Energy & Buildings 246 (2021) 111044

Contents lists available at ScienceDirect

Energy & Buildings

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

A semi-supervised approach to fault detection and diagnosis for building HVAC systems based on the modified generative adversarial network

Bingxu Li [a,b], Fanyong Cheng [a,c], Hui Cai [d], Xin Zhang [e], Wenjian Cai [a,*]

[a] School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore
[b] Energy Research Institute @ NTU (ERI@N), Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore
[c] Key Laboratory of Advanced Perception and Intelligent Control of High-end Equipment, Anhui Polytechnic University, Wuhu 241000, China
[d] College of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou 310018, China
[e] College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 5 January 2021
Revised 20 March 2021
Accepted 19 April 2021
Available online 26 April 2021

Keywords:
Semi-supervised learning
Fault detection and diagnosis
Generative adversarial network
Building HVAC system
Imbalanced learning

ABSTRACT

Developing efficient fault detection and diagnosis (FDD) techniques for building HVAC systems is important for improving buildings' reliability and energy efficiency.

The existing FDD methods can achieve satisfying results only if there are sufficient labeled training data. However, labelling the data is often costly and laborious, and most data collected in practice are unlabeled. Most of the existing FDD methods cannot leverage the unlabeled dataset which contains much information beneficial to fault classification, and this will impede the improvement of the FDD performance. To deal with this problem, a semi-supervised FDD approach is proposed for the building HVAC system based on the modified generative adversarial network (modified GAN). The binary discriminator in the original GAN is replaced with the multiclass classifier. After the modification, both the unlabeled and labeled datasets can be utilized simultaneously: the modified GAN can learn the data distribution information present in unlabeled samples and then combine this information with the limited number of labeled data to accomplish a supervised learning task. Additionally, a novel self-training scheme is proposed for the modified GAN to correct the class imbalance in both labeled and unlabeled data. With the self-training scheme, the modified GAN can still efficiently exploit the information contained in unlabeled data to enhance the FDD performance even if the class distribution is highly imbalanced. Experimental results demonstrate the effectiveness of the proposed modified GAN-based approach and the self-training scheme.

1. Introduction

Buildings are responsible for about 20–40% of the total final energy consumption in developed countries [1]. The heating, ventilation, and air conditioning (HVAC) system is the most important part of the building system and accounts for a large portion of energy consumption in commercial or residential buildings [2]. Unfortunately, many HVAC systems in real buildings may operate below their rated efficiency due to various operational faults. The faults may affect many key variables of the system, such as, the coefficient of performance (COP), the power consumption of the compressor, etc. [3]. Literature [3] presents detailed description of the common faults and their effects on the system. Some faults are difficult to diagnose, and the system may operate with faulty states for a long time, which will waste significant energy. It is estimated that about 10%~40% of the energy can be saved by applying fault detection and diagnosis (FDD) strategies to HVAC systems [4]. In addition, FDD strategies can also help to reduce the maintenance costs and the system downtime, which can improve the reliability of the building systems. Hence, it is of great significance to develop accurate, reliable, and automated fault detection and diagnosis strategies for the building HVAC systems.

Many studies have been devoted to developing FDD methods for HVAC systems recently. Among these methods, data-driven methods receive the most attention since data-driven methods are based on the historical operational data and do not require developing the physical models which may be complex and time-consuming [5]. Another reason for the popularity of data driven methods is that buildings are becoming more data rich with the development of internet of things (IoT) technologies and

* Corresponding author.
E-mail address: ewjcai@ntu.edu.sg (W. Cai).

https://doi.org/10.1016/j.enbuild.2021.111044
0378-7788/

Nomenclature

Abbreviations

BMS	Building management system
D	Discriminator network
FDD	Fault detection and diagnosis
G	Generator network
GAN	Generative adversarial network
HVAC	Heating, ventilation, and air conditioning system
NN	neural network
PCA	Principle component analysis
RTU	Rooftop unit
SVM	Support vector machine
t-SNE	t-distributed stochastic neighbor embedding
VAV	Variable air volume

Symbols

$D(\cdot)$	the function of the discriminator
$G(\cdot)$	the function of the generator
K	the number of classes
$o_j(x)$	the output value of the j^{th} neuron in the output layer
x	the input data sample
y	one-hot label of the sample x
z	the random noise vector input to the generator
$zp(\tilde{z})$	the random noise vector drawn from the Gaussian distribution
χ^L	the labeled dataset
χ^U	the unlabeled dataset
θ_c	parameters of the multiclass classifier network
θ_d	parameters of the discriminator
θ_g	parameters of the generator

building management systems (BMS), which means that more and more building operational data are available. The data-driven methods can recognize the patterns of different faults from the massive data and then diagnose the faults accurately and reliably. Recent related works on the data driven FDD methods of HVAC systems are reviewed in the following part.

Most existing FDD methods on HVAC systems are supervised learning methods, which utilize the labeled data samples to train the fault classifiers. Labeled data is a group of data samples which have been tagged with labels. The label of a data sample indicates whether this sample belongs to the normal class or the faulty classes, and which of the faulty classes this sample belongs to. In 2013, Zhao *et al* [6] developed a data driven FDD method for the chiller systems based on the Bayesian Belief Network theory. Posterior probabilities of different faults are calculated based on the probability analysis and graph theories. In [7], Zhao *et al* proposed a data-driven fault detection method for the chiller system using the support vector data description (SVDD) algorithm. The minimum-volume hypersphere is found to enclose most data under the normal condition and the data sample lying outside of this hypersphere will be detected as the faulty sample. In 2014, Du *et al* [8] proposed a FDD approach for building HVAC systems using the combined neural networks and subtractive clustering analysis. The combined networks are used to detect the abnormalities while the clustering analysis is used to classify different faulty conditions. Yan *et al* [9] presented a hybrid data-driven FDD method for chiller systems which incorporates auto-regressive model with exogenous variables and support vector machines. Zhao *et al* extended their work in [7] and proposed a chiller FDD method in [10] based on the SVDD algorithm. Different hyperspheres are trained to enclose most data of the corresponding faulty class. In 2016, Li *et al* [11] proposed a FDD approach for chiller systems using the principle component analysis (PCA) and SVDD. A SVDD model is developed in the residual subspace of the PCA. Li *et al* [12] designed a chiller FDD approach based on linear discriminant analysis (LDA). The data is projected into a lower dimensional space to achieve as large separation as possible between different faulty classes. In 2017, Wang *et al* [13] designed a FDD scheme for chiller systems by fusing distance rejection and multi-source information into the Bayesian network. This method can diagnose both the known faults and new types of faults. In 2018, Guo *et al* [14] proposed a fault diagnosis method for the variable refrigerant flow air-conditioning (VRF) system based on the deep belief network. The unsupervised layer can extract salient features, which can help to enhance the FDD performance. Huang *et al* [15] proposed a FDD approach for centrifugal chillers using associative classification algorithm. An associative classifier can be constructed by finding strong rules between fault classes and physical attributes. In 2019, Shahnazari *et al* [16] developed a fault detection and isolation (FDI) methods for HVAC systems using recurrent neural networks. The plant data are utilized to build predictive models and input/output estimators which are embedded within FDI filters. Lee *et al* [17] proposed a fault diagnostic model for air handling units (AHUs) based on deep neural networks. This model can achieve high diagnostic accuracy and can be used to conduct real-time maintenance and diagnosis of AHUs. In 2020, Guo *et al* [18] proposed a FDD method based on improved Gaussian mixture model for the VRF system. The PCA is used for dimensionality reduction to reduce large model complexity and long running time. Han *et al* [19] developed a FDD strategy for chiller systems, which merges simulated annealing into the deep neural network. This strategy can improve the diagnostic accuracy and the model stability. Zhou *et al* [20] proposed a comparison study on data driven FDD methods for VRF systems. Different data driven models were compared and analyzed. They found that SVM method is preferred for single fault diagnosis while the deep neural network is preferred for multiple fault diagnosis. In 2021, Yun *et al* [21] proposed a data driven FDD scheme for AHUs considering undefined states. The FDD model performs fault diagnosis only when it can perform significant inferences on input variables; otherwise, the model will consider the system state as undefined and feedback of the FDD model is performed by retraining of the inputs.

The above FDD methods are developed based on the supervised learning methods. Some researchers began to apply the unsupervised learning algorithms to this field and combined them with the supervised learning methods. In 2016, a semi-supervised FDD framework was proposed in [22] for HVAC water chillers. This framework exploits PCA to distinguish anomalies and uses a reconstruction-based contribution method to isolate variables related to faults. Then the fault diagnosis is tackled by the decision table. In 2018, a self-training-based FDD method was proposed in [23] for AHUs to cope with the sample imbalance issue that the faulty training samples are not sufficient. This method can enrich the training pool by iteratively inserting confidently labeled testing samples. In 2020, an emerging unsupervised deep learning model called generative adversarial network (GAN) was used by Yan *et al* to generate faulty training samples and re-balance the training dataset for the FDD problem [24,25].

Although current research works have achieved success in the FDD for HVAC systems, there still exists some problems which are not well addressed:

1) The existing FDD methods cannot make use of the unlabeled dataset. Most of the building operational data collected from BMS are completely unlabeled. We do not know whether these data belong to the faulty class or the normal class. Labelling these data is laborious and time-consuming and only the technicians with the expertise can finish the data labelling task in most circumstances. The most common scenario in practice is that the unlabeled dataset contains enough samples but the samples in the labeled dataset are insufficient to train an accurate classification model. Actually, the unlabeled dataset contains much information which is beneficial to fault classification. The existing methods cannot leverage this information, which significantly impedes the improvement of the FDD performance.
2) The existing methods cannot deal with the class imbalance in unlabeled data. Even if some research works have developed some novel FDD methods [23–25] which can mitigate the data imbalance, these methods only focus on the data imbalance in labeled data. In addition, most of existing semi-supervised algorithms assume a balanced class distribution for both labeled and unlabeled datasets. The imbalanced class distribution in unlabeled data will impair the ability of these semi-supervised algorithms to make use of the useful information contained in unlabeled data. The predictions will be heavily biased toward the majority class when the class distribution is highly imbalanced.

