Building and Environment 203 (2021) 108057

Contents lists available at ScienceDirect

Building and Environment

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

An explainable one-dimensional convolutional neural networks based fault diagnosis method for building heating, ventilation and air conditioning systems

Guannan Li [a], Qing Yao [a], Cheng Fan [b,*], Chunlin Zhou [c], Guanghai Wu [c], Zhenxin Zhou [d], Xi Fang [e]

[a] School of Urban Construction, Wuhan University of Science and Technology, Wuhan, PR China
[b] Department of Construction Management and Real Estate, Shenzhen University, Shenzhen, PR China
[c] The Third Construction Co.Ltd of China Construction Third Engineering Bureau, Wuhan, PR China
[d] School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan, PR China
[e] College of Civil Engineering, Hunan University, Changsha, PR China

ARTICLE INFO

Keywords:
Deep learning
Convolutional neural networks
Fault diagnosis
Building energy systems
Fault class-discriminative feature
Model visualization

ABSTRACT

Due to the frequently changed outdoor weather conditions and indoor requirements, heating, ventilation and air conditioning (HVAC) experiences faulty operations inevitably throughout its lifespan. Therefore, it is important to monitor and diagnose HVAC fault operations. Recently, deep learning methods have attracted more attentions for their guarantee of better diagnosis performance under various system configurations and operating conditions. However, these methods are black-box models which though highly accurate for fault diagnosis but are extremely hard to explain. To overcome the disadvantage of poor interpretability of deep learning black-box models, this study therefore proposes a novel explainable deep learning based fault diagnosis method that is suitable for HVACs. To maintain HVAC operational information and variable locations of all chiller input data samples, proposed method is established with three characteristics: 1) the pooling layer is excluded, 2) the size of convolution filter kernel is set as 1, and 3) use softsign as an activation function. Considering the resulting impacts of HVAC faults on system operating variables, a new Absolute Gradient-weighted Class Activation Mapping (Grad-Absolute-CAM) method is proposed to visualize the fault diagnosis criteria and make the model explainable by providing the fault-discriminative information. The proposed method is validated using fault experimental dataset of a typical building HVAC system (i.e., chiller) from the ASHRAE research project 1043 (RP-1043). The fault diagnosis accuracy is over 98.5% for seven chiller faults. Results indicates that it is capable of interpreting the model work mechanism by activation feature maps and explaining the fault diagnosis criteria by Grad-Absolute-CAM.

1. Introduction

Heating, ventilation and air conditioning (HVAC) is an energy consumption system that used for generating controlled indoor environment and air quality in building spaces. It is widely applied in many indoor circumstances where the certain indoor temperature, humidity, CO2 concentration conditions are required by consumers. For different building circumstances, HVAC is employed to work under a variety of operating conditions. Due to the frequently changed indoor requirements and outdoor weather, part load operating conditions and high system nonlinearity, HVAC has great risks of suffering various unhealthy working processes and abnormal energy consumption operations in the long-term lifespan [1]. Therefore, it is very important to diagnose and monitor the operating conditions of HVAC since the resulting faults can lead to poor indoor environment, increasing energy penalty and maintenance cost, and even system damages [1–3].

Many researchers have attempted to solve the fault diagnosis problems through various methods [4–6]. For HVAC systems, fault diagnosis methods can be divided into quantitative methods, qualitative methods and historical process data-based methods [7,8]. Compared with the quantitative and qualitative model-based methods, the data-based fault diagnosis methods have proven to be of more generalized diagnosis

* Corresponding author.
E-mail address: fancheng@szu.edu.cn (C. Fan).

https://doi.org/10.1016/j.buildenv.2021.108057
Received 10 March 2021; Received in revised form 13 June 2021; Accepted 14 June 2021
Available online 17 June 2021

performance especially for large-scale HVAC systems [9]. From the perspective of data analytics, fault diagnosis is a type of data classification problem. Mainly, its objective is to classify the HVAC input data samples, to decide whether they are from the normal class or fault class and to further identify which fault class they belong to Ref. [10].

Various data based analytics algorithms have applied to monitor and diagnose faults in the HVAC system [11,12]. To obtain a generalized model, these data-based methods often aim to learn fault-discriminative information from a large variety of HVAC data to detect and isolate faults from the normal operating conditions [10]. Of these data-based methods, some are unsupervised learning based methods such as Principal Component Analysis (PCA) [13–17], Exponentially Weighted Moving Average (EWMA) [18–20], Clustering Analysis (CA) [21–24] and Association Rules Mining (ARM) [25–27]; some others are supervised learning based methods such as Decision Tree (DT) [28], Support Vector Machine (SVM) [29–31], Bayesian Network (BN) [32–39], Artificial Neural Networks (ANNs) [21,40], Fuzzy Neural Network [41,42], Random Forest (RF) [43] and Ensemble Learning (EL) [44–46]; and there are also some semi-supervised learning based methods [47–50].

For the coming era of big data in buildings, various data-based fault diagnosis using deep learning have been studied for the HVAC system due to the rapid developments of building intelligence [51,52] along with the increasing data accumulation ability and calculation speed. Deep learning algorithms such as Deep Neural Network (DNN) [53], Autoencoder (AE) [54,55], Deep Belief Network (DBN) [56,57], Recurrent Neural Network (RNN) [58,59] and Convolutional Neural Network (CNN) [60–65] are mainly explored for HVAC fault diagnosis. Recently, these deep learning methods have received more and more attention because they guarantee better diagnosis performance under various system configurations and operating conditions. Owing to their advantages of extracting deep fault-discriminative features and realizing high-precision nonlinear fitting, various deep learning models are widely applied in computer vision, natural language processing, industrial process monitoring and mechanical fault diagnosis, etc.

However, most of these deep learning methods are black-box models. Although they are highly accurate for system healthy state monitoring and fault diagnosis, these methods are extremely hard to interpret or even explain. It is difficult to understand the working mechanism and diagnosis criteria of a deep learning model that classifies the HVAC input data samples as fault or normal. From the perspective of practical applications, model interpretation on deep learning methods is of large significance. For many practical issues, system operators not only requires higher diagnosis accuracy, but also a detailed report of further explanations on the deep learning based fault diagnosis process to the system managers [66]. In this way, managers and operators can better understand and trust these deep learning methods.

To overcome the disadvantage of poor interpretability and explainability of deep learning black-box models, interpretable and explainable artificial intelligence research [67,68] has become an active topic on the applications of deep learning methods. Especially for image classification research area, the visualization technique is adopted to make deep learning models interpretable and explainable. Visualization aims to figure out which part of the input image contributes most to the final model classification output. In other words, it makes the criteria of the deep learning model based image classification process visible. Recently, the popular used back-propagation-based visualization method can provide activation map to illustrate the locations where the pixels in the input image have important influences on the classification output by propagating output to previous layers [69,70]. Similarly, data based fault diagnosis problem is also a type of data classification task. In terms of the fault diagnosis task, visualization helps understand the diagnosis criteria of the deep learning model based on learned fault-discriminative features.

Unlike images with two-dimensional size and only non-negative pixel values in the range of 0–255, the building HVAC operational data contain both positive and sub-zero values of various units. These data variables are generated from built-in sensor measured thermo-physical parameters, electrical signatures (i.e. temperatures, pressures, flow rates, electrical current, voltage and power input, etc.), along with other operational information (i.e., control signals, operating status, valve opening, control commands, feedback and alarms, etc.) recorded in the building management system [10]. Previous studies have proven that some HVAC data variables can be acted as fault indicators [29,71,72] to separate the fault from the normal. Consequently, the fault indicative variables can be utilized to establish the diagnosis criteria when determining normal and fault. Many fault indicators based on virtual sensors [73,74], grey box modeling [75–77], features selection [72,78] are developed for the purpose of fault diagnosis in the HVAC system. However, using either mathematical equations or optimization functions to convert the original data into virtual sensor-based and selected fault indicators will result in additional computation cost. In order to improve the computational efficiency, the one-dimensional (1D) deep CNN model can be used directly for fault diagnosis and feature extraction without converting the raw data to two-dimensional images [64,79,80]. Actually, it is not difficult to apply the traditional 1D deep CNN model to diagnose and monitor HVAC faults. However, if an interpretable and explainable CNN model is required for HVAC fault diagnosis, problems will arise. There is a lack of researches on explaining the diagnosis mechanism of deep learning models for HVAC fault diagnosis. Currently, most deep learning based faults diagnosis studies focus only on the diagnosis performance and none of them consider visualization. Therefore, in order to address the research gap, this study aims to propose a novel explainable deep learning based fault diagnosis method that is suitable for HVAC systems.

Main contributions of this paper are as follows:

1) A novel deep neural network that adopts softsign as activation function, excludes Pooling layer and filter with >1 size kernel is established to maintain HVAC operational information and variable locations of all chiller input data samples, which as a result can further visualize the diagnosis criteria in determining the fault and the normal.
2) As a modification of Gradient-weighted Class Activation Mapping (Grad-CAM), the new Absolute Gradient-weighted Class Activation Mapping (Grad-Absolute-CAM) method is proposed to visualize the fault diagnosis criteria using the learned network by considering the resulting impacts of HVAC faults on normalized variables.
3) Using the proposed Grad-Absolute-CAM, the diagnosis criteria of the proposed deep neural network model are explained, and validated using the widely used fault experimental data from a typical HVAC system, chillers.

The rest of the paper is organized as follows. Section 2 presents the basic principle of the CNN for data classification and Grad-CAM for visualization. Section 3 describes the proposed deep learning method for fault diagnosis and Grad-Absolute-CAM for model explanation. In Section 4, the proposed method is firstly validated using the widely used chiller fault data from the public RP 1043 experiments [81]. Next, the model working mechanism is interpreted for fault diagnosis and feature learning. Further, the model diagnosis mechanism is explained for the criteria to classify the fault and the normal. The last section concludes the paper.

2. Principle of 1D-CNN and Grad-CAM

2.1. One-dimensional convolutional neural network

A typical CNN [82] is a multilayer network consisting of the input layer, alternative convolutional (CONV) layers and pooling layers, fully connected layer, and output layer as depicted in Fig. 1. The 1D HVAC operational data are first inputted into the input layer; Next, they are used for feature extraction in the core module of CNN, the CONV layer;

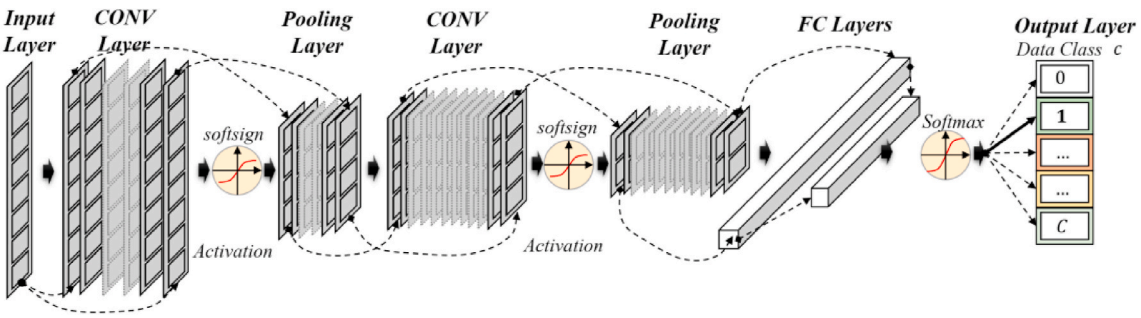

Fig. 1. Illustration of CNN structure for a data classification task.

Further, extracted features of the CONV layer are flattened and then inputted into the fully-connected (FC) layers which work in a similar manner as the traditional back-propagating (BP) neural network; Last, the output layer uses the classifier for data classification.

Mathematically, for an 1-D HVAC input with length N (number of variables), its output after 1-D convolution operation using an 1-D filter of fixed size Q in the lth CONV layer can be expressed as in Eq. (1).

$$a_i^l = \sum_{p=i}^{i+Q-1} u_{p-i}^l x_p^{l-1} \ (i = 1, 2, \ldots, N) \tag{1}$$

where u_p denotes the pth coefficient of 1-D filter and x_p denotes the pth element of 1-D input.

For the kth channel, its output $a_i^{l,k}$ of 1-D CONV layer can be expressed as in Eq. (2).

$$a_i^{l,k} = \sigma\left(\sum_{j=1}^{J} \sum_{p=i}^{i+Q-1} u_{p-i}^{l,j} x_p^{l-1,j} + b^{l,j}\right) \tag{2}$$

where $u_p^{l,j}$ denotes the pth coefficient of jth channel of 1-D convolutional filter in the lth CONV layer, $x_p^{l-1,j}$ denotes pth input element that received by the jth channel of the lth CONV layer, J denotes the total number of channels in the previous CONV layer, $b^{l,j}$ is the bias of convolution, and $\sigma(x) = signsoft(x) = \frac{1}{1+|x|}$ is the softsign activation function used in this study for CONV layer de-linearization. Using Eq. (2), the feature map activations A_i^k can be calculated as expressed in Eq. (3).

$$A_i^k = \sigma\left(\sum_{j=1}^{J} \sum_{p=i}^{i+Q-1} u_{p-i}^{j} x_p^{j} + b^{j}\right) \tag{3}$$

Further, the FC layers are usually employed to process the activation features learned by the last convolutional layer. Given a flattened 1-D HVAC activation feature data after data flatten operation, its mathematical calculation of the FC layer can be expressed as:

$$y_h^l = \sigma\left(\sum_{g=1}^{G} a_g^{l-1} v_{g,h}^l + r_h^l\right) \tag{4}$$

where $v_{g,h}^l$ denotes the weight coefficient located at (g, h) position of the connection weight matrix v, r_h^l denotes the hth element of bias vector in the lth FC layer, respectively. and a_g^{l-1} denotes gth element of the activation feature that received by the lth FC layer. Using Eq. (4), the score vector for fault class c before the softmax function can be calculated as expressed in Eq. (5).

$$y^c = \sigma\left(\sum av + r\right) \tag{5}$$

Lastly, the FC layer outputs are inputted into the output layer that employs a softmax function as shown in Eq. (6) to predict classification results.

$$f_{softmax}(y^c) = Prob(y^c) = \frac{e^{y^c}}{\sum_{c=1}^{C} e^{y^c}} \tag{6}$$

where $Prob(y^c)$ denotes the probability function that y^c belongs to fault class $c = 0, 1, 2, \ldots, C$.

