Energy & Buildings 210 (2020) 109689

Contents lists available at ScienceDirect

Energy & Buildings

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

Unsupervised learning for fault detection and diagnosis of air handling units

Ke Yan[a], Jing Huang[b], Wen Shen[c], Zhiwei Ji[b,*]

[a] Department of Building, School of Design and Environment, National University of Singapore, 4 Architecture Drive, Singapore 117566
[b] School of Information & Electronic Engineering, Zhejiang Gongshang University, 18 Xuezheng Road, Hangzhou, China, 311300
[c] Computer Science Department, School of Science and Engineering, Tulane University, 6823 St. Charles Ave, New Orleans, LA, United States, 70118

ARTICLE INFO

Article history:
Received 29 June 2019
Revised 5 December 2019
Accepted 8 December 2019
Available online 24 December 2019

Keywords:
Air handling unit
Fault detection and diagnosis
Unsupervised learning
Generative adversarial network

ABSTRACT

Supervised learning techniques have witnessed significant successes in fault detection and diagnosis (FDD) for heating ventilation and air-conditioning (HVAC) systems. Despite the good performance, these techniques heavily rely on balanced datasets that contain a large amount of both faulty and normal data points. In real-world scenarios, however, it is often very challenging to collect a sufficient amount of faulty training samples that are necessary for building a balanced training dataset. In this paper, we introduce a framework that utilizes the generative adversarial network (GAN) to address the imbalanced data problem in FDD for air handling units (AHUs). To this end, we first show the necessary procedures of applying GAN to increase the number of faulty training samples in the training pool and re-balance the training dataset. The proposed framework then uses supervised classifiers to train the re-balanced datasets. Finally, we present a comparative study that illustrates the advantages of the proposed method for FDD of AHU with various evaluation metrics. Our work demonstrates the promising prospects of performing robust FDD of AHU with a limited number of faulty training samples.

1. Introduction

Artificial intelligence (AI) techniques, including machine learning (ML) and deep learning (DL) methods, are important data-driven methods for fault detection and diagnosis (FDD) of engineering systems, such as the heating ventilation and air-conditioning (HVAC) systems in buildings. Recent publications in the literature have shown the successfulness of applying the supervised learning techniques to detect and diagnose faults for different crucial HVAC components, such as the air handling units (AHUs). The FDD classification accuracy was reported as high as over 93% for typical AHU faults [1–4].

The above mentioned high FDD classification accuracy indeed depends on a well-shaped training dataset in the training phase [5]. Imbalanced training datasets invalidate most of the supervised learning based FDD systems [6,7]. In real-world scenarios, historical data is collected through remote sensors. The HVAC system operates under normal conditions most of the time. Faulty data samples are hard to be collected with the following reasons:

- The chance of any particular HVAC system becoming faulty is much less than the chance of the HVAC system working normally.
- Any faulty HVAC system is usually fixed within a very short period before a sufficient amount of faulty data is collected.

Generative adversarial network (GAN), proposed by Goodfellow et al. in 2014 [8], is capable of generating synthetic faulty training samples that are 'very close/similar' to the real-world faulty data samples to re-balance the original training dataset. As a result, the traditional supervised learning based FDD system is revised by adding GAN as a pre-processing step to re-balance the training dataset (Fig. 1). While GAN and its extensions are considered

Abbreviations: AFDD, Automated Fault Detection and Diagnosis; AHU, Air Handling Unit; ARX, Auto-Regressive Model with Exogenous Variables; BBN, Bayesian Belief Network; CGAN, Conditional Generative Adversarial Network; CWGAN, Conditional Wasserstein Generative Adversarial Network; CWGAN-ELQCP, Conditional Wasserstein Generative Adversarial Network with Ensemble Learning Quality Control Protocol; DBN, Dynamic Bayesian Network; DR, Distance Rejection; DT, Decision Tree; ELQCP, Ensemble Learning Quality Control Protocol; EKF, Extended Kalman Filter; FDD, Fault Detection and Diagnosis; FP, False Positive; FN, False Negative; GAN, Generative Adversarial Network; HVAC, Heating Ventilation and Air-Conditioning; KF, Kalman Filter; KNN, K-Nearest-Neighbor; LSSVM, Least Square Support Vector Machine; MLP, Multi-Layer Perceptron; PCA, principal component analysis; ORL, Online Reinforcement Learning; QCP, Quality Control Protocol; RF, Random Forest; SVM, Support Vector Machine; TP, True Positive; TN, True Negative; WGAN, Wasserstein Generative Adversarial Network; XGBoost, Extreme Gradient Boosting.

[*] Corresponding author.

E-mail addresses: keddiyan@gmail.com (K. Yan), zhji@zjgsu.edu.cn (Z. Ji).

https://doi.org/10.1016/j.enbuild.2019.109689
0378-7788/© 2019 Elsevier B.V. All rights reserved.

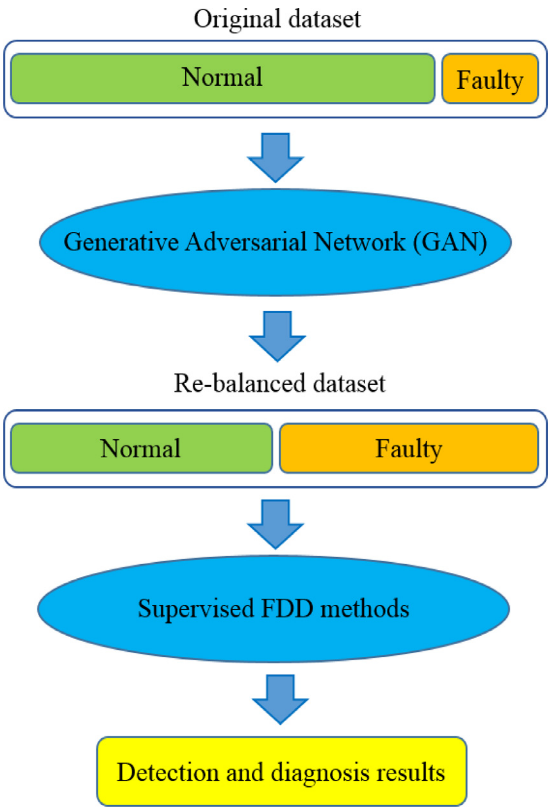


Fig. 1. The general flowchart of the revised AHU FDD method. The original imbalanced dataset is re-balanced by GAN, and then fed into the traditional supervised learning based FDD methods to enhance the classification performance.

as important solutions for the imbalanced training dataset problem of AHU FDD, there exist several questions in the literature:

- What is the minimum number of real-world faulty data samples for the revised AHU FDD framework to achieve acceptable performance?
- As an unsupervised learning method, how are GAN and its extensions applied to generate various types of faulty training samples?
- How much improvement does the revised AHU FDD framework provide compared with traditional supervised learning FDD approaches?

