Energy & Buildings 259 (2022) 111893

Contents lists available at ScienceDirect

Energy & Buildings

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

Outlier detection via multiclass deep autoencoding Gaussian mixture model for building chiller diagnosis

Viet Tra [a,*], Manar Amayri [b], Nizar Bouguila [a]

[a] Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada
[b] G-SCOP lab / Grenoble Institute of Technology, Grenoble, France

ARTICLE INFO

Article history:
Received 19 July 2021
Revised 7 January 2022
Accepted 20 January 2022
Available online 26 January 2022

Keywords:
Building chiller
Fault diagnosis
Outlier detection
Pretraining
Supervised multiclass deep autoencoding Gaussian mixture model (S-DAGMM)
Deep neural network (DNN)

ABSTRACT

In HVAC systems, the application of fault detection and diagnosis (FDD) for chillers has brought many benefits to building utilities, i.e., indoor thermal comfort, energy saving, and energy consumption management. Although recent studies have achieved great progress on chiller fault diagnostic problem, some serious issues have been raised in these studies and need to be resolved comprehensively. In that context, two problems are listed and studied in this paper namely outliers detection and the insufficiency of labeled data. To deal with outliers mixed in chiller data, this paper proposes a supervised multiclass deep autoencoding Gaussian mixture model (S-DAGMM) algorithm which is an ensemble model of individual unsupervised DAGMMs. This mechanism helps S-DAGMM to detect ambiguous outliers in both training and testing data. In addition, this paper proposes to use DAGMM to pretrain a deep neural network (DNN). Since DAGMM can learn the data distribution from the unlabeled data, it can prevent DNN from getting stuck on local optima. Comprehensive experimental results in this study have proved the outstanding performance of the proposed approach, compared with state-of-the-art rival methods.

1. Introduction

Large energy consumption is one of the significant concerns for heating, ventilation and air conditioning (HVAC) systems in modern days [1,2]. Statistics show that while HVAC systems consume about 33% of the total energy consumption of commercial buildings in the USA, this number can be up to 60% or more, especially in tropical countries with hot and humid weather, i.e., Singapore, etc. [3,4]. Beside the essential usage, a number of contributing factors to this problem can be listed: underqualified maintenance, underperformed components, installation and control failure. Since chillers are the largest consumer of HVAC systems in constructions with a rate of 35–40% of total building energy consumption [5,6], they can also possibly be the one with the greatest potential in terms of energy efficiency refinement. Besides the obvious discomfort caused by chiller faults, energy waste is a serious matter which needs to be addressed. To tackle this problem, an accurate fault detection/diagnosis (FDD) technique is crucial.

FDD approaches [7–9] can be roughly divided into three different categories: rule-based, model-based, and data-driven methods. The first family of approaches, rule-based [10,11], detects faults based on a set of rules established by expert's knowledge and experience. The second category, model-based [12,13], focuses on developing a mathematical model reliable enough to mirror the real physical process. This category of approaches requires high complexity and time for model establishment, especially for large-scale, nonlinear systems like HVAC. The final group of approaches, data-driven, is capable of working without constructing a complex chiller model first, but by learning fault patterns with machine learning techniques. With the growth of machine learning algorithms and tools, and the availability of many historical monitoring data, data-driven FDD approaches are now the most favored in the field not just for HVAC, but for other systems in general.

In the earliest days, statistical approaches were the main go-to of data-driven models and principal component analysis (PCA) was a typical solution for FDD [14–16]. The rapid growth of machine learning has introduced a large number of advanced data-driven models ranging from classic machine learning techniques: Bayesian belief networks [17,18], association rule mining [10,19], support vector machines (SVM) [20,21], decision trees [22,23], clustering analysis [24], to deep learning algorithms: generative adversarial networks [3,4,25], convolution neural networks [26,27], and deep neural networks [28,29]. Han et al. [23] inte-

* Corresponding author.
E-mail addresses: viet.tra@mail.concordia.ca (V. Tra), manar.amayri@grenoble-inp.fr (M. Amayri), nizar.bouguila@concordia.ca (N. Bouguila).

https://doi.org/10.1016/j.enbuild.2022.111893
0378-7788/

grated K-nearest neighbors (KNN), SVM and random forest (RF) into an ensemble diagnostic mode by the majority voting method to identify the chiller faults. Zhu et al. [6] introduced a method for the diagnosis model construction based on a generic framework which transfers the prior knowledge of an information-rich source chiller. Yan et al. [4] resolved the imbalanced-class classification problem of chiller data by using a new GAN-integrated framework.

In general, reliable diagnosis is achievable with data-driven approaches, given the assumptions that normal and fault data are available at a large number and the data set is not contaminated by outliers. An outlier is an observation which is distinct from the data's major population. In chiller data, outliers can be caused by errors in recording process due to noise, faults in measurement or missing values, etc. Such outliers have an adverse impact on the fault diagnosis model's reliability. Therefore, outlier detection is a crucial component in constructing a reliable machine learning model.

Several approaches have been proposed for anomaly detection, which can be divided into four categories as following: (1) distance-based, (2) model-based, (3) reconstruction-based, and (4) density-based [30].