2.2. Gradient-weighted Class Activation Mapping (Grad-CAM)

As a generalized form of Class Activation Mapping (CAM) [70], Grad-CAM [69] is a widely used visualization method for CNN models. For a classification task, as shown in Fig. 2, it focuses on visualizing the important feature map activations A_i^k using the gradient information that are transferred from the input layer to the last CONV layer of the 1-D CNN model for identifying a target class y^c. The degree of importance to A_i^k is assigned based on a global average pooling (GAP) operation on all BP gradients. The importance weights α_k^c of the kth channel in the target CONV layer can be obtained as in Eq. (7)

$$\alpha_k^c = \frac{1}{N} \sum_{i}^{N} \frac{\partial y^c}{\partial A_i^k} \tag{7}$$

where, N denotes the length of the 1-D input, $\frac{1}{N}\sum_{i}^{N}$ denotes the GAP operation and $\frac{\partial y^c}{\partial A_i^k}$ denotes the BP gradients, respectively.

The linear combination of feature map importance weights α_k^c and

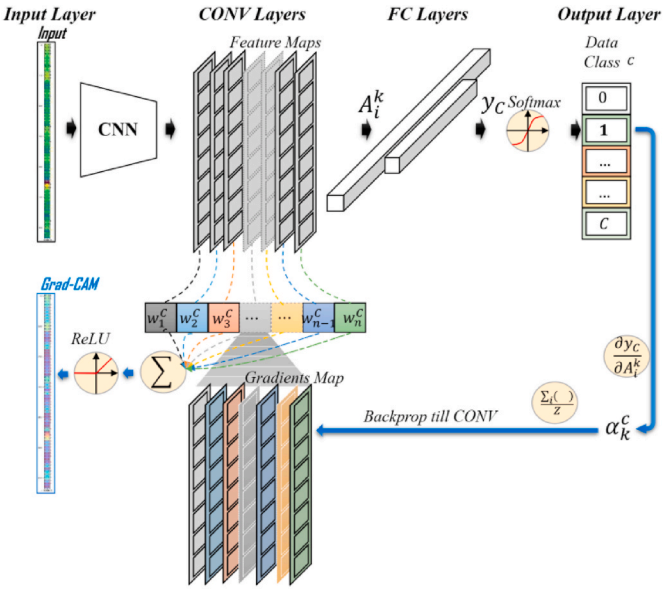

Fig. 2. Illustration of Grad-CAM for visualizing a CNN based data classification task.

the forward activation feature maps is the basis of Grad-CAM outputs as expressed in Eq. (8). To identify the target class y^c, Grad-CAM process uses a ReLU function after the linear combination to focus on features that have positive influences on the classification results. $\Omega^c_{Grad-CAM}$ is expressed as in Eq. (8).

$$\Omega^c_{Grad-CAM} = ReLU\left(\sum_k^K \alpha_k^c A^k\right) \tag{8}$$

where, K is the number of channel in the target CONV layer.

From the deduction process, Grad-CAM attempts to make the CNN model interpretable by conducting a reversed modeling process. What's more important is that the weight αc k in Eq. (8) denotes a partial linearization of the deep network downstream from A, and captures the importance of feature map k for a target class c which can be used to visualize the criteria for classifying the target class c [69].

Fig. 3 shows an example of visualizing the classification criteria for a single-layer CNN based data classification model using the Grad-CAM method. The model is designed to classify data from three different classes. Each data sample has three variables which means the input data is a size of $[3, 1]$. The data examples and their corresponding class labels (class $c = 1, 2, 3.$) can be seen in Fig. 3. It is obvious that the three classes differ from each other for their individual unique variable information. For instance, the class 2 data sample can be distinguished from the other classes for its positive value localized at the position of Variable No.1. In other words, data information from the Variable No.1 can be used as the criteria for classifying the target class $c = 2$.

To make the visualization process simply, the CNN model is defined with only a single channel. For an one-dimensional input data $(x, c) = ([0, 5.5, 0], [2])$ from data class 2, it is transformed to feature map activations $A_i^k = [0.53, -0.81, 0.53]$ after the convolutional and activation operation. Through the two FC layers, the class scores can be obtained as $y^c = [-1.9, -1.8, 3.1]$. Clearly, the CNN model correctly classifies the input data as Class 2. For the target Class 2, the importance weights in the CONV layer can be obtained using α_k^c in Eq. (7). Followed with the ReLU function, Grad-CAM returns the importance of each variable in the feature map from the channel by multiplying the importance weights with the feature map activations A_i^k. It can be found that the all the three variables in the input data have positive effects on the data classification process. And the most important variable for classification localizes at Variable No.1 of which has the largest $\Omega^c_{Grad-CAM} = 0.034$ than the other two variables. From Fig. 3, Grad-CAM captures the importance of each variable for class $c = 2$ and provides the classification criteria as located in Variable No.1.

3. Proposed methodology for HVAC health monitoring and fault diagnosis

This section presents the framework of the proposed interpretable deep learning method for HVAC health state monitoring and fault diagnosis. The framework contains the deep learning based model which utilizes 1D CNN to learn HVAC operating data and visualize the classification criteria based on a modification of the traditional Grad-CAM. Section 3.1 presents the proposed CNN structure with three characteristics: 1) the pooling layer is excluded, 2) the size of convolution filter kernel is set as 1, and 3) use softsign as an activation function. Section 3.2 provides the modified Grad-CAM for the visualization of diagnosis criteria of the proposed model.

3.1. Proposed 1D CNN structure for HVAC health monitoring and fault diagnosis

The proposed deep learning based fault diagnosis model is illustrated in Fig. 4. The model is constructed by three CONV layers using filter with size = 1 kernel. The pooling operation is not allowed in the model. The softsign is employed as activation function behind the each CONV layer. Further, two FC layers are used to receive the flatten data after convolution operation. Last, the softmax function is adopted as the classifier to classify the normal and fault data.

The reason why pooling layers and filters with size>1 kernel are excluded as follows: Prior to the CONV operation, the model inputs -HVAC operational data samples - are normalized 1D tensors. The HVAC input tensors are not as understandable as the digital images from the ImageNet dataset. It is required that the order of each variable in the input tensor should not be permuted so that the locations of activation features for diagnosis can be traced. Using fixed variable order, the critical fault class discriminative features that act as diagnosis criteria

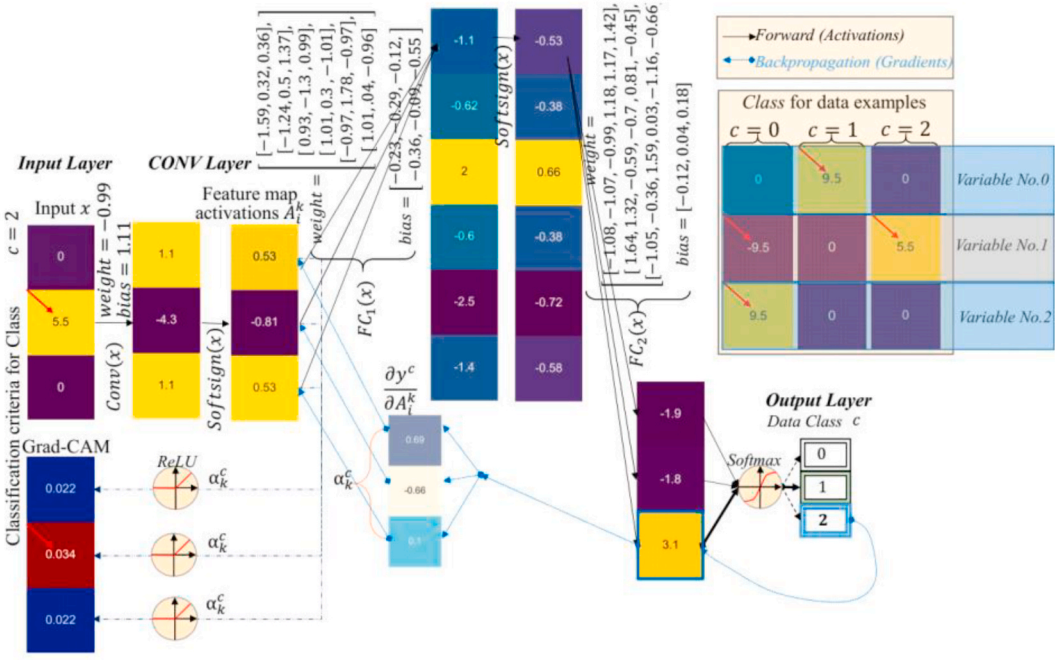

Fig. 3. An example of the Grad-CAM based classification (diagnosis) criteria for a single-layer CNN based fault data classification (Softsign as activation function).

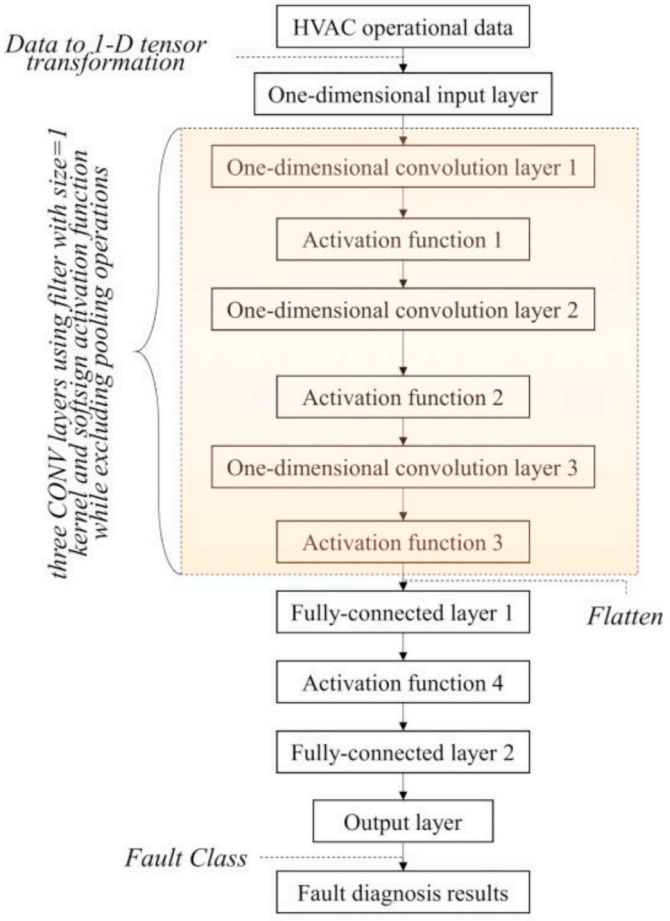

Fig. 4. Proposed structure for learning the 1D input HVAC data.

can be visualized from the Grad-CAM. Therefore, the pooling layer and the size>1 kernel that may cause order disruption and information loss are excluded from our proposal. By doing this, the features generated during the learning process of the CNN model maintain the system variable information (i.e., the order or location of each variable, positive and negative values, etc) in the input HVAC data. Using the learned features and comparing with fault indicators, the fault diagnosis criteria of the model can be visualized and interpreted reasonably.

The reason of choosing softsign as activation function is as follow. The proposed model uses the CONV layer to convolute HVAC input tensor with the size = 1 filter and transfer the convolutional to the nonlinear activation function. Generally, the activation layer is connected with the former CONV layer to learn non-linearity deep learning models. Although ReLU is widely used in deep learning, it has a problem of eliminating all negative values after the convolution operation. The elimination of negative values can cause unknown information changes. Unlike ReLU, softsign activation can learn nonlinearity from both positive and negative values [80] of the HVAC data samples. Hence, this study chooses softsign as activation function for maintaining the system operational information (i.e., the order or location of each variable, positive and negative values, etc.) in the HVAC input data as much as possible during the CNN modelling process.

3.2. Visualization of the diagnosis criteria using Grad-Absolute-CAM

The conventional Grad-CAM is primarily used for providing visual explanations from deep neural networks for image classification in the computer vision research area. For a color image, it consists of many pixels in two-dimensional and three-channel. Normally, all pixel values ranges from 0 to 255 which means an image data has no negative pixel. Grad-CAM utilizes the ReLU to remain the positive linear combination part in Eq. (8) that has positive influence on the target class c. In other words, it only put the interest on the pixels whose intensity should be increased in order to increase the score of target class [69].

However, the HVAC operational data sample is considerably different in two aspects. For one aspect, as can be seen later in Fig. 5, a normalized HVAC data sample has both positive and subzero values. This means that the negative feature map activations should be remained as much as possible. With concerns of this point, a modified Grad-Negative-CAM is proposed to focus on the ignored negative feature map activations in the traditional Grad-CAM. The Grad-Negative-CAM is expressed as in Eq. (9).

$$\Omega^{c}_{Grad-Negative-CAM} = ReLU\left(\sum_{k}^{K} \alpha^{c}_{k}\left(-A^{k}\right)\right) \tag{9}$$

For another aspect, previous studies indicated that the resulting impacts of HVAC faults on normalized faults indicators can be either positive or negative [83–87]. This indicates that both positive and negative gradients should be seriously considered. In order to combine the both aspects of the positive and the negative, instead of ReLU, this study applies an absolute function to components of the linear combination part. The modified Grad-Absolute-CAM that can preserve the possible lost negative information is expressed as in Eq. (10).

$$\Omega^{c}_{Grad-Absolute-CAM} = ReLU\left(\sum_{k}^{K} Abs\left(\alpha^{c}_{k}\right) Abs\left(A^{k}\right)\right) \tag{10}$$

Using the Grad-Absolute-CAM for HVAC fault data, the diagnosis criteria can be located for a particular fault class.