Therefore, it would be necessary and of great significance to develop a method which can fully leverage the unlabeled dataset and further reduce the reliance on the labeled samples. Additionally, the method is expected to correct the data imbalance in both labeled and unlabeled sets and can efficiently learn useful information from labeled and unlabeled datasets even if the class distribution of these datasets is highly imbalanced. Motivated by this, we propose a semi-supervised FDD approach for building HVAC systems based on the modified generative adversarial network (GAN). The binary discriminator in the original GAN is replaced with the multiclass classifier. After the modification, the modified GAN can learn the distribution information present in unlabeled data and then combine this information with the limited number of labeled data to accomplish a supervised learning task. Then, a novel self-training scheme is proposed for the modified GAN to correct the class imbalance in both labeled and unlabeled data. The main contributions of this study can be summarized and highlighted as follows:

(1) The proposed FDD method can efficiently make use of the labeled and unlabeled datasets. It can efficiently learn the data distribution information present in unlabeled data, which can help enhance the FDD performance. The proposed method fills the current research gap in the field of FDD of HVAC systems on how to leverage massive unlabeled samples.
(2) By learning the underlying patterns of the data from unlabeled samples, only a small number of labeled samples are required to assign the correct class labels. That is, the proposed method can mitigate the reliance on labeled training samples which are relatively difficult to obtain in the practical scenario.
(3) The proposed modified GAN with the self-training scheme can correct the class imbalance in both labeled and unlabeled datasets. It can still efficiently utilize labeled and unlabeled datasets even if the class distribution of these datasets is highly imbalanced. The self-training scheme can efficiently improve the robustness against the class imbalance in training data.

The remainder of this article is organized as follows. The background on semi-supervised learning is included in Section 2. The proposed FDD framework based on the modified generative adversarial network is described in Section 3. The proposed self-training scheme for the modified GAN is also presented in Section 3. In Section 4, the descriptions of the experimental data are presented. In addition, Section 4 shows the fault detection and diagnosis results of the proposed method and compares them with other methods. Finally, Section 5 concludes this article.

2. Background on semi-supervised learning

Supervised learning methods learn the function to map an input to an output based on the input–output sample pairs (x, y). x is the input of the sample and y is the label of the sample. Although very powerful supervised learning methods have been developed, they need enough labeled data samples to train. Supervised learning mainly includes two categories of algorithms: classification and regression. Common classification algorithms include support vector machine, neural networks, decision tree, etc. Unsupervised learning is a type of machine learning methods which learns the underlying hidden structure of the data. It only utilizes the data samples without any labels, i.e., the unlabeled data. Common unsupervised learning methods include cluster analysis, density estimation, dimensionality reduction, etc.

Semi-supervised learning falls between supervised learning and unsupervised learning. It can make use of both labeled data and unlabeled for training. The motivation of developing semi-

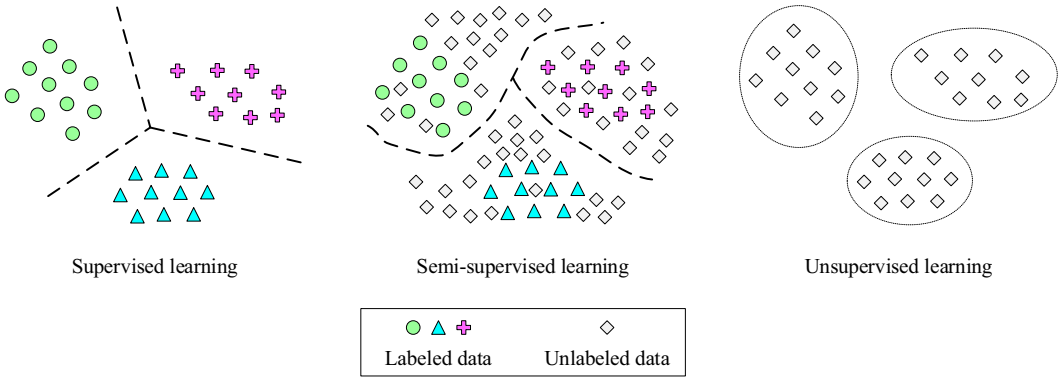

Fig. 1. Illustration of the difference between supervised, semi-supervised, and unsupervised learning.

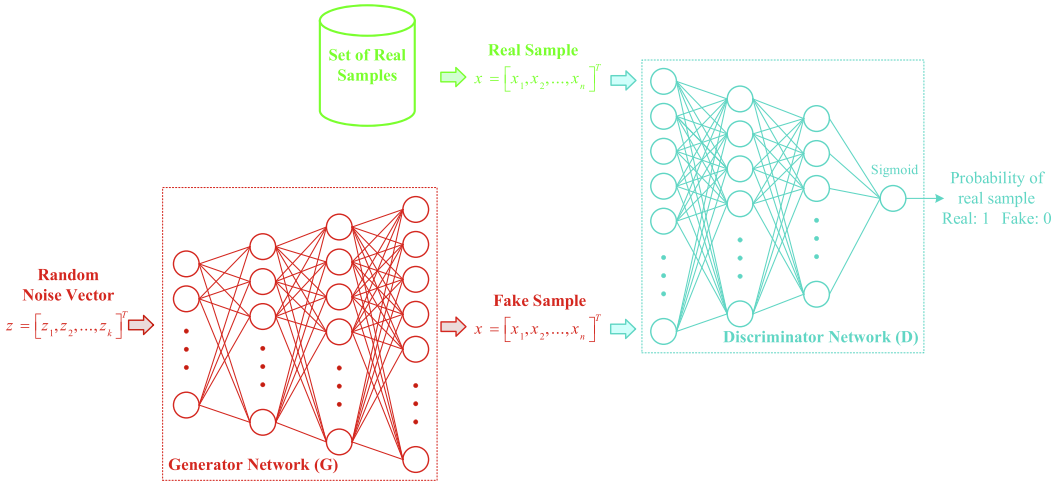

Fig. 2. Structure of a typical generative adversarial network (generating 1-D samples).

supervised learning methods is that labelling data is usually expensive and most data samples in the practical system are unlabeled. Under this scenario, the supervised learning methods cannot achieve good performance due to the lack of labeled data. Semi-supervised learning methods provide an efficient solution by learning the data patterns present in unlabeled data and combining this knowledge with those limited number of labeled training samples to accomplish a supervised learning task. Fig. 1 shows an illustration of the difference between supervised, semi-supervised, and unsupervised learning. The shapes colored in green, blue and pink belong to the labeled data of the 1st class, the 2nd class, and the 3rd class, respectively. The shapes colored in grey belong to the unlabeled data. For the unlabeled data, we do not know which class these data belong to. Compared with the supervised learning, the semi-supervised learning can utilize the unlabeled data and obtain more accurate boundary among different classes. By utilizing the unlabeled data with semi-supervised learning methods, the performance of machine learning models can be significantly improved compared with that if only the labeled data are used. On the one hand, from the unlabeled data samples, semi-supervised learning methods learn the data distribution or patterns. On the other hand, semi-supervised learning learns to assign correct labels from the labeled data samples. Semi-supervised learning methods can reduce the dependency on the labeled samples which are relatively difficult to obtain in the practical scenario.

In this study, we proposed a semi-supervised learning FDD approach based on the modified GAN to make use of the useful information contained in unlabeled samples, which is illustrated in detail in Section 3. We merge the multiclass classifier into the GAN framework. The binary discriminator in the original GAN is replaced with the multiclass classifier. The modified GAN can efficiently learn the data distribution information from the unlabeled set.

3. Proposed fault detection and diagnosis approach based on the modified generative adversarial network

3.1. Generative adversarial network (GAN)

Generative adversarial network (GAN) is a powerful unsupervised deep learning model which implicitly learns complex, high-dimensional data distribution from the training data. GAN was firstly proposed in 2014 [26] and has been generally utilized to create images in the computer vision field [27]. It can generate images whose distribution is similar with that of the training image samples. The state-of-the-art generative model based on GAN framework can generate realistic facial images [28]. In the field of fault diagnosis, GAN has been used to generate faulty training samples and help to cope with imbalanced data problems [24]. In this study, we borrow the idea of learning data distribution in an adversarial manner from GAN and propose a semi-supervised learning framework for the FDD of HVAC systems.

The basic principles of GAN are introduced as follows: GAN contains two neural networks, the generator network and the discriminator network which are denoted as G and D, respectively. The structure of a particular GAN is shown in Fig. 2. G is used to generate fake samples. Specifically, G takes the random noise vectors as input and outputs fake samples. D is used to distinguish between the real samples and the fake samples. It takes the data sample as input and outputs 0 for fake samples and 1 for real samples. The training objective of D is to accurately distinguish fake data samples from real data samples while the training objective of G is to try to fool the discriminator by producing samples as similar as possible to the real ones. $x^{(i)}$ denotes the i^{th} real data sample and $z^{(i)}$ denotes the i^{th} random noise vector input to the generator. $D(\cdot)$ and $G(\cdot)$ represent the function of discriminator and generator, respectively. Thus, the training objective of D can be formulated as:

$$\max_{\theta_d} \left[\mathrm{E}_{x\sim p_{data}} \log D_{\theta_d}(x) + \mathrm{E}_{z\sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))\right] \quad (1)$$

where θ_d and θ_g refer to the parameters of D and G, respectively. $x \sim p_{data}$ means that the sample $\times$ is drawn from the set of real data samples. $z \sim p(z)$ denote the noise vector drawn from the random distribution (e.g. Gaussian distribution). The first term in (1) is the expectation of $\log D_{\theta_d}(x)$ over many real samples. The second term in (1) is the expectation of $\log(1 - D_{\theta_d}(G_{\theta_g}(z)))$ over many random noise vectors. The first term is to make the output of D close to 1 when the real samples are present to D, while the second term is to make the output of D close to 0 when the fake samples are present to D.

The training objective of G can be formulated by:

$$\min_{\theta_g} \mathrm{E}_{z\sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \quad (2)$$

This term is to make the output of D close to 1 when the fake samples are present to D. θ_d in Eq. (2) are fixed when updating θ_g according to Eq. (2).

