Building and Environment 201 (2021) 107982

Contents lists available at ScienceDirect

Building and Environment

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

Chiller fault detection and diagnosis with anomaly detective generative adversarial network

Ke Yan

Department of Building, National University of Singapore, 4 Architecture Drive, 117566, Singapore

ARTICLE INFO

Keywords:
Chiller
Fault detection and diagnosis
Generative adversarial network
Anomaly detection
GANomaly

ABSTRACT

Data augmentation is one of the necessary steps in the process of automated data-driven fault detection and diagnosis (FDD) for chillers, while real-world operational training samples are usually imbalanced. Faulty data samples are usually more difficult for collection than normal operation data. Existing works show that the generative adversarial networks (GAN) are useful generating synthetic faulty data samples to enrich the training dataset. However, it remains a problem for the automated FDD applications to select high-quality synthetic faulty samples generated by GAN. The FDD accuracy becomes unstable when the quality of synthetic fault data samples cannot be controlled entirely. In this study, we proposed to use the classic definition of anomaly detection to select high-quality synthetic fault data samples with the generative adversarial networks. Two anomaly detection methods were investigated, including the traditional variational auto-encoder (VAE) and the GANomaly. Through a series of experiments, it is justified that, with a small amount of real fault data, the proposed GAN-based chiller FDD framework with GANomaly achieves the highest FDD accuracy than all compared methods.

1. Introduction

A chiller of a heating ventilation and air-conditioning (HVAC) system, consisting of a condenser, a compressor, an evaporator and an expansion valve, represents the most complex and energy consuming subsystem [1]. Fault detection and diagnosis (FDD) of various chiller faults are necessary and crucial for total energy performance of buildings, since the energy consumption contribution of the HVAC systems usually occupies from 40% to 60% for developed countries. Faults in chillers statistically reduce the total energy efficiency of HVAC systems by 20–30% [2,3]. Under the above-mentioned context, the automated chiller FDD technology plays an essential role in improving energy efficiency, facility management levels, HVAC operational reliability and indoor comfort [4–8].

Existing automated chiller FDD approaches can be typically divided into two categories, namely, model-based methods and data-driven methods [9–11]. Model based methods utilize physics model and compare the simulation data with actual data to detect and diagnose possible faults [12–14]. Data-driven methods collect data from remote sensors and build computational models for automatic classifications [15–18]. In 2010, Han et al. [19] studied a hybrid model of chiller FDD application, which combines support vector machine (SVM), genetic algorithm (GA) and parameter tuning technology. In this model, GA is responsible for searching potential feature subsets and SVM is used as FDD tool and feature selection evaluation method. The results show that the accuracy of fault classification after feature selection is higher than that of simple SVM model. Sun [20] et al. proposed a hybrid RCA fault diagnosis model combining support vector machine (SVM) with wavelet de-noising (WD), and improved the maximum correlation and minimum redundancy (mRMR) algorithm, which has outstanding diagnosis performance for RCA fault. Guo et al. [21] put forward a novel fault diagnosis approach for building energy saving based on the deep learning method which is deep belief network. Through parameter optimization selection strategy, the accuracy of fault diagnosis of the optimized model is 97.7%. Guo et al. [22] proposed a fault diagnosis strategy for the variable refrigerant flow (VRF) system based on expert rules for the first time. The VRF fault diagnosis rule (VFDR) is obtained through the expertise and characteristics of the VRF system. The results show that the fault diagnosis strategy based on expert rules can diagnose the faults of VRF system well.

The traditional FDD methods based on supervised learning already achieve high classification accuracy with various chiller faults [23–26]. However, the FDD methods based on supervised learning require sufficient training data. In the real-world practices, the amount of data is usually insufficient, since it is always difficult to obtain a well-shaped training dataset for each fault type. Yan et al. [27] proposed a semi-supervised fault detection and diagnosis method, which only uses a

E-mail address: keddiyan@gmail.com.

https://doi.org/10.1016/j.buildenv.2021.107982
Received 26 January 2021; Received in revised form 19 April 2021; Accepted 17 May 2021
Available online 26 May 2021

Nomenclature

D	Discriminator of GAN
E	Expectation
f	Output function of decoder
G	Generator of GAN
k	The Ideal number of anomalies
h	Encoded data
L	Objective function
n (m)	Number of (selected) data samples
p	Prior distribution
q	Posterior probability
t	The number of iterations
V_G	*V*Loss function of generator
$_D$	Loss function of discriminator
x	The sensor data
$x^{(i)}$	The *i*th real fault sample
$\widetilde{x}$	Generated data
$\widetilde{x}^i$	High quality generated data
$\widehat{x}$	Label/fault type
y	Linear interpolation
y'	Predicted label
z	Latent variable
θ	Parameter of decoder
ϕ	Parameter of encoder
α	Coefficient for reconstruction error
β	Coefficient for encoder loss
γ	Coefficient for decoder loss
ε	Constant between 0 and 1
λ	Penalty coefficient
$\nabla_{\widetilde{x}} D(\widehat{x})$	Partial derivative of $D(\widehat{x})$to $\widehat{x}$

small amount of fault data to diagnose the fault of air conditioning unit. Semi-supervised FDD method inserts high confidence fault test samples into the training pool to enrich the fault training sample set. However, the work in Ref. [27] is only effective when the same fault occurs again. Zhong et al. [28] proposed a data augmentation method using the generative adversarial network (GAN) for air handling units (AHUs) [29]. A large number of synthetic data is generated to enrich the fault data part in the training dataset. However, GAN itself is well-known as an extremely unstable neural network. Further screening of data is demanded selecting the high-quality synthetic data samples. In 2020, Yan et al. applied the GAN method to chiller FDD and showed that, by increasing the number of faulty samples in the training dataset, the supervised learning chiller FDD performance improved significantly [30]. However, the quality of synthetic data samples are again not guaranteed. It remains a problem for the chiller FDD methods to select the high-quality synthetic faulty samples to optimize the performance.