4. Results and discussion

Using the proposed deep learning based method, this section provides the fault diagnosis results and visualization of fault diagnosis criteria for a typical HVAC system, chillers. Section 4.1 describes the primary information of the chiller fault experimental data from the ASHRAE research project 1043 [81]. Section 4.2 presents the data pre-processing results. In Section 4.3, the proposed model is validated to be of high health state monitoring and fault diagnosis performance. In Section 4.4, the proposed model is used to extract fault features for specific fault. The model working mechanism is interpreted by providing visualization of activation feature maps. In addition, Section 4.5 shows the model diagnostic mechanism by providing visualization of diagnosis criteria and fault class discriminative features using Grad-Absolute-CAM. In this study, all experiments are conducted on an Intel(R) Core TM i5–8250H CPU @1.60 GHz with 8 GB RAM and 512 GB SSD. The proposed neural network model is implemented using Python 3.6 and Pytorch 1.4.0. The random seed is fixed to make the model results repeatable.

4.1. Description of the fault experimental data - ASHRAE research project 1043

To explore the model performance, interpretability and explainability, the fault experimental dataset of a typical building HVAC system (i.e., chiller) from the ASHRAE research project 1043 (RP-1043) [81] is used for the model validation. The fault experiment was conducted on a R134a refrigerant centrifugal chiller with 90 tons cooling capacity. The condenser and evaporator are shell and tube type. For the tested chiller system, Table 1 lists the one normal healthy state (NRM) and seven typical faulty states used for validation in this study. The seven chiller fault states are condenser fouling (CdF), non-condensate gas in refrigerant (NCdG), reduced water flow in evaporator (RWEv), reduced water flow in condenser (RWCd), excessive oil (ExO), refrigerant leakage (RfL) and refrigerant overcharge (RfO). Each fault is tested with four fault

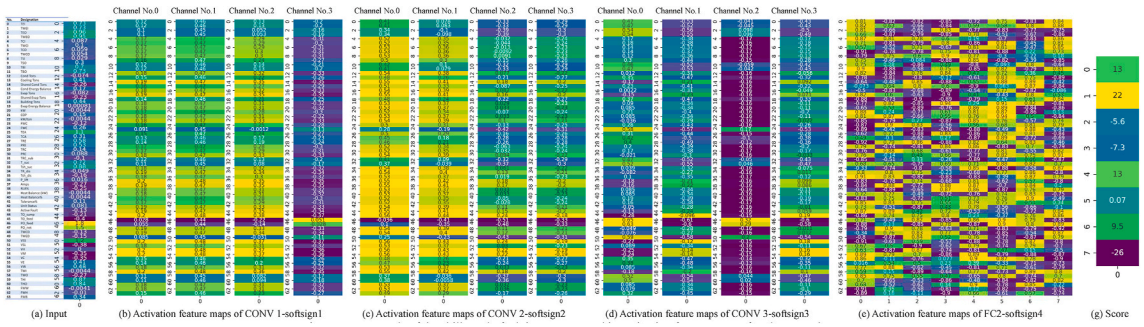

Fig. 5. An example of the chiller CdF fault input tensor and its activation feature maps of each CONV layer.

Table 1
Description of the RP-1043 chiller faults.

Health state	Fault experiment approach for four severity levels	Fault class
Normal (NRM)	–	0
Condenser fouling (CdF)	Plugging tubes by 10%, 20%, 30%, 40%	1
Non-condensate gas in refrigerant (NCdG)	Adding nitrogen volume by 1.0%, 1.7%, 2.4%, 5.7%	4
Reduced water flow in evaporator (RWEv)	Reducing water flow rate by 10%, 20%, 30%, 40%	5
Reduced water flow in condenser (RWCd)	Reducing water flow rate by 10%, 20%, 30%, 40%	6
Excessive oil (ExO)	Increasing lubricant in charge by 14%, 32%, 50%, 68%	7
Refrigerant leakage (RfL)	Discharging refrigerant weight by 10%, 20%, 30%, 40%	2
Refrigerant overcharge (RfO)	Overcharging refrigerant weight by 10%, 20%, 30%, 40%	3

severity levels under 27 different operating conditions. The severity level increases from SL-1 to SL-4 as the fby changing three control variables: chilled water supply temperature, condenser water entering temperature and chiller cooling load. In each operating condition, the experimental test is first carried out to reach the steady state for about 30 min at most. Next, the steady state test is continuously performed for 15–25 min for each operating condition. Finally, the experimental test data are collected by the chiller controller and transferred to the computer for further system health state analyses. Each data sample consists 64 variables of which 48 are directly measured variables and 16 are indirect calculated variables.

4.2. Data pre-processing results

The data pre-processing consists of four sub-steps: data normalization, steady-state data filter, data split and data-to-tensor transformation.

First, Z-score standardization is adopted to conduct data normalization for each of the 64 numerical variables. In this study, only the NRM data are used to calculate the means and unit variances for each variable. All fault data are converted to standardization formats for latter data analyses. The normalization process can greatly reduce the effects of different units and different orders of magnitude on the performance of monitoring and diagnosing faults.

Second, the steady-state data filter [88] is used to remove the transient-state data in RP-1043 fault data sets. This is because transient-state data may deteriorate the model performance. Further, the remain data samples are kept the same size of 170 in each fault class to avoid the unknown effect caused by data imbalance between classes [89,90]. The *Reduced Dataset* of RP-1043 is chosen and the number of raw data samples is 433. After the steady-state data filter is implemented on three state variables, i.e., temperatures of evaporator water in (*TWEI*), evaporator water out (*TWEO*) and condenser water in (*TWCI*), the number of steady-state data samples drops to 170 for each fault severity level [72].

The third sub-step is dataset split. 60% of the steady-state *Reduced Dataset* is randomly selected as model training set and the rest 40% is used for model validation and testing (20% for validation and 20% for testing). The data split is performed in a stratified and non-repetitive sampling manner, which randomly chooses 60% of the data samples for each severity of each fault class. This means that the training data set contains 102 samples for each fault severity level. For the left 68 samples, half are the validation set and half are the testing set.

Last, data-to-tensor transformation is conducted for each sample in the three sets. The CNN model inputs are required to be the format of image tensors. So each chiller data sample is transformed to its 1D tensor with size of [1,64]. An example of the chiller CdF input tensor is shown in Fig. 5(a). The descriptions of each variables corresponding to the variable number can be found in Fig. 5(a).

4.3. Health monitoring and fault diagnosis results

For the proposed CNN model, the model training process uses the CrossEntropyLoss and the Adam optimizer to develop a classification model reduce the fault diagnosis error rate. The batch size is set as 64. The training epoch is set as 50 to reach the convergence in searching for

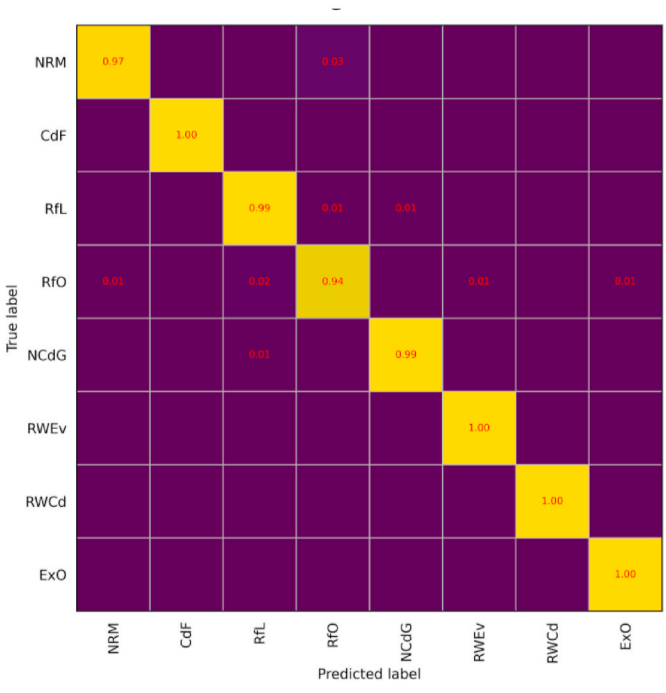

Fig. 6. Normalized confusion matrix of the three-layer CNN for testing set.

the near optimal classification model. The model loss curves of diagnosis error rates for 50 training epochs can be seen from Fig. 7. The model performance is evaluated by confusion matrix in Fig. 6 using the testing set. For health state monitoring, the model correctly identifies over 97% of the NRM data in the testing set. The rest 3% are wrongly classified as RfO. For fault diagnosis, the CNN model successfully identifies over 98.5% of all seven fault data in the testing set. Specifically, it shows

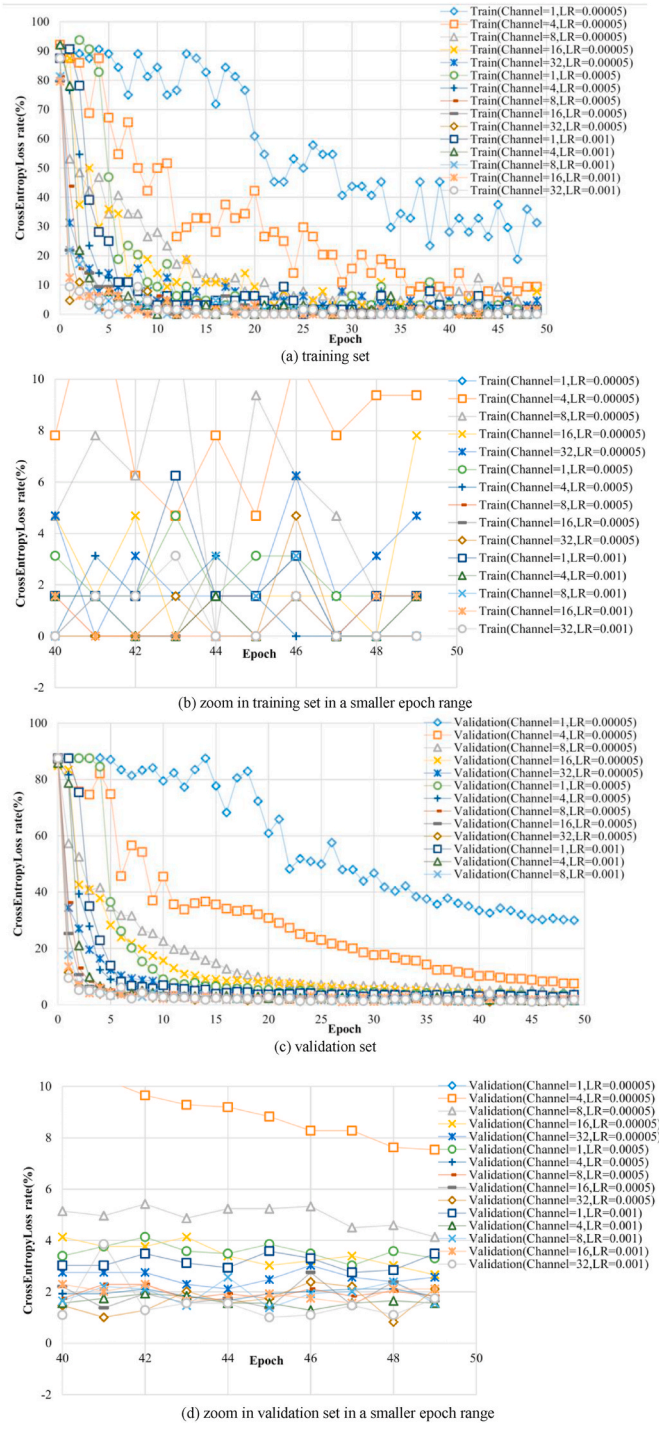

Fig. 7. Impacts of the network channels and the learning rate (LR) on model performance: (a) full screen of CrossEntropyLoss curves of the training set for all 50 epochs; (b) zoom-in visualization of the CrossEntropyLoss curves of the training set for the last 10 epochs; (c) full screen of CrossEntropyLoss curves of the validation set for all 50 epochs; (d) zoom-in visualization of the CrossEntropyLoss curves of the validation set for the last 10 epochs.

larger than 99% diagnosis accuracies for component faults, CdF, NCdG, RWEv, RWCd and ExO. While for the two system working fluid faults, RfL and RfO, the diagnosis accuracies are relatively smaller 99% and 94%. Due to data similarity, the model confuses the two system faults at less severity levels with NRM. But for higher severity levels, the model can distinguish the two faults.

Generally, a CNN model with less layers, less channels and fewer training epochs reduces the model training cost. As described before, batch size and maximum epoch are set to be 64 and 50 respectively. Also, softsign and Adam are chosen as the activation function and optimization function, respectively. So parameter tuning is conducted on three factors, such as the learning rate, the number of channels in each convolutional layer and the volume of training data used for model training. Fig. 7 shows the impact of the learning rate and the channel number on the model training process. More channels in the convolutional layers (no more than 64 channels) and larger learning rate can reduce the model diagnosis error rate (%) to a smaller level at a relatively smaller epoch number. So, the learning rate is selected as 0.0005 and the number of channels is set as 64.

Further, the impact of the volume of training data used for model development on fault diagnosis performance. Fig. 8 illustrates the fault diagnosis performance of the proposed model that trained with various training datasets. As mentioned in Section 4.2, all training data samples are randomly selected from the whole steady-state *Reduced dataset* in a stratified and non-repetitive sampling manner. The training data samples are required to be different from those in the validation and testing sets. As the proportion of training set to the whole steady-state *Reduced dataset* increases from 10% to 60%, the diagnosis accuracy of the testing set rises from 92.1% to 98.6%. The CNN model successfully identifies about 92% of all normal and fault data even when the proportion is only 10%. The largest diagnosis accuracy can be over 98.6% if the proportion rises up to 60%.

