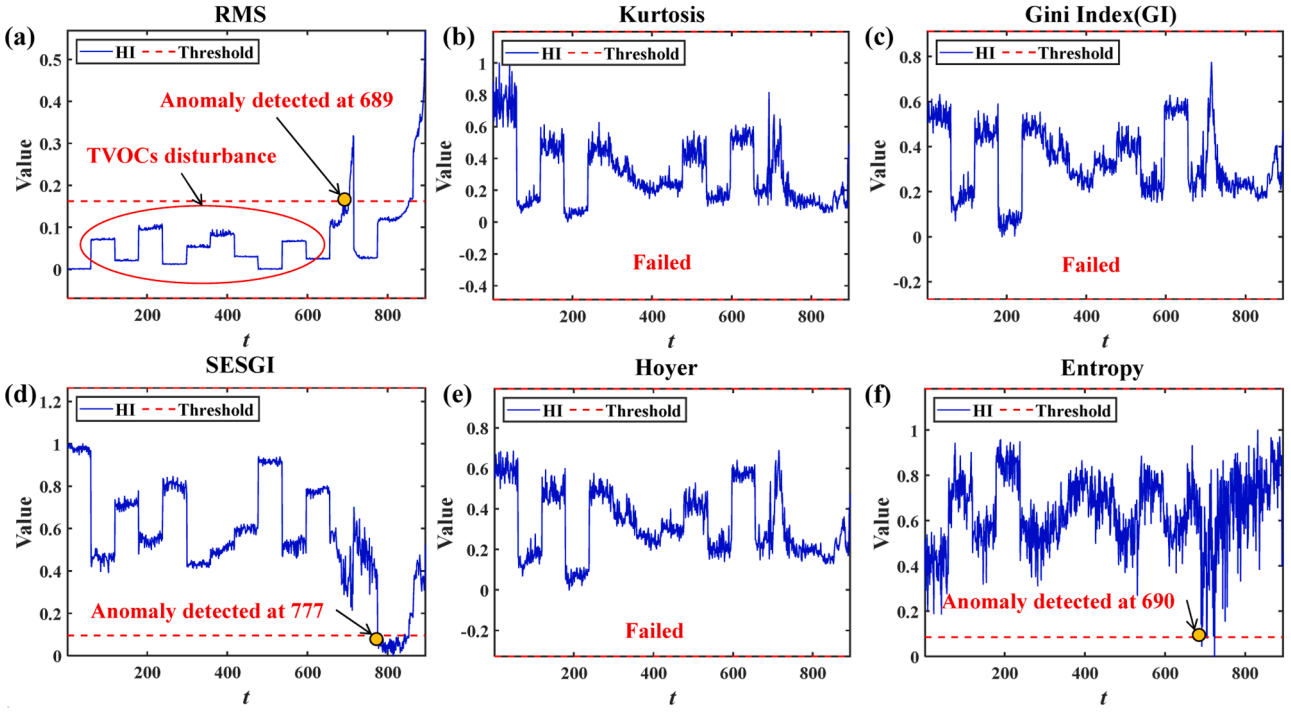

Fig. 33. Comparative results of different HIs belonging to the expert-knowledge-based group. Each name of a HI is marked on the top of each subplot.