Following the work in Ref. [30], in this study, we propose two evaluation models selecting the high-quality synthetic data samples generated by GAN. The two evaluation models include a variational auto-encoder (VAE) [31,32] and an anomaly detection algorithm implemented using GAN (GANomaly) [33]. The evaluation models screen the synthetic data set generated by GAN and select the high-quality synthetic data samples, following the fundamental concept of anomaly detection [34,35]. The main contribution of the current study is, compared with existing works, such as [28,30], the synthetic data is trained with the evaluation models. And the limited real-world data samples are used as testing data. In the testing phase, if the number of anomalies is small, the distribution of the synthetic fault samples is close to that of the real-world samples. In other words, the synthetic training dataset is of high-quality. The quality of the training dataset is low, when the number of anomalies exceeds a certain threshold. Experimental results show that traditional supervised learning methods can be greatly enhanced with the data augmentation methods proposed in this study. The overall chiller FDD framework implemented using GANomaly is better than that implemented using VAE.

2. Methodology

2.1. Data description

The experimental data of chiller used in this paper is collected by ASHRAE project 1043-RP "Development of analysis tools for the evaluation of fault detection and diagnostics for chillers" [36]. The project used a 90 ton centrifugal chiller for experimental research.

The experimental data of the typical faults of chillers are collected in the database, including: F1. Reduced condenser water flow; F2. Reduced evaporator water flow; F3. Refrigerant leak; F4. Refrigerant overcharge; F5. Excess oil; F6. Condenser fouling; F7. Non-condensables in refrigerant. All tested fault types are also listed in Table 1. According to the ASHRAE report of the project ASHRAE 1043-RP, all fault types are further tested in four severity levels. The percentage in the table represents the percentage value higher or lower than the normal rating.

Table 1
Description of the typical fault types for chiller. All fault types are further tested in four severity levels (SL-1, SL-2, SL-3 and SL-4).

Fault	Description	Simulation method	SL-1	SL-2	SL-3	SL-4
F1	Reduced Condenser Water Flow;	Reduce flow	10%	20%	30%	40%
F2	Reduced Evaporator Water Flow;	Reduce flow	10%	20%	30%	40%
F3	Refrigerant Leak;	Reduce refrigerant	10%	20%	30%	40%
F4	Refrigerant Overcharge;	Add refrigerant	10%	20%	30%	40%
F5	Excess Oil;	Add oil	14%	32%	50%	68%
F6	Condenser Fouling	Plug the pipe	12%	20%	30%	45%
F7	Non-condensables in Refrigerant	Add nitrogen	1%	2%	3%	5%

2.2. Conditional Generative Adversarial Network

The generative adversarial network (GAN) consists of a generator and a discriminator that are essentially neural networks [37]. The generator converts the input random noise to the synthetic data samples under the supervision of discriminator. Both synthetic data samples and the real-world data samples are inputted into the discriminator for the discriminator training process. When the discriminator cannot determine the authenticity of the input sample, the entire training process terminates. The objective function of GAN is shown in Eq. (1).

$$\min_G \max_D V_{(G,D)} = E_{x\sim P_{data}}[\log D(x)] + E_{x\sim P_G}[\log(1-D(x))], \quad (1)$$

where G stands for the generator and D stands for the discriminator. $V_{(G,D)}$ is the objective function. For G, the objective function $V_{(G,D)}$ is minimized. For D, $V_{(G,D)}$ is expected to be maximized. The convergence is reached when G and D reach Nash equilibrium [37]. E stands for the expected value of a distribution. $x \sim P_{data}$ denotes the distribution of a random sampling from the real data. And $x \sim P_G$ denotes the distribution of a random sampling from the generated data.

In 2014, Mirza et al. proposed CGAN (Conditional Generative

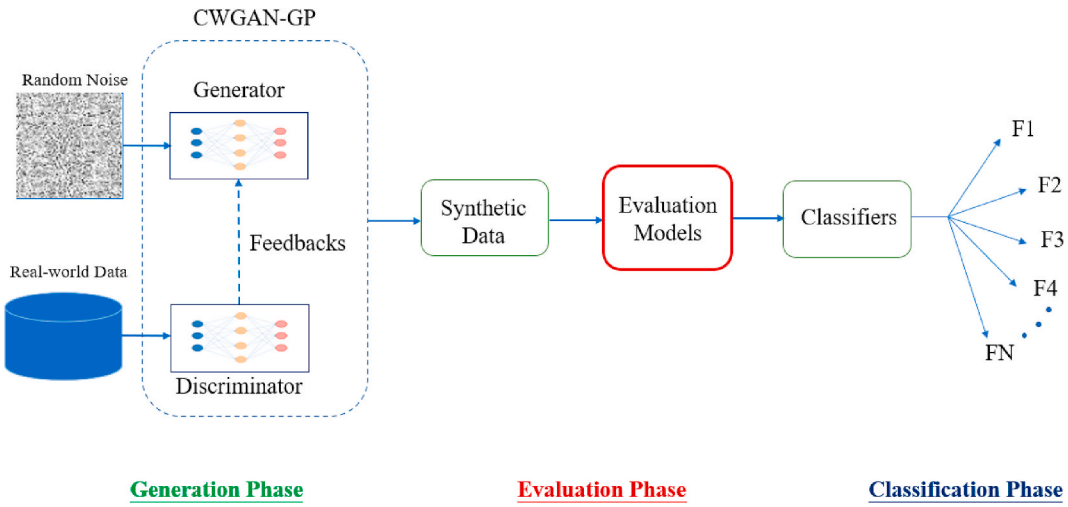

Fig. 1. Flowchart of chiller FDD framework based on CWGAN-GP and GANomaly.

Adversarial Network), which added conditional variables to the training process of GAN to guide the convergence process between the generator and the discriminator [38]. The conditional variables can be treated as labels (fault types). The objective function of CGAN is shown in Eq. (2).

$$\min_G \max_D V_{(G,D)} = E_{x\sim P_{data}}\left[\log D(x|y)\right] + E_{x\sim P_G}\left[\log(1 - D(x|y))\right], \tag{2}$$

where y is the label information for x.