The first approach, distance-based, aims to identify anomalous data points through their proximity relationships [31,32]. The authors in [33] introduced a method based on k-nearest neighbor graph, which is used to construct an in-degree number algorithm. The authors in [34] presented a novel Local Distance-based Outlier Factor which is used in scattered datasets to measure the level of distinction an object possesses. Nevertheless, the curse of dimensionality still remains the bane of this approach, since in high dimensional spaces, the distance between data points becomes more and more similar.

The second approach, model-based, consists of a various range of methods. The models' purpose is to harness the feature relationship along with data's normal pattern. A popular trend often converses from the original unsupervised setting into a supervised one. The authors in [30] proposed Out-of-Bag which is based on the decomposition of the unsupervised problem into a supervised problem of an ensemble models set. Other notable examples of model-based approaches are One-Class SVM (OC-SVM) [35] and Isolation Forest (IF) [36]. OC-SVM aims to find a decision boundary where the distance between data points and the margin in the feature space is maximized, from which the anomalies are differentiated from normal data. Meanwhile, an ensemble of random trees is harnessed in IF, where anomalies are isolated based on the partition of random features and values.

The third approach, reconstruction-based, relies on learning the principal factors of the data, from which the data reconstruction is performed later. This process defines the points with high reconstruction error to be anomaly. Principal Component Analysis (PCA)-based methods [37,38] are capable of learning the linear relationship between features, whereas the non-linear ones can be handled efficaciously by Autoencoder (AE)-based methods [39,40].

The fourth approach, density estimation, is the core of outlier detection. A probability distribution is utilized to fit the data in density-based approach. The data distribution can be estimated by either a parametric model (i.e., Gaussian Mixture Models [41], etc.) or a non-parametric model (i.e., feature histograms in Histogram-based Outlier Score [42], etc.). With a large number of samples, it can be observed that low-probability-density areas are where anomalies can be located.

Despite all the advances in recent years concerning density-based outlier detection approach, building a robust anomaly detection system without human supervision is still a formidable obstacle, especially with multi- or high-dimensional data. The higher the dimension of input data is, the higher the probability any input sample can be observed as a rare event. Two-step approaches are a popular solution to this problem, which consists of the first step in which dimensionality reduction is conducted and the second step in which the density estimation is performed in a lower dimensional space [43]. The problem with this method is that sub-optimal performance is likely to happen due to the unawareness of the first step in regard to the second one. In addition, there is a possibility that key information for anomaly detection can be lost during dimensionality reduction step. Therefore, a join optimization for both dimensionality reduction and density estimation is crucial to this process. A deep learning framework called Deep Autoencoding Gaussian Mixture Mode (DAGMM) that addresses the aforementioned challenges has been proposed by Zong et al. [44]. A compression network and an estimation network are the core of DAGMM construction. Dimensionality reduction is firstly conducted by an autoencoder in the compression network. This compression network thereby prepares low-dimensional representations for the estimation network by concatenating reduced low-dimensional features from encoding and reconstruction error from decoding. The estimation network acts as GMM. It inputs the compression network's output and returns the mixture membership prediction of each sample. Based on that result, GMM parameters are estimated next, which facilitates input samples' statistical energy calculation. The sample's statistical energy value models the probability that we can observe the sample. If the sample has low statistical energy, there is a high probability that the sample follows the GMM (i.e., an inlier sample). Conversely, if the sample has high statistical energy, then it does not follow the GMM (i.e., an outlier sample). With the reconstruction error from the compression network and sample statistical energy from the estimation network minimized, a join optimization for both dimensionality reduction and density estimation is possible.

However, DAGMM was originally designed as an unsupervised model that is mainly used for anomaly detection and some fault detection applications. For multiclass data like chiller data with different types of faults, the original version of DAGMM is ineffective for outlier detection. In this study, we propose a supervised multiclass version of DAGMM (S-DAGMM) that can work well on multiclass data. This supervised version undergoes the training phase in which labeled training data is first divided into sub-datasets, according to the data samples' labels. Each sub-dataset is used to train one individual DAGMM so that the number of trained individual DAGMMs is proportional to the number of data classes. DAGMMs are then used to detect outliers of their according sub-datasets based on sample statistical energy values. Clean sub-datasets are obtained after removing outliers and then are grouped to form the clean labeled training data for a chiller diagnosis model. During the online phase, new samples (unlabeled or labeled samples) are checked for outliers by trained DAGMMs and a voting scheme. Votes are increased by 1 if one of the trained DAGMMs identifies the sample as an outlier. After being checked by all DAGMMs, if the number of votes is larger than a pre-defined value, the sample is considered as an outlier and vice versa.