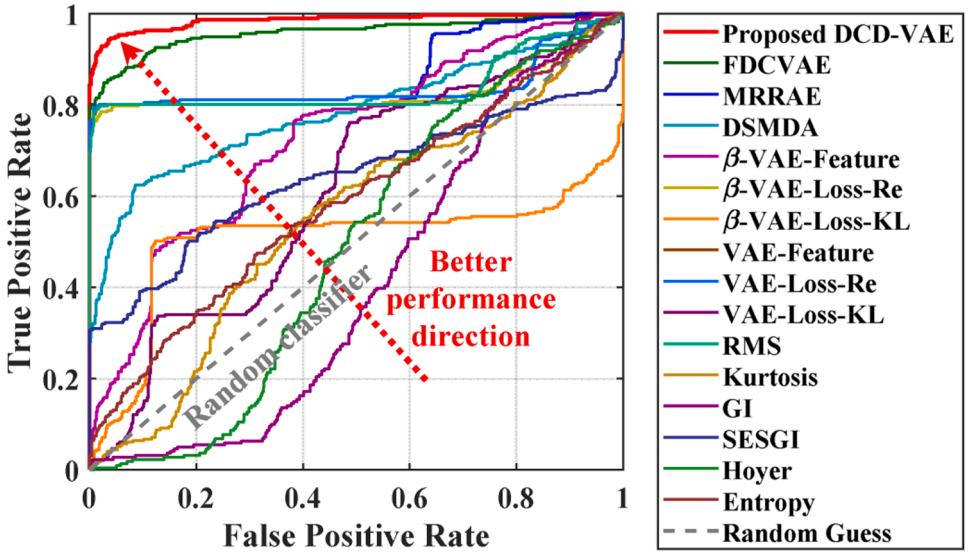

Fig. 34. ROC of the proposed method and comparative methods in Case 2.

time of anomalies based on visual inspection alone. Furthermore, after calculating the threshold value using the 3-sigma rule, Kurtosis, Gini index, and Hoyer completely fail to detect anomalies in this scenario. Additionally, even though methods such as RMS, SESGI, and Entropy are capable of detecting so-called anomalies, they also exhibit noticeable lags in detecting the initial anomalies, Thus the missing alarms are remained in these cases.

To effectively demonstrate the AD performance of the proposed method in Case 2, Fig. 34 similarly presents the ROC curves of various methods as in Case 1. It is evident that the proposed DCD-VAE method consistently outperforms the comparative methods, maintaining higher true positive rates across nearly all false positive rate levels. This superior performance is particularly noticeable compared to methods such as MRRAE, β-VAE-Loss_*Re*, and VAE-Loss_*Re*, which exhibit characteristics akin to random classifiers, as indicated by their proximity to the random classifier line. The DCD-VAE's ability to achieve significantly better ROC characteristics, demonstrated by its curve closely hugging the top-left corner, underscores its effectiveness in minimizing false positives while maximizing true positives. This robust performance validates the DCD-VAE's efficacy in AD and its superiority over other comparative methods in Case 2.

To further quantitatively compare the AD performance of comparative methods in Case 2 against the proposed method, this section continues to employ the evaluation metrics from Table 3 in Case 1. Considering the analysis around data point 674 in Section 4.2.2, and acknowledging that bearing faults do not exhibit self-healing, we define point 674 as the normal-to-anomalous transition boundary for the entire dataset (i.e., data from 1 to 674 as normal and 675–894 as anomalous). The final results are presented in Table 5.

Results from Table 5 reveal that the proposed DCD-VAE model achieves the best outcomes on multiple evaluation metrics compared to its competitors. Even in metrics where it does not reach the best results, it still performs satisfactorily. Notably, the proposed method achieves the highest F1 score, a comprehensive measure of AD performance. Combined with other evaluation metrics, these results demonstrate the superiority of the DCD-VAE and its corresponding AD methodology over its rivals under TVOCs.

Table 5
Evaluation results of the proposed method and other rivals in Case 2. The up and down arrows indicate the optimal corresponding metric value. The bolded numbers in the table indicate the best results in each column of metrics.

Metrics Models/ HIs	Acc (%)↑	Pr (%)↑	*Re* (%)↑	F1 (%)↑	FPR (%)↓	FNR (%)↓	AUROC (%)↑
FDCVAE	93.40	94.98	82.82	88.48	1.93	17.17	97.01
MRRAE	92.47	97.86	77.10	86.25	0.74	22.89	87.36
DSMDA	77.15	**100**	25.92	41.17	**0.00**	74.07	77.54
β-VAE-Feature	75.16	**100**	0.00	0.00	**0.00**	100	71.28
β-VAE-Loss-*Re*	82.68	99.56	76.43	86.47	0.14	23.56	83.16
β-VAE-Loss-KL	75.16	**100**	0.00	0.00	**0.00**	100	54.85
VAE-Feature	75.16	**100**	0.00	0.00	**0.00**	100	59.78
VAE-Loss-*Re*	92.88	97.89	78.45	87.10	0.74	21.54	83.61
VAE-Loss-KL	75.16	**100**	0.00	0.00	**0.00**	100	59.78
RMS	79.58	**100**	33.55	50.25	**0.00**	66.44	84.07
Kurtosis	75.16	**100**	0.00	0.00	**0.00**	100	60.86
GI	75.16	**100**	0.00	0.00	**0.00**	100	40.86
SESGI	74.22	**100**	16.10	27.74	**0.00**	83.89	63.82
Hoyer	75.16	**100**	0.00	0.00	**0.00**	100	48.59
Entropy	70.10	**100**	2.68	5.22	**0.00**	97.31	58.58
DCD-VAE	**95.77**	98.85	**87.24**	**92.69**	0.44	**12.75**	**98.55**