Training of the GAN can be implemented by alternating between the gradient ascent on D according to Eq. (1) and the gradient descent on G according to Eq. (2). This training process can be understood as

the competition between G and D. G tries to make fake samples to fool D while D tries to identify the fake samples generated by G. The race between G and D drives both to boost their performances until the fake samples made by G cannot be distinguished from the real ones. After the training, G can be used to generate fake samples whose distribution is similar with that of real samples. That is, GAN implicitly learns the data distribution or patterns of real samples.

3.2. The modified GAN: Replace the binary discriminator with a multiclass classifier

In this part, we introduce a new technique which can not only use the labeled data, but also use the unlabeled data to learn the information of data distribution. On the one hand, the conventional multiclass classifier can only use the labeled samples. On the other hand, as illustrated in Sec. 3.1, GAN is an unsupervised learning framework and can implicitly learn the data distribution in an adversarial manner, with the competition between the generator and the discriminator. This means that GAN cannot utilize the labeled data. Thus, to leverage both the unlabeled and labeled datasets, we merge the multiclass classifier into the GAN framework and propose a semi-supervised FDD approach for building HVAC systems. Specifically, the binary discriminator in the original GAN is replaced with the multiclass classifier. Fig. 3 shows the structure of the modified GAN which can leverage both unlabeled and labeled dataset. In the original GAN, the output layer of the discriminator has only one neuron, with the Sigmoid function as its activation function. After the modification, this output layer is replaced with a "classification layer" which outputs the probabilities that the sample is assigned to different classes (shown in orange color in Fig. 3). For the modified GAN, the generator network and the other parts of discriminator network remain the same as the original GAN. Similar idea has also been applied to accomplish the semi-supervised image classification tasks in [29,30].

1) *Training with labeled samples*

When the labeled samples are present to modified GAN, the parameters of multiclass classifier should be updated to maximize the probability that the sample is assigned to its corresponding class. The loss function for these labeled samples is given by:

$$\max_{\theta_c} \mathrm{E}_{(x,y)\sim\chi^L}\left[\sum_{j=1}^{K} y_j \log\left(e^{o_j(x)} / \sum_{k=1}^{K} e^{o_k(x)}\right)\right] \tag{3}$$

where χ^L denotes the set of labeled samples, and K is the number of classes. $o_j(x)$ is the output value of the j^{th} neuron in the linear layer of multiclass classifier when sample $\times$ is input to the classifier, which is also shown in Fig. 3. y is the one-hot label of sample x. For example, for a classification problem with 5 classes in total, if a sample $\times$ belongs to the 2nd class, then its one-hot label is: [0, 1, 0, 0, 0]. y_j in Eq. (3) is the j^{th} element of the one-hot label y. θ_c is the parameters of multiclass classifier network and E denotes the expectation.

For the modified GAN, training with the labeled samples is the same as those conventional supervised learning methods. Only the parameters of the classifier are updated to decrease the difference between the actual labels and the predicted labels of labeled samples. Among the output values $o_1, o_2, \ldots, o_K$, the maximal one corresponds to the class the sample will be assigned to.

2) *Training with unlabeled samples*

As for the training with unlabeled samples, we expect that the classifier can learn the data distribution information from the unlabeled dataset. We borrow the idea of learning data distribution in an adversarial manner from GAN. The unlabeled samples are the real samples while the samples generated by G are the fake ones. When these samples are present to the multiclass classifier:

- For the fake sample generated by the generator network, this sample does not belong to any class. Thus, we hope that all the output values of the classifier $o_1, o_2, \ldots, o_K$ should be as small as possible. For the input sample $\times$, the predicted probability of the sample $\times$ being the fake one can be calculated by:

$$p_{predict}(x \in fake) = \frac{1}{1 + e^{o_1(x)} + e^{o_2(x)} + \ldots + e^{o_K(x)}} \tag{4}$$

- For the real sample (unlabeled sample), this samples belong to one of K classes although we do not know the exact label. Thus, we hope that all the output values of the classifier $o_1, o_2, \ldots, o_K$ should be as large as possible. The predicted probability of the sample $\times$ being the real one by the classifier can be calculated by:

$$p_{predict}(x \in real) = \frac{e^{o_1(x)} + e^{o_2(x)} + \ldots + e^{o_K(x)}}{1 + e^{o_1(x)} + e^{o_2(x)} + \ldots + e^{o_K(x)}} \tag{5}$$

where $p_{predict}(x \in real)$ equals $1 - p_{predict}(x \in fake)$.

The training objective of the multiclass classifier is similar with Eq. (1):

$$\max_{\theta_c} \left[\mathrm{E}_{x\sim\chi^U}\left[\log p_{predict}(x \in real)\right] + \mathrm{E}_{z\sim p(z)}\left[\log\left(1 - p_{predict}(G_{\theta_g}(z) \in real)\right)\right]\right] \tag{6}$$

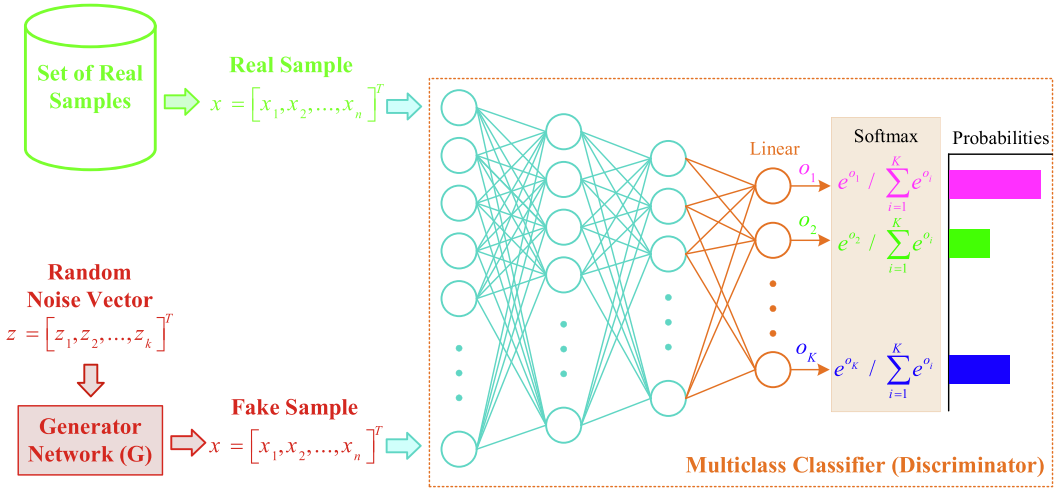

Fig. 3. Structure of the modified GAN (replace the binary discriminator with a multi-class classifier). The modified GAN comprises a generator network and a multi-class classifier network.

The training objective of G is similar with Eq. (2):

$$\min_{\theta_g} \mathrm{E}_{z\sim p(z)}\left[\log(1-p_{predict}(G_{\theta_g}(z)\in real))\right] \tag{7}$$

To minimize (maximize) the corresponding objective function, gradient descent (ascent) algorithm can be used to update the network parameters. Gradient ascent can be converted to the gradient descent by multiplying the objective function by−1. For instance, to maximize the objective function in Eq. (6), the parameters of the classifier θ_c can be updated according to Eq. (2):

$$\theta_c = \theta_c - \eta\nabla_{\theta_c}(-L_6) \tag{8}$$

where η is the learning rate and L_6 is the objective function in Eq. (6)

The term $D_{\theta_d}(x)$ in Eq. (1) is just the counterpart of $p_{predict}(G_{\theta_g}(z)\in real)$ in Eq. (6). After the training, the discriminator (classifier) can accurately differentiate real samples from fake samples, which means that the classifier has learnt the distribution information of unlabeled samples. The information of data distribution can help to enhance the performance of the classifier.

The objectives of Eq. (3) and Eq. (6) can be combined together, which yields the following whole training objective for the multi-class classifier:

$$\begin{aligned}\max_{\theta_c}\Bigg\{&\mathrm{E}_{(x,y)\sim\chi^L}\left[\sum_{j=1}^{K} y_j\log\left(e^{o_j(x)}/\sum_{k=1}^{K}e^{o_k(x)}\right)\right]\\&+\mathrm{E}_{x\sim\chi^U}\left[\log p_{predict}(x\in real)\right]\\&+\mathrm{E}_{z\sim p(z)}\left[\log(1-p_{predict}(G_{\theta_g}(z)\in real))\right]\Bigg\}\end{aligned} \tag{9}$$

3.3. Details of training for the modified GAN

3.3.1. Training procedures

The modified GAN can leverage the labeled and unlabeled datasets to train the fault classifier. The unlabeled dataset χ^U is divided into N^U_{batch} minibatch sets $\{\chi^U_1,\chi^U_2,\ldots,\chi^U_{N^u_{batch}}\}$. Each minibatch set has b_u unlabeled samples. Since the number of labeled samples is often limited and much less than the number of unlabeled samples, all labeled samples are presented to the classifier as a whole batch χ^L. The detailed training procedures for the modified GAN are shown in Fig. 4.

3.3.2. Optimization algorithm

The stochastic gradient descent algorithm (Eq. (8)) can be utilized to train the modified generative adversarial network. However, the stochastic gradient descent algorithm may suffer from the problems of local minima or slow convergence. To tackle with these problems, a modified version of the gradient descent algorithm called Adam algorithm is proposed in [31]. The Adam algorithm is chosen as the optimization algorithm for training the modified GAN in this study. The updating rule of the Adam algorithm is given by:

$$\begin{aligned}\theta_t &= \theta_{t-1} - \eta\frac{m_t/(1-\beta_1^t)}{\sqrt{v_t/(1-\beta_2^t)}+\varepsilon}\\ \begin{cases} m_t = \beta_1\cdot m_{t-1}+(1-\beta_1)\cdot g_t\\ v_t = \beta_2\cdot v_{t-1}+(1-\beta_2)\cdot(g_t\odot g_t)\\ g_t = \nabla_\theta f_t(\theta_{t-1})\end{cases}\end{aligned} \tag{10}$$

where t is the update timestep, g_t is the gradient of objective function f, η is the learning rate, m_t is the estimate of the first moment, v_t is the estimate of the second moment. $\odot$ is an element-wise multiplication operator. β_1,β_2 are the exponential decay rates for the moment estimate. In this study, the learning rate η is chosen as 0.003 and β_1,β_2 are chosen as 0.5 and 0.999, respectively.