Recently, deep learning techniques have witnessed a lot of successful applications in many areas, including natural language processing, computer vision, and robotics [45,46]. End-to-end classification models have displayed a promising competence to extract features automatically during the requisite training period. Therefore, this study uses a deep neural network (DNN) as the chiller diagnostic model. Like other supervised classifiers, the high classification accuracy of DNN-based chiller FDD relies on the sufficiency of the labeled training data. In practice, however, while unlabeled data can be handily collected by remote sensors, the cost to obtain labeled fault data is expensive. There are two effective approaches to handle this obstacle. The first approach is to generate artificial labeled training samples of minority classes using data

augmentation methods such as synthetic minority over-sampling technique (SMOTE) [47]. The upgraded version of SMOTE called PCA-SMOTE in which principal component analysis (PCA) technique is first used to extract principal features from high-dimensional original features before applying SMOTE to generate synthetic minority samples [48,49]. Another effective method is to employ the concept of function approximator from neural networks like generative adversarial network (GAN) [4] to scale-up the minority dataset. The advantage of these methods is that they can generate an unlimited number of artificial data samples from a small number of initial minority seeds. However, generated artificial data samples are often distributed close to the seeds in the feature space, so if the number of initial minority seeds is too small, the generated samples will not more likely fully cover the inherent data space of the minority class in testing online data.

The second approach is to take advantage of unlabeled data by using unsupervised layer-wise pretraining. One of the most used pretrained networks is a stacked autoencoder (SAE) [46]. Since pretrained network can learn the data distribution from the unlabeled data, it can prevent DNN from getting stuck on local optima. By pretraining, moreover, sophisticated and abstract features with hierarchical structures can be efficiently learned because the technique offers layer-by-layer, high-level feature extraction from lower-level ones. Instead of using SAE, this research work utilizes the encoding layers of DAGMM's compression network to pretrain the DNN in order to prevent it from being trapped in local optimum due to random initialization. DNN's hidden layers are first pretrained successively in a layer-wise, unsupervised learning strategy. After a soft-max output layer is added, the whole DNN is fine-tuned in a supervised manner with the labeled data. The studies in [44] have shown that with regularization brought by the estimation network, the autoencoder is more capable of escaping from less attractive local optima. In addition, DAGMM's autoencoder reconstruction error is significantly lower than that of an autoencoder without the regularization part from the estimation network.

The proposed fault diagnosis framework for chiller is schematically illustrated in Fig. 1. First, S-DAGMM is used to detect and remove outliers in both labeled and unlabeled chiller data. Next, the unlabeled data are utilized to train DAGMM which plays a role as a pretrained network for DNN. After fine-tuning using the labeled data, the DNN classifier is used as the fault diagnostic model in the online application. The DNN classifier inputs real-time data and outputs the corresponding fault type.

In summary, this paper presents the following contributions:

1. The supervised multiclass version of DAGMM is proposed. This method has outstanding performance to detect outliers of multiclass data by profiling the label information of training data.
2. The encoding layers of the DAGMM's compression network are suggested as a pre-trained network for the DNN-based diagnostic model. The encoding layers of DAGMM can compress the input well during end-to-end training, so that their compressed representation can help DNN resolve generic classification tasks better.
3. The comprehensive comparative study with state-of-the-art algorithms to validate the effectiveness of the proposed methods.

The following sections of this paper are arranged as: Section 2 describes the supervised multiclass version of DAGMM and the DNN pretraining procedure. Section 3 introduces chiller datasets used in the study. Experimental results to validate the proposed method's performance are shown in Section 4. Section 5 summarizes conclusions.

2. Methodologies

2.1. Deep autoencoding Gaussian mixture model

As previously described, DAGMM has two central parts: a compression network and an estimation network, which are depicted in Fig. 2. The estimation network inputs the representations returned by the dimensionality reduction process in the compression network and then outputs the statistical energy in GMM framework.

There are two sources of features provided by the compression network's output: 1) the deep autoencoder's reduced low-dimensional representations; 2) reconstruction error features. In this study, a sample $\boldsymbol{x}$ is the vector of chiller features. Each feature is a sensor measurement from a chiller. The low-dimensional representation $\boldsymbol{z}$ can be computed from the sample $\boldsymbol{x}$ as follows:

$$\boldsymbol{z}_c = h(\boldsymbol{x};\boldsymbol{\theta}_e)\boldsymbol{x}' = g(\boldsymbol{z_c};\boldsymbol{\theta}_d) \tag{1}$$

$$\boldsymbol{z}_r = f(\boldsymbol{x},\boldsymbol{x}'), \boldsymbol{z} = [\boldsymbol{z}_c,\boldsymbol{z}_r] \tag{2}$$

where $\boldsymbol{z}_c$ is the reduced low-dimensional representation learnt by the deep autoencoder; $\boldsymbol{z}_r$ represents reconstruction error features (i.e., relative Euclidean distance, absolute Euclidean distance, cosine similarity, etc.); $\boldsymbol{\theta}_e$ and $\boldsymbol{\theta}_d$ are the parameters of the deep autoencoder; $\boldsymbol{x}'$ is the reconstructed counterpart of $\boldsymbol{x}$; $h(\cdot)$ is the encoding function; $g(\cdot)$ is the decoding function; and $f(\cdot)$ is the reconstruction error feature calculation function. Afterwards, the estimation network is fed with the computed $\boldsymbol{z}$.