4.4. Activation feature maps

The model working mechanism is interpreted by presenting the extracted activation feature maps. Fig. 5(a)–(g) present the feature maps from the input layer to convolutional layers and the class score before the Softmax classifier. To reduce the interpretation complexity, this study continuously chooses the CdF fault sample as an instance for visualization of feature maps. For all 64 channels (Channel No.0 to No.63) of each convolutional layer, only the former 4 channels (Channel No.0 to No.3) are presented for visualization. This makes the activation feature maps much easier to understand. The 64-channel activation feature maps of each CONV layer can be found in Appendix A. As shown in Fig. 5(b)–(d), the former 4-channel activation feature maps can be divided into two types according to their largest activations. In Fig. 5(a), the activation feature maps are very similar for channels No.0 to No.2 in the CONV 1-softsign 1 layer while channel No.3 is the opposite. For the former type of activation feature map (channel No.0 to No.2), the largest activation is located at *45 TO_feed*. Actually, other four activations have approximate values to *45 TO-feed*. They are located at *51 VSL* (Large Steam Valve Position), *53 VM* (3-way Mixing Valve Position), *54 VC* (Condenser Valve Position) and *58 TWO* (Temperature of City Water Out). But for the latter type of activation feature map (channel No.3), *46 PO_feed*, *47 PO_net* and *50 VSS* (Small Steam Valve Position) show relatively larger feature map activations than the other 61 variables in the activation feature map. Fig. 5(e) illustrates the activation feature map of the FC2-softsign4 layer. Unfortunately, it is hard to interpret this layer for its 512 neuron units and non-linear property. Lastly, the class scores are presented in Fig. 5(f). Class 1 obtains the largest score 22 before inputting into the Softmax classifier. This indicates that the CdF sample is correctly classified as the Class 1 which denotes the CdF fault. The t-distributed stochastic neighbor embedding (t-SNE) [91] based scatter plot of class score data for all seven chiller faults can be seen in Fig. 10.

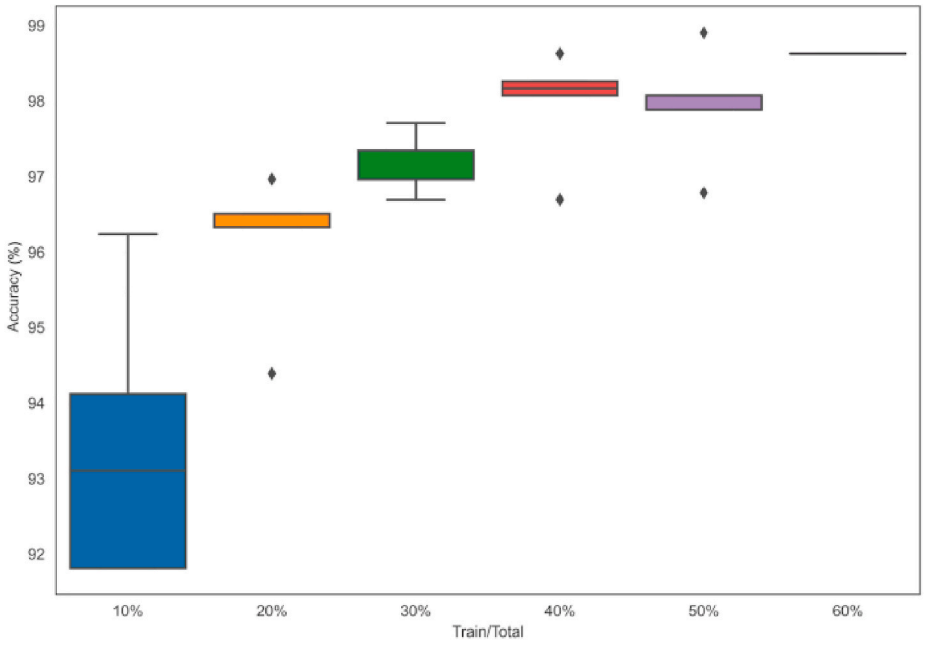

Fig. 8. Diagnosis accuracy of testing dataset for various training datasets.

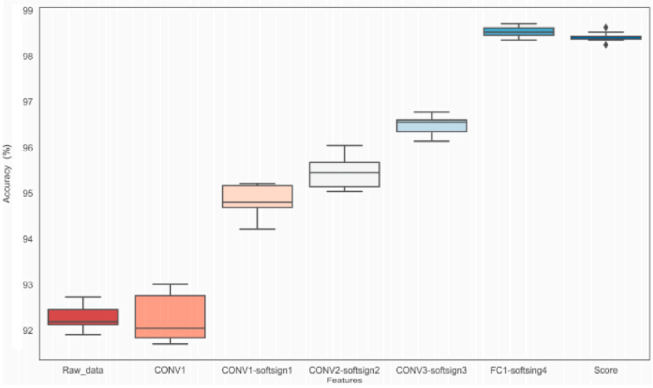

Fig. 9. Diagnosis accuracy of testing datasets for each probe on different layers.

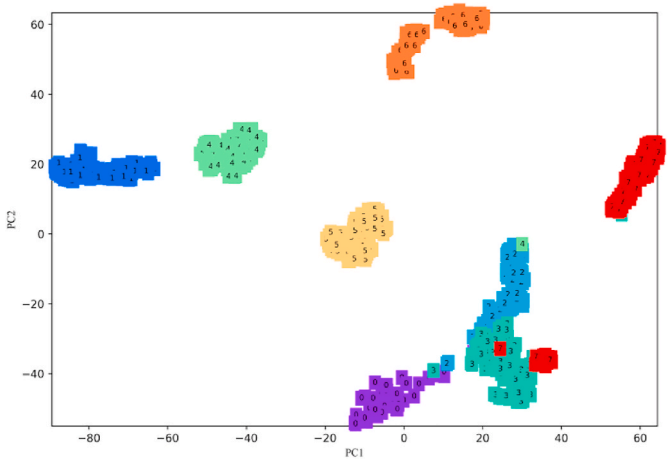

Fig. 10. T-SNE based two-dimensional scatter plot of class scores for the testing data in the last layer of the proposed CNN model. (For each sample, the number text represents the true fault class label as described in Table 1 in Section 4.1).

Further, a linear classifier probe method is adopted to understand the intermediate layers by monitoring the features at each CONV layer in the proposed model and evaluating how suitable they are for fault diagnosis [92]. Fig. 9 presents the testing diagnosis accuracy for each linear classifier probe at the end of training. Each probe uses the extracted features from the training set to train the Logistical Linear Regression classifier and test the corresponding testing features. The features are learned from the proposed model after 50 training epochs. The probes are repeatedly evaluated for ten times without setting the random state. Parameters tuning is conducted using a grid searching method to find a near-optimal pair of parameters for each probe. From Fig. 9, it can be found that the testing diagnosis accuracy grows from 92% to 98% using the features learned from each layer. The diagnosis accuracy becomes increasingly higher as the network moves to later softsign activated convolutional layers for the three-layer CNN. The biggest impact comes from the first softsign activated CONV layer. Although the testing diagnosis accuracy improves at each layer, the first softsign activated CONV layer is the more significant than the deeper layers (the second and the third) for feature extraction and fault diagnosis process.

4.5. Fault discriminative feature map activations and visualization of diagnosis criteria

4.5.1. Comparison results of feature map activations using Grad-CAM, Grad-Negative-CAM and Grad-Absolute-CAM

As described in Section 3.2, the impacts of both positive and negative types of gradients and activation feature maps should be investigated. Fig. 11 compares the outputs of the three visualization methods including the traditional Grad-CAM, Grad-Negative-CAM and Grad-Absolute-CAM. From Fig. 5(a), the chiller input tensor is actually a vector with one column and sixty-four rows. Each row represents one of the sixty-four chiller measured variables. The position of each variable is fixed in the chiller input tensor. For each chiller input tensor, Fig. 11 shows the feature activation heat maps based on the traditional Grad-CAM, Grad-Negative-CAM and Grad-Absolute-CAM. The feature activation heat maps are drawn with the colormap = 'jet' and alpha = 0.5 through matplotlib tools. According to the rule of colormap = 'jet', as the feature activation Ω^c increases from zero to its maximum positive, the color of the corresponding variable (row) changes from dark blue to

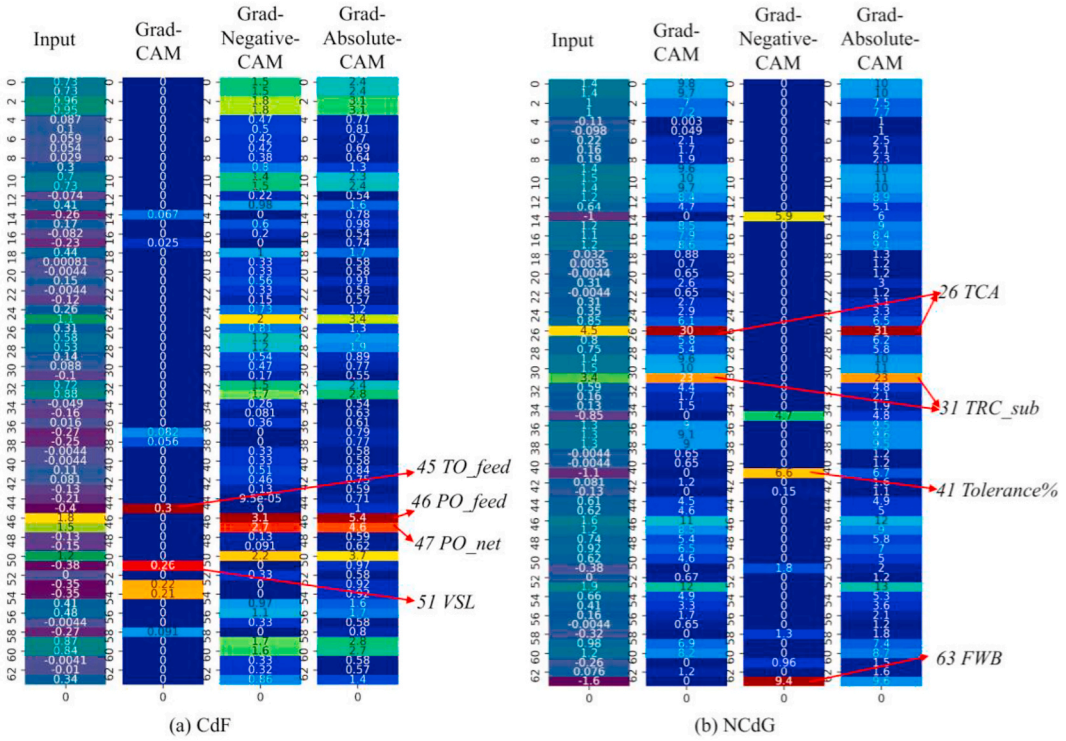

Fig. 11. Fault class discriminative information locations using the traditional Grad-CAM, Grad-Negative-CAM and Grad-Absolute-CAM for seven chiller faults at their forth severity levels: (a) CdF; (b) NCdG; (c) RWEv; (d) RWCd; (e) ExO; (f) RfL and (g) RfO. Note: The red arrows provide the locations of key variables that are top feature activations identified by the model visualization method. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

blue, green, light green, yellow, orange, red and dark red finally. From Fig. 3 in Section 2.2, the variables that are marked in dark red have larger feature activations Ω^c should be used as the fault indicators for classifying the target class c. So the variables colored in dark red contains the most important fault class discriminative information should be considered as the diagnosis criteria for the target fault.

Still, the CdF fault input tensor sample is taken as an example, both Grad-Negative-CAM and Grad-Absolute-CAM correctly identify the top two activations which are located at *46 PO_feed* and *47 PO_net*. From the RP-1043 report of the centrifugal chiller fault tests [81], *PO_feed* and *PO_net* are pressure of oil feed and oil feed minus oil vent pressure, respectively. Fig. 13 depicts the positions of the two oil pressures in the chiller system of RP-1043. In accordance with the professional knowledge and previous feature selection studies [72], the two variables *PO_feed* and *PO_net* are fault indicative features for the CdF fault. As shown in Fig. 12(a) and (b), the two variables are critical variables for separating the CdF fault data from the Normal data. However, the Grad-CAM outputs indicate that *51 VSL* (Large Steam Valve Position) and *45 TO_feed* (Temperature of Oil Feed) are top two activated features for the CdF fault. This means that the traditional Grad-CAM fails to find out the correct locations of CdF fault class activations. As indicated in Fig. 12(i) and (j), *51 VSL* and *45 TO_feed* do not contribute most for the classification of the CdF fault and the Normal statue. Moreover, Grad-Absolute-CAM is less sensitive than Grad-Negative-CAM since the positive values are also activated in the output. Unlike Grad-Negative-CAM and Grad-Absolute-CAM, Grad-CAM fails to locate the physically reasonable fault class activations for ExO and RfO faults. In contrast, Grad-CAM identifies the correct fault class activations for four faults including NCdG, RWEv, RWCd and RfL while Grad-Negative-CAM fails. Obviously, for the RWCd fault, the fault indicator is *23 FWC* (Flow Rate of Condenser Water) due to its fault generation approach. From Fig. 12(g), it can be found that *23 FWC* alone can separate the RWCd fault data from the Normal data and other fault data. Both Grad-CAM and Grad-Absolute-CAM shows the correct fault class discriminative information location. Whereas, Grad-Negative-CAM recognizes the top activated variable as *50 VSS* (Small Steam Valve Position) which is very hard to explain based on professional knowledge of the chiller system. Also, the boxplot of *50 VSS* in Fig. 12(m) shows no obvious difference between the RWCd fault and the Normal. To sum up, Grad-Absolute-CAM correctly identifies the locations of fault class activations for all seven faults although Grad-Absolute-CAM is less sensitive than Grad-CAM or Grad-Negative-CAM for those correctly identified faults.