In this study, the importance of unsupervised learning techniques in the existing typical supervised FDD framework is shown by comparing the FDD results before and after applying GAN to the imbalanced training datasets mimicking the real-world scenarios. A semi-supervised learning FDD framework is designed that combines GAN with traditional supervised learning based FDD methods to detect and diagnose typical air handling unit faults. Both fault detection and diagnosis accuracy rates are drastically improved after applying extended GAN to re-balance the original training dataset. The main contributions of the current study include:

1. **Detailed steps of applying GAN to generate synthetic faulty training samples.** As an unsupervised learning technique, GAN is capable of generating synthetic faulty training samples with only a few real-world samples available. First, the available real-world faulty data samples are collected to form the initial dataset. Second, we illustrate the detailed steps of applying GAN to generate synthetic faulty samples with random noise. Last, we provide a quality check protocol to validate the quality of synthetic samples.
2. **A complete FDD framework based on extended GAN methods.** Extended GAN techniques have been utilized to re-balance the training dataset in both detection and diagnosis processes. We complete the work in [9] by designing a more sophisticated framework using extended GAN methods to enhance the classification accuracy.
3. **Quality control protocol optimization.** The quality control protocol proposed in [9] has been adopted and optimized to achieve better FDD performance. The original ensemble learning quality control protocol with the combination of support vector machine (SVM), decision tree (DT) and random forest (RF) has been evaluated and replaced by a combination of SVM, RF and multi-layer perceptron (MLP).
4. **Comprehensive experimental results.** The importance of applying GAN to the supervised FDD framework is shown by illustrating the experimental results with/without the GAN process. With 5, 10, 20, 30 and 40 faulty training samples available for each fault type, both fault detection and fault diagnosis results are improved significantly.

2. Related works

Automatic FDD methods in intelligent buildings can be roughly divided into two categories: model-based methods and data-driven methods [10]. Different from the model-based approach, the data-driven methods build models purely based on historical sensor data without any prior knowledge about the physics system. The models are hardly interpretable using mathematical formulas and also known as black models. Fig. 2 shows a typical supervised learning data-driven FDD approaches, where historical data is divided into training and testing datasets to perform binary classification (fault detection) and multi-class classification (fault diagnosis). A pre-processed step that involves feature extraction/statistical models, such as principal component analysis (PCA), Kalman filters (KFs) and regression models, is optional to improve the classification performance [11].

In the current decade, there are numerous data-driven FDD methods proposed for HVAC subsystems faults. Zhao et al. [12] proposed a three-layer Bayesian belief network (BBN) to detect and diagnose chiller faults. Yan et al. [11] combined autoregressive model with exogenous variables (ARX) model with support vector machine (SVM) to perform chiller FDD. More recently, Wang et al. [13] employed Bayesian network with distance rejection (DR) to confirm new chiller fault types. Shi [14] performed AHU FDD using Kalman filter (EKF) and dynamic Bayesian network (DBN). Chakraborty and Elzarka [15] proposed to use extreme gradient boosting (XGBoost) detecting HVAC subsystems faults.

In 2018, Kim and Katipamula [16] reviewed close to 200 automated FDD (AFDD) studies in the field of building HVAC systems and categorized all surveyed works into three groups: process history-based, qualitative model-based and quantitative model-based methods. A few existing gaps between theoretical FDD works and real-world applications have been identified, including the class imbalanced problem. A more recent survey work published by Shi and O'Brien [6] suggests applying active learning to constantly update the AFDD performance. Essentially, active learning is still a supervised learning method, which requires external experts' efforts for manual labeling. Yan et al. [17] proposed a semi-supervised learning approach to solve the problem of imbalanced class problem, which is a more straightforward and fully automated approach compared to active learning. The unlabeled testing data samples were automatically labeled using a semi-supervised SVM to enrich the training dataset. However, the labeling process only happens when the same fault type happens again.

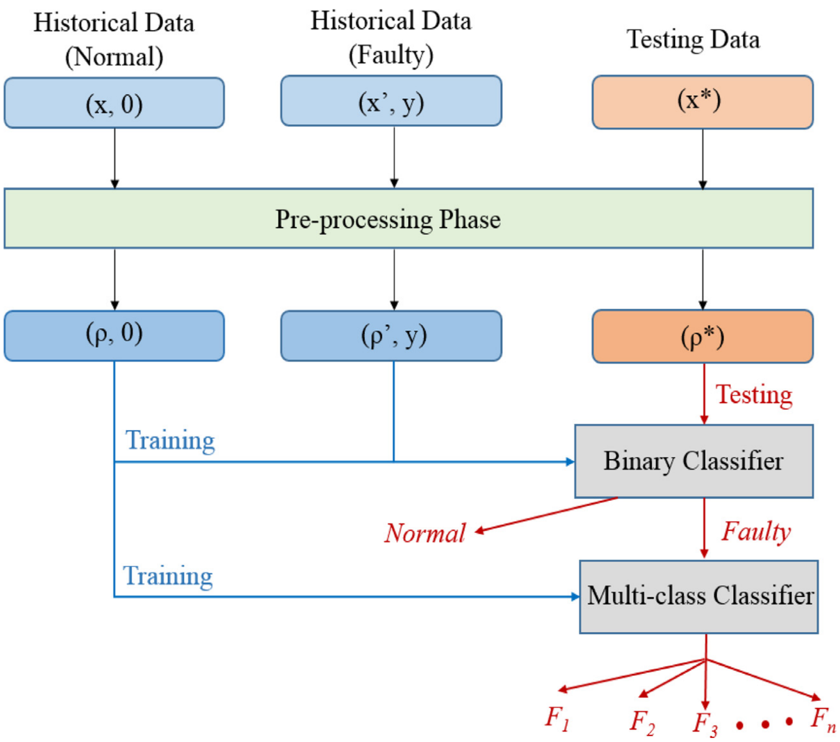


Fig. 2. A typical supervised learning FDD framework for various faults in HVAC subsystems.

3. Preliminary study

In this section, we introduce the necessary background of this study, which includes the original data description and the feature selection process.