3.3.3. Dropout techniques

Dropout technique is proposed to alleviate the overfitting problem when training a neural network with limited data samples [32]. The dropout technique is implemented by randomly dropping out some neurons so that these neurons are not considered during the forward or backward pass of the training process. As for the training of the modified GAN in this study, we found that applying the dropout technique can yield slightly better performance. The testing accuracy can be improved by 1%–3% after applying the dropout technique. The dropout rate is chosen as 20% in this study.

3.3.4. Hyperparameters settings

The modified GAN is comprised of two neural networks, the classifier network, and the generator network. The hyperparameters setting for these two networks in this study are illustrated as follows:

The data sample has 50 features (details of the selected variables for fault diagnosis are illustrated in Sec 4.1) so the number of the input neurons of the classifier is 50. The number of the output neurons of the classifier is set as 9, which equals the number of classes (8 faulty classes and 1 normal class). The generator takes in the random noise vector and generates the fake samples. The generated fake samples should have the same number of features as the real samples in the training dataset. Thus, the output layer of the generator network should have 50 neurons. The size of the random noise vector is chosen as 8 in this study, so the number of input neurons of the generator is 8.

As for the settings of the hidden layers, the number of the hidden layers of the generator and classifier is both set as 2. We found from many numerical tests that the combination of a generator with relatively complex structure and a discriminator with relatively simple structure can yield better performance. This may be because a complex generator can generate fake samples more similar to real samples and this will force the classifier to enhance its ability to recognize the data distribution of real samples so that the fake samples can be discriminated. Therefore, the unlabeled samples can be utilized more efficiently, and the final classification performance can be improved. In this study, more neurons are included for the hidden layers of the generator. Both the hidden layers of the generator have 64 neurons while the classifier has two hidden layers with 32 and 16 neurons, respectively.

Although the proposed method can efficiently leverage unlabeled samples to enhance the classification performance, it has some disadvantages: Training for the modified GAN is unstable sometimes. It occasionally gets stuck in the local minima. Thus, relevant hyperparameters of training modified GAN (such as learning rates of the generator and discriminator, dropout rate, etc.) should be carefully tuned to maximize the performance of modified GAN. Inappropriate hyperparameter settings will impair the performance.

3.4. Improved training scheme for the modified GAN: Correct the imbalance in labeled and unlabeled data

The success of the modified GAN relies on the condition that the class distribution of the training data is balanced. However, in the realistic scenarios, the probability of the system in the faulty condition is much less than that in the normal condition. Thus, the number of the faulty data samples collected in the practical system is much less than that of the normal data samples. It is very likely that the class distribution of the collected training data is highly imbalanced. Such an imbalanced class distribution will impair

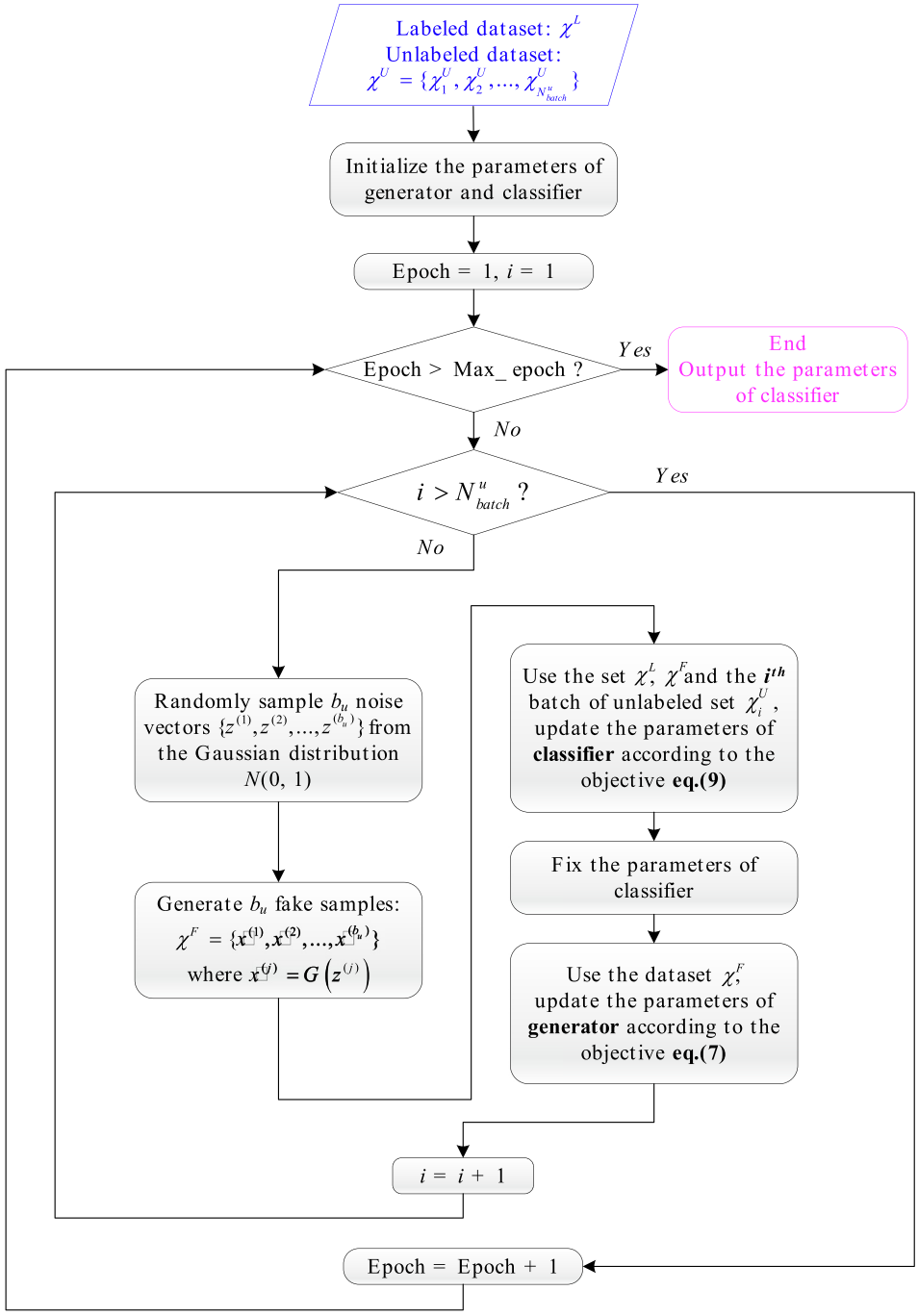

Fig. 4. Flowchart of training procedures of the modified GAN.

the performance of the classifier, which makes its predictions to be biased toward the majority class, i.e., the normal class.

For the labeled dataset, many efficient approaches have been proposed to tackle the class-imbalance problem. The SMOTE algorithm, which was proposed in [33], can efficiently balance the imbalanced datasets by oversampling. According to the sample labels, the synthetic samples are generated for each minority class. These synthetic samples and the original samples can be combined to create a balanced training dataset. It is very easy to incorporate the SMOTE algorithm into the proposed FDD approach, to tackle the class imbalance in labeled data.

However, the current balance correction methods cannot be used to deal with the data imbalance problem of the unlabeled set since we do not know which class each sample in the unlabeled set belongs to. In this part, we propose a simple and effective self-training scheme for modified GAN, to correct the data imbalance in the unlabeled set.

Based on modified GAN, an initial classifier can be firstly trained using the balanced labeled dataset (obtained by the SMOTE algorithm) and the imbalanced unlabeled dataset. Then, this classifier can be used to predict labels for the unlabeled data. We refer to these newly predicted labels as pseudo-labels. From these pseudo-labels, the class distribution of unlabeled data $\{M_k\}_{k=1}^K$ can be obtained. M_k denotes the number of unlabeled samples in class k. It should be noted that $\{M_k\}_{k=1}^K$ is the predicted class distribution of the unlabeled data, instead of the true class distribution.

The data imbalance in the unlabeled set can be corrected by weighting the terms relevant to unlabeled data in the objective function Eq. (6) in a per sample basis. Higher weights should be assigned to the samples in the minority class while lower weights should be assigned to the samples in the majority class. We introduce a new term called normalized correction weight coefficient w. $w^{(i)}$ denotes the normalized correction weight coefficient for the i^{th} unlabeled sample, which is calculated by:

$$w^{(i)} = \frac{\frac{1}{M_j}}{\sum_{k=1}^{K}\frac{1}{M_k}},\ j = c^{(i)} \tag{11}$$

where $c^{(i)}$ is the pseudo-label (class index) of the i^{th} unlabeled sample. Then, the objective function Eq. (6) can be modified as:

$$\max_{\theta_c}\left[\frac{1}{m_u}\sum_{i=1}^{m_u} w^{(i)}\log p_{predict}(x^{(i)}\in real) + \bar{w}\,\mathrm{E}_{zp(z)}\left[\log\left(1-p_{predict}(G_{\theta_g}(z)\in real)\right)\right]\right] \tag{12}$$

where m_u is the number of samples in unlabeled set χ^U. It should be noted that the terms relevant to unlabeled samples will shrink after multiplied by the weight coefficient $w^{(i)}$. For the sake of fairness, the terms relevant to labeled samples and fake samples should also shrink with the same degree. These terms can be multiplied by the mean of the normalized correction weighted coefficients $\bar{w}$:

$$\bar{w} = \frac{1}{m_u}\sum_{i=1}^{m_u} w^{(i)} \tag{13}$$

Thus, the original training objective for the multiclass classifier Eq. (9) can be modified as:

$$\max_{\theta_c}\left\{\bar{w}\,\mathrm{E}_{(x,y)\sim\chi^L}\left[\sum_{j=1}^{K} y_j\log\left(e^{o_j(x)}/\sum_{k=1}^{K}e^{o_k(x)}\right)\right] + \frac{1}{m_u}\sum_{i=1}^{m_u} w^{(i)}\log p_{predict}(x^{(i)}\in real) + \bar{w}\,\mathrm{E}_{z\sim p(z)}\left[\log\left(1-p_{predict}(G_{\theta_g}(z)\in real)\right)\right]\right\} \tag{14}$$

Similarly, the training objective for the generator is modified as:

$$\min_{\theta_g}\bar{w}\,\mathrm{E}_{z\sim p(z)}\left[\log\left(1-p_{predict}(G_{\theta_g}(z)\in real)\right)\right] \tag{15}$$

The modified GAN is then trained again based on the new objective functions Eq. (14) and (15).