4.5.2. *Grad-Absolute-CAM results of visualizing the diagnosis criteria using the last CONV layer*

Hence, the Grad-Absolute-CAM is used to explain the model diagnostic mechanism by visualization of the diagnosis criteria and localization of fault class-discriminative regions. Fig. 14 shows the outputs of Grad-Absolute-CAM for seven chiller fault classes at their forth severity levels. The locations of fault class-discriminative information for each chiller input tensor can be clearly observed in the Grad-Absolute-CAM based visualization maps. The proposed Grad-Absolute-CAM method is highly fault class-discriminative which contributes a lot on understanding the diagnosis criteria for distinguishing the seven chiller faults from the Normal. As the fault severity level increases, the locations of fault class-discriminative feature activations for component faults are more accurate. But for the two system faults RfL and RfO, their fault class-discriminative feature activations can be clearly visualized at the forth severity level.

Specifically, the Grad-Absolute-CAM visualization maps of all testing chiller data samples at CdF SL-4 are presented in Fig. 14(a). For CdF SL-4, the top two feature map activations are located at variable No. *46 PO_feed* (pressure of oil feed) and No. *47. PO_net* (oil feed minus oil vent

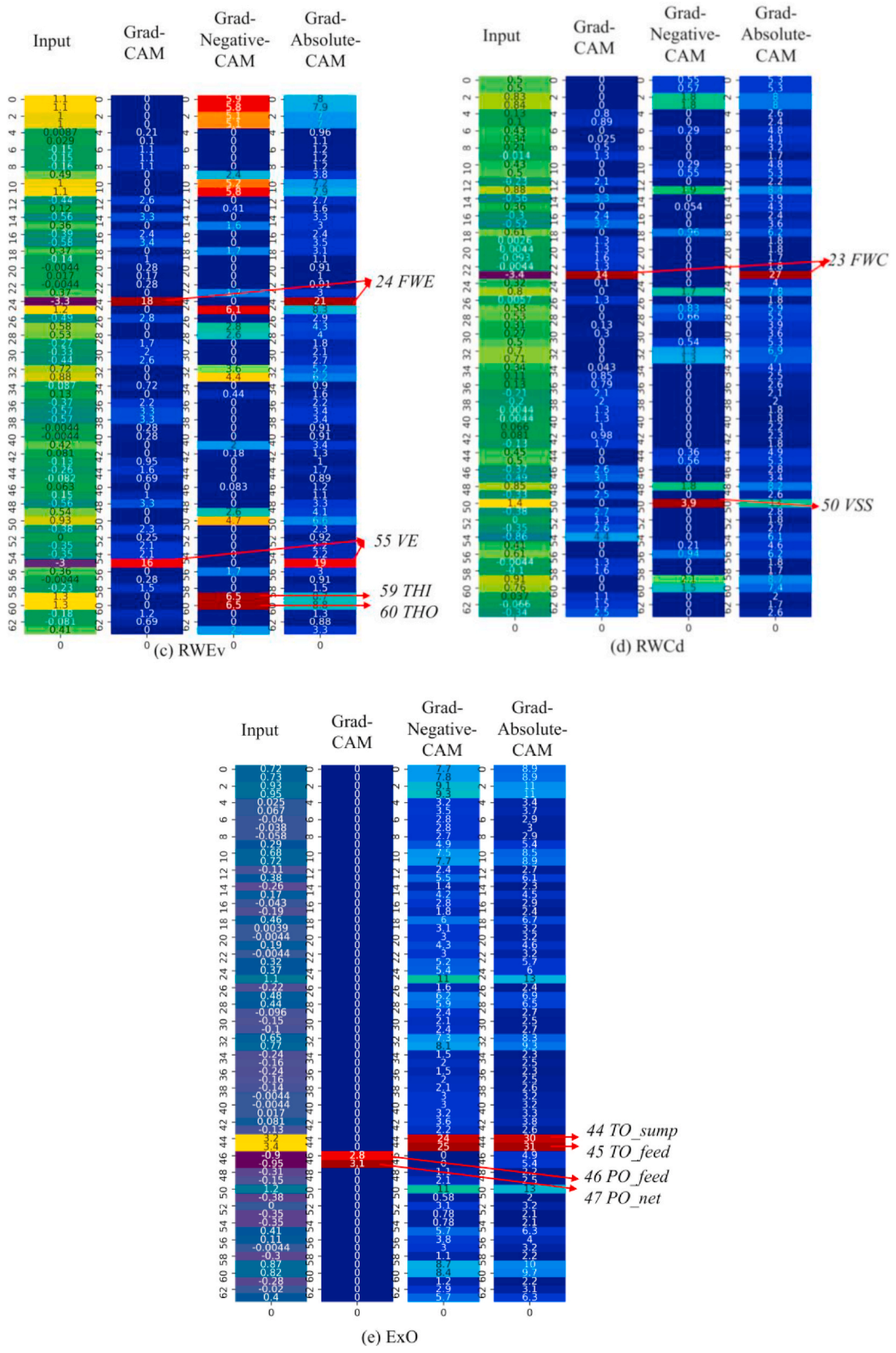

Fig. 11. (*continued*).

pressure). This means that the two variables *PO_feed* and *PO_net* are fault indicative features for the chiller CdF fault. The localization of fault-discriminative regions is reasonable based on professional knowledge about the chiller system. Further, Fig. 14(b)–(g) show the fault-discriminative regions for the other six faults. Obviously, Grad-Absolute-CAM is more sensitive to the five component faults (CdF, NCdG, RWEv, RWCd and ExO) than the other two system faults (RfL and RfO). All the five component faults have at most two activated variables for the local impacts on only an individual chiller component. Since the two system faults are related to the working fluid refrigerant that

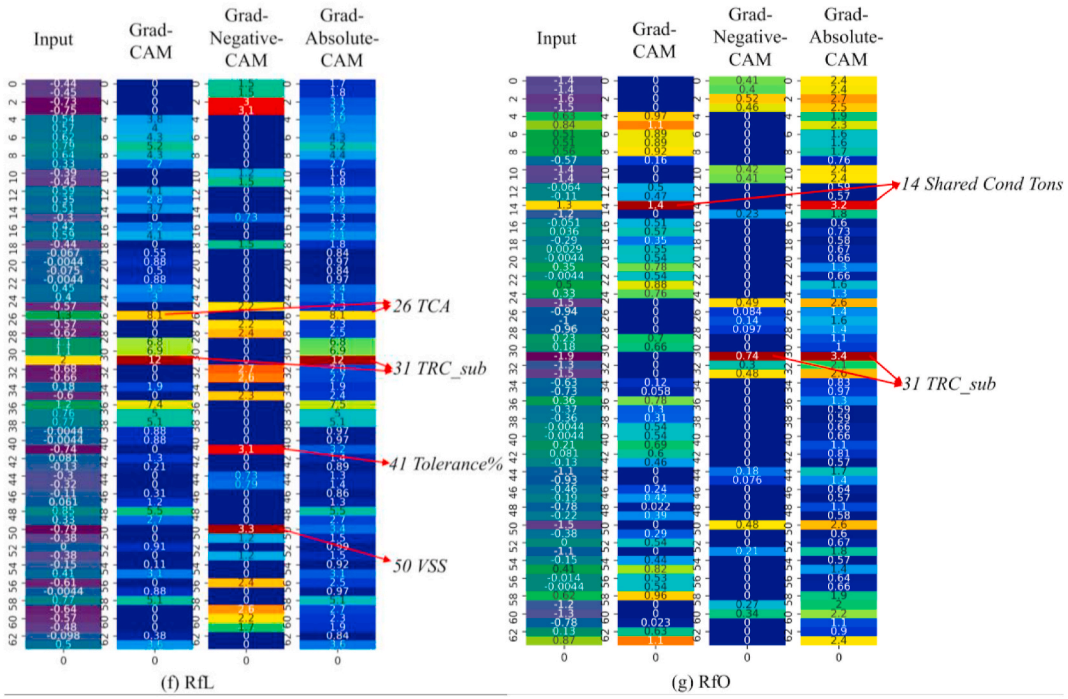

Fig. 11. (*continued*).

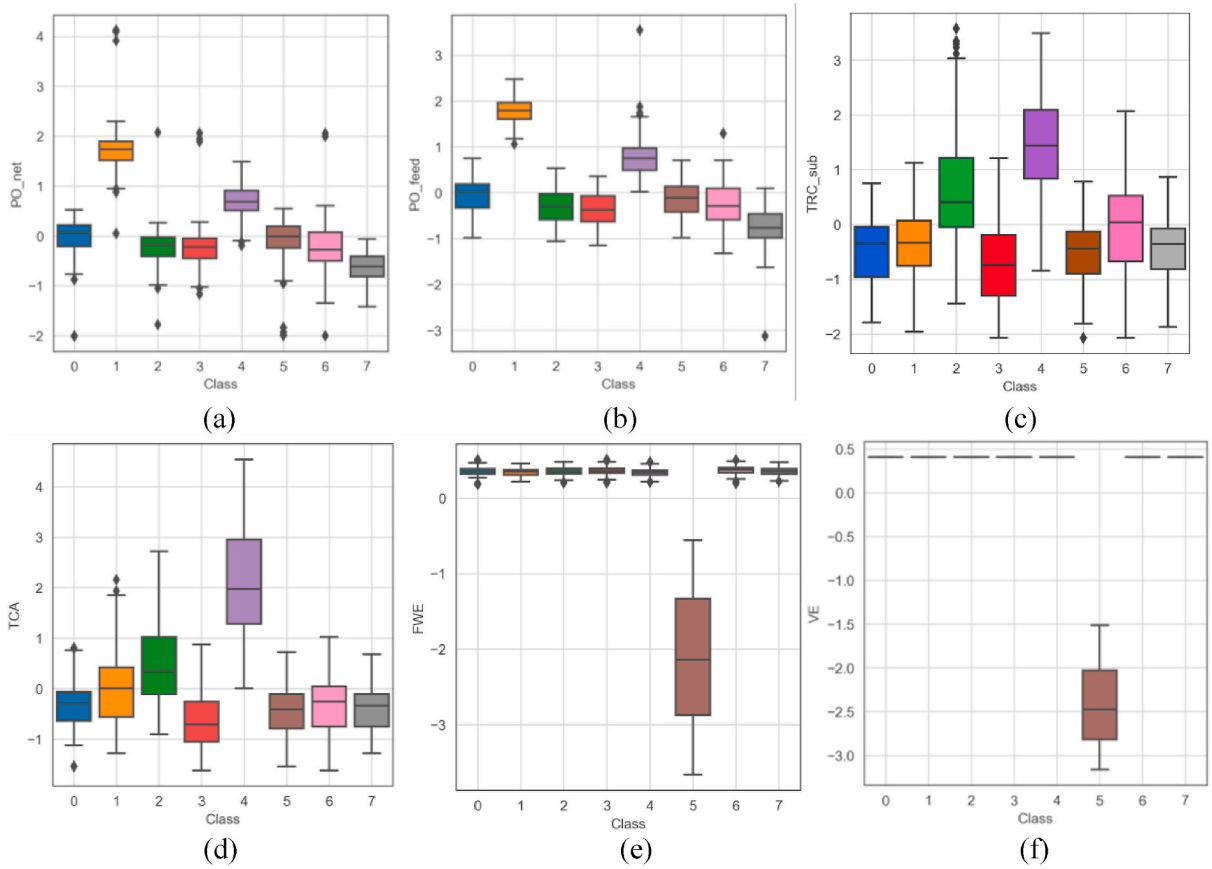

Fig. 12. Data distribution results for the RP-1043 chiller operating data after data pre-processing: (a) *47 PO_net* v.s. Fault Class; (b) *46 PO_feed* v.s. Fault Class; (c) *31 TRC_sub* v.s. Fault Class; (d) *26 TCA* v.s. Fault Class; (e) *24 FWE* v.s. Fault Class; (f) *55 VE* v.s. Fault Class; (g) *23 FWC* v.s. Fault Class; (h) *44 TO_sump* v.s. Fault Class; (i) *45 TO_feed* v.s. Fault Class; (j) *51 VSL* v.s. Fault Class; (k) *14 Shared Cond Tons* v.s. Fault Class; (l) *61 FWW* v.s. Fault Class; (m) *50 VSS* v.s. Fault Class; (n) *31 TRC_sub* v.s. RfO Fault Severity Level; (o) *61 FWW* v.s. RfL Fault Severity Level.
Note: The horizontal axis is the fault class/RfO fault severity level (SL)/RfL fault severity level (SL)as introduced in Table 1.