2.3. Wasserstein Generative Adversarial Network with gradient penalty

In 2017, Arjovsky et al. [39] proposed WGAN (Wasserstein Generative Adversarial Network) that adopts Wasserstein distance to replace the Jensen-Shannon (JS) divergence of the original GAN [40]. When the generation feature space is similar to the real space, the gradient disappears with the JS divergence. The gradient disappearance can be avoid using Wasserstein distance in the training process of WGAN. The objective functions of WGAN are shown in the following formulas (Eqs. (3)–(5)).

$$L = \max_D \left\{E_{x\sim P_{data}} D(x) - E_{x\sim P_G} D(x)\right\}, \tag{3}$$

$$V_G = -E_{x\sim P_G}(D(x)), \tag{4}$$

$$V_D = E_{x\sim P_G}(D(x)) - E_{x\sim P_{data}}(D(x)), \tag{5}$$

where L is the overall objective function of WGAN. L was divided into two parts: V_G and V_D, where V_G is the objective function of generator. And V_D is the objective function of discriminator. P_{data} is the distribution of the real-world data and P_G is the distribution of the generated data.

Although WGAN has a great advantage in stabilizing the training proves, it faces sample quality problems when it forces weight clipping strategy to meet the Lipschitz constraints [41]. In 2017, Gulrajani et al. proposed another constraint condition of discriminator, which uses gradient penalty to guide the training process [42]. The improved WGAN model is named WGAN-GP, and the objective function is shown in Eqs. (6) and (7).

$$L = \max_D \left\{D(\tilde{x}) - D(x) + \lambda\left(\nabla_{\hat{x}} D(\hat{x})_2 - 1\right)^2\right\}, \tag{6}$$

$$\hat{x} = \varepsilon x + (1-\varepsilon)\tilde{x}, \tag{7}$$

where $\tilde{x}$ represents generated data and x represents real data. ε is a constant between 0 and 1 and λ is the penalty coefficient.

In this study, we combine CGAN with WGAN-GP to generate synthetic faulty samples improving the traditional supervised learning chiller FDD framework. The combined neural network is named as conditional Wasserstein generative adversarial network with gradient penalty (CWGAN-GP). The overall chiller FDD framework is divided into three phases, namely the generation phase, the evaluation phase and the classification phase (Fig. 1). The CWGAN-GP is employed to generate a large number of synthetic fault samples in the generation phase. In the evaluation phase, high-quality synthetic data samples are selected by evaluation models. In the following subsections, two evaluation models are introduced, including the variational auto-encoder (VAE) and GANomaly.

2.4. Variational auto-encoder as an evaluation model

Auto-encoder (AE) is a tool that represents and learns input infor-

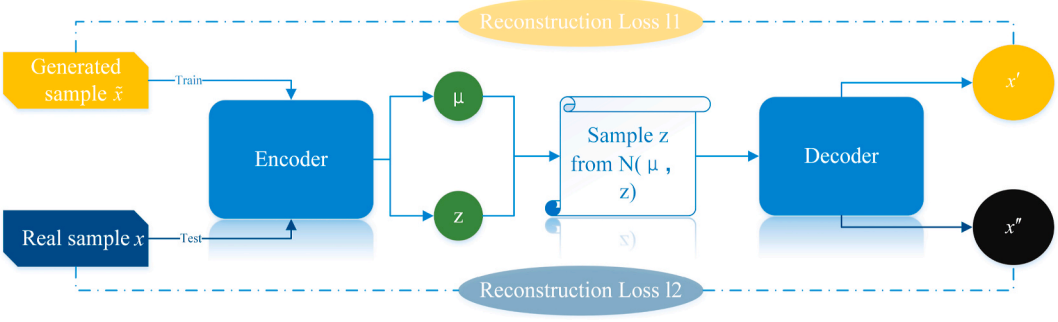

Fig. 2. The VAE based evaluation model structure.

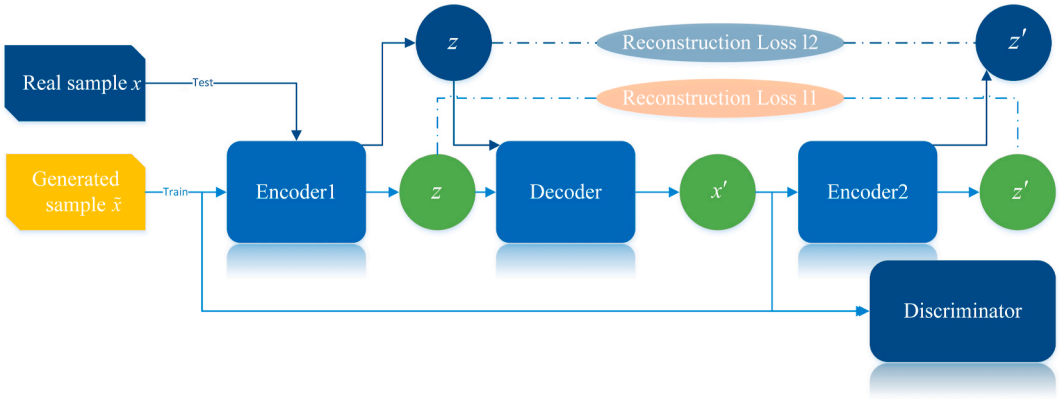

Fig. 3. The GANomaly based evaluation model structure.

mation through artificial neural network (ANN). Variational auto-encoder (VAE) extends AE, which includes an encoder $q_\theta(z|x_i)$ and a decoder $p_\phi(x_i|z)$, generating encoded vector obeying the Gaussian distribution [31]. With VAE, data samples are encoded and decoded following the Gaussian distribution with re-construction errors [31]. The re-construction errors between the original samples and the decoded samples can be used to infer anomaly. The objective function of VAE is as follows:

$$L_i(\theta, \varnothing) = -E_z \sim q_\theta(z|x_i)[\log(p_\phi(x_i|z))] + KL(q_\theta(z|x_i)||p(z)), \tag{8}$$

where ϕ and θ are the hyperparameters for the encoder and the decoder respectively. $q_\theta(z|x_i)$ is a posterior probability approximated by decoder. $p_\phi(z)$ is the prior distribution of the hidden variable z. $p_\phi(x_i|z)$ is the likelihood of data x_i given a latent variable z. $KL(q_\theta(z|x_i)||p_\phi(z))$ stands for the Kullback–Leibler (KL) distance [32] between priori distribution and posteriori probability.

The VAE implementation of the evaluation model for anomaly detection is shown in Fig. 2. The reconstruction error is calculated between the original sample and the reconstructed sample to infer whether the testing data is an anomaly. After training VAE with synthetic samples, we will get reconstruction loss *l1*. Then real samples are used for testing and reconstruction loss *l2* will be obtained. The testing samples with $l2 > l1$ is evaluated as abnormal samples.