Given the low-dimensional representations computed by the compression network, the process follows with density estimation performed by GMM framework. The parameters of GMM and the sample statistical energy are then estimated by the estimation network with unknown mixture component distribution φ , mixture means μ , and mixture covariance Σ in the training phase. This is achievable through a multi-layer neural network which predicts each sample's mixture membership as follow:

$$\boldsymbol{p} = MLP(\boldsymbol{z},\boldsymbol{\theta}_m), \hat{\gamma} = \text{softmax}(\boldsymbol{p}) \tag{3}$$

where $\boldsymbol{z}$ is the low-dimensional representation; K is the number of mixture components; $\hat{\gamma}$ is a K - dimensional vector utilized for the prediction of soft mixture component membership; $\boldsymbol{p}$ is the multi-layer network's output; and $\boldsymbol{\theta}_m$ is the multi-layer network's parameter.

The learning of GMM can be taken further with a N -sample batch and its membership prediction as follow:

$$\begin{aligned}\hat{\phi}_k &= \sum_{i=1}^{N}\frac{\hat{\gamma}_{ik}}{N}, \hat{\mu}_k = \frac{\sum_{i=1}^{N}\hat{\gamma}_{ik}\boldsymbol{z}_i}{\sum_{i=1}^{N}\hat{\gamma}_{ik}}, \hat{\Sigma}_k \\ &= \frac{\sum_{i=1}^{N}\hat{\gamma}_{ik}(\boldsymbol{z}_i-\hat{\mu}_k)(\boldsymbol{z}_i-\hat{\mu}_k)^T}{\sum_{i=1}^{N}\hat{\gamma}_{ik}} \forall 1<k<K\end{aligned} \tag{4}$$

where $\hat{\gamma}_i$ is the low-dimensional representation $\boldsymbol{z}_i$'s membership prediction; $\hat{\phi}_k, \hat{\mu}_k, \hat{\sum}_k$ are the k^{th} mixture component's weight, mean and covariance, respectively.

Afterwards, sample statistical energy can be calculated with the estimated parameters:

$$E(\boldsymbol{z}) = -\log\left(\sum_{k=1}^{K}\hat{\phi}_k\frac{\exp\left(-1/2(\boldsymbol{z}-\hat{\mu}_k)^T\hat{\sum}_k^{-1}(\boldsymbol{z}-\hat{\mu}_k)\right)}{\sqrt{2\pi\left|\hat{\Sigma}_k\right|}}\right) \tag{5}$$

with $|\cdot|$ indicating the matrix's determinant.

The construction of DAGMM objective function is given as follow with a N-sample dataset:

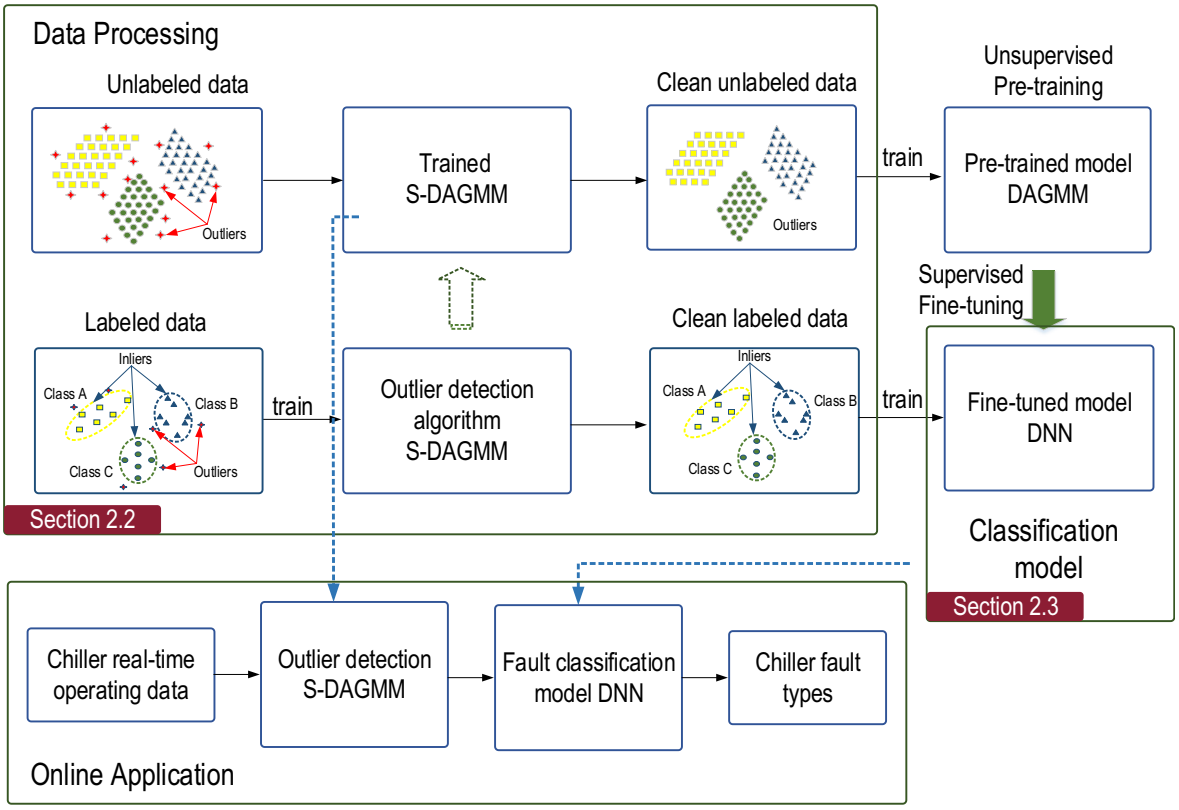

Fig. 1. Graphical diagram of the proposed chiller fault diagnosis framework.