3.1. Data description

This study employs a real-world AHU faulty operational dataset that was collected by Li et al. from 2007 to 2008 in an ASHRAE (the American society of heating, refrigerating and air-conditioning engineers) project indexed 1312-RP, through a series experiments been carried out with two identical AHUs, namely, AHU-A and AHU-B, located in Philadelphia, USA [18]. One AHU simulated data under normal condition and the other AHU simulated data under various fault circumstances. There are in total 13 different types of faults were tested under summer season, where we select six typical faults and denote them from F1 to F6:

- F1: Exhausted air (EA) damper stuck (fully open);
- F2: Return fan at fixed speed;
- F3: Cooling coil valve control unstable;
- F4: Cooling coil valve partially closed (15% open);
- F5: Outdoor air damper leak;
- F6: AHU duct leaking (after supply fan (SF)).

The general structure of the AHU used in ASHRAE project No. RP-1312 is shown in Fig. 3 with critical features marked. The collection time interval was 1 min. Each AHU was equipped with 102 sensors, i.e., in each minute, there were 102 features data collected from each AHU. In general, each fault was tested for only one day, which means that there are in total 1440 faulty samples for each fault.

In the fault detection phase, we randomly select from 5 to 40 samples from each fault dataset and 7200 samples from the normal dataset. Different classifiers were trained, firstly using the imbalanced training dataset, and then re-balanced training dataset with synthetic samples generated by GAN. The testing data is a balanced dataset consisting of 1800 faulty data samples and 1800 normal data samples.

In the fault diagnosis phase, we again randomly selected from 5 to 40 samples from each fault dataset to form the initial training dataset. The initial training dataset was enhanced by GAN. The number of samples of each fault type increased to 3000 in the enhanced training dataset. Both original and enhanced training datasets were used to train different classifiers to show the performance difference before and after applying GAN. The testing dataset consists of 300 faulty data samples for each fault type.

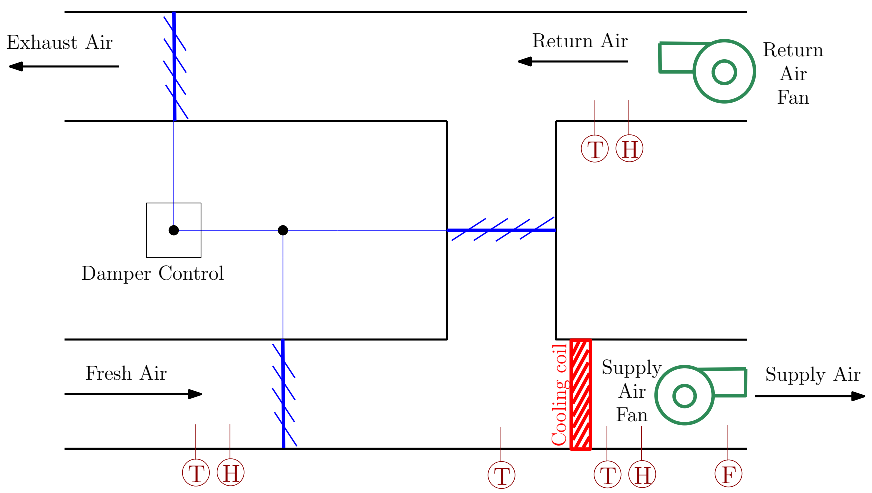
All experiments were repeated 30 times and averaged to ensure the validity of the final classification results.

3.2. Important features selection

Feature selection is an important step for the machine learning techniques to filter out redundant, noisy data in the training process. In this study, a recently proposed cost-sensitive sequential forward feature selection (CS-SFS) algorithm is employed to select the top eleven most important features using SVM as a base classifier [19]. The original data size with 102 features has been shrunken into almost 1/10 of the original size. Both synthetic data generation speed and classification speed are increased. The CS-SFS algorithm selects features from a minimal set that contains a baseline feature. In AHU FDD scenarios, power consumption by the cooling coil is usually the most important feature among all features. Therefore, E_{ccoil} is selected as the baseline feature; and the top eleven important features selected from the real-world AHU FDD dataset are listed in Table 1.

4. Methodology

In this study, we improved the work in [17] by applying generative adversarial network (GAN) to generated synthetic training samples to re-balance the training dataset. Detailed procedures of



T: Temperature sensor; H: Humidity sensor; F: Fan speed (air flow rate) sensor

Fig. 3. The general structure of the AHUs used in ASHRAE project No. RP-1312.

Table 1
Top eleven important feature variables for unsupervised AHU fault diagnosis.

Index	Variable	Description
1	E_{ccoil}	Cooling coil energy consumption
2	T_{sa}	Supply air temperature
3	T_{ra}	Return air temperature
4	T_{oa}	Outside air temperature
5	H_{ra}	Mixed air temperature
6	H_{sa}	Supply air humidity
7	T_{ma}	Return air humidity
8	T_{chwc}	Chilled Water Coil Discharge Air Temperature
9	E_{sf}	Supply fan energy consumption
10	F_{sa}	Supply air flow rate
11	F_{ra}	Return air flow rate

applying generative adversarial network (GAN) are shown in this section. It is noted that a pre-assumption is made, such that the number of faulty training samples is much less than the normal training samples and not adequate to support conventional supervised AHU FDD approaches.

4.1. Generative adversarial network

Generative adversarial network (GAN) is an unsupervised deep learning technique, which was first proposed by Goodfellow et al. in 2014 [8]. Since GAN is a relatively newly developed artificial intelligence (AI) technology, its main applications were mainly focused in the field of computer vision, including image processing [20,21], text-image conversion [22], face recognition [23] and auto-encoding [24]; the effectiveness and robustness of GAN in time series analysis are still questionable. In 2018, Zhong et al. [9] utilized GAN to generate synthetic faulty training samples to perform AHU FDD. The synthetic samples were passed through a quality control protocol to guarantee the consistency between original and generated data. Nevertheless, Zhong et al. focused only on fault diagnosis, while the fault detection also faces very serious imbalanced data problem. In the fault detection phase, the number of normal training samples is usually much bigger than the number of faulty training samples, according to our pre-assumption, e.g., in Fig. 1.

A typical GAN framework consists of two neural networks, namely, a generator and a discriminator (Fig. 4). The generator G produce synthetic samples from random noises; and the discriminator D compares the generated sample with real-world data and provides feedback to G with the similarity percentage. The iterative adversarial procedure slowly makes G learn the real-world data pattern. And a well trained G maximizes the probability that D can hardly distinguish between real and synthetic samples. The general formula of GAN is given in Eq. (1), where $G(z)$ creates synthetic samples with random noise z; and $D(x)$ calculates the discriminative probability that x is a real-world sample.

$$\min_{G} \max_{D} V(D,G) = E_{x\sim P_{data}}[\log D(x)] + E_{z\sim P_z}[\log(1-D(G(z)))]. \tag{1}$$

The term $E_{x\sim P_{data}}[\log D(x)]$ represents the expectation of $D(x) = 1$, while x is approximately distributed following the real world data pattern. The term $E_{z\sim P_z}[\log(1-D(G(z)))]$ calculates the expectation of discriminative results on synthetic data, given the distribution of noise P_z. In each iteration, the discriminator D labels the synthetic data samples using true (1) or false (0), indicating real-world or fake data, in order for the generator to produce better quality samples in the next iteration.