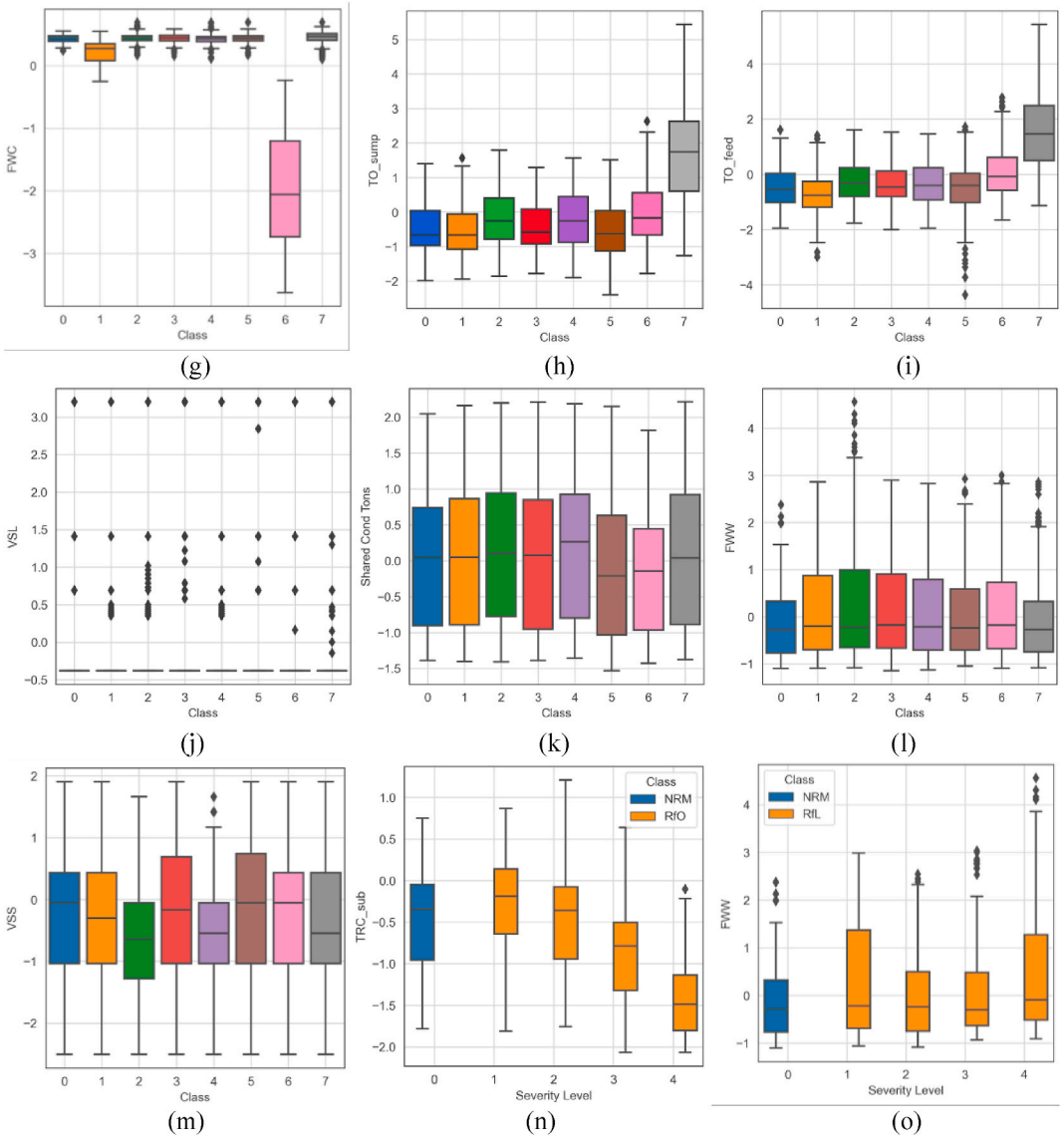

Fig. 12. (*continued*).

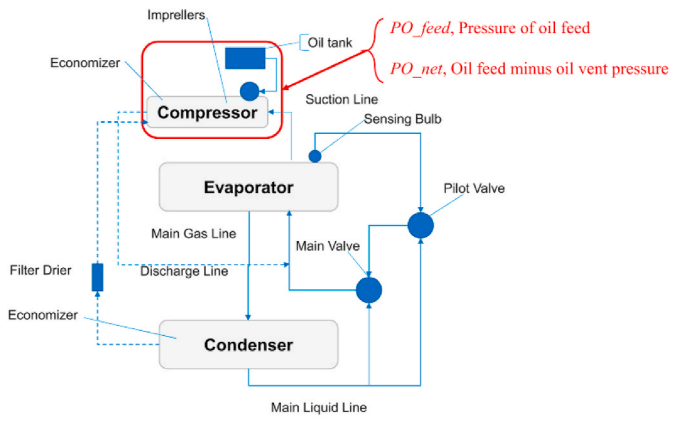

Fig. 13. Illustration of positions to monitor the two oil pressures in the chiller compressor [81].

circulated in the system, they have global effects on the entire system. There is fault propagation [93] effect resulted from the refrigerant flows to nearly all components. The system fault discriminative information may show a scattered distribution of variables from no less than one components of the chiller system. Accordingly, the diagnosis criteria of system faults are more complex than the component faults. From the Grad-Absolute-CAM visualization results, fault discriminative features of seven chiller faults are listed as in Table 2. As explained in Section 4.5.1, the criteria for CdF and RWCd faults are in accordance with the professional knowledge. Validation results for the fault-discriminative features of other five chiller faults are as follows.

NCdG (Class = 4): From Fig. 12(c) and (d), it is obviously that NCdG has extremely larger *26 TCA* and *31 TRC_sub* values than the Normal statue and the other six faults. So the two variables can be considered as fault indicators to separate the NCdG fault from the Normal statue and the other six faults.

RWEv (Class = 5): In RP-1043 experimental tests, the generation approach of the RWEv fault is to reduce the water flow rate at the evaporator side by 10%, 20%, 30%, 40% as shown in Table 1. It is no doubt that *24 FWE* (water flow rate at the evaporator side) is the fault indicator for the RWEv fault. Also, Reference [72] indicated that *55*

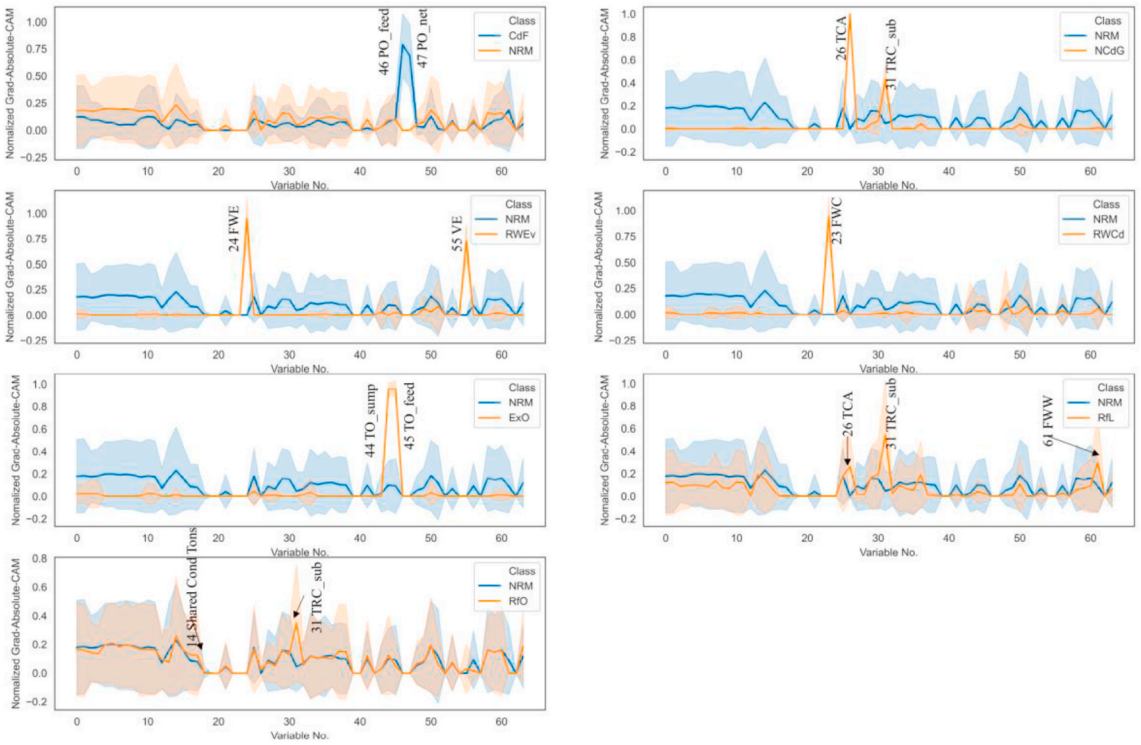

Fig. 14. Grad-Absolute-CAM based visualization of seven faults at their forth severity levels.

Table 2
The diagnosis criteria for seven faults in the chiller system.

Fault types	Fault monitoring and diagnosis criteria
Condenser fouling (CdF)	*46 PO_feed; 47 PO_net*
Non-condensate gas in refrigerant (NCdG)	*26 TCA; 31 TRC_sub*
Reduced water flow in evaporator (RWEv)	*55 VE; 24 FWE*
Reduced water flow in condenser (RWCd)	*23 FWC*
Excessive oil (ExO)	*44 TO_sump; 45 TO_feed*
Refrigerant leakage (RfL)	*31 TRC_sub; 26 TCA; 61 FWW*
Refrigerant overcharge (RfO)	*31 TRC_sub; 14 Shared Cond Tons*

VE can be used as fault indicator to the RWEv fault. Further, Fig. 12 (e) and (f) indicate that *24 FWE* and *55 VE* can separate the data samples of the RWEv fault from those of the Normal statue and the other six faults.

ExO (Class = 7): For the ExO fault, its generation approach is to increase oil charge amount in the system. So the oil temperatures are the fault-related variables which are also explained in Xiao et al.'s study [94]. Further, Fig. 12(h) and (i) indicate that *44 TO_sump* and *45 TO_feed* can separate the data samples of the ExO fault from those of the Normal statue and the other six faults.

RfL (Class = 2): As explained in Reference [94], refrigerant improper charges and non-condensable gas faults display nearly the same measurement trends based on the RP-1043 datasets. From Fig. 12(c) and (d), RfL fault data samples have larger *26 TCA* and *31 TRC_sub* values than the Normal data. The two variables should be the fault indicators of RfL fault. In Fig. 12(l) and (o), it can be seen that *61 FWW* is helpful for classifying some RfL fault data samples in fault SL-1 and SL-4. So *61 FWW* is also chosen as the fault indicator of RfL fault.

RfO (Class = 3): RfO shows very similar measurement trends as NCdG based on the RP-1043 datasets [94]. From Fig. 12(c), RfO fault data samples have smaller *31 TRC_sub* values than the Normal statue and the other six faults. Fig. 12(n) further indicates that *31 TRC_sub* tends to decrease as the fault SL increases. This is the reason why *31 TRC_sub* should be the RfO fault diagnosis criteria. Although *14 Shared Cond Tons* in Fig. 12(k) shows no obvious distinguishing measurement trend for separating the RfO fault, it is an important feature for improving the fault diagnosis accuracy when only a small amount of chiller data samples are used for CNN model training.

4.5.3. Impacts of different CONV layers on Grad-Absolute-CAM based visualization of diagnosis criteria

Further, Fig. 15 show the Grad-Absolute-CAM visualization results for the CdF fault at different CONV layers in the proposed three-layer CNN network. As described in Section 4.3, *46 PO_feed* and *47 PO_net* are two critical features for diagnosing the CdF fault. From Fig. 15(a)–(c), it can be found that fault diagnosis criteria can be more accurately visualized in feature map activations from the last CONV layer. In Fig. 15(a), the percentage of correctly activated chiller input tensor samples is 44% (15/34). This percentage increases to 79% (27/34) at the second CONV layer in Fig. 15(b) and finally grows up to 85% (29/34) at the last CONV layer in Fig. 15(c). This indicates that fault feature map activations localization becomes progressively worse as the network moves to earlier CONV layers for the three-layer CNN. This is because later CONV layers can better capture fault dependence information and the spatial locations than earlier CONV layers, that have smaller receptive fields and only focus on local features.

5. Conclusions

In this paper, a novel interpretable deep learning based method is proposed for building HVAC system fault diagnosis. The method contains the deep convolutional neural model structure for learning features and the visualization of the diagnosis criteria for diagnosing the target fault class from operational data samples of the HVAC system. Since the model is developed by maximally preserving the system operational data information (i.e., the location of each variable, positive and negative data, etc.) and resulting impacts of HVAC faults on normalized variables, it is capable of interpreting the model work mechanism by activation feature maps and explaining the fault diagnosis criteria by

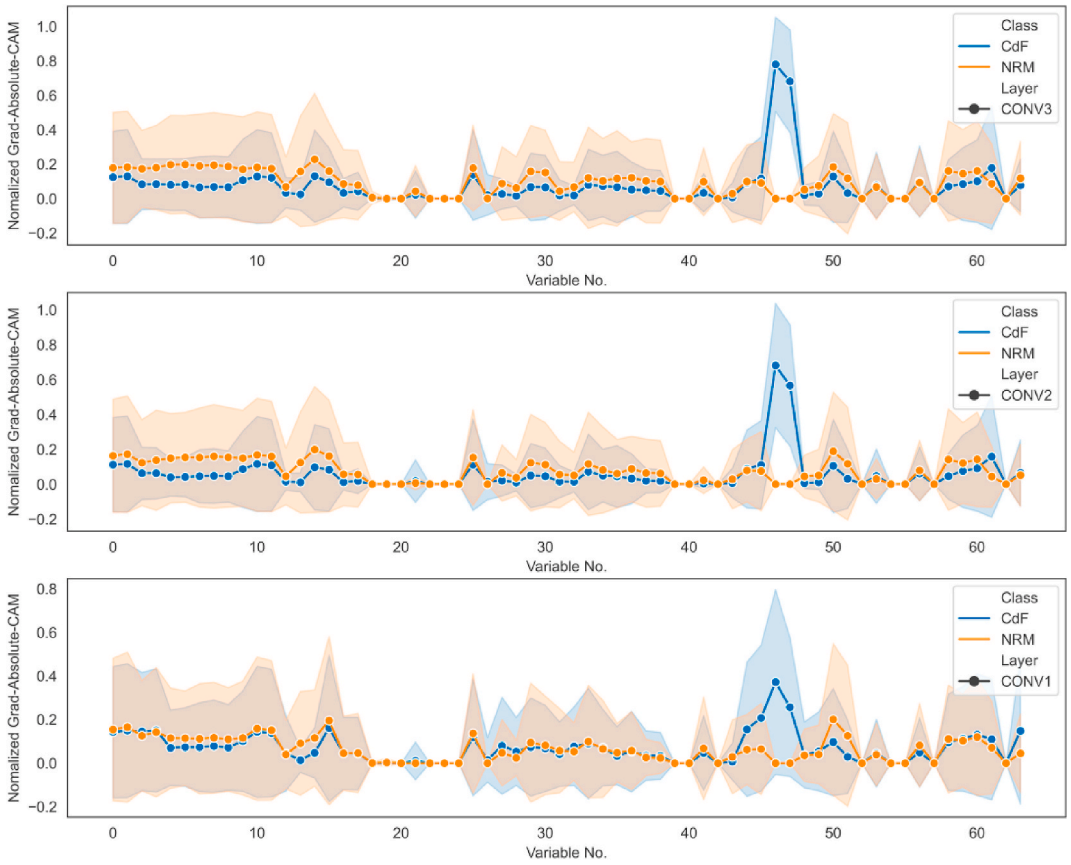

Fig. 15. Grad-Absolute-CAM at different convolutional layers for the CdF fault.