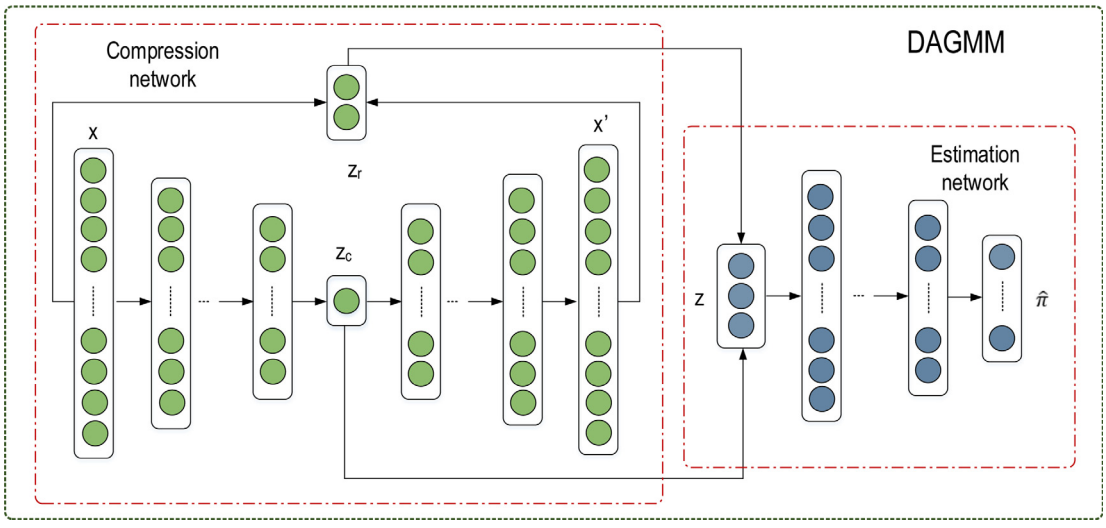

Fig. 2. The pictorial diagram of Deep Autoencoding Gaussian Mixture Model.

$$J(\theta_e, \theta_d, \theta_m) = \frac{1}{N}\sum_{i=1}^{N} L(\boldsymbol{x}_i, \boldsymbol{x}'_i) + \frac{\lambda_1}{N}\sum_{i=1}^{N} E(\boldsymbol{z}_i) + \lambda_2 P(\hat{\Sigma}) \quad (6)$$

This objective function includes three components:: 1) $L(\boldsymbol{x}_i, \boldsymbol{x}'_i)$ as the loss function which characterizes the deep autoencoder's reconstruction error; 2) $E(\boldsymbol{z}_i)$ as the component that models the probability that an input sample can be observed; 3) $P(\hat{\Sigma}) = \sum_{k=1}^{K}\sum_{j=1}^{d} 1/\hat{\Sigma}_{kjj}$ is the component utilized to penalize small values on diagonal entries in order to minimize the singularity problem that is inherent in DAGMM. In here, λ_1, λ_2 are DAGMM's meta parameters and d is the dimension of the low-dimensional representations provided by the compression network.

2.2. Supervised multiclass deep Autoencoding Gaussian mixture model

DAGMM has shown outstanding performance for anomaly detection by distinguishing observations that have statistical energy values distinct from the majority of the data population. However, this method only works well for binary classification where anomalies are considered as the minority class. In case of multi-class data such as the chiller data, which is being under investigation here, DAGMM is not capable of completely resolving the outlier detection problem. Since multiclass data often follow different distributions and an outlier can follow one of them, therefore, the DAGMM cannot distinguish a given outlier using data dis-

tributions based on its statistical energy. In this study, we have proposed the supervised multiclass version of DAGMM called Supervised Multiclass Deep Autoencoding Gaussian Mixture Model (S-DAGMM). This method divides the multiclass outlier detection problem into small anomaly detection tasks by utilizing the label information of samples during training process. Fig. 3 illustrates the proposed S-DAGMM.

In the offline phase, labeled training data are first divided into sub-datasets, according to their labels. Each sub-dataset is used to train one DAGMM. The number of trained DAGMMs is proportional to the number of data classes. Each DAGMM is then used to detect outliers in the sub-dataset used to train that DAGMM based on samples' statistical energy values. Clean sub-datasets are obtained after removing outliers and then grouped to create the clean labeled training data for chiller diagnosis model. In the online phase, a new test sample (a labeled sample or an unlabeled sample) is checked (to be an outlier or not) by trained DAGMMs. The test sample's statistical energy is firstly computed by trained DAGMMs. Votes are incremented by 1 if one of trained DAGMMs identifies the sample as an outlier. At the end, if the pre-defined value n is exceeded by the number of votes, the sample is then considered as an outlier (and vice versa). The voting scheme enforces the reliability and ensures that S-DAGMM can detect efficiently outliers, even ambiguous ones.