2.5. GANomaly as an evaluation model

GANomaly is a model that uses GAN for anomaly detection [33]. It uses the potential encoder error between the original sample and the reconstructed sample as the standard to measure whether the sample is abnormal [19]. GANomaly introduces the idea of confrontation, and uses encoder1-decoder-encoder2 structure as the generator [33]. And the generator and discriminator are trained alternately until the end of iteration. The objective function of GANomaly is shown in Eqs. (9)–(13).

$$L_{enc} = h - h_2^{'} \tag{9}$$

$$L_{con} = x - x_1^{'} \tag{10}$$

$$L_{adv} = f(x) - f(x^{'})_2 \tag{11}$$

$$L_{G-Net} = \alpha L_{rec} + \beta L_{enc} + \gamma L_{adv} \tag{12}$$

$$L_{D-Net} = L_{adv} \tag{13}$$

where h, $h^{'}$ represent the potential variables encoded by encoder1 and encoder2 respectively, and x, $x^{'}$ represent the original samples and synthetic samples respectively. $f(x)$ and $f(x^{'})$ represent the output value of a certain layer of the discriminator for the original sample and the generated sample respectively, L_{G-Net} represents generator loss, L_{D-Net} represents the discriminator loss. α, β, γ are coefficients for reconstruction error, encoder loss and decoder loss, respectively. GANomaly uses reconstruction loss to judge whether the sample is abnormal, and the sample with high reconstruction loss is regarded as abnormal sample.

The implementation of the evaluation model using GANomaly is shown in Fig. 3. GANomaly is composed of three subnetworks, namely, the encoder1, the encoder2 and the decoder. The first subnetwork is a common autoencoder. The synthetic data sample $\tilde{x}$ is encoded into a n-dimensional vector z through encoder1. The reconstructed sample $x^{'}$ is obtained through the decoder (Fig. 3). The second subnetwork is another encoding network named encoder2, which encodes the reconstruction sample output from the first sub network to an n-dimensional vector z'. The reconstruction loss *l1* between the n-dimensional vector z' and the n-dimensional vector z output by the encoder in the first subnetwork is considered as the inferential exception. The real-world data samples are used to obtain the reconstruction loss *l2* (Fig. 3). The testing samples with $l2 > l1$ is evaluated as abnormal samples. Instead of comparing the difference between the original sample and the reconstructed sample, which is usually used by the VAE, GANomaly uses the difference between the original sample and the reconstructed sample in a high dimensional feature space. The additional feature extraction in the high dimensional feature space makes the whole structure anti-noise and robust as an anomaly detection model. In this study, the encoder1 and the encoder2 combined together can be viewed as a generator. And the decoder is considered as a discriminator, which introduces the idea of confrontation to judge whether the samples are real or artificial.

2.6. The overall flowchart of the proposed method

In this paper, CWGAN-GP is utilized to generate a large number of synthetic fault samples. The two evaluation models are used to evaluate and filter out the high-quality generated synthetic samples. The overall flowchart of the proposed chiller FDD framework is shown in Fig. 1.

The overall flowchart can be divided into three phases. In Fig. 1, from left to right, the three frames are the generation phase, the evaluation phase and the classification phase. The three phases are executed sequentially from left to right. In the generation phase, the CWGAN-GP generates a large number of fault samples according to the input noise and labels. The generated fault samples is put into the evaluation models, which can be actual implementations of VAE or GANomaly, in the evaluation phase, to select high-quality fault samples. Lastly, the high-quality fault samples are used to train different classifier, mimicking the process of FDD.

In the generation phase, CWGAN-GP uses a limited number of real fault samples to generate a large number of fault samples. After a series of iterative training, the final generator will generate a large group of synthetic faulty samples, n samples of each fault type in every severity

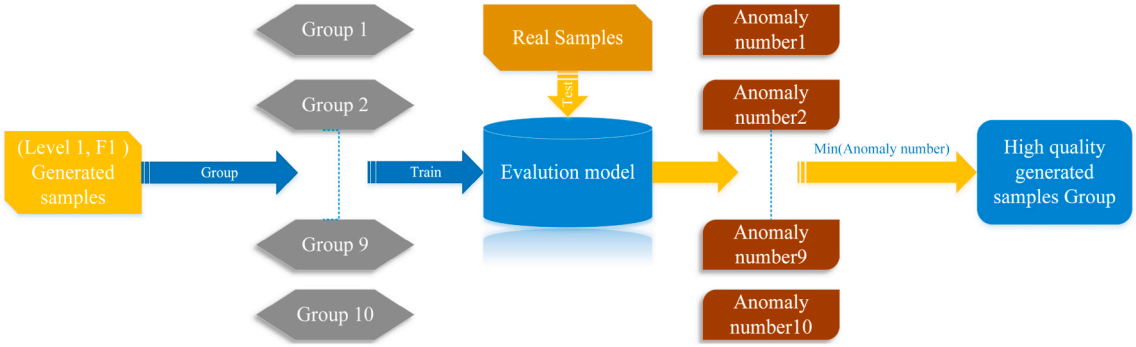

Fig. 4. The evaluation process selecting the high-quality synthetic samples from a large pool of samples generated by GAN.

Table 2
The confusion matrix.

	Predicted class		
		Class = True	Class = False
Actual Class	Class = True	True Positive (TP)	False Negative (FN)
	Class = False	False Positive (FP)	True Negative (TN)

level. Perform the same operation as above for levels 2, 3, and 4.

In the evaluation phase, for the n generated synthetic samples introduced from the CWGAN-GP model at each severity level, every type of fault trains its own evaluation model (a VAE or a GANomaly model) and selects k high-quality synthetic samples. The evaluation process is as follows: for each fault type, k samples will be randomly selected from n generated samples. The selection process repeats m times, producing m groups of data. Each group of data is used to train a VAE or a GANomaly model. The real-world data samples of the same fault type is used for testing. For each data group DG, if many of the real-world samples are evaluated as anomaly with the reconstruction errors, the DG is considered as low-quality data group. Otherwise, DG is considered high-quality. The number of the anomalies is used as a standard to measure the quality level of each DG. Among the m groups of DG, the group with the least number of anomalies is selected as the high-quality synthetic samples group. Fig. 4 shows the generalized evaluation process selecting the high-quality synthetic samples from a large pool of samples generated by CWGAN-GP, taking the generated samples are severity level 1 for fault type F1 as an example. The value of m is 10 as a demonstration in Fig. 4.