Grad-Absolute-CAM. The proposed method is validated using RP-1043 experimental data of a typical chiller system. Main conclusions are as follows:

1) The model can correctly classify more than 98.6% of the steady-state normal and fault data.
2) The first activated convolutional layer is a significant layer for improving the fault diagnosis performance.
3) Grad-Absolute-CAM on the last convolutional layer is suitable for fault criteria visualization.

Further, it can be seen that the diagnosis mechanism of the proposed deep learning method is very similar to our professional knowledge and previous experimental results for fault detection and isolation especially for the component faults. So this study may be helpful for providing evidences to prove the reasonability of applying deep learning based fault diagnosis method for a typical building HVAC system, chillers. It can be expected that the proposed explainable deep learning method can be used in other HVAC systems. In addition, CNN is the base of many deep learning based methods such as, deep transfer learning [95], deep reinforcement learning [96], deep generative adversarial networks [89,97], etc. It can also be expected that the proposed approach in this study can provide model interpretability and reasonability on these methods for practical applications in building HVAC systems.

For the practical application issues, the real building sensing environments including sensor faults, sensor absences, and the limited operational datasets of normal and faulty conditions should be seriously considered. For the sensor faults and absences problems, future work can apply the online virtual in-situ sensor calibration [98,99] to guarantee the field measured data accuracy and the virtual sensors [73,100] to make up the possible information loss, respectively. For the limited operational datasets problems, it is suggested to use generative adversarial networks [97,101] to enlarge the diversity and complexity of data and operations. Furthermore, when multiple faults occur simultaneously [102], future work can combine the proposed method with recurrent neural networks [103–105] to solve the multi-label classification problems.

Author statement

Guannan Li: Conceptualization, Methodology, Writing and Funding acquisition. Qing Yao: Data analysis, Software, Original draft preparation. Cheng Fan: Supervision, Reviewing and Editing. Chunlin Zhou: Reviewing and Editing. Guanghai Wu: Reviewing and Editing. Zhenxin Zhou: Data analysis and Editing. Xi Fang: Data analysis and Editing.

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.

Acknowledgements

This work is jointly supported by National Natural Science Foundation of China (51906181), Excellent Young and Middle-aged Talent in Universities of Hubei, China (Q20181110).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.buildenv.2021.108057.

References

[1] S. Ginestet, D. Marchio, O. Morisot, Evaluation of faults impacts on energy consumption and indoor air quality on an air handling unit, Energy Build. 40 (2008) 51–57. https://doi.org/10.1016/j.enbuild.2007.01.012.

[2] G. Lin, H. Kramer, J. Granderson, Building fault detection and diagnostics: achieved savings, and methods to evaluate algorithm performance, Build. Environ. 168 (2020) 106505. https://doi.org/10.1016/j.buildenv.2019.106505.

[3] J. Verhelst, G.V. Ham, D. Saelens, L. Helsen, Economic impact of persistent sensor and actuator faults in concrete core activated office buildings, Energy Build. 142 (2017) 111–127. https://doi.org/10.1016/j.enbuild.2017.02.052.

[4] 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. https://doi.org/10.1016/j.rser.2019.04.021.

[5] A.P. Rogers, F. Guo, B.P. Rasmussen, A review of fault detection and diagnosis methods for residential air conditioning systems, Build. Environ. 161 (2019) 106236. https://doi.org/10.1016/j.buildenv.2019.106236.

[6] W. Kim, S. Katipamula, A review of fault detection and diagnostics methods for building systems, Sci. Technol. Built Environ. 24 (2018) 3–21. https://doi.org/10.1080/23744731.2017.1318008.

[7] S. Katipamula, M.R. Brambley, Review article: methods for fault detection, diagnostics, and prognostics for building systems—a review, Part I, HVAC R Res. 11 (2005) 3–25. https://doi.org/10.1080/10789669.2005.10391123.

[8] S. Katipamula, M.R. Brambley, Review article: methods for fault detection, diagnostics, and prognostics for building systems—a review, Part II, HVAC R Res. 11 (2005) 169–187. https://doi.org/10.1080/10789669.2005.10391133.

[9] 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. 229 (2020) 110492. https://doi.org/10.1016/j.enbuild.2020.110492.

[10] G. Li, Y. Hu, J. Liu, X. Fang, J. Kang, Review on fault detection and diagnosis feature engineering in building heating, ventilation, air conditioning and refrigeration systems, IEEE Access 9 (2021) 2153–2187. https://doi.org/10.1109/ACCESS.2020.3040980.

[11] 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. https://doi.org/10.1016/j.autcon.2019.04.002.

[12] C. Fan, D. Yan, F. Xiao, A. Li, J. An, X. Kang, Advanced data analytics for enhancing building performances: from data-driven to big data-driven approaches, Build. Simul. 14 (2021) 3–24. https://doi.org/10.1007/s12273-020-0723-1.

[13] S. Wang, Y. Chen, Sensor validation and reconstruction for building central chilling systems based on principal component analysis, Energy Convers. Manag. 45 (2004) 673–695. https://doi.org/10.1016/S0196-8904(03)00180-8.

[14] Y. Hu, H. Chen, J. Xie, X. Yang, C. Zhou, Chiller sensor fault detection using a self-Adaptive Principal Component Analysis method, Energy Build. 54 (2012) 252–258. https://doi.org/10.1016/j.enbuild.2012.07.014.

[15] S. Li, J. Wen, A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform, Energy Build. 68 (2014) 63–71. https://doi.org/10.1016/j.enbuild.2013.08.044.

[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. https://doi.org/10.1016/j.conengprac.2016.04.018.

[17] Y. Chen, L. Lan, A fault detection technique for air-source heat pump water chiller/heaters, Energy Build. 41 (2009) 881–887. https://doi.org/10.1016/j.enbuild.2009.03.007.

[18] D.A.T. Tran, Y. Chen, H.L. Ao, H.N.T. Cam, An enhanced chiller FDD strategy based on the combination of the LSSVR-DE model and EWMA control charts, Int. J. Refrig. 72 (2016) 81–96. https://doi.org/10.1016/j.ijrefrig.2016.07.024.

[19] J. Liu, G. Li, H. Chen, J. Wang, Y. Guo, J. Li, A robust online refrigerant charge fault diagnosis strategy for VRF systems based on virtual sensor technique and PCA-EWMA method, Appl. Therm. Eng. 119 (2017) 233–243. https://doi.org/10.1016/j.applthermaleng.2017.03.074.

[20] M. Horrigan, W.J.N. Turner, J. O'Donnell, A statistically-based fault detection approach for environmental and energy management in buildings, Energy Build. 158 (2018) 1499–1509. https://doi.org/10.1016/j.enbuild.2017.11.023.

[21] Z. Du, B. Fan, X. Jin, J. Chi, Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis, Build. Environ. 73 (2014) 1–11. https://doi.org/10.1016/j.buildenv.2013.11.021.

[22] Z. Du, L. Chen, X. Jin, Data-driven based reliability evaluation for measurements of sensors in a vapor compression system, Energy 122 (2017) 237–248. https://doi.org/10.1016/j.energy.2017.01.055.

[23] X.J. Luo, K.F. Fong, Y.J. Sun, M.K.H. Leung, Development of clustering-based sensor fault detection and diagnosis strategy for chilled water system, Energy Build. 186 (2019) 17–36. https://doi.org/10.1016/j.enbuild.2019.01.006.

[24] M. Yuwono, Y. Guo, J. Wall, J. Li, S. West, G. Platt, et al., Unsupervised feature selection using swarm intelligence and consensus clustering for automatic fault detection and diagnosis in Heating Ventilation and Air Conditioning systems, Appl. Soft Comput. 34 (2015) 402–425. https://doi.org/10.1016/j.asoc.2015.05.030.

[25] J. Liu, D. Shi, G. Li, Y. Xie, K. Li, B. Liu, et al., Data-driven and association rule mining-based fault diagnosis and action mechanism analysis for building chillers, Energy Build. (2020) 109957. https://doi.org/10.1016/j.enbuild.2020.109957.

[26] C. Zhang, X. Xue, Y. Zhao, X. Zhang, T. Li, An improved association rule mining-based method for revealing operational problems of building heating, ventilation and air conditioning (HVAC) systems, Appl. Energy 253 (2019) 113492. https://doi.org/10.1016/j.apenergy.2019.113492.

[27] J. Liu, G. Li, B. Liu, K. Li, H. Chen, Knowledge discovery of data-driven-based fault diagnostics for building energy systems: a case study of the building variable refrigerant flow system, Energy 174 (2019) 873–885. https://doi.org/10.1016/j.energy.2019.02.161.

[28] 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. https://doi.org/10.1016/j.enbuild.2016.09.039.

[29] J. Liang, R. Du, Model-based fault detection and diagnosis of HVAC systems using Support vector machine method, Int. J. Refrig. 30 (2007) 1104–1114. https://doi.org/10.1016/j.ijrefrig.2006.12.012.

[30] A. Ebrahimifakhar, A. Kabirikopaei, D. Yuill, Data-driven fault detection and diagnosis for packaged rooftop units using statistical machine learning classification methods, Energy Build. 225 (2020) 110318. https://doi.org/10.1016/j.enbuild.2020.110318.

[31] 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. https://doi.org/10.1016/j.enbuild.2014.10.069.

[32] 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. https://doi.org/10.1016/j.applthermaleng.2015.09.121.

[33] K. Verbert, R. Babuška, B. De Schutter, Combining knowledge and historical data for system-level fault diagnosis of HVAC systems, Eng. Appl. Artif. Intell. 59 (2017) 260–273. https://doi.org/10.1016/j.engappai.2016.12.021.

[34] Y. Zhao, J. Wen, S. Wang, Diagnostic Bayesian networks for diagnosing air handling units faults – Part II: faults in coils and sensors, Appl. Therm. Eng. 90 (2015) 145–157. https://doi.org/10.1016/j.applthermaleng.2015.07.001.

[35] B. Cai, Y. Liu, Q. Fan, Y. Zhang, Z. Liu, S. Yu, et al., Multi-source information fusion based fault diagnosis of ground-source heat pump using Bayesian network, Appl. Energy 114 (2014) 1–9. https://doi.org/10.1016/j.apenergy.2013.09.043.

[36] F. Xiao, Y. Zhao, J. Wen, S. Wang, Bayesian network based FDD strategy for variable air volume terminals, Autom. ConStruct. 41 (2014) 106–118. https://doi.org/10.1016/j.autcon.2013.10.019.

[37] Y. Zhao, F. Xiao, S. Wang, An intelligent chiller fault detection and diagnosis methodology using Bayesian belief network, Energy Build. 57 (2013) 278–288. https://doi.org/10.1016/j.enbuild.2012.11.007.

[38] M. Najafi, D.M. Auslander, P.L. Bartlett, P. Haves, M.D. Sohn, Application of machine learning in the fault diagnostics of air handling units, Appl. Energy 96 (2012) 347–358. https://doi.org/10.1016/j.apenergy.2012.02.049.

[39] 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. https://doi.org/10.1016/j.apenergy.2016.11.130.

[40] Z. Du, X. Jin, Y. Yang, Fault diagnosis for temperature, flow rate and pressure sensors in VAV systems using wavelet neural network, Appl. Energy 86 (2009) 1624–1631. https://doi.org/10.1016/j.apenergy.2009.01.015.

[41] W.H. Allen, A. Rubaai, R. Chawla, Fuzzy neural network-based health monitoring for HVAC system variable-air-volume unit, IEEE Trans. Ind. Appl. 52 (2016) 2513–2524. https://doi.org/10.1109/TIA.2015.2511160.

[42] N. Kocyigit, Fault and sensor error diagnostic strategies for a vapor compression refrigeration system by using fuzzy inference systems and artificial neural network, Int. J. Refrig. 50 (2015) 69–79. https://doi.org/10.1016/j.ijrefrig.2014.10.017.

[43] H. Wang, D. Feng, K. Liu, Fault detection and diagnosis for multiple faults of VAV terminals using self-adaptive model and layered random forest, Build. Environ. 193 (2021) 107667. https://doi.org/10.1016/j.buildenv.2021.107667.

[44] H. Han, Z. Zhang, X. Cui, Q. Meng, Ensemble learning with member optimization for fault diagnosis of a building energy system, Energy Build. 226 (2020) 110351. https://doi.org/10.1016/j.enbuild.2020.110351.

[45] D.B. Araya, K. Grolinger, H.F. ElYamany, M.A.M. Capretz, G. Bitsuamlak, An ensemble learning framework for anomaly detection in building energy consumption, Energy Build. 144 (2017) 191–206. https://doi.org/10.1016/j.enbuild.2017.02.058.

[46] 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. https://doi.org/10.1016/j.enbuild.2018.12.032.

[47] C. Fan, X. Liu, P. Xue, J. Wang, Statistical characterization of semi-supervised neural networks for fault detection and diagnosis of air handling units, Energy Build. 234 (2021) 110733. https://doi.org/10.1016/j.enbuild.2021.110733.