2.3. Pretraining DNN classifier using DAGMM

The multiple hidden layers concept has been available since the early years of deep learning. This approach was initially disappointing because its performance was even worse than shallow networks. The reason behind this inferior performance is that conventional back-propagation with random initialization often causes the training process to get stuck in unoptimized local solutions. In order to deal with the existing limitations of DNN optimization, authors in [50] proposed unsupervised layer-wise pretraining. This technique is especially powerful in case the number of labeled training samples is not sufficient for DNN to generalize a classification task, while the number of unlabeled training samples is abundant and available. By pretraining, sophisticated and abstract features with hierarchical structures can be efficiently learned because the technique offers layer-by-layer, high-level feature extraction from lower-level ones.

In this study instead of using Stacked Autoencoder (SAE), DAGMM is used to pretrain DNN due to its friendliness to end-to-end training. The pretraining and supervised fine-tuning of the DNN classifier using DAGMM are illustrated in Fig. 4. First, DAGMM is pretrained using abundant unlabeled training chiller data. Following the unsupervised pretraining, in order to fine tune the weights and biases to the DNN classifier, the output layer with the size equal to the number of data categories is added to the top of the DAGMM's encoding layers. While the initialization of DNN classifier's hidden layer weights is done with pretrained encoding weights $\theta_e = \{\omega_e, b_e\}$, the output layer parameter $\{\omega_o, b_o\}$ can be initialized randomly. Afterwards, entire DNN is fine-tuned via back propagation using labeled training data to achieve better weights. With regularization brought by the estimation network, the DAGMM's autoencoder is more capable of escaping from less attractive local optima [44].

3. Chiller system and datasets

The chiller operational data used in this study is the ASHRAE RP-1043 Dataset [5,51]. In this data, the authors artificially introduced various common chiller faults, which were then utilized

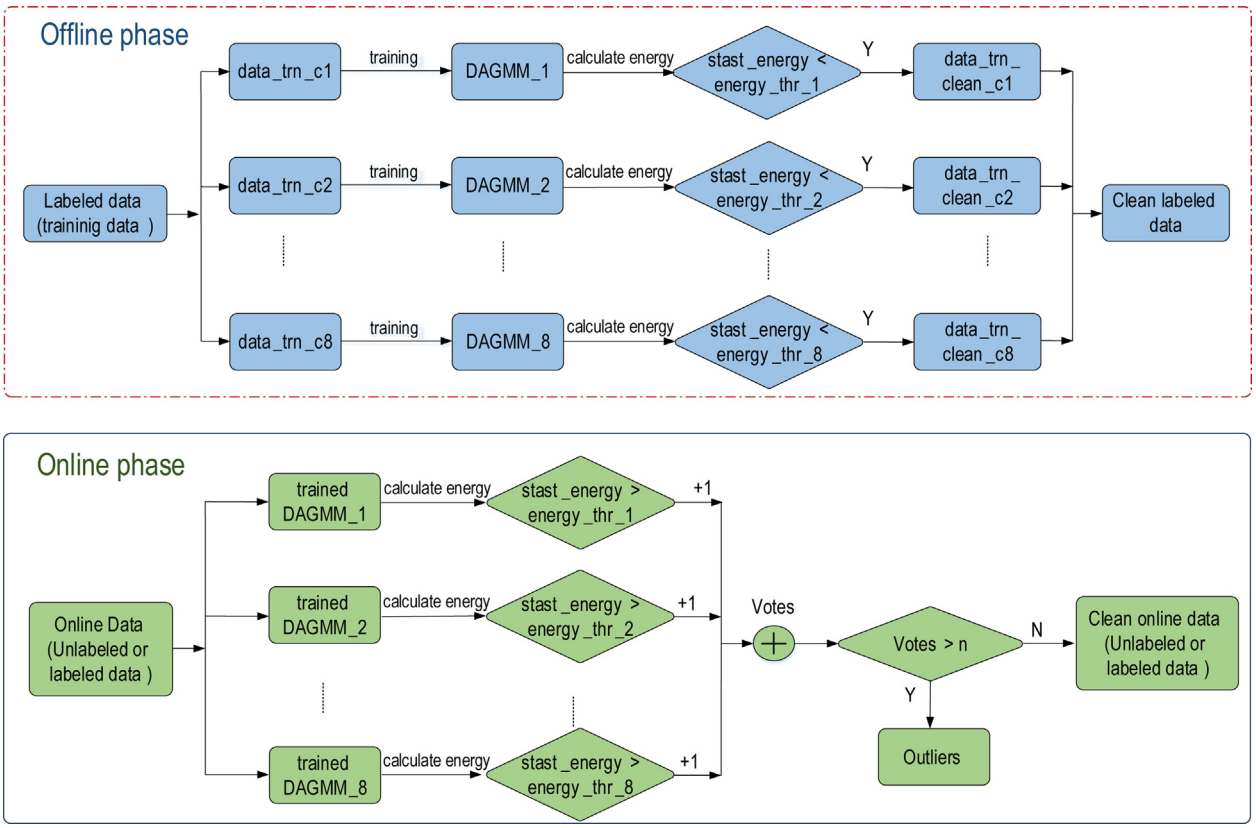

Fig. 3. An illustrative diagram of Supervised Multiclass Deep Autoencoding Gaussian Mixture Model.