4.2. Extensions of GAN

The first extension of GAN allows the real-world data to have different conditions (categories) before inserting into the discriminator, which is named as conditional GAN (CGAN) [25]. In the situation of AHU FDD, since the original GAN is an unsupervised learning method, the fault type is ignored for the generator. Synthetic samples for different types of faults have to be generated separately. In contrast, CGAN generates synthetic samples of different fault types simultaneously by considering the fault type information. In CGAN, the target formula of GAN is modified to:

$$\min_{G} \max_{D} V(D,G) = E_{x\sim P_{data}}[\log D(x|c)] + E_{z\sim P_z}[\log(1-D(G(z|c)))], \tag{2}$$

where c indicates the fault type. And both generator G and discriminator D consider c as one of the inputs. We further demonstrate the extension of CGAN in Fig. 5. CGAN is extremely useful in the fault diagnosis step of the FDD process. For example, the patterns of different types of real-world AHU faults can be learned simultaneously by CGAN; and synthetic samples are generated together with the fault type labels. Moreover, It is worthwhile to mention that, by considering the label information, CGAN is able to

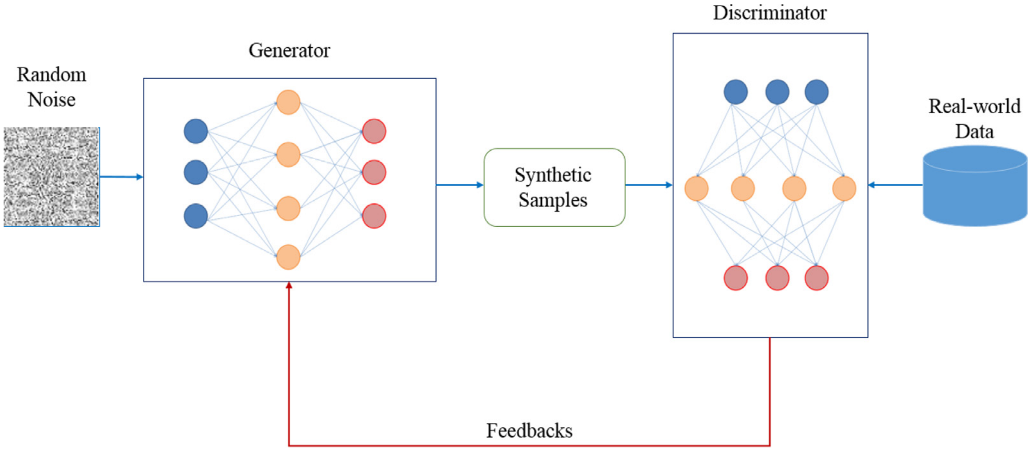


Fig. 4. A typical generative adversarial network (GAN) framework.

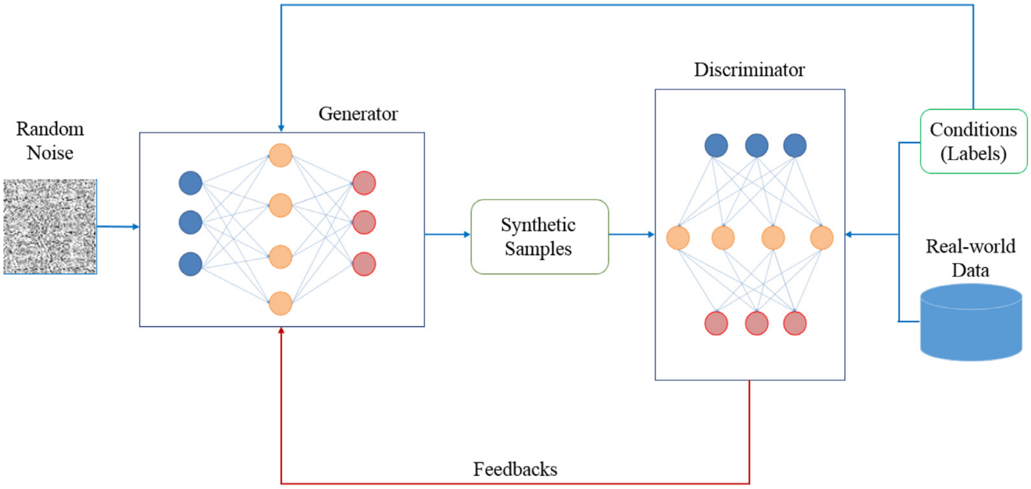


Fig. 5. The flowchart of conditional generative adversarial network (CGAN) framework.

generate more accurate synthetic samples compared to the original GAN.

The main problem of the original GAN model is the lack of variety of the generated samples. Moreover, the training process usually can hardly converge because of the gradient disappearance [26]. The second important extension of GAN utilize the Wasserstein distance to measure the similarity of two distributions, which is consequently named as Wasserstein GAN (WGAN) [27]. The Wasserstein distance between the distribution of real data P_{data} and the distribution of generated data P_G is defined as:

$$W(P_{data}, P_G) = \max_{D \in 1-Lipschitz} \{E_{x\sim P_{data}}[D(x)] - E_{x\sim P_G}[D(x)]\}. \quad (3)$$

The purpose of WGAN is to minimize $W(P_{data}, P_G)$. However, the calculation of $W(P_{data}, P_G)$ is difficult according to Eq. (3), while the discriminator D has to follow 1-Lipschitz [28]. Arjovsky et al. further approximated Eq. (3) using gradient penalty [27]. The target of WGAN becomes minimizing:

$$\max_{D}\{E_{x\sim P_{data}}[D(x)] - E_{x\sim P_G}[D(x)] - \lambda E_{x\sim P_{penalty}}[(\|\nabla_x D(x)\| - 1)^2]\},$$

where λ is a constant; and $P_{penalty}$ is the distribution of a dataset sampled between *data* and *G*.

Zhong et al. [9] applied the definition of WGAN to CGAN and introduced conditional Wasserstein GAN (CWGAN) to AHU FDD. A quality control protocol (QCP) was designed to select the high-quality synthetic samples and insert them into the training dataset. In [9], there were two types of QCPs introduced: with a single SVM and with an ensemble learning framework. In this study, we inherit the QCP with ensemble learning framework to generate synthetic faulty samples in the experiment section (Section 5).