Back to Fig. 1, the classification phase implements six different classifiers, including random forest (RF), support vector machine (SVM), K-nearest neighbors (KNN), decision tree (DT), Naïve Bayes (NB) and multi-layer perceptron (MLP). In total, 1000 high-quality generated samples of each fault type are used to train the above-mentioned different classifiers. A total of 500 real samples of each fault type are used as the testing dataset.

3. Experimental process and results

3.1. Evaluation metrics

In the experimental process and results stage, we tested the proposed chiller FDD strategy with data augmentation methods using different numbers of initial real-world samples. Two different evaluation metrics are used for experimental comparison, namely classification accuracy and F1-score. We define the two evaluation metrics with the help of confusion matrix shown in Table 2. In Table 2, we show the confusion matrix with a two-class classification. Suppose that there are N data samples for testing. Table 2 divides N into N_{TP}, N_{FN}, N_{FP} and N_{TN}, where N_{TP}, N_{FN}, N_{FP} and N_{TN} are the numbers of true positive, false negative, false positive and true negative samples, respectively.

Table 3
Selected features for classifications.

Selected Feature	Description
TCI	Temperature of Condenser Water In
TEO	Temperature of Evaporator Water Out
kW/ton	Chiller efficiency
TCO	Temperature of Condenser Water Out
PO_feed	Pressure of Oil Feed
TEI	Temperature of Evaporator Water In
PRC	Pressure of Condenser
EvapTons	Evaporator Cooling Rate
TCA	Condenser Approach Temperature
TRC_sub	Subcooling Temperature

The classification accuracy is defined as:

$$\text{Classification accuracy } (\%) = \frac{N_{TP} + N_{TN}}{N} \times 100. \tag{14}$$

Then, we define the precision and recall as:

$$\text{Precision} = \frac{N_{TP}}{N_{TP} + N_{FP}}; \tag{15}$$

$$\text{Recall } = \frac{N_{TP}}{N_{TP} + N_{FN}}. \tag{16}$$

The F1-score is defined using precision and recall:

$$\text{F} - \text{score } = 2 \cdot \frac{\text{Precisio} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

3.2. Selected features for classifications

The original dataset collected from ASHRAE project 1043-RP involves 65 features, where we select 10 top important features for the comparative study conducted in this section (Table 3). The 10 top important features are selected by the cost-sensitive sequential feature selection (CSSFS) algorithm proposed by Yan et al., in 2018 [43]. In this study, we used this feature subset for all classifications performed with different classifiers.

3.3. Results

The initial fault training samples numbers of each fault type are set as 10, 30 and 50 at each severity level for experimental comparison. Both classification accuracy and F1-score are calculated for all compared methods. The hyperparameters of all classifiers, including RF, SVM, KNN, DT, NB and MLP, are tuned for each augmented training dataset. The classification accuracy are optimized for the given classifier. The kernel function of SVM is selected to be the radial basis function (RBF).

In addition to the two proposed data augmentation evaluation

Table 4
The classification accuracy and F1-score using ensemble algorithm, VAE and GANomaly models as evaluation methods for the data augmentation step in chiller FDD. Initial 10 real-world fault samples from different severe levels are used for validations.

10 real-world fault samples		Accuracy			F1-score		
		Ensemble	VAE	GANomaly	Ensemble	VAE	GANomaly
Severe level 1	RF	69.73	76.81	80.20	70.48	77.15	80.47
	SVM	**76.24**	**82.52**	**82.66**	**76.68**	**82.73**	**82.81**
	DT	70.87	75.38	78.54	71.41	75.35	78.53
	NB	35.82	43.48	47.60	37.29	43.36	50.37
	MLP	32.99	40.14	62.37	24.75	33.37	53.86
	KNN	48.84	52.05	59.20	50.57	53.66	61.40
Severe level 2	RF	77.63	78.67	86.88	77.71	78.57	86.98
	SVM	**82.57**	**85.33**	**87.81**	**82.86**	**85.60**	**88.20**
	DT	76.96	77.24	82.51	76.89	77.12	82.77
	NB	42.04	43.43	50.98	43.31	45.42	52.67
	MLP	63.11	66.76	69.17	58.50	63.04	64.39
	KNN	58.14	62.00	61.95	58.83	61.94	63.33
Severe level 3	RF	86.53	88.62	93.42	86.50	88.54	93.25
	SVM	**90.30**	**92.05**	**95.95**	**90.23**	**89.94**	**95.91**
	DT	86.03	88.86	92.12	85.82	83.27	91.86
	NB	54.72	55.95	54.10	56.75	54.43	55.89
	MLP	64.15	64.81	62.68	60.58	59.48	58.60
	KNN	63.64	69.57	66.58	64.48	60.36	66.82
Severe level 4	RF	90.10	91.90	95.05	94.13	91.63	95.05
	SVM	**92.07**	**95.33**	**97.10**	**96.03**	**93.05**	**97.10**
	DT	88.35	89.05	89.04	88.17	88.62	89.08
	NB	65.86	60.95	70.34	66.92	61.93	71.00
	MLP	88.55	92.38	78.23	87.80	92.27	76.43
	KNN	72.96	73.33	74.65	73.86	73.29	75.13

Table 5
The classification accuracy and F1-score using ensemble algorithm, VAE and GANomaly models as evaluation methods for the data augmentation step in chiller FDD. Initial 30 real-world fault samples from different severe levels are used for validations.