4.2.4. *Parameters sensitivity analysis*

This section continues to analyze the sensitivity of the key parameters β_s and β_o in the proposed DCD-VAE model to AD results. We first reintroduce the parameter settings from Section 4.1.4 in Case 1, as shown in Fig. 18, while keeping the training and network hyperparameters of the DCD-VAE consistent with those in Section 4.2.2 of Case 2. Consequently, the heatmap of the F1 Score for AD as a function of β_s and β_o in this case is presented in Fig. 35.

Same as observed in Case 1, when $\beta_s > \beta_o$, the proposed model achieves a desirable F1 Score. Additionally, as illustrated in Fig. 35(b), there is a significant jump in the F1 Score at $\frac{\beta_s}{\beta_o}=1$, indicating poorer AD performance when $\beta_s < \beta_o$. This is consistent with the results obtained in Case 1, further validating the effectiveness of the proposed DCD-based feature disentanglement in improving AD performance under TVOCs.

Fig. 36 presents a two-dimensional perspective, showing that for various parameter combinations satisfying $\beta_s > \beta_o$, the maximum F1 Score is 0.9625, and the minimum is 0.8789, with a range of 0.0836. The minimum value of 0.8789 is only slightly lower than the result of the comparative method FDCVAE in Table 5. Overall, all possible parameter combinations for the proposed method maintain a high level of F1 Score

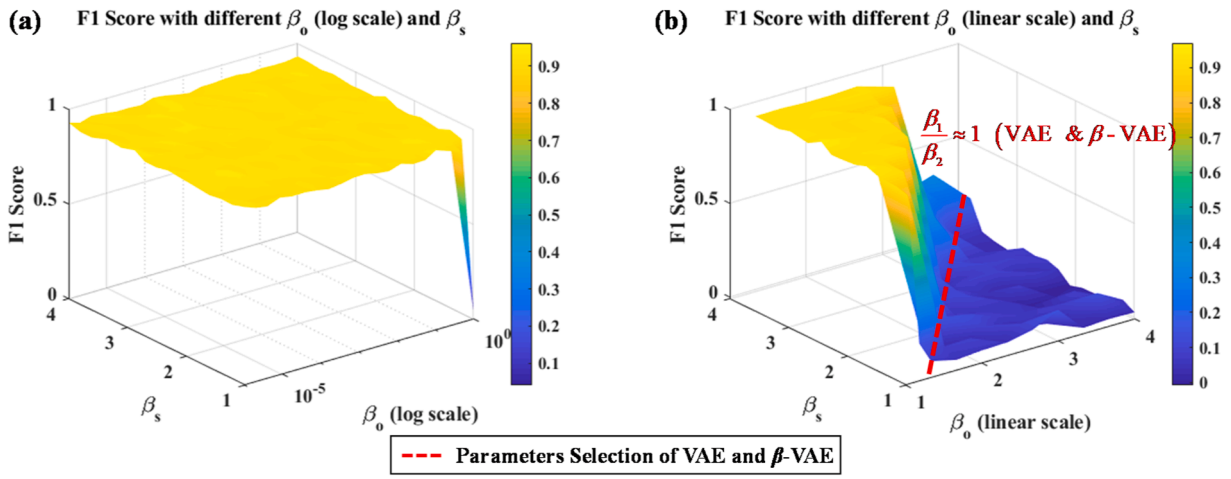

Fig. 35. CE Performance (F1 Scores) for different combinations of parameters β_s and β_o in Case 2.