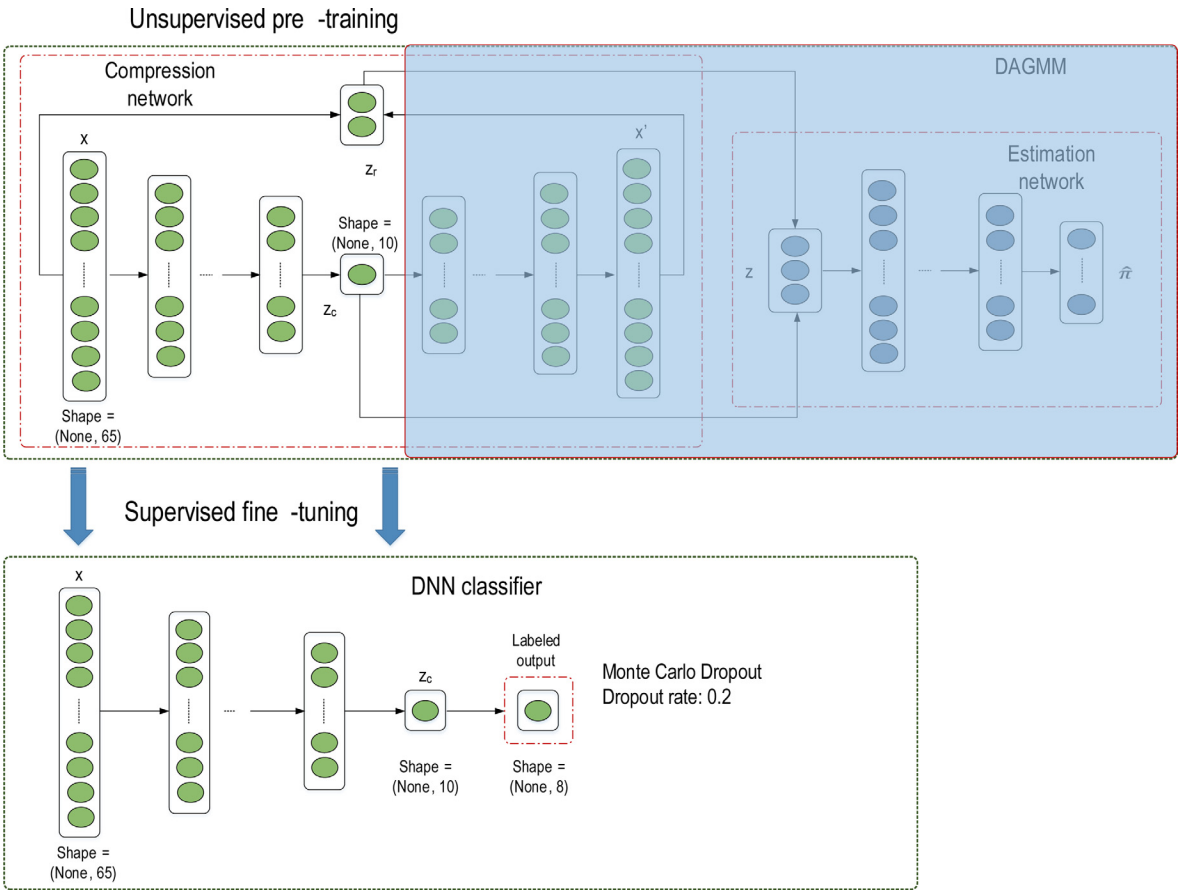

Fig. 4. The pre-training and fine-tuning procedure of the DNN classifier.

for operation recording at 10-second and 2-minute intervals. The proposed method was verified on the 10-second-interval data.

In the project RP-1043, a 90-ton centrifugal water-cooled chiller was utilized for normal and fault data acquisition, which can be considered the representative of chillers commonly found in larger installations [51]. The schematic of a typical, four-component, single-stage vapor compression chiller is shown in Fig. 5. The four main components of a typical chiller include (1) compressor, (2) condenser, (3) expansion valve, and (4) evaporator. Condenser operation at high pressure (high temperature) and evaporator operation at low pressure (low temperature) are enabled for heat rejection and absorption.