In summary, the self-training scheme for the modified GAN mainly consists of three parts as shown in Fig. 5. The first part is to train the modified GAN based on the objective functions Eq. (7) and (9). The obtained discriminator (i.e., the multiclass classifier) is used as the initial fault classifier. The second part is to calculate the correction weight coefficient for each unlabeled sample. Using the initial fault classifier obtained in part 1, the pseudo-labels for unlabeled samples can be predicted. Then, the correction weight coefficients can be calculated based on the pseudo-labels. These coefficients are used to weight the terms in the original objective functions to correct the data imbalance. The objective functions for the multiclass classifier and the generator can be changed to Eq. (14) and (15), respectively. In the third part, the modified GAN is trained from scratch using the weighted training objectives Eqs. (14) and (15). After the training, the multiclass classifier is used as the final fault classifier.

It should be noted that the training scheme in Fig. 5 can be implemented iteratively: the discriminator obtained in part 3 can be used to predict the pseudo-labels and calculate the new weight coefficients; then, the weight coefficients in Eqs. (14) and (15). can be updated; the modified GAN is trained from scratch using the newly updated training objectives Eqs. (14) and (15). However, we found that it is not necessary to implement this training scheme iteratively since there is almost no increase in the testing accuracy after implementing for many times. Implementing the training scheme in Fig. 5 once is enough to obtain the classifier with satisfying performance, while ensuring the computation burden is not too high.

3.5. The FDD framework for the building HVAC systems based on the modified GAN

Based on the modified GAN, we proposed a semi-supervised FDD framework for the building HVAC systems. The fault diagnosis problem can be considered as a multiclass classification problem. The data of different faults correspond to different classes. It should be noted that the data of normal state belong to the class 0. Fig. 6. shows the schematic of the proposed fault diagnosis framework, which consists of the following 3 steps:

1) ***Collect unlabeled/labeled operating data of building HVAC systems.*** With the development of internet of things (IoT) and building management system (BMS), it is much easier to collect plenty of historical operating data of HVAC systems. However, most data collected are unlabeled. A small amount of data can be labeled manually by experts based on the prior knowledge. These labeled and unlabeled data are the input of the proposed fault diagnosis framework.
2) ***Offline training: train the modified generative adversarial network.*** Using the unlabeled and labeled datasets, the modified GAN is trained according to the training procedures in Fig. 4. The generator serves as the adversary of the discriminator and forces the discriminator to learn the data distribution from unlabeled data. After the training,

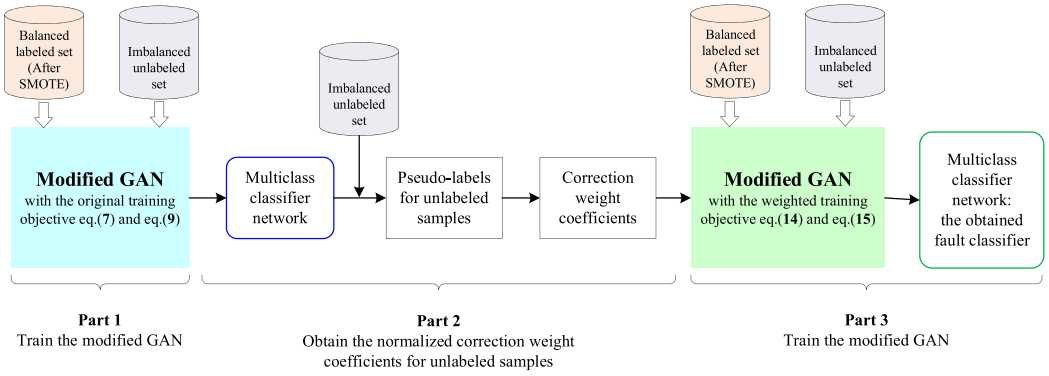

Fig. 5. Self-training scheme for the modified GAN for imbalanced semi-supervised FDD.

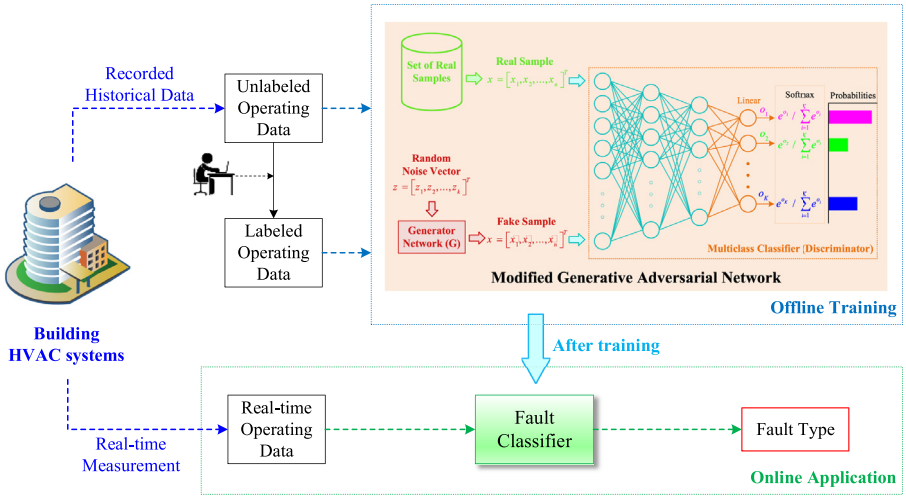

Fig. 6. Schematic of the FDD framework for building HVAC systems based on the modified GAN.

the generator is discarded, and the discriminator is used as the fault classifier to recognize different faults. If the class distribution of training data is imbalanced, the self-training scheme can be adopted to correct the class imbalance in training data, as shown in Fig. 5.

3) ***Online application.*** The obtained classification network is utilized as the fault classifier to recognize different faults. The classifier takes in the real-time operating data sample recorded in BMS and outputs the probability of different faults. The sample will be assigned to the class with the maximal probability (class 0 indicates the normal state and other classes are fault classes). The diagnostic system will report the corresponding fault alarm if a fault is diagnosed.

4. Validation and results

4.1. Experiment data

The proposed semi-supervised fault diagnosis framework is firstly validated using the data collected from an office building in Oak Ridge National Laboratory in Tennessee, United States from 2017 to 2018 [34]. This office building is reserved for experiments and is not occupied, but internal load is emulated. This building is equipped with a single packaged RTU (rooftop unit) connected to a multi-zone VAV (variable air volume) system. The RTU is a Trane YCD150 12500-kg unit with an energy efficiency rating of 9.6. The RTU provides cooling and heating for this building. The connected VAV system serves 10 zones including 8 perimeter zones and 2 core zones. The reheating coil is installed at each VAV box. The outdoor air intake of the RTU is blocked so there is no outside air introduced during the experiments. Four kinds of physical quantity were measured in this system: temperature, humidity, airflow, and the energy consumption. In this study, 50 variables (features) are selected for the purpose of fault diagnosis. These variables can be classified into 5 types as listed in Table 1.

Eight typical faults of the HVAC system are included in this study. For each faulted scenario, the data were collected for one day, with the interval of 1 min. The faulted and normal scenarios used in this study are summarized in Table 2, including the label for each fault, fault type, fault intensity, the method of fault imposition and the experiment date [34]. The data in the occupied mode are used in this study. The dataset of each faulted scenario includes about 900 observations. Since the ranges of different features are different, the data for each feature are standardized to the same range [−1, 1].

In addition to the above dataset, we use another dataset to verify the effectiveness of the proposed approach under the self-training scheme, under the scenario in which the class distribution is highly imbalanced. This is because the above dataset does not contain enough samples to create the scenario in which the class distribution is highly imbalanced in both labeled and unlabeled data (e.g., the ratio of normal samples to faulty samples in labeled and unlabeled data is 100:1).

Table 1
Selected variables for the fault diagnosis of the building HVAC system.

Type	Number of variables	Description of the variables	Unit
Temperature	1	Measured RTU supply air temperature	K
	1	Measured RTU return air temperature	
	10	Measured ambient temperature in each room	
	10	VAV box supply air temperature for each room	
Humidity	10	Measured ambient relative humidity in each room	%
Air flow rate	1	Measured RTU volumetric air flow rate	m^3/s
Energy consumption	1	RTU electricity consumption	J
	1	Total electricity consumption of HVAC system including RTUs and VAV terminal reheat	
	1	Total electricity consumption of lighting system	
	1	RTU supply air fan electricity consumption	
Status/Control command	2	RTU compressor status; 0—off, 1—on	\
	1	Control command of lighting system; 0—off, 1—on	
	10	VAV box reheat status for each room; 0—off, 1—on	

Table 2
Details of the faulted scenarios.

Fault Label	Fault type	Fault intensity	Method of fault imposition	Experiment date (yy-mm-dd)
1	Condenser fouling	25% reduction in condenser coil air flow full load	Cover the condenser face using screen, mesh, or cloth	17/08/27
		50% reduction in condenser coil air flow full load		17/08/29
2	HVAC setback error: delayed onset	3-hour onset delay	Modifying the control programming	17/12/01
3	HVAC setback error: early Termination	3-hour early termination	Modifying the control programming	17/12/03
4	Excessive infiltration	+20% infiltration	Open windows to achieve target infiltration area	17/12/07
		+40% infiltration		17/12/14
5	Lighting setback error: delayed onset	3-hour onset delay	Modifying the control programming	18/02/07
6	Lighting setback error: early termination	3-hour early termination	Modifying the control programming	18/02/09
7	No overnight HVAC setback	No setback	Modifying the control programming	17/12/20
8	No overnight lighting setback	No setback	Modifying the control programming	18/02/18
0	Normal (no faults)		/	17/09/0117/11/30

Another dataset is the dataset from the ASHRAE Research Project 1043 (ASHRAE rp-1043) [35]. This dataset contains enough samples, which has more than 10,000 observations for each class. This dataset is collected from a practical building chiller system. This chiller dataset can be used to test the effectiveness of the proposed approach under the self-training scheme. We study 7 typical faults in this study: condenser fouling (*cf*), excess oil (*eo*), reduced condenser water flow (*fwc*), reduced evaporator water flow (*fwe*), non-condensable gas in refrigerant (*nc*), refrigerant overcharge (*ro*), and refrigerant leak (*rl*). The experimental data contains 65 features. We removed 4 features since the values of these features remain constant with various faults and cannot help to discriminate different faults.