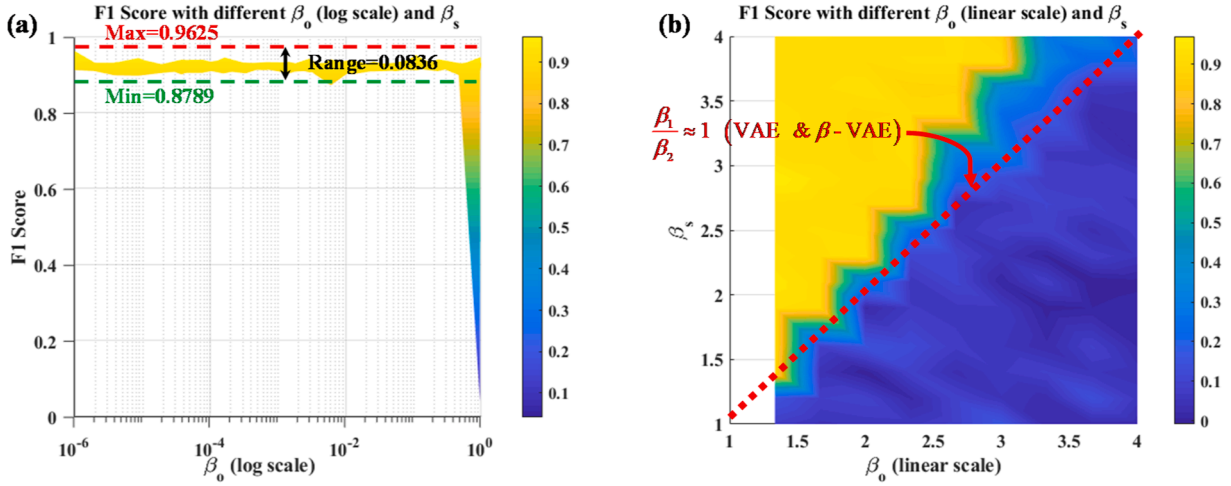

Fig. 36. CE Performance (F1 Scores) for Different Combinations of Parameters β_s and β_o (2D version).

compared to the comparative methods, indicating that the proposed method can sustain excellent AD stability across different parameter settings within a certain range. Fig. 36(b) also clearly demonstrates the F1 Score boundary at $\frac{\beta_s}{\beta_o} = 1$, confirming that the model achieves good AD performance when $\beta_s > \beta_o$.

5. Conclusion

This paper starts from the perspective of mechanical equipment CM and introduces the challenges of AD in machines under TVOCs due to the existence of the MPN problem within the context of unsupervised AD techniques. It explores and investigates a novel AD approach that does not require the introduction of OCs' information to address this challenge. To achieve this, DRL is incorporated into this study. After analyzing the strengths and limitations of disentanglement learning based on β-VAE, a novel unsupervised representation learning model called DCD-VAE is proposed. The model aims to disentangle the OCs' information from the OSs' information in the machine monitoring data. ANI for AD based on the disentangled OSs features, is constructed to detect anomalous states in machines. Simulation and full-lifetime experiments on bearing validate the effectiveness of the proposed model and methods. In comparison to conventional data-driven methods and expert-knowledge-based approaches, the proposed method demonstrates significant superiority. It provides an effective technical approach for unsupervised AD in mechanical equipment under TVOCs, especially in scenarios with limited OCs' information. Future research can explore the application of the proposed model in predicting the remaining useful life of machinery under the influence of TVOCs. Additionally, since the experimental case introduced in this study involves basic rotating components, further investigation into the application of this model for AD in more complex machinery using multiple monitoring data is also warranted.

CRediT authorship contribution statement

Haoxuan Zhou: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Bingsen Wang:** Writing – review & editing, Software, Methodology, Data curation, Conceptualization. **Enrico Zio:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zihao Lei:** Writing – review & editing, Methodology, Investigation, Data curation, Conceptualization. **Guangrui Wen:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Xuefeng Chen:** Writing – review & editing, Supervision, Resources, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This research was partly supported by The National Key Research and Development Program of China (No.2020YFB1710002). The work of Bingsen Wang presented in this article is supported by the Centre for Advances in Reliability and Safety (CAiRS), Hong Kong SAR, China admitted under AIR@InnoHK Research Cluster.

Data availability

Data will be made available on request.