30 real-world fault samples		Accuracy			F1-score		
		Ensemble	VAE	GANomaly	Ensemble	VAE	GANomaly
Severe level 1	RF	81.22	86.81	89.94	81.20	86.74	90.00
	SVM	**86.26**	**93.95**	**94.35**	**86.30**	**94.07**	**94.32**
	DT	77.80	83.48	83.34	77.76	83.29	83.57
	NB	47.46	52.05	46.10	48.84	53.17	48.82
	MLP	46.30	48.24	53.67	40.03	42.62	49.47
	KNN	69.15	72.68	73.29	69.62	73.31	73.73
Severe level 2	RF	87.87	89.57	91.34	87.89	89.51	91.32
	SVM	**94.77**	**95.13**	**96.62**	**94.76**	**95.07**	**96.60**
	DT	84.36	86.87	87.65	84.33	86.92	87.59
	NB	52.47	54.33	56.21	54.48	55.30	57.33
	MLP	67.58	67.19	73.60	65.42	66.63	70.27
	KNN	76.02	76.87	76.22	76.42	76.98	76.62
Severe level 3	RF	96.27	95.87	98.27	96.27	95.88	98.29
	SVM	**97.50**	**97.30**	**98.93**	**97.51**	**97.30**	**98.94**
	DT	91.10	89.68	91.30	91.14	89.78	91.12
	NB	59.50	60.79	59.42	59.99	61.00	59.91
	MLP	57.37	49.37	94.74	53.37	44.21	94.74
	KNN	82.02	81.27	82.97	82.16	81.45	83.06
Severe level 4	RF	97.93	98.41	98.61	97.94	98.41	98.61
	SVM	**98.31**	**98.57**	**99.11**	**98.32**	**98.57**	**99.12**
	DT	96.95	95.87	97.16	96.95	95.89	97.17
	NB	76.86	76.98	70.81	77.24	76.97	71.42
	MLP	66.83	67.78	85.59	63.24	64.15	84.00
	KNN	89.22	88.73	84.76	89.31	88.68	84.92

methods, namely VAE and GANomaly, the ensemble learning evaluation method that was used in Refs. [28,30] is adopted for the comparative study. Due to the instability of GAN training, the above process will be executed in 30 cycles. The averaged outputs of the 30 repeated experiment cycles are recorded as the final results (Tables 4-6).

In Tables 4-6, each table represents a different number (either 10, 30 or 50) of real fault samples at different levels and horizontally represent the classifier used in the classification model. The best performance classification accuracy and F1-score among all six classifiers are shown in bold font. Among all the six tested classifiers, SVM outperforms the rest of the classifiers in most of the cases. The performance between the data augmentation methods using different evaluation functions for high-quality synthetic data selection is visualized in Figs. 5 and 6, for classification accuracy and F1-score, respectively. The performance of the data augmentation methods adopting the evaluation functions inheriting from anomaly detection generally better than the traditional

Table 6
The classification accuracy and F1-score using ensemble algorithm, VAE and GANomaly models as evaluation methods for the data augmentation step in chiller FDD. Initial 50 real-world fault samples from different severe levels are used for validations.

50 real-world fault samples		Accuracy			F1-score		
		Ensemble	VAE	GANomaly	Ensemble	VAE	GANomaly
Severe level 1	RF	84.03	86.71	90.22	83.83	86.54	90.17
	SVM	**89.60**	**92.43**	**94.96**	**89.50**	**92.42**	**94.93**
	DT	81.56	80.90	84.63	81.43	80.85	84.73
	NB	49.02	49.86	51.60	47.92	49.26	52.42
	MLP	70.20	75.38	73.74	68.28	74.17	72.75
	KNN	74.37	77.76	77.82	75.01	78.27	78.35
Severe level 2	RF	92.05	93.43	93.57	92.10	93.52	93.61
	SVM	**96.02**	**96.48**	**96.99**	**96.05**	**96.51**	**96.99**
	DT	87.43	88.29	89.34	87.37	88.27	89.48
	NB	57.59	57.81	58.59	58.13	58.11	60.02
	MLP	46.70	72.29	66.72	40.96	70.14	63.61
	KNN	84.82	84.39	83.01	84.96	84.68	83.30
Severe level 3	RF	95.70	97.05	98.07	95.62	97.03	98.06
	SVM	**97.80**	**97.97**	**98.46**	**97.80**	**97.72**	**98.46**
	DT	92.63	92.76	92.55	92.39	92.70	92.54
	NB	54.92	57.05	58.70	56.51	58.28	59.56
	MLP	55.09	71.24	61.77	48.51	69.03	56.87
	KNN	84.47	83.14	87.30	84.34	83.17	87.43
Severe level 4	RF	98.74	97.9	98.68	98.74	97.91	98.68
	SVM	**98.98**	**98.19**	**99.04**	**98.98**	**98.19**	**99.05**
	DT	97.68	96.19	98.22	97.68	96.20	98.22
	NB	80.15	79.71	80.50	80.26	79.72	80.70
	MLP	87.29	61.24	84.90	85.99	56.39	83.22
	KNN	94.87	92.86	95.17	94.88	92.92	95.18

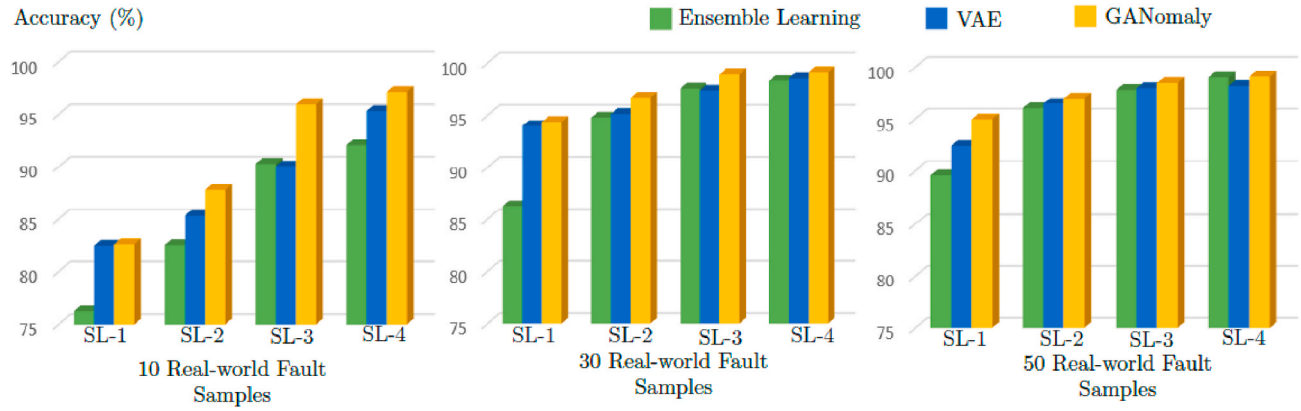

Fig. 5. A comparison of the classification accuracy using different data augmentation evaluation models and SVM as the classifier for different severe levels (SLs).