The ASHRAE rp-1043 dataset has been widely used to test various newly proposed FDD methods for chiller systems / HVAC systems in the past decade [6,9,24]. Thus, details about this dataset (such as the description of the building chiller system, description of different faults, methods of fault imposition, etc.) are omitted in this article. More details about the ASHRAE rp-1043 dataset can be found in [35].

It should be noted that the chiller system is the key component of the HVAC system. The ASHRAE rp-1043 dataset can be used to validate the effectiveness of the proposed FDD approach for building HVAC systems. In addition, the proposed approach is data-driven, which means that it only requires the data to train the fault classifier and does not need to build the physical model. The proposed approach can be easily applied to different types of systems as long as the training data are available.

4.2. Results and discussions

4.2.1. Verification of the proposed FDD approach: The ability to utilize unlabeled data to enhance the performance

As for the first dataset, there are 9 classes in the studied fault diagnosis problem, including a normal class and 8 faulted classes. The size of the original dataset is 10782. There are four kinds of datasets in this study: the labeled training dataset, unlabeled training dataset, validation dataset, and the testing dataset. These datasets can be formed by randomly selecting samples from the original dataset. The labeled training dataset only includes a small number of samples while the unlabeled training dataset includes many samples. This setting is to simulate the practical scenario that most data collected from the building system are unlabeled. We conducted several testing experiments by changing the sizes of the labeled training dataset (from 80 to 800). And the size of the unlabeled dataset is set as 8000. We remove the labels of the samples in the unlabeled dataset and assume that these labels are unknown.

The diagnostic accuracy is used as a simple measure for evaluating the FDD performance of different methods. The diagnostic accuracy equals the proportion of the correctly diagnosed samples in the testing set. After the training process, we tested the obtained fault classifier using a testing dataset containing 1000 samples. The samples in the testing dataset are different from those in the training dataset.

It should be noted that using different random seeds to generate the training and testing datasets will result in slightly different results. Thus, we repeatedly generate different datasets for 10 times. Different methods are evaluated by averaging the diagnostic accuracy over 10 experiments. In addition, the standard deviation of the diagnostic accuracy over 10 experiments is calculated. The relevant results are shown in the following parts.

1) Comparison results with other methods

We evaluate and compare the performances of the proposed approach with several supervised-learning-based methods. Plenty of supervised machine learning techniques have been developed and widely used in fault diagnosis for a wide range of engineering applications. We chose the following supervised models as the comparison methods: 1) logistic regression (LR), 2) decision tree (DT), 3) K-nearest neighbors (KNN), 4) random forest (RF), 5) support vector machine (SVM), and 6) neural network (NN). These models are used as the fault classifiers which are trained by the labeled dataset.

Although there are some literatures which proposed semi-supervised FDD methods in the field of HVAC, most of them do not report how to efficiently utilize the unlabeled data. Only the literature [23] proposes a FDD method based on the self-training scheme to use the unlabeled data. Specifically, the classifier is firstly trained with a small number of labeled samples, and then iteratively retrained with its own most confident predictions over the unlabeled samples. Thus, this self-training method is also included in the comparison study to further demonstrate the effectiveness of the proposed approach. Among the above supervised classifiers, SVM and NN exhibit better performance than others. Thus, the proposed method is compared with two self-training classifiers, the self-training SVM (abbreviated as self-SVM) and the self-training NN (abbreviated as self-NN). To make fair comparison, all methods use the same training dataset and testing dataset in each experiment. For each method, the hyperparameters are chosen using the validation dataset. To avoid the influence of different network architectures on the results, the neural networks used in all the above methods share the same architecture.

Fig. 7 shows the diagnostic accuracies on the testing dataset of the proposed methods compared with other methods under the scenario with 8000 unlabeled training samples and 80 labeled training samples. The labeled and unlabeled datasets are randomly selected from the original dataset. We repeatedly generate the datasets for 10 times using 10 different random seeds. The accuracy results corresponding to different random seeds are shown separately in Fig. 7. The supervised-learning-based methods (LR,

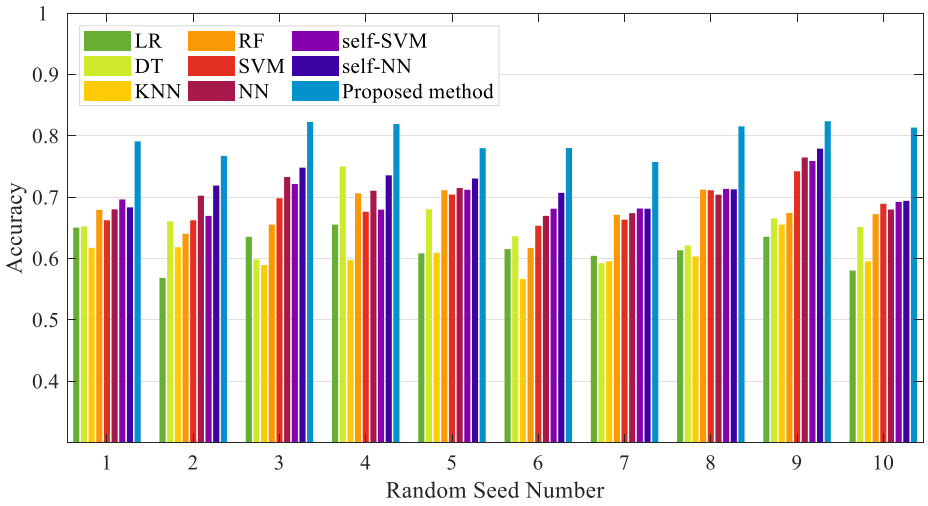

Fig. 7. Testing diagnostic accuracies for different methods under the scenario with 8000 unlabeled training samples and 80 labeled training samples.

DT, KNN, RF, SVM, NN) can only leverage the labeled dataset to train the corresponding models while the semi-supervised learning based methods (self-SVM, self-NN, proposed) can use both the labeled and unlabeled datasets. We can see from the Fig. 7 that the NN-based method achieves the best performance among the supervised methods, with the accuracies ranging from 65% to 75%. The self-NN method can slightly increase the diagnostic accuracy by 1%~2% compared with the NN-based method. The diagnostic accuracy of the proposed method is the highest under the scenario of each random seed, which is about 10% higher than the NN-based method. The results in Fig. 7 demonstrate that the proposed method can leverage the unlabeled dataset more efficiently, which significantly improves the FDD performance.

We conducted experiments with different number of samples in labeled dataset from 80 to 800. Table 3 shows the results of testing diagnostic accuracies of the proposed FDD method compared with other supervised FDD methods. Each accuracy value is the average of the diagnostic accuracies under ten random seeds. The value in the parentheses is the standard deviation of the diagnostic accuracies under ten random seeds. For each method, the diagnostic accuracy increases with the number of the labeled samples. The proposed FDD method can achieve much higher accuracy especially when there are limited number of labeled samples. When the number of samples in the labeled dataset is below 200, the diagnostic accuracy can be increased by almost 10% by the proposed method, compared with the best results obtained by the supervised methods. By leveraging the unlabeled dataset, the distribution information of the unlabeled data can be learnt and utilized to help improve the classification performance.

Fig. 8 presents the comparison results of the proposed FDD method with the existing semi-supervised FDD methods. The number of the labeled samples varies from 80 to 800 and the number of the unlabeled samples is fixed at 8000. It can be seen from Fig. 8 that the accuracy rises rapidly in the beginning and then tend to be flat with the increase of the size of labeled dataset. The accuracy of the proposed method is higher than the other semi-supervised FDD methods. The diagnostic accuracy can be improved to about 80% by the proposed method when there are only 80 labeled samples. However, the self-SVM and self-NN method can only reach the accuracy of 70% and 72%, respectively. When there are 800 labeled samples, the proposed method can achieve the diagnostic accuracy of 98% while the accuracies of self-SVM and self-NN method are 96% and 95%, respectively.

Fig. 9 shows the results of the accuracy improvement (compared with the best results obtained by the supervised methods) of different semi-supervised FDD methods. The accuracy can be slightly increased by 1%~2% with the self-SVM and self-NN methods. The proposed method can enhance the diagnostic accuracy by about 3%~10%. We found that the increase in accuracy obtained by the proposed approach is more significant when we reduce the size of the labeled dataset. The reason for this is that the classifier can acquire enough ability to discriminate different faults when

Table 3
Testing diagnostic accuracy of the proposed method and other supervised-learning-based FDD methods with different number of labeled samples (N_l denotes the number of samples in the labeled dataset). The number of samples in the unlabeled dataset is fixed at 8000.