There are other GAN extensions available, such as deep convolution GAN (DCGAN) and long short term memory (LSTM) GAN, where the multi-layer neural networks in the generator and discriminator are replaced by deep learning techniques, such as convolutional neural network and long short term memory (LSTM) neural network [29,30]. However, deep learning techniques usually require a relative long sequence of historical data to reach acceptable performance [31]. And the pre-assumption of this study is that the number of faulty training samples is much less than the normal samples. Therefore, in this study, CWGAN is more suitable than the deep learning GANs.

4.3. The flowchart of the proposed framework

The flowchart of the complete framework of fault detection and diagnosis for air handling units is depicted in Fig. 6. The system separates the fault detection and diagnosis phases. For each phase, there is a training process and a testing process. The initial training dataset consists of from 5 to 40 faulty samples for each fault type and 7200 normal samples. The detection testing dataset consists of 1800 normal samples and 1800 faulty samples (300 samples of each fault type). And the diagnosis training and testing datasets contain only the faulty samples from the detection training and testing datasets, respectively.

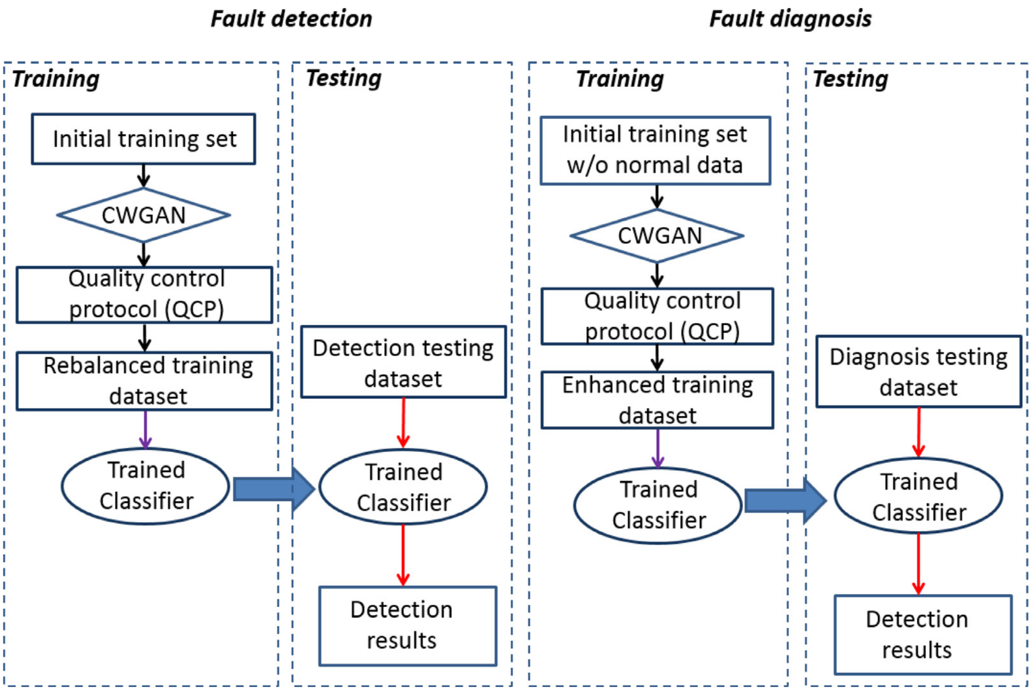

Fault detection
Fault diagnosis
Training
Testing
Training
Testing
Initial training set
CWGAN
Quality control protocol (QCP)
Rebalanced training dataset
Trained Classifier
Detection testing dataset
Trained Classifier
Detection results
Initial training set w/o normal data
CWGAN
Quality control protocol (QCP)
Enhanced training dataset
Trained Classifier
Diagnosis testing dataset
Trained Classifier
Detection results

Fig. 6. The complete framework of fault detection and diagnosis for air handling units. The initial training dataset consists of from 5 to 40 faulty samples for each fault type and 7200 normal samples. The detection testing dataset consists of 1800 normal samples and 1800 faulty samples (300 samples of each fault type). And the diagnosis testing dataset contains only the faulty samples from the detection testing dataset.

Table 2
The fault diagnosis classification accuracy rates of different combinations of ELQCPs with 5, 10, 15, ..., 40 faulty training samples for each fault type in the initial training dataset.

Combinations	5	10	15	20	25	30	35	40
SVM-DT-RF	63.17	76.98	79.02	80.17	81.58	83.70	84.35	90.44
SVM-KNN-RF	67.26	77.38	82.16	83.70	83.76	**86.25**	86.31	91.15
SVM-DT-KNN	66.98	76.74	82.44	83.77	83.52	85.73	87.54	91.31
SVM-DT-MLP	68.22	78.30	82.03	83.34	83.49	84.87	85.66	90.65
SVM-RF-MLP	**67.66**	**78.34**	**83.19**	**83.97**	**84.78**	84.67	**87.90**	**91.53**
DT-KNN-RF	67.53	77.67	83.10	83.56	84.47	85.82	87.31	90.62
DT-KNN-MLP	67.34	77.04	82.66	82.99	82.33	83.94	86.82	90.41
DT-RF-MLP	68.33	78.15	82.87	82.45	83.54	84.39	84.72	88.89
KNN-RF-MLP	66.93	77.16	82.60	83.3	83.65	84.4	87.52	90.33

In the detection phase, the CWGAN is utilized to re-balance the numbers between normal training data samples and faulty training data samples. In the diagnosis phase, the CWGAN increases the number of samples in each fault type to enhance the diagnosis capability.

4.4. Quality control protocol optimization

The quality control protocol (QCP) proposed in [9] is a necessary step to guarantee the quality of synthetic faulty training samples generated by CWGAN. In this study, we adopt the ensemble learning QCP (ELQCP) in [9] to perform the quality check. Generally speaking, ELQCP is an ensemble classifier that is trained using initial faulty training samples. Each synthetic sample is classified by the trained ELQCP to check whether the generated label is consistent with the label given by ELQCP. The original ELQCP in [9] consists of support vector machine (SVM), decision tree (DT) and random forest (RF). In this section, the original ELQCP is evaluated using different combinations of various base classifiers, including k-nearest neighbor (KNN), SVM, DT, multi-layer perceptron (MLP) and RF. With 5, 10, 15, ..., 40 faulty training samples for each fault type in the initial training dataset, CWGAN combining various ELQCPs are evaluated. A testing set containing 300 faulty samples of each fault type (in total 1800 faulty samples) is used.