[48] B. Li, F. Cheng, X. Zhang, C. Cui, W. Cai, A novel semi-supervised data-driven method for chiller fault diagnosis with unlabeled data, Appl. Energy 285 (2021) 116459. https://doi.org/10.1016/j.apenergy.2021.116459.

[49] 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. https://doi.org/10.1016/j.enbuild.2018.10.016.

[50] M. Dey, S.P. Rana, S. Dudley, Semi-Supervised Learning Techniques for Automated Fault Detection and Diagnosis of HVAC Systems, 2018, pp. 872–877. IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI) 2018.

[51] M.W. Ahmad, M. Mourshed, B. Yuce, Y. Rezgui, Computational intelligence techniques for HVAC systems: a review, Build. Simul. 9 (2016) 359–398. https://doi.org/10.1007/s12273-016-0285-4.

[52] B. Dong, Z. O'Neill, Z. Li, A BIM-enabled information infrastructure for building energy Fault Detection and Diagnostics, Autom. ConStruct. 44 (2014) 197–211. https://doi.org/10.1016/j.autcon.2014.04.007.

[53] K.-P. Lee, B.-H. Wu, S.-L. Peng, Deep-learning-based fault detection and diagnosis of air-handling units, Build. Environ. 157 (2019) 24–33. https://doi.org/10.1016/j.buildenv.2019.04.029.

[54] C. Fan, F. Xiao, Y. Zhao, J. Wang, Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data, Appl. Energy 211 (2018) 1123–1135. https://doi.org/10.1016/j.apenergy.2017.12.005.

[55] D. Li, D. Li, C. Li, L. Li, L. Gao, A novel data-temporal attention network based strategy for fault diagnosis of chiller sensors, Energy Build. 198 (2019) 377–394. https://doi.org/10.1016/j.enbuild.2019.06.034.

[56] Y. Guo, Z. Tan, H. Chen, G. Li, J. Wang, R. Huang, et al., Deep learning-based fault diagnosis of variable refrigerant flow air-conditioning system for building energy saving, Appl. Energy 225 (2018) 732–745. https://doi.org/10.1016/j.apenergy.2018.05.075.

[57] X. Zhu, S. Zhang, X. Jin, Z. Du, Deep learning based reference model for operational risk evaluation of screw chillers for energy efficiency, Energy 213 (2020) 118833. https://doi.org/10.1016/j.apenergy.2018.05.075.

[58] H. Shahnazari, P. Mhaskar, J.M. House, T.I. Salsbury, Modeling and fault diagnosis design for HVAC systems using recurrent neural networks, Comput. Chem. Eng. 126 (2019) 189–203. https://doi.org/10.1016/j.compchemeng.2019.04.011.

[59] Z. Sun, H. Jin, J. Gu, Y. Huang, X. Wang, H. Yang, et al., Studies on the online intelligent diagnosis method of undercharging sub-health air source heat pump water heater, Appl. Therm. Eng. 169 (2020) 114957. https://doi.org/10.1016/j.applthermaleng.2020.114957.

[60] Y.H. Eom, J.W. Yoo, S.B. Hong, M.S. Kim, Refrigerant charge fault detection method of air source heat pump system using convolutional neural network for energy saving, Energy 187 (2019) 115877. https://doi.org/10.1016/j.energy.2019.115877.

[61] Z. Sun, H. Jin, J. Gu, Y. Huang, X. Wang, X. Shen, Gradual fault early stage diagnosis for air source heat pump system using deep learning techniques, Int. J. Refrig. 107 (2019) 63–72. https://doi.org/10.1016/j.ijrefrig.2019.07.020.

[62] F. Cheng, W. Cai, X. Zhang, H. Liao, C. Cui, Fault detection and diagnosis for Air Handling Unit based on multiscale convolutional neural networks, Energy Build. 236 (2021) 110795. https://doi.org/10.1016/j.enbuild.2021.110795.

[63] J. Gao, H. Han, Z. Ren, Y. Fan, Fault diagnosis for building chillers based on data self-production and deep convolutional neural network, J. Build. Eng. 34 (2021) 102043. https://doi.org/10.1016/j.jobe.2020.102043.

[64] H. Cheng, H. Chen, Z. Li, X. Cheng, Ensemble 1-D CNN diagnosis model for VRF system refrigerant charge faults under heating condition, Energy Build. (2020) 224. https://doi.org/10.1016/j.enbuild.2020.110256.

[65] S. Miyata, J. Lim, Y. Akashi, Y. Kuwahara, K. Tanaka, Fault detection and diagnosis for heat source system using convolutional neural network with imaged faulty behavior data, Sci. Technol. Built Environ. 26 (2020) 52–60. https://doi.org/10.1080/23744731.2019.1651619.

[66] D. Li, Y. Zhou, G. Hu, C.J. Spanos, Identifying unseen faults for smart buildings by incorporating expert knowledge with data, IEEE Trans. Autom. Sci. Eng. 16 (2019) 1412–1425. https://doi.org/10.1109/TASE.2018.2876611.

[67] M. Madhikermi, A.K. Malhi, K. Främling, Explainable artificial intelligence based heat recycler fault detection in air handling unit, in: D. Calvaresi, A. Najjar, M. Schumacher, K. Främling (Eds.), Explainable, Transparent Autonomous Agents and Multi-Agent Systems, Springer International Publishing, Cham, 2019, pp. 110–125.

[68] W. Samek, T. Wiegand, K.-R. Müller, Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, 2017, 08296 arXiv:1708.

[69] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, 2016 arXiv:1610.02391.

[70] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning Deep Features for Discriminative Localization, 2016, pp. 2921–2929. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)2016.

[71] Y. Li, M. Liu, J. Lau, B. Zhang, Experimental study on electrical signatures of common faults for packaged DX rooftop units, Energy Build. 77 (2014) 401–415. https://doi.org/10.1016/j.enbuild.2014.04.008.

[72] H. Han, B. Gu, T. Wang, Z.R. Li, Important sensors for chiller fault detection and diagnosis (FDD) from the perspective of feature selection and machine learning, Int. J. Refrig. 34 (2011) 586–599. https://doi.org/10.1016/j.ijrefrig.2010.08.011.

[73] H. Li, J.E. Braun, Decoupling features and virtual sensors for diagnosis of faults in vapor compression air conditioners, Int. J. Refrig. 30 (2007) 546–564. https://doi.org/10.1016/j.ijrefrig.2006.07.024.

[74] X. Zhao, M. Yang, H. Li, Development, evaluation, and validation of a robust virtual sensing method for determining water flow rate in chillers, HVAC R Res. 18 (2012) 874–889. https://doi.org/10.1080/10789669.2012.667036.

[75] P. Wang, X. Tang, R. Gao, Automated Performance Tracking for Heat Exchangers in HVAC, 2015, pp. 949–954. IEEE International Conference on Automation Science and Engineering (CASE)2015.

[76] M. Bonvini, M.D. Sohn, J. Granderson, M. Wetter, M.A. Piette, Robust on-line fault detection diagnosis for HVAC components based on nonlinear state estimation techniques, Appl. Energy 124 (2014) 156–166. https://doi.org/10.1016/j.apenergy.2014.03.009.

[77] B. Sun, P.B. Luh, Q. Jia, Z.O. Neill, F. Song, Building energy doctors: an SPC and kalman filter-based method for system-level fault detection in HVAC systems, IEEE Trans. Autom. Sci. Eng. 11 (2014) 215–229. https://doi.org/10.1109/TASE.2012.2226155.

[78] 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 (2007) 469–473. https://doi.org/10.1109/TASE.2006.888053.

[79] H. Wang, Z. Liu, D. Peng, Y. Qin, Understanding and learning discriminant features based on multiattention 1DCNN for wheelset bearing fault diagnosis, IEEE Transact. Industr. Inform. 16 (2020) 5735–5745. https://doi.org/10.1109/TII.2019.2955540.

[80] M.S. Kim, J.P. Yun, P. Park, An explainable convolutional neural network for fault diagnosis in linear motion guide, IEEE Transact. Industr. Inform. (2020) 1. https://doi.org/10.1109/TII.2020.3012989.

[81] M.C.B.J.E. Comstock, R. Bernhard, Development of Analysis Tools for the Evaluation of Fault Detection and Diagnostics in Chillers, Purdue University, 1999.

[82] J. Jiao, M. Zhao, J. Lin, K. Liang, A comprehensive review on convolutional neural network in machine fault diagnosis, Neurocomputing 417 (2020) 36–63. https://doi.org/10.1016/j.neucom.2020.07.088.

[83] J.M. Cho, J. Heo, W.V. Payne, P.A. Domanski, Normalized performance parameters for a residential heat pump in the cooling mode with single faults imposed, Appl. Therm. Eng. 67 (2014) 1–15. https://doi.org/10.1016/j.applthermaleng.2014.03.010.

[84] D.P. Yuill, J.E. Braun, Effect of the distribution of faults and operating conditions on AFDD performance evaluations, Appl. Therm. Eng. 106 (2016) 1329–1336. https://doi.org/10.1016/j.applthermaleng.2016.06.149.

[85] M. Mehrabi, D. Yuill, Generalized effects of faults on normalized performance variables of air conditioners and heat pumps, Int. J. Refrig. 85 (2018) 409–430. https://doi.org/10.1016/j.ijrefrig.2017.10.017.

[86] S.H. Yoon, W.V. Payne, P.A. Domanski, Residential heat pump heating performance with single faults imposed, Appl. Therm. Eng. 31 (2011) 765–771. https://doi.org/10.1016/j.applthermaleng.2010.10.023.

[87] M. Kim, W.V. Payne, P.A. Domanski, S.H. Yoon, C.J.L. Hermes, Performance of a residential heat pump operating in the cooling mode with single faults imposed, Appl. Therm. Eng. 29 (2009) 770–778. https://doi.org/10.1016/j.applthermaleng.2008.04.009.

[88] G. Li, Y. Hu, H. Chen, L. Shen, H. Li, M. Hu, et al., An improved fault detection method for incipient centrifugal chiller faults using the PCA-R-SVDD algorithm, Energy Build. 116 (2016) 104–113. https://doi.org/10.1016/j.enbuild.2015.12.045.

[89] K. Yan, J. Huang, W. Shen, Z. Ji, Unsupervised learning for fault detection and diagnosis of air handling units, Energy Build. 210 (2020) 109689. https://doi.org/10.1016/j.enbuild.2019.109689.

[90] Z. Zhou, H. Chen, G. Li, H. Zhong, M. Zhang, J. Wu, Data-driven fault diagnosis for residential variable refrigerant flow system on imbalanced data environments, Int. J. Refrig. 125 (2021) 34–43. https://doi.org/10.1016/j.ijrefrig.2021.01.009.

[91] V.D.M. Laurens, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[92] G. Alain, Y. Bengio, Understanding Intermediate Layers Using Linear Classifier Probes, 2016, 01644 arXiv:1610.

[93] Y. Yan, P.B. Luh, K.R. Pattipati, fault diagnosis of HVAC air-handling systems considering fault propagation impacts among components, IEEE Trans. Autom. Sci. Eng. 14 (2017) 705–717. https://doi.org/10.1109/TASE.2017.2669892.

[94] F. Xiao, C. Zheng, S. Wang, A fault detection and diagnosis strategy with enhanced sensitivity for centrifugal chillers, Appl. Therm. Eng. 31 (2011) 3963–3970. https://doi.org/10.1016/j.applthermaleng.2011.07.047.

[95] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, A.K. Nandi, Applications of machine learning to machine fault diagnosis: a review and roadmap, Mech. Syst. Signal Process. 138 (2020) 106587. https://doi.org/10.1016/j.ymssp.2019.106587.

[96] Z. Zou, X. Yu, S. Ergan, Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network, Build. Environ. 168 (2020) 106535. https://doi.org/10.1016/j.buildenv.2019.106535.

[97] K. Yan, A. Chong, Y. Mo, Generative adversarial network for fault detection diagnosis of chillers, Build. Environ. 172 (2020) 106698. https://doi.org/10.1016/j.buildenv.2020.106698.

[98] S. Yoon, Y. Yu, Hidden factors and handling strategies on virtual in-situ sensor calibration in building energy systems: Prior information and cancellation effect, Appl. Energy 212 (2018) 1069–1082. https://doi.org/10.1016/j.apenergy.2017.12.077.

[99] S. Yoon, Y. Yu, Strategies for virtual in-situ sensor calibration in building energy systems, Energy Build. 172 (2018) 22–34. https://doi.org/10.1016/j.enbuild.2018.04.043.

[100] Y. Choi, S. Yoon, Virtual sensor-assisted in situ sensor calibration in operational HVAC systems, Build. Environ. 181 (2020) 107079. https://doi.org/10.1016/j.buildenv.2020.107079.

[101] K. Yan, J. Su, J. Huang, Y. Mo, chiller fault diagnosis based on VAE-enabled generative adversarial networks, IEEE Trans. Autom. Sci. Eng. (2020) 1–9. https://doi.org/10.1109/TASE.2020.3035620.

[102] H. Han, B. Gu, Y. Hong, J. Kang, Automated FDD of multiple-simultaneous faults (MSF) and the application to building chillers, Energy Build. 43 (2011) 2524–2532. https://doi.org/10.1016/j.enbuild.2011.06.011.

[103] Y.Y. Yang, Y.A. Lin, H.M. Chu, H.T. Lin, Deep Learning with a Rethinking Structure for Multi-Label Classification, 2018.
[104] H. Jiang, J. Xu, R. Shi, K. Yang, W. Qian, A multi-label deep learning model with interpretable grad-CAM for diabetic retinopathy classification, 2020, 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) in conjunction with the 43rd Annual Conference of the Canadian Medical and Biological Engineering Society2020.
[105] W. Jiang, Y. Yi, J. Mao, Z. Huang, X. Wei, CNN-RNN: A unified framework for multi-label image classification, 2016. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)2016.