Method	N_l = 80	N_l = 160	N_l = 240	N_l = 320	N_l = 400	N_l = 800
LR	0.6167	0.6849	0.7271	0.7481	0.7690	0.7949
	(0.0268)	(0.0219)	(0.0161)	(0.0175)	(0.0207)	(0.0144)
DT	0.6508	0.7196	0.7587	0.7888	0.8130	0.8522
	(0.0428)	(0.0306)	(0.0242)	(0.0285)	(0.0225)	(0.0193)
KNN	0.6048	0.6808	0.7230	0.7511	0.7751	0.8143
	(0.0220)	(0.0251)	(0.0235)	(0.0221)	(0.0209)	(0.0170)
RF	0.6740	0.7660	0.8332	0.8496	0.8760	0.9068
	(0.0294)	(0.0335)	(0.0292)	(0.0266)	(0.0219)	(0.0132)
SVM	0.6863	0.7953	0.8537	0.8851	0.9141	0.9469
	(0.0266)	(0.0264)	(0.0257)	(0.0199)	(0.0145)	(0.0165)
NN	0.7033	0.8099	0.8607	0.8902	0.9130	0.9488
	(0.0282)	(0.0283)	(0.0196)	(0.0192)	(0.0146)	(0.0101)
Proposed	**0.7971**	**0.8815**	**0.9258**	**0.9434**	**0.9598**	**0.9781**
	(0.0236)	**(0.0236)**	**(0.0164)**	**(0.0113)**	**(0.0085)**	**(0.0052)**

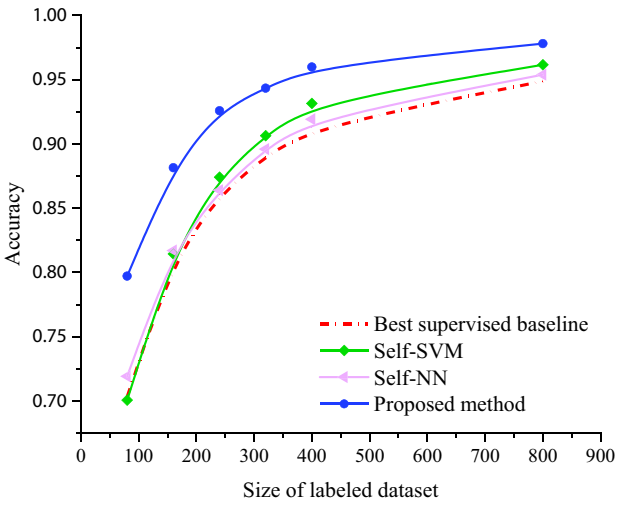

Fig. 8. Testing diagnostic accuracies of the proposed FDD method and the existing semi-supervised FDD methods, with different number of labeled samples.

there are enough labeled samples. However, when the labeled samples are insufficient, the information contained in the unlabeled dataset can help to compensate the information deficiency caused by the lack of labeled samples. The proposed semi-supervised FDD method shows its superiority especially in the practical scenario when there are fewer labeled samples and a large number of unlabeled samples.

2) Qualitative results: visualization of the features

To compare the feature extraction ability of different methods, the t-distributed stochastic neighbor embedding (t-SNE) technique [36] is used to reduce the high-dimensional output features of the 2nd hidden layer into two-dimensional vectors. The distribution of the reduced features for testing samples is shown in Fig. 10. We can see from Fig. 10 that the testing samples are clustered into 9 groups using the proposed semi-supervised methods (corresponding to 9 classes), with clearer boundaries which separate different groups. In contrast, some learned features of the NN-based method are mixed together, and the boundaries are not clear for some classes. This result demonstrates that the information learned from the unlabeled dataset by the proposed method can help to shape clearer class boundaries and results in better class separation.

4.2.2. Verification of the proposed FDD approach under the self-training scheme: The robustness against class imbalance

We used the ASHRAE rp-1043 dataset to validate that the proposed FDD approach under the self-training scheme is robust against the class imbalance. It can efficiently utilize both the labeled and unlabeled data even if the class distribution is highly imbalanced. In this study, we use the imbalance ratio γ to measure the level of class imbalance. γ is defined as the ratio of the number of samples in the majority class (i.e., normal class) to the number of samples in each minority class (i.e., faulty class). Test scenarios with different levels of class imbalance ($\gamma = 5, 20, 50, 100, 150, 200$) were established. In each test scenario, there are 16,000 unlabeled data samples. For the labeled data, the number of samples in each minority class is set as 5 for each test scenario. We assume that the imbalance ratio of labeled data is the same as the imbalance ratio of unlabeled data.

Four different approaches are studied: the supervised NN (without SMOTE), the supervised NN (with SMOTE), modified GAN, modified GAN under the self-training scheme. The 1st and 2nd methods are based on the supervised neural network (NN), which only use the labeled dataset to train the fault classifier. The 1st method directly utilizes the original imbalanced labeled dataset while the 2nd method utilizes the balanced labeled dataset processed by the SMOTE algorithm. The 3rd and 4th methods are based on modified GAN, which can utilize both labeled data and unlabeled data. The modified GAN of the 4th method is trained under the self-training scheme to tackle the class imbalance problem. For both the 3rd and 4th methods, the SMOTE algorithm is incorporated into the modified GAN, which means that the labeled dataset has been processed by the SMOTE algorithm before fed into the modified GAN. The obtained fault classifiers by the above approaches were tested by the testing dataset containing 10,000 samples. The samples in the testing dataset are different from those in the training dataset. The class distribution of the testing dataset is balanced (with 1250 samples in each class).

Table 4 summarizes the testing performance of four different approaches under different imbalance ratios. Each accuracy value is the average of classification accuracies under ten different random seeds. The value in the parentheses is the standard deviation of classification accuracies under ten different random seeds. For fair comparison, the architecture of the supervised NN is the same as that of the discriminator (multi-class classifier) of the modified GAN.

In Table 4, one can observe that the class imbalance affects the performance of all methods. The accuracy of each method decreases

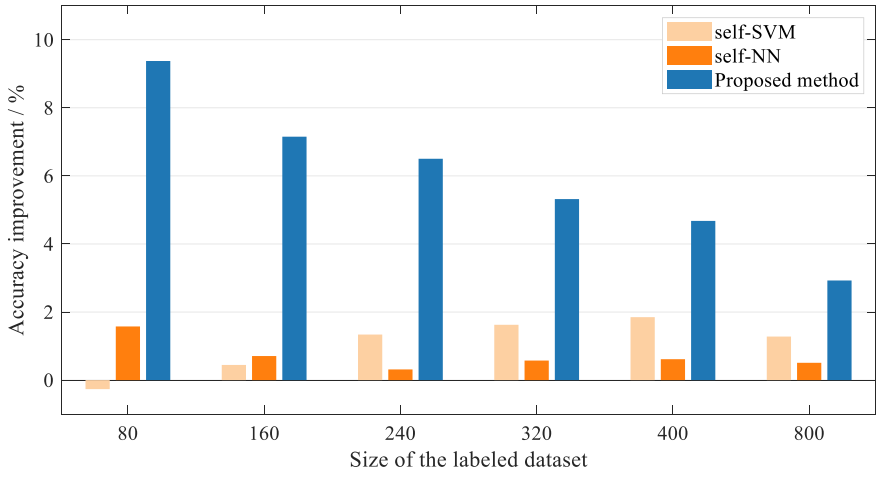

Fig. 9. Testing diagnostic accuracy improvement (compared with the best results obtained by supervised methods) of the proposed FDD method and the existing semi-supervised FDD methods, with different number of labeled samples.

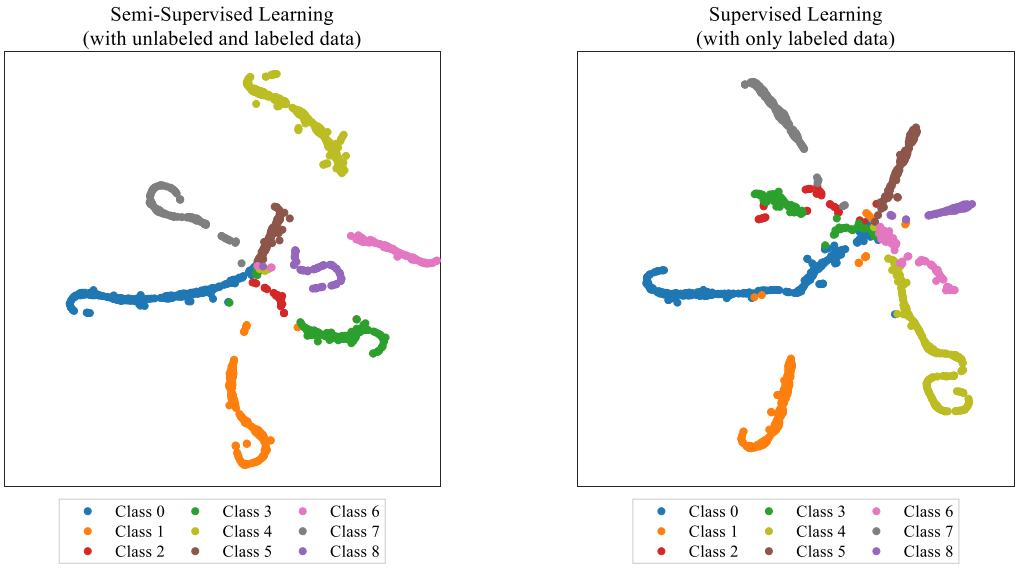

Fig. 10. t-SNE visualization of learned features of testing data by the proposed semi-supervised learning FDD method and the best supervised learning FDD method (the NN-based method).

when the imbalance ratio increases. It is noticeable that the SMOTE algorithm can efficiently correct the class imbalance in labeled data and improve the performance of the supervised NN-based method. As for the 3rd method, although the SMOTE algorithm is incorporated to correct the class imbalance in labeled data, there still exists class imbalance in unlabeled data, which will impair the ability of modified GAN to make use of the information contained in unlabeled data. As expected, the accuracy drops significantly when the imbalance ratio increases, from 84.3% ($\gamma = 5$) to 70.7% ($\gamma = 200$). In contrast, if the modified GAN is trained under the self-training scheme, the performance can be improved under the scenario with higher imbalance ratio, compared with the 3rd method. For example, the 4th method can achieve the test accuracy of 76.6% under the scenario with $\gamma = 200$, which is higher than the accuracy of the 3rd method (70.7%). The performance improvement achieved by the self-training scheme is more significant under the scenario with higher imbalance ratio. This means that the proposed FDD approach under the self-training scheme has robustness against the class imbalance in labeled and unlabeled data.

The imbalanced class distribution will make predictions of the classifier to be biased toward the majority class, i.e., the normal class. Thus, the class-specific performance of different approaches should be studied. The false positive rate (FPR) can be used to measure the performance of a classifier for each class. For a given class C_p, the FPR is defined as the ratio of the observations which do not belong to C_p but are predicted as C_p to the total observations which are predicted as C_p [37]. Fig. 11 shows the FPR values for each class. From Fig. 11, one can observe that the FPR value for the normal class is much higher than the FPR value for the faulty classes. Since the number of normal samples in training data is more than that of faulty samples, the obtained classifier is biased toward the normal class and wrongly predicts many faulty samples as normal ones. The modified GAN under the self-training scheme can reduce the FPR for the normal class and the obtained classifier is less biased toward the normal class.