References

[1] Zhou H, Huang X, Wen G, Lei Z, Dong S, Zhang P, et al. Construction of health indicators for condition monitoring of rotating machinery: a review of the research. Expert. Syst. Appl. 2022:203.
[2] Zio E. Prognostics and Health Management (PHM): where are we and where do we (need to) go in theory and practice. Reliab. Eng. Syst. Saf. 2022;218.
[3] Xiao Y, Shao H, Wang J, Yan S, Liu B. Bayesian Variational Transformer: a generalizable model for rotating machinery fault diagnosis. Mech. Syst. Signal. Process. 2024;207:110936.
[4] González-Muñiz A, Diaz I, Cuadrado AA, García-Pérez D. Health indicator for machine condition monitoring built in the latent space of a deep autoencoder. Reliab. Eng. Syst. Saf. 2022;224:108482.
[5] Zhou H, Wang B, Zio E, Wen G, Liu Z, Su Y, et al. Hybrid system response model for condition monitoring of bearings under time-varying operating conditions. Reliab. Eng. Syst. Saf. 2023:239.
[6] Zhang C, Hu D, Yang T. Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost. Reliab. Eng. Syst. Saf. 2022;222:108445.
[7] Yan S, Shao H, Min Z, Peng J, Cai B, Liu B. FGDAE: a new machinery anomaly detection method towards complex operating conditions. Reliab. Eng. Syst. Saf. 2023;236:109319.
[8] Jiao J, Zhao M, Lin J, Liang K. Hierarchical discriminating sparse coding for weak fault feature extraction of rolling bearings. Reliab. Eng. Syst. Saf. 2019;184:41–54.
[9] Ma C, Li Y, Wang X, Cai Z. Early fault diagnosis of rotating machinery based on composite zoom permutation entropy. Reliab. Eng. Syst. Saf. 2023;230:108967.
[10] Zhou H, Li H, Liu T, Chen Q. A weak fault feature extraction of rolling element bearing based on attenuated cosine dictionaries and sparse feature sign search. ISA Trans. 2020;97:143–54.
[11] Zhang X, Feng Y, Chen J, Liu Z, Wang J, Huang H. Knowledge distillation-optimized two-stage anomaly detection for liquid rocket engine with missing multimodal data. Reliab. Eng. Syst. Saf. 2024;241:109676.
[12] Yan H, Li F, Chen J, Liu Z, Wang J, Feng Y, et al. A Graph embedded in graph framework with dual-sequence input for efficient anomaly detection of complex equipment under insufficient samples. Reliab. Eng. Syst. Saf. 2023:109418.
[13] Yang Z, Baraldi P, Zio E. A method for fault detection in multi-component systems based on sparse autoencoder-based deep neural networks. Reliab. Eng. Syst. Saf. 2022;220.
[14] Zheng M, Man J, Wang D, Chen Y, Li Q, Liu Y. Semi-supervised multivariate time series anomaly detection for wind turbines using generator SCADA data. Reliab. Eng. Syst. Saf. 2023:235.
[15] Lu S, Dong H, Yu H. Abnormal condition detection method of industrial processes based on cascaded bagging-PCA and CNN classification network. IEEe Trans. Industr. Inform. 2023:1–11.
[16] Pang G, Shen C, Cao L, Hengel AVD. Deep Learning for Anomaly Detection, 54. ACM Computing Surveys; 2021. p. 1–38.
[17] Le Cun Y., Fogelman-Soulié F.J.I. Modèles connexionnistes de l'apprentissage. 1987;2:114–43.
[18] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., et al. Generative Adversarial Networks, 1–9. 2014.
[19] Yang Z, Baraldi P, Zio E. A method for fault detection in multi-component systems based on sparse autoencoder-based deep neural networks. Reliab. Eng. Syst. Saf. 2022;220:108278.
[20] Qin Y., Zhou J., Chen DJIAToM. Unsupervised health indicator construction by a novel degradation-trend-constrained variational autoencoder and its applications. 2021;27:1447–56.
[21] Li Z, Sun Y, Yang L, Zhao Z, Chen X. Unsupervised machine anomaly detection using autoencoder and temporal convolutional network. IEEe Trans. Instrum. Meas. 2022;71:1–13.
[22] Xu F, Yang F, Fan X, Huang Z, Tsui KL. Extracting degradation trends for roller bearings by using a moving-average stacked auto-encoder and a novel exponential function. Measurement 2020;152.