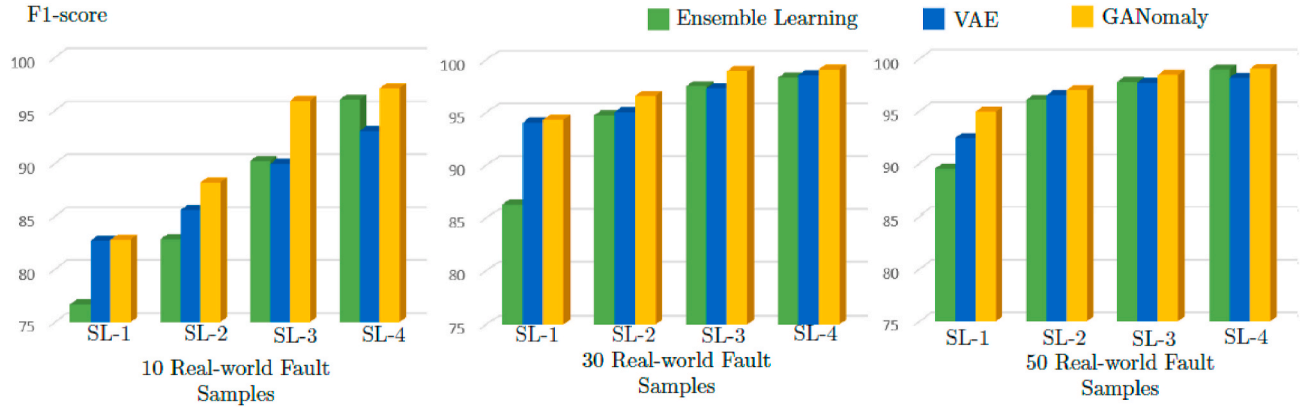

Fig. 6. A comparison of the F1-scores using different data augmentation evaluation models and SVM as the classifier for different severe levels (SLs).

ensemble learning method in Refs. [28,30]. And among the two anomaly detection algorithm based evaluation models, the performance of GANomaly is better than that of VAE, as shown in Figs. 5 and 6.

4. Conclusion

After generating a large number of synthetic fault samples from a small amount of real-world fault data for chillers, two new synthetic data evaluation models are proposed to select the high-quality synthetic samples for better performance of chiller FDD. Both proposed evaluation models adopt the idea of anomaly detection. A large number of generated samples are randomly grouped and input to the evaluation models for training. Then, the real-world data is used for anomaly detection in the testing phase. The reconstruction error between the original sample and the reconstructed sample and the reconstruction error of the encoders are used as the criteria determining whether the samples are abnormal. The generated samples group with the least number of anomalies is considered as the highest quality sample group. Two particular evaluation models are implemented, namely, the VAE model and the GANomaly model. The experimental results show that the performances of the two newly proposed synthetic data evaluation models are better than that of the ensemble algorithm evaluation model. And the performance of GANomaly is in distinctly better than that of VAE.

The main contribution of this paper is that high-quality chiller FDD is achieved with limited number of real-world faulty training data samples available. In real-world situations, we assume that the number of normal operational data dominates the training dataset. The number of real-world fault training samples are relative difficult for data collection. Traditional supervised learning model becomes bias in distinguishing fault samples from normal samples. The data augmentation method re-balances the training dataset and improves the classification accuracy of traditional supervised learning based chiller FDD methods. Manpower, material and financial resources can be saved. It is also of great significance to increase the reliability of the whole HVAC, reduce economic losses, and improve the comfort and health conditions of the indoor environment.

One obvious limitation of the current study is that we only apply the proposed data augmentation framework to the fault diagnosis part of the chiller FDD process. In fact, the fault detection part also demands the data augmentation technique to re-balance between normal operation data and fault data. In our future works, the proposed GANomaly framework will be applied to the chiller fault detection part to enhance the fault detection performance of the supervised learning based chiller FDD.

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 work was supported by the Ministry of Education (MoE) Singapore Tier 1 research grant under grant number R-296-000-208-133.