According to [9], considering ELQCP, the fault diagnosis classification accuracy rates of all different combinations of the base classifiers are listed in Table 2. The relatively more optimal combination of ELQCP consists of base classifier: SVM, RF and MLP.

5. Results

5.1. Performance evaluation metrics

Classification accuracy, precision, recall and F-score are four important performance evaluation metrics for HVAC FDD methods. Before we show the experimental results of the proposed

Table 3
Confusion matrix for a binary classification problem.

	Actual class True	False
Testing result		
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

Table 4
The fault detection accuracy rates obtained from the proposed semi-supervised framework comparing with the conventional supervised fault detection method based on different traditional machine learning classifiers, including RF, SVM, MLP, KNN and DT. The number of faulty training samples varies from 5 to 40 for each fault type together with 7200 normal samples to form the training set.

Detect.	5		10		20		30		40	
Accu.(%)	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.
RF	**97.93**	58.56	**97.49**	60.11	**97.08**	58.77	**98.24**	59.45	**99.96**	58.17
SVM	**99.54**	52.05	**99.59**	57.13	**99.72**	59.82	**99.65**	60.31	**99.92**	60.49
MLP	**97.53**	50.00	**96.26**	50.00	**97.10**	50.00	**98.45**	50.00	**99.83**	50.00
KNN	**97.27**	60.69	**97.60**	62.57	**97.35**	61.38	**98.61**	61.30	**99.75**	60.58
DT	**95.49**	58.91	**91.78**	58.90	**94.47**	58.62	**91.20**	58.33	**99.97**	60.57

framework, formal mathematical formulas of the four evaluation metrics are revisited in this subsection.

Referring to a confusion matrix as shown in Table 3, the true positive (TP) value counts the number of samples that are correctly classified and belonging to the current class. The true negative (TN) value refers to the number of samples that are correctly classified and belonging to the other classes. The false positive (FP) value counts the number of samples that are incorrectly classified and belonging to the other classes. And false negative (FN) value refers to the number of samples that are incorrectly classified and belonging to the current class. The classification accuracy, precision, recall and F-score values of Table 3 are defined as following:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{4}$$

$$Presicion = \frac{TP}{TP+FP}, \tag{5}$$

$$Recall = \frac{TP}{TP+FN}, \tag{6}$$

$$F{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{7}$$

In the fault detection phase, all faulty samples are considered as positive samples. All normal samples are negative samples. In the fault diagnosis phase, since we always use an equal number of testing samples for each fault class, the final accuracy, precision, recall and F-score values are calculated by taking the average value among all different fault classes.

5.2. Fault detection results with/without CWGAN for AHU faults

The experimental results were obtained by applying the proposed framework (as shown in Fig. 6) to the AHU operational data collected by ASHRAE project number 1312-RP. In the fault detection phase, we randomly selected 5, 10, 20, 30 and 40 faulty samples for each fault type and combined with 7200 normal samples to form the training set, mimicking the situation that the specific fault has been repaired within 40 min. The CWGAN with optimized ELQCP (CWGAN-ELQCP) was employed to generate synthetic faulty samples to re-balance the training dataset. In the testing phase, a binary dataset containing 1800 normal data samples and 1800 faulty data samples (with samples of various fault types) was utilized. The fault detection accuracy, precision, recall and F-score rates obtained from the proposed semi-supervised framework and the conventional supervised fault detection method are shown in Tables 4–7, respectively. To complete the experiment, we tested different base classifiers on both proposed framework and conventional FDD framework, referring to the 'trained classifier' shown in Fig. 6, including random forest (RF), support vector machine (SVM), multi-layer perceptron (MLP), k-nearest-neighbor (KNN) and decision tree (DT). All the experiments are repeated 30 times and taking the average values to ensure the generalization of the final results.

Since the original training dataset is extremely imbalanced (consisting of 5 to 40 faulty samples for each fault type and 7200 normal samples), without CWGAN-ELQCP, the conventional fault detection methods always tend to classify the faulty samples as normal samples, since the normal samples have dominated the training pool. The F-score values are less than 0.2 with 7200 normal samples and less than 40 faulty samples in the training dataset. After re-balancing training dataset using CWGAN-ELQCP framework, all accuracy, precision, recall and F-score values achieve almost 1 for all types of base classifiers (Tables 4–7).

5.3. Fault diagnosis results comparison

In the fault diagnosis phase, the normal samples are removed from the training dataset that was utilized in the fault detection phase. As a result, the training set contains 5, 10, 20, 30 and 40 faulty samples from each fault type. And a testing dataset containing 300 faulty samples of each fault type (in total 1800 faulty samples) is constructed. All testing samples are randomly selected from the original ASHRAE 1312-RP dataset. All experiments are repeated 30 times and averaged to ensure the generalization of the results.

The diagnosis results using the original diagnosis training dataset are again compared with results obtained after applying CWGAN with ELQCP to show the effectiveness and importance of the unsupervised learning technique in the particular problem. The F-score rates increase more than 0.1 in all cases while the faulty

Table 5
The fault detection precision rates obtained from the proposed semi-supervised framework comparing with the conventional supervised fault detection method based on different traditional machine learning classifiers.

Detect.	5		10		20		30		40	
Precis.	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.
RF	**0.986**	0.276	**0.985**	0.283	**0.983**	0.284	**0.987**	0.284	**0.999**	0.285
SVM	**0.995**	0.229	**0.996**	0.247	**0.997**	0.275	**0.996**	0.279	**0.999**	0.279
MLP	**0.982**	0.225	**0.975**	0.237	**0.979**	0.241	**0.987**	0.248	**0.998**	0.256
KNN	**0.978**	0.260	**0.981**	0.276	**0.980**	0.282	**0.987**	0.282	**0.997**	0.282
DT	**0.970**	0.284	**0.957**	0.285	**0.969**	0.285	**0.937**	0.285	**0.999**	0.285

Table 6
The fault detection recall rates obtained from the proposed semi-supervised framework comparing with the conventional supervised fault detection method based on different traditional machine learning classifiers.

Detect.	5		10		20		30		40	
Recall	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.
RF	**0.979**	0.167	**0.974**	0.182	**0.970**	0.176	**0.982**	0.169	**0.999**	0.169
SVM	**0.995**	0.186	**0.995**	0.163	**0.997**	0.179	**0.996**	0.172	**0.999**	0.177
MLP	**0.975**	0.168	**0.962**	0.175	**0.971**	0.182	**0.984**	0.181	**0.998**	0.183
KNN	**0.972**	0.173	**0.976**	0.178	**0.973**	0.175	**0.986**	0.175	**0.997**	0.178
DT	**0.954**	0.168	**0.917**	0.168	**0.944**	0.167	**0.912**	0.166	**0.999**	0.173

Table 7
The fault detection F-score rates obtained from the proposed semi-supervised framework comparing with the conventional supervised fault detection method based on different traditional machine learning classifiers.