[23] Qin Y., Yang J., Zhou J., Pu H., Mao YJAEI. A new supervised multi-head self-attention autoencoder for health indicator construction and similarity-based machinery RUL prediction. 2023;56:101973.
[24] Han P, Ellefsen AL, Li G, Holmeset FT, Zhang H. Fault Detection with LSTM-based variational autoencoder for maritime components. IEEe Sens. J. 2021;21:21903–12.
[25] Chen J, Li J, Chen W, Wang Y, Jiang T. Anomaly detection for wind turbines based on the reconstruction of condition parameters using stacked denoising autoencoders. Renew. Energy 2020;147:1469–80.
[26] Zhou H, Lei Z, Zio E, Wen G, Liu Z, Su Y, et al. Conditional feature disentanglement learning for anomaly detection in machines operating under time-varying conditions. Mech. Syst. Signal. Process. 2023:191.
[27] Higgins I., Amos D., Pfau D., Racaniere S., Matthey L., Rezende D., et al. Towards a definition of disentangled representations. 2018.
[28] Wang X., Chen H., Tang Sa, Wu Z., Zhu W. Disentangled representation learning..
[29] Tran L, Yin X, Liu X. Disentangled representation learning gan for pose-invariant face recognition. Proc. IEEE Conference Computer Vision Pattern Recognition 2017:1415–24.
[30] Deng W, Zhao L, Liao Q, Guo D, Kuang G, Hu D, et al. Informative feature disentanglement for unsupervised domain adaptation. IEEe Trans. Multimedia 2021;24:2407–21.
[31] Liu D, Zhang C, Song Y, Huang H, Wang C, Barnett M, et al. Decompose to adapt: cross-domain object detection via feature disentanglement. IEEe Trans. Multimedia 2022;25:1333–44.
[32] Kingma D.P., Welling MJapa. Auto-encoding variational bayes. 2013.
[33] Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. beta-vae: learning basic visual concepts with a constrained variational framework. Int. Confer. Learning Representations 2017.
[34] Peng X, Huang Z, Sun X, Saenko K. Domain agnostic learning with disentangled representations. In: International Conference on Machine Learning. PMLR; 2019. p. 5102–12.
[35] Odaibo SJapa. Tutorial: deriving the standard variational autoencoder (vae) loss function. 2019.
[36] Chen R.T., Li X., Grosse R.B., Duvenaud DKJAinips. Isolating sources of disentanglement in variational autoencoders. 2018;31.
[37] Shwartz-Ziv R., Tishby NJapa. Opening the black box of deep neural networks via information. 2017.
[38] Wang R, Huang W, Wang J, Shen C, Zhu Z. Multisource domain feature adaptation network for bearing fault diagnosis under time-varying working conditions. IEEe Trans. Instrum. Meas. 2022;71:1–10.
[39] Ding Y, Jia M, Zhuang J, Cao Y, Zhao X, Lee CG. Deep imbalanced domain adaptation for transfer learning fault diagnosis of bearings under multiple working conditions. Reliab. Eng. Syst. Saf. 2023;230:108890.
[40] Masci J, Meier U, Cireşan D, Schmidhuber J. Stacked convolutional auto-encoders for hierarchical feature extraction. International Conference on Artificial Neural Networks. Springer; 2011. p. 52–9.
[41] Huang X, Wen GR, Dong SZ, Zhou HX, Lei ZH, Zhang ZF, et al. Memory residual regression autoencoder for bearing fault detection. IEEe Trans. Instrum. Meas. 2021;70:1–12.
[42] Ou XL, Wen GR, Huang X, Su Y, Chen XF, Lin HL. A deep sequence multi-distribution adversarial model for bearing abnormal condition detection. Measurement 2021;182.
[43] Cannarile F., Compare M., Baraldi P., Yang Z., Zio EJRfE-Phwe-poACp. The aramis challenge: prognostics and health management in evolving environments. 2020.
[44] Miao Y, Wang J, Zhang B, Li H. Practical framework of Gini index in the application of machinery fault feature extraction. Mech. Syst. Signal. Process. 2022:165.
[45] Hoyer POJJomlr. Non-negative matrix factorization with sparseness constraints. 2004;5.