References

[1] C.L. Chen, Y.C. Chang, T.S. Chan, Applying smart models for energy saving in optimal chiller loading, Energy Build. 68 (2014) 364–371.
[2] B.E.D. Book, Energy Efficiency and Renewable Energy, US department of energy, 2011.
[3] R. Lapisa, E. Bozonnet, P. Salagnac, M.O. Abadie, Optimized design of low-rise commercial buildings under various climates–Energy performance and passive cooling strategies, Build. Environ. 132 (2018) 83–95.
[4] A. Abid, M.T. Khan, J. Iqbal, A review on fault detection and diagnosis techniques: basics and beyond, Artif. Intell. Rev. (2020) 1–26.
[5] Z. Shi, W. O'Brien, Development and implementation of automated fault detection and diagnostics for building systems: a review, Autom. ConStruct. 104 (2019) 215–229.
[6] Z. Wang, L. Wang, K. Liang, Y. Tan, Enhanced chiller fault detection using Bayesian network and principal component analysis, Appl. Therm. Eng. 141 (2018) 898–905.
[7] T. Hong, Z. Wang, X. Luo, W. Zhang, State-of-the-art on research and applications of machine learning in the building life cycle, Energy Build. 212 (2020) 109831.
[8] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, J. Li, A review of data mining technologies in building energy systems: load prediction, pattern identification, fault detection and diagnosis, Energy and Built Environment 1 (2) (2020) 149–164.
[9] K. 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] S. Katipamula, M.R. Brambley, Methods for fault detection, diagnostics, and prognostics for building systems—a review, part I, HVAC R Res. 11 (1) (2005) 3–25.
[11] S. Katipamula, M.R. Brambley, Methods for fault detection, diagnostics, and prognostics for building systems—a review, part II, HVAC R Res. 11 (2) (2005) 169–187.
[12] R. Isermann, P. Ballé, Trends in the application of model based fault detection and diagnosis of technical processes, IFAC Proceedings Volumes 29 (1) (1996) 6325–6336.
[13] Q. Zhou, S. Wang, Z. Ma, A model-based fault detection and diagnosis strategy for HVAC systems, Int. J. Energy Res. 33 (10) (2009) 903–918.
[14] X. Dai, Z. Gao, From model, signal to knowledge: a data-driven perspective of fault detection and diagnosis, IEEE Transactions on Industrial Informatics 9 (4) (2013) 2226–2238.
[15] 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.
[16] A. Beghi, R. Brignoli, L. Cecchinato, G. Menegazzo, M. Rampazzo, F. Simmini, Data-driven fault detection and diagnosis for HVAC water chillers, Contr. Eng. Pract. 53 (2016) 79–91.
[17] S.M. Namburu, M.S. Azam, J. Luo, K. Choi, K.R. Pattipati, Data-driven modeling, fault diagnosis and optimal sensor selection for HVAC chillers, IEEE Trans. Autom. Sci. Eng. 4 (3) (2007) 469–473.
[18] M.S. Mirnaghi, F. Haghighat, Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: a comprehensive review, Energy Build. (2020), 110492.
[19] H. Han, B. Gu, J. Kang, et al., Study on a hybrid SVM model for chiller FDD applications[J], Appl. Therm. Eng. 31 (4) (2011) 582–592.
[20] K. Sun, G. Li, H. Chen, et al., A novel efficient SVM-based fault diagnosis method for multi-split air conditioning system's refrigerant charge fault amount, Appl. Therm. Eng. (2016). S1359431116312406.
[21] Y. Guo, Z. Tan, H. Chen, et al., Deep learning-based fault diagnosis of variable refrigerant flow air-conditioning system for building energy saving, Appl. Energy 225 (2018) 732–745.
[22] Y. Guo, J. Wang, H. Chen, et al., An expert rule-based fault diagnosis strategy for variable refrigerant flow air conditioning systems, Appl. Therm. Eng. 149 (2019) 1223–1235.
[23] K. Yan, Z. Ji, H. Lu, J. Huang, W. Shen, Y. Xue, Fast and accurate classification of time series data using extended ELM: application in fault diagnosis of air handling units, IEEE Transactions on Systems, Man, and Cybernetics: Systems 49 (7) (2017) 1349–1356.
[24] Y. Zhao, T. Li, X. Zhang, C. Zhang, Artificial intelligence-based fault detection and diagnosis methods for building energy systems: advantages, challenges and the future, Renew. Sustain. Energy Rev. 109 (2019) 85–101.
[25] D. Li, Y. Zhou, G. Hu, C.J. Spanos, Fault detection and diagnosis for building cooling system with a tree-structured learning method, Energy Build. 127 (2016) 540–551.
[26] S. He, Z. Wang, Z. Wang, X. Gu, Z. Yan, Fault detection and diagnosis of chiller using Bayesian network classifier with probabilistic boundary, Appl. Therm. Eng. 107 (2016) 37–47.
[27] K. Yan, C. Zhong, Z. Ji, et al., Semi-supervised learning for early detection and diagnosis of various air handling unit faults, Energy Build. 181 (2018) 75–83.
[28] C. Zhong, K. Yan, Y. Dai, et al., Energy efficiency solutions for buildings: automated fault diagnosis of air handling units using generative adversarial networks, Energies 12 (3) (2019) 527.
[29] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, S.K. Ng, MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks, in: International Conference on Artificial Neural Networks, Springer, Cham, 2019, September, pp. 703–716.
[30] K. Yan, A. Chong, Y. Mo, Generative adversarial network for fault detection diagnosis of chillers, Build. Environ. 172 (2020), 106698.
[31] Z. Niu, K. Yu, X. Wu, LSTM-based VAE-GAN for time-series anomaly detection, Sensors 20 (13) (2020) 3738.
[32] Q. Zheng, Z. Lu, W. Yang, M. Zhang, Q. Feng, W. Chen, A robust medical image segmentation method using KL distance and local neighborhood information, Comput. Biol. Med. 43 (5) (2013) 459–470.
[33] J. Du, L. Guo, L. Song, H. Liang, T. Chen, Anomaly Detection of Aerospace Facilities Using Ganomaly, in: Proceedings of the 2020 5th International Conference on Multimedia Systems and Signal Processing, 2020, May, pp. 40–44.
[34] J. Wu, Z. Zhao, C. Sun, R. Yan, X. Chen, Fault-attention generative probabilistic adversarial autoencoder for machine anomaly detection, IEEE Transactions on

Industrial Informatics 16 (12) (Dec. 2020) 7479–7488, https://doi.org/10.1109/TII.2020.2976752.
[35] M. Hu, Z. Ji, K. Yan, Y. Guo, X. Feng, J. Gong, L. Dong, Detecting anomalies in time series data via a meta-feature based approach, IEEE Access 6 (2018) 27760–27776.
[36] C. Comstock Mathew, E. Braun James, Development of analysis tools for the evaluation of fault detection and diagnostics for chillers, ASHRAE Research Project 1043-RP (1999). HL 99–20, Report #4036-3.
[37] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A.A. Bharath, Generative adversarial networks: an overview, IEEE Signal Process. Mag. 35 (1) (2018) 53–65.
[38] H. Zhang, V. Sindagi, V.M. Patel, Image De-raining Using a Conditional Generative Adversarial Network, IEEE Transactions on Circuits and Systems for Video Technology 30 (11) (Nov. 2019) 3943–3956, https://doi.org/10.1109/TCSVT.2019.2920407.
[39] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein Generative Adversarial Networks, in: International Conference on Machine Learning, 2017, July, pp. 214–223.
[40] R. Turner, J. Hung, E. Frank, Y. Saatchi, J. Yosinski, May). Metropolis-hastings generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6345–6353.
[41] K. Yan, H.L. Cheng, J. Huang, Representing implicit surfaces satisfying Lipschitz conditions by 4-dimensional point sets, Appl. Math. Comput. 354 (2019) 42–57.
[42] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, S.K. Ng, MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks, in: International Conference on Artificial Neural Networks, Springer, Cham, 2019, September, pp. 703–716.
[43] K. Yan, L. Ma, Y. Dai, W. Shen, Z. Ji, D. Xie, Cost-sensitive and sequential feature selection for chiller fault detection and diagnosis, Int. J. Refrig. 86 (2018) 401–409.