Detect.	5		10		20		30		40	
F-score	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.
RF	**0.977**	0.178	**0.969**	0.188	**0.971**	0.184	**0.981**	0.187	**0.999**	0.182
SVM	**0.995**	0.137	**0.995**	0.155	**0.997**	0.184	**0.996**	0.188	**0.999**	0.188
MLP	**0.973**	0.167	**0.957**	0.174	**0.969**	0.182	**0.984**	0.187	**0.998**	0.193
KNN	**0.971**	0.178	**0.975**	0.194	**0.972**	0.193	**0.986**	0.193	**0.997**	0.190
DT	**0.944**	0.183	**0.892**	0.184	**0.931**	0.184	**0.888**	0.182	**0.999**	0.192

Table 8
The fault diagnosis accuracy rates of the proposed framework with CWGAN-ELQCP (prop.) comparing with ordinary supervised learning method (conv.) using different classifiers.The number of faulty training samples varies from 5 to 40 for each fault type.

Diagno.	5		10		20		30		40	
Accu.(%)	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.
RF	**49.47**	40.22	**61.76**	55.63	**73.24**	67.08	**78.86**	72.88	**87.35**	78.36
SVM	**51.04**	40.17	**61.17**	57.06	**71.41**	65.84	**78.14**	72.58	**86.08**	76.71
MLP	**31.14**	25.17	**36.82**	36.30	**46.66**	43.13	**50.64**	46.81	**61.30**	49.45
KNN	**47.25**	31.56	**60.44**	49.60	**71.81**	63.54	**79.33**	72.20	**86.53**	77.87
DT	**44.69**	40.50	**52.54**	45.75	**63.36**	54.86	**68.39**	61.40	**78.47**	65.59

Table 9
The fault diagnosis precision rates of the proposed framework with CWGAN-ELQCP (prop.) comparing with ordinary supervised learning method (conv.) using different classifiers.

Diagno.	5		10		20		30		40	
Precis.	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.
RF	**0.562**	0.498	**0.670**	0.610	**0.767**	0.719	**0.816**	0.783	**0.884**	0.814
SVM	**0.582**	0.516	**0.674**	0.633	**0.757**	0.716	**0.816**	0.772	**0.878**	0.804
MLP	**0.329**	0.311	**0.386**	0.357	**0.486**	0.452	**0.521**	0.490	**0.640**	0.516
KNN	**0.488**	0.397	**0.560**	0.482	**0.658**	0.578	**0.705**	0.643	**0.796**	0.683
DT	**0.530**	0.431	**0.646**	0.548	**0.745**	0.667	**0.811**	0.743	**0.874**	0.794

training dataset size is increased to 3000 samples for each fault type (i.e., from 5 to 3000, from 10 to 3000, ..., from 40 to 3000). The classification accuracy, precision, recall and F-score rates of the proposed framework with CWGAN-ELQCP comparing with conventional supervised learning approaches using different classifiers in the fault diagnosis phase are shown in Tables 8–11, respectively.

From Tables 8–11, with less than 40 faulty training samples available for each fault type and without CWGAN-ELQCP, it is difficult for conventional supervised learning based methods to achieve acceptable diagnosis accuracy. The highest classification accuracy, precision, recall and F-score rates are 78.36%, 0.814, 0.792 and 0.780, respectively, using random forest as the classifier. The CWGAN technique has shown its significance by improving the highest classification accuracy, precision, recall and F-score rates to 87.35%, 0.884, 0.841 and 0.868, respectively, using the proposed framework. With both higher F-score rates for fault detection and

Table 10
The fault diagnosis recall rates of the proposed framework with CWGAN-ELQCP (prop.) comparing with ordinary supervised learning method (conv.) using different classifiers.

Diagno.	5		10		20		30		40	
Recall	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.
RF	**0.433**	0.441	**0.552**	0.552	**0.654**	0.687	**0.709**	0.713	**0.841**	0.792
SVM	**0.439**	0.463	**0.548**	0.570	**0.636**	0.658	**0.697**	0.725	**0.799**	0.767
MLP	**0.290**	0.316	**0.357**	0.368	**0.449**	0.431	**0.488**	0.468	**0.582**	0.494
KNN	**0.406**	0.363	**0.498**	0.457	**0.586**	0.548	**0.631**	0.614	**0.737**	0.655
DT	**0.447**	0.369	**0.576**	0.496	**0.673**	0.635	**0.736**	0.722	**0.820**	0.778

Table 11
The fault diagnosis F-score rates of the proposed framework with CWGAN-ELQCP (prop.) comparing with ordinary supervised learning method (conv.) using different classifiers.

Diagno.	5		10		20		30		40	
F-score	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.	prop.	conv.
RF	**0.463**	0.413	**0.597**	0.530	**0.721**	0.668	**0.808**	0.691	**0.868**	0.780
SVM	**0.500**	0.401	**0.606**	0.563	**0.711**	0.648	**0.778**	0.713	**0.861**	0.753
MLP	**0.257**	0.272	**0.315**	0.329	**0.437**	0.399	**0.480**	0.439	**0.602**	0.469
KNN	**0.460**	0.354	**0.603**	0.494	**0.719**	0.635	**0.794**	0.721	**0.865**	0.778
DT	**0.427**	0.318	**0.513**	0.416	**0.625**	0.515	**0.677**	0.586	**0.781**	0.631

diagnosis, the proposed framework perfectly solves the problem of imbalanced training datasets for conventional supervised learning based FDD approaches, and bridges the gap between theoretical FDD approaches and real-world industrial applications.

6. Conclusion, limitation and future works

This work demonstrates the importance of unsupervised learning techniques, more specifically, the generative adversarial networks (GANs), in the field of data-driven FDD for air handling units (AHUs). In real-world scenarios, there is always the case that the number of faulty data samples is not enough for the training process. For example, a specific fault might be fixed within 40 min; and only less than or equal to 40 faulty data samples were collected, along with thousands of normal operational data samples. Conventional supervised learning based FDD methods cannot handle the above situation well, since the training dataset is extremely imbalanced. In contrast, the proposed CWGAN-ELQCP method is able to perform AHU FDD with less than or equal to 40 faulty training samples. According to the experimental results, the CWGAN-ELQCP method outperforms the conventional supervised learning based methods in both detection and diagnosis phases. For example, in the fault detection phase, for conventional supervised learning based FDD frameworks, the fault detection accuracy rates are always around 50%, since the normal training samples dominate the training dataset. After re-balancing training dataset using the CWGAN-ELQCP framework, fault detection accuracy reaches almost 1 for all types of base classifiers (Tables 4–6).