5. Conclusion

A semi-supervised fault detection and diagnosis approach is proposed for the building HVAC system based on the modified generative adversarial network (modified GAN) in this article. The modified GAN can leverage both the labeled and unlabeled datasets simultaneously: it learns the information of data distribution or patterns present in unlabeled data and then combine this information with the limited number of labeled data to accomplish a supervised learning task. Then, a novel self-training scheme is proposed for the modified GAN to correct the class imbalance in both labeled and unlabeled data. The self-training scheme can efficiently improve the robustness against the class imbalance in training data. The ability of the proposed approach to utilize unlabeled data is verified using the experimental data collected from a real building HVAC system. In addition, the experimental data collected from a real building chiller system is used to verify that the modified GAN with

Table 4
Comparison of classification performance on ASHRAE rp-1043 dataset under six different imbalance ratios with different approaches.

Method	Imbalance ratio					
	5:1	20:1	50:1	100:1	150:1	200:1
Supervised NN (without SMOTE)	0.689 (0.019)	0.648 (0.015)	0.631 (0.019)	0.594 (0.018)	0.578 (0.017)	0.551 (0.015)
Supervised NN (with SMOTE)	0.693 (0.016)	0.656 (0.015)	0.653 (0.013)	0.654 (0.011)	0.655 (0.011)	0.661 (0.014)
Modified GAN	0.843 (0.027)	0.796 (0.025)	0.758 (0.018)	0.726 (0.030)	0.726 (0.014)	0.707 (0.014)
Modified GAN under self-training scheme	0.860 (0.020)	0.813 (0.026)	0.794 (0.025)	0.788 (0.016)	0.767 (0.034)	0.766 (0.015)

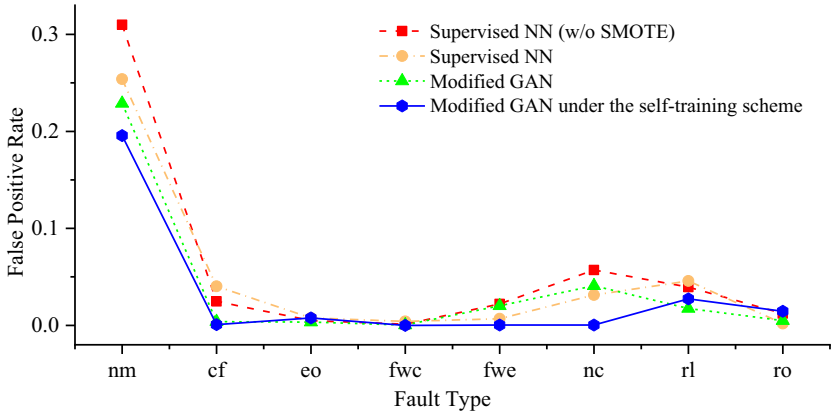

Fig. 11. False positive rate values for each class under $\gamma = 100$ using four different approaches.

self-training scheme is robust against the class imbalance in training data. The main findings of this paper can be summarized as follows:

- With the proposed modified GAN-based method, the data distribution information present in unlabeled samples can be leveraged to enhance the FDD performance. The diagnostic accuracy can be improved by almost 10% compared with the best results obtained by supervised methods in the scenario that the number of labeled samples is below 200.
- The proposed modified GAN-based method can obtain higher diagnostic accuracies than the existing semi-supervised FDD methods. This means that the unlabeled samples can be leveraged more efficiently by the proposed method.
- The increase in accuracy obtained by the proposed approach is more significant for the scenario where there are fewer labeled samples available for training. The reason is that the classifier can acquire enough capability to recognize various faults from training with sufficient labeled data. When the labeled samples are insufficient, the information contained in unlabeled samples can help to compensate the information deficiency caused by the lack of labeled samples.
- The distribution information learned from the unlabeled dataset by the proposed method can help to shape clearer class boundaries and results in better class separation.
- The proposed self-training scheme can help to enhance the robustness against the class imbalance in training data. With the self-training scheme, the modified GAN can achieve the test accuracy of 76.6% under the highly class imbalanced scenario (imbalance ratio = 200:1), which is higher than the accuracy of the modified GAN under the original training scheme (70.7%).
- The imbalanced class distribution will make predictions of the classifier to be biased toward the majority class, i.e., the normal class. The modified GAN under the self-training scheme can reduce the false positive rate for the normal class and the obtained classifier is less biased toward the normal class.

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.

References

[1] L. Pérez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, Energy Build. 40 (3) (2008) 394–398.
[2] K.J. Chua, S.K. Chou, W.M. Yang, J. Yan, Achieving better energy-efficient air conditioning–a review of technologies and strategies, Appl. Energy 104 (2013) 87–104.
[3] I. Bellanco, E. Fuentes, M. Vallès, J. Salom, "A review of the fault behavior of heat pumps and measurements, detection and diagnosis methods including virtual sensors, J. Build. Eng. (2021) 102254.
[4] J. Schein, S.T. Bushby, N.S. Castro, J.M. House, A rule-based fault detection method for air handling units, Energy Build. 38 (12) (2006) 1485–1492.
[5] M. S. Mirnaghi and F. Haghighat, "Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review," Energy and Buildings, p. 110492, 2020.
[6] Y. Zhao, F.u. Xiao, S. Wang, An intelligent chiller fault detection and diagnosis methodology using Bayesian belief network, Energy Build. 57 (2013) 278–288.
[7] Y. Zhao, S. Wang, F. Xiao, Pattern recognition-based chillers fault detection method using support vector data description (SVDD), Appl. Energy 112 (2013) 1041–1048.
[8] Z. Du, B.o. Fan, X. Jin, J. Chi, Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis, Build. Environ. 73 (2014) 1–11.
[9] K.e. Yan, W. Shen, T. Mulumba, A. Afshari, ARX model based fault detection and diagnosis for chillers using support vector machines, Energy Build. 81 (2014) 287–295.
[10] Y. Zhao, F.u. Xiao, J. Wen, Y. Lu, S. Wang, A robust pattern recognition-based fault detection and diagnosis (FDD) method for chillers, HVAC&R Res. 20 (7) (2014) 798–809.
[11] G. Li, Y. Hu, H. Chen, L. Shen, H. Li, M. Hu, J. Liu, K. Sun, An improved fault detection method for incipient centrifugal chiller faults using the PCA-R-SVDD algorithm, Energy Build. 116 (2016) 104–113.
[12] D. Li, G. Hu, C.J. Spanos, A data-driven strategy for detection and diagnosis of building chiller faults using linear discriminant analysis, Energy Build. 128 (2016) 519–529.
[13] Z. Wang, Z. Wang, S. He, X. Gu, Z.F. Yan, Fault detection and diagnosis of chillers using Bayesian network merged distance rejection and multi-source non-sensor information, Appl. Energy 188 (2017) 200–214.
[14] Y. Guo, Z. Tan, H. Chen, G. Li, J. Wang, R. Huang, J. Liu, T. Ahmad, Deep learning-based fault diagnosis of variable refrigerant flow air-conditioning system for building energy saving, Appl. Energy 225 (2018) 732–745.
[15] R. Huang, J. Liu, H. Chen, Z. Li, J. Liu, G. Li, Y. Guo, J. Wang, An effective fault diagnosis method for centrifugal chillers using associative classification, Appl. Therm. Eng. 136 (2018) 633–642.
[16] H. Shahnazari, P. Mhaskar, J.M. House, T.I. Salsbury, Modeling and fault diagnosis design for HVAC systems using recurrent neural networks, Comput. Chem. Eng. 126 (2019) 189–203.
[17] K.-P. Lee, B.-H. Wu, S.-L. Peng, Deep-learning-based fault detection and diagnosis of air-handling units, Build. Environ. 157 (2019) 24–33.
[18] Y. Guo, H. Chen, Fault diagnosis of VRF air-conditioning system based on improved Gaussian mixture model with PCA approach, Int. J. Refrig 118 (2020) 1–11.
[19] H. Han, L. Xu, X. Cui, Y. Fan, Novel Chiller Fault Diagnosis Using Deep Neural Network (DNN) with Simulated Annealing (SA), Int. J. Refrig 121 (2021) 269–278.
[20] Z. Zhou, G. Li, J. Wang, H. Chen, H. Zhong, Z. Cao, A comparison study of basic data-driven fault diagnosis methods for variable refrigerant flow system, Energy Build. 224 (2020) 110232.
[21] W.-S. Yun, W.-H. Hong, H. Seo, "A data-driven fault detection and diagnosis scheme for air handling units in building HVAC systems considering undefined states," *Journal of Building*, Engineering (2020) 102111.
[22] A. Beghi, R. Brignoli, L. Cecchinato, G. Menegazzo, M. Rampazzo, F. Simmini, Data-driven fault detection and diagnosis for HVAC water chillers, Control Eng. Pract. 53 (2016) 79–91.

[23] K.e. Yan, C. Zhong, Z. Ji, J. Huang, Semi-supervised learning for early detection and diagnosis of various air handling unit faults, Energy Build. 181 (2018) 75–83.
[24] K. Yan, A. Chong, Y. Mo, Generative adversarial network for fault detection diagnosis of chillers, Build. Environ. 172 (2020) 106698.
[25] K. Yan, J. Huang, W. Shen, Z. Ji, Unsupervised learning for fault detection and diagnosis of air handling units, Energy Build. 210 (2020) 109689.
[26] I. Goodfellow et al., "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672-2680.
[27] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical imaging: A review, Med. Image Anal. 58 (2019) 101552.
[28] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.
[29] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in Advances in neural information processing systems, 2016, pp. 2234-2242.
[30] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," arXiv preprint arXiv:1511.06390, 2015.
[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[33] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357.
[34] J. Granderson, G. Lin, A. Harding, P. Im, Y. Chen, Building fault detection data to aid diagnostic algorithm creation and performance testing, Sci. Data 7 (1) (2020) 1–14.
[35] M. Comstock and J. E. Braun, "ASHRAE 1043-RP: Fault detection and diagnostic (FDD) requirements and evaluation tools for chillers," Amer. Soc. Heating, Refrigerating Air-Conditioning Eng. (ASHRAE), Atlanta, GA, USA2006.
[36] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of machine learning research, vol. 9, no. Nov, pp. 2579-2605, 2008.
[37] A. Ebrahimifakhar, A. Kabirikopaei, D. Yuill, Data-driven fault detection and diagnosis for packaged rooftop units using statistical machine learning classification methods, Energy Build. 225 (2020) 110318.