This work extends our previous work in [9] from three perspectives. First, a more sophisticated framework of CWGAN has been designed, compared to the framework shown in [9], to demonstrate a more concise way of generating synthetic faulty samples. Second, we complete the semi-supervised AHU fault detection and diagnosis framework using GAN, whereas [9] only deals with fault diagnosis of AHUs. According to our experimental results listed in Section 5, the proposed method has significant improvements on classification accuracy, precision, recall and F-score rates in the fault detection stage compared to the conventional supervised learning based methods. The importance of applying GAN in both detection and diagnosis processes has been emphasized. Second, we optimize the ensemble learning quality control protocol (ELQCP) by evaluating all combinations of base classifiers. The original ensemble learning quality control protocol with the combination of support vector machine (SVM), decision tree (DT) and random forest (RF) has been replaced by a combination of SVM, RF and multi-layer perceptron (MLP). In summary, this work shows the importance of applying unsupervised learning techniques in the data-driven FDD process and the possibility of performing reliable AHU FDD with a limited number of faulty training samples.

One limitation of the existing data-driven AHU FDD models is that the optimized parameters become invalid from one AHU to another. Parameter tuning is necessary whenever the configuration of the AHU changes. Future works targeting the above problem include utilizing online reinforcement learning approaches to learn different system configurations during the training process. And to further verify the robustness and generalization of the proposed method, another future work direction of this study is to apply the proposed method on more comprehensive real-world datasets.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Ke Yan: Conceptualization, Methodology, Writing - original draft, Formal analysis, Writing - review & editing. **Jing Huang:** Software, Data curation, Investigation. **Wen Shen:** Formal analysis, Writing - review & editing. **Zhiwei Ji:** Supervision, Validation, Software, Data curation, Investigation.

Acknowledgment

This study is supported by faculty start-up funding from National University of Singapore under grant number R-296-000-208-133.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.enbuild.2019.109689.

References

[1] Z. Du, B. Fan, J. Chi, X. Jin, Sensor fault detection and its efficiency analysis in air handling unit using the combined neural networks, Energy Build. 72 (2014) 157–166.

[2] T. Mulumba, A. Afshari, K. Yan, W. Shen, L.K. Norford, Robust model-based fault diagnosis for air handling units, Energy Build. 86 (2015) 698–707.

[3] R. Yan, Z. Ma, Y. Zhao, G. Kokogiannakis, A decision tree based data-driven diagnostic strategy for air handling units, Energy Build. 133 (2016) 37–45.

[4] Y. Zhao, J. Wen, F. Xiao, X. Yang, S. Wang, Diagnostic Bayesian networks for diagnosing air handling units faults–part i: faults in dampers, fans, filters and sensors, Appl. Therm. Eng. 111 (2017) 1272–1286.

[5] 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, Renewable Sustainable Energy Rev. 109 (2019) 85–101.

[6] Z. Shi, W. O'Brien, Development and implementation of automated fault detection and diagnostics for building systems: a review, Autom. Constr. 104 (2019) 215–229.

[7] Y. Fan, X. Cui, H. Han, H. Lu, Chiller fault diagnosis with field sensors using the technology of imbalanced data, Appl. Therm. Eng. (2019) 113933.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[9] C. Zhong, K. Yan, Y. Dai, N. Jin, B. Lou, Energy efficiency solutions for buildings: automated fault diagnosis of air handling units using generative adversarial networks, Energies 12 (3) (2019) 527.

[10] S. Shaw, L. Norford, D. Luo, S. Leeb, Detection and diagnosis of HVAC faults via electrical load monitoring, HVAC&R Res. 8 (1) (2002) 13–40.

[11] 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.

[12] Y. Zhao, F. Xiao, S. Wang, An intelligent chiller fault detection and diagnosis methodology using Bayesian belief network, Energy Build. 57 (2013) 278–288.

[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] Z. Shi, Building Operation Specialist: A Probabilistic Distributed Fault Detection, Diagnostics and Evaluation Framework for Building Systems, Carleton University, 2018 Ph.D. thesis.

[15] D. Chakraborty, H. Elzarka, Early detection of faults in HVAC systems using an XGboost model with a dynamic threshold, Energy Build. 185 (2019) 326–344.

[16] W. Kim, S. Katipamula, A review of fault detection and diagnostics methods for building systems, Sci. Technol. Built Environ. 24 (1) (2018) 3–21.

[17] K. 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.

[18] J. Wen, S. Li, Tools for Evaluating Fault Detection and Diagnostic Methods for Air-Handling Units, ASHRAE 1312-RP (2011) 1–173.

[19] 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.

[20] E.L. Denton, S. Chintala, R. Fergus, et al., Deep generative image models using a Laplacian pyramid of adversarial networks, in: Advances in Neural Information Processing Systems, 2015, pp. 1486–1494.

[21] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.

[22] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, Stackganbv√bv√: Realistic image synthesis with stacked generative adversarial networks, IEEE Trans. Pattern Anal. Mach. Intell. 41 (8) (2018) 1947–1962 IEEE.

[23] G. Antipov, M. Baccouche, J.-L. Dugelay, Face aging with conditional generative adversarial networks, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 2089–2093.

[24] X. Yan, B. Cui, Y. Xu, P. Shi, Z. Wang, A Method of Information Protection for Collaborative Deep Learning under GAN Model Attack, IEEE/ACM Trans. Computational Biol. Bioinformatics (2019) Early Access IEEE.

[25] G. Douzas, F. Bacao, Effective data generation for imbalanced learning using conditional generative adversarial networks, Expert Systems with applications 91 (2018) 464–471 Elsevier.

[26] Y. Zhang, X. Wang, H. Tang, An Improved Elman Neural Network with Piecewise Weighted Gradient for Time Series Prediction, Neurocomputing 359 (2019) 199–208 Elsevier.

[27] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, 2017, pp. 214–223.

[28] 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.

[29] K. Yan, X. Wang, Y. Du, N. Jin, H. Huang, H. Zhou, Multi-Step short-term power consumption forecasting with a hybrid deep learning strategy, Energies 11 (11) (2018) 3089 Multidisciplinary Digital Publishing Institute.

[30] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[31] M. Hu, W. Li, K. Yan, Z. Ji, H. Hu, Modern machine learning techniques for univariate tunnel settlement forecasting: a comparative study, Math. Prob. Eng. 2019 (2019).