Table 1 shows the data details of seven process fault types introduced, each at four severity levels (ranging from the lowest SL1 to the highest SL4), into the chiller of RP-1043 project. Fault injections were performed as follow: reduced condenser water flow rate (ReduCF) and reduced evaporator water flow rate (ReduEF) faults by directly lowering condenser and evaporator's water flow rate; the refrigerant leakage (RefLeak) fault by reducing refrigerant charge; refrigerant overcharge (RefOver) fault by increasing refrigerant charge; excess oil (ExcsOil) fault by charging more oil than normal; condensed fouling (ConFoul) fault by plugging tubes into condenser and non-condensable in refrigerant (NonCon) fault by adding nitrogen to the refrigerant. Among typical faults, RefLeak, RefOver, ExcsOil are considered system-level ones, in contrast to element-level faults in which the fault and major symptom can be confined to a certain location such as ReduCF, ReduEF, ConFoul, and NonCon.

A total of 65 features were monitored in the processed ASHRAE RP-1043 dataset, which contains in/out water temperature from condenser, evaporator, oil feed, etc., recorded in every 10-second interval. A set of approximately 5191 samples were collected at each severity level of each fault. This means that for each severity level, the total number of samples is 5191*8 = 41,528 (i.e., seven faults and normal). In this study, four primary datasets according to four severity levels of faults are used. Each dataset contains 41,528 chiller data samples (i.e., both steady samples and transient samples) of normal operation and seven typical faults.

4. Experimental results

4.1. The efficacy of S-DAGMM

In this experiment, each severity level will be tested separately. To mimic the practical scenario that outliers in chiller data are generated by errors made in data recording process such as wrong measurements, or noises, for each dataset 20% of data is manually modified by randomly changing some features of each sample record to the value of 500 (i.e., 2 out of 65 original features). The order of the changed features in the records is not the same and is chosen at random. The reason for choosing a value of 500 is that except for some features with very high values (i.e., TSO, Shared Cond Tons, Cond Energy Balance, THO, etc.), most features have mean values in approximate range from 14 to 270, so the value of 500 can be considered the wrong feature value of most features and should be detected with outlier detection algorithm. If we change features with lower values (i.e., 300), even if the outlier detection algorithms fail to detect simulated outliers, the classifiers still correctly classify these outliers. In contrast, if we change features with higher values (i.e., 1000), simulated outliers in this

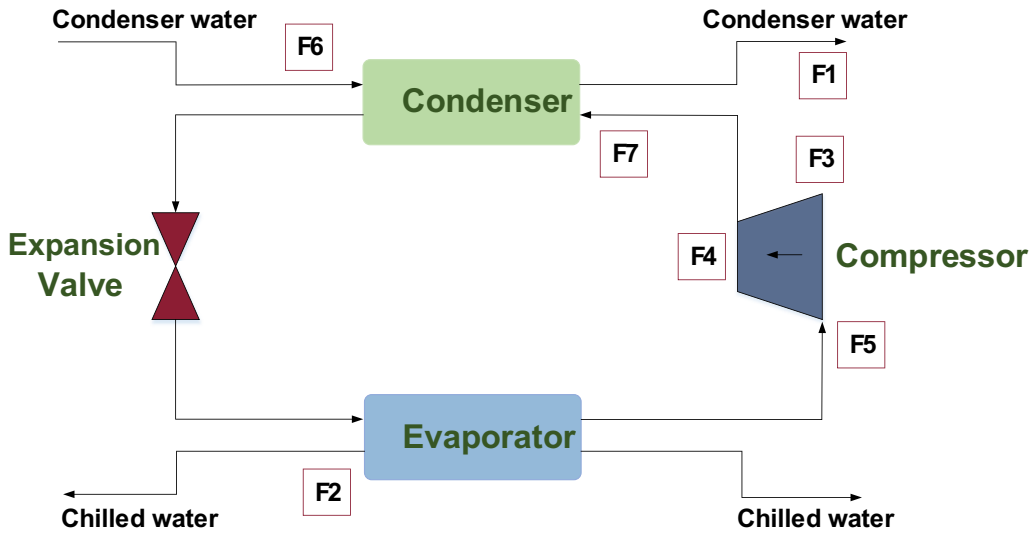

Fig. 5. The diagram of a 90-ton chiller internal structure.

Table 1
Chiller faults definition and severity levels.

Fault type	Description	Severity Level 1	Severity Level 2	Severity Level 3	Severity Level 4
F1	Reduced Condenser Water Flow	10% reduced in flow	20% reduced in flow	30% reduced in flow	40% reduced in flow
F2	Reduced Evaporator Water Flow	10% reduced in flow	20% reduced in flow	30% reduced in flow	40% reduced in flow
F3	Refrigerant Leak	10% reduced in charge	20% reduced in charge	30% reduced in charge	40% reduced in charge
F4	Refrigerant Overcharge	10% increase in charge	20% increase in charge	30% increase in charge	40% increase in charge
F5	Excess Oil	14% increase in charge	32% increase in charge	50% increase in charge	68% increase in charge
F6	Condenser Fouling	12% reduced in tubes	20% reduced in tubes	30% reduced in tubes	45% reduced in tubes
F7	Non-condensables in Refrigerant	1% by volume Nitrogen	2% by volume Nitrogen	3% by volume Nitrogen	5% by volume Nitrogen

study will be easily detected by most of outlier detection algorithms. Then, the comparison between algorithms will not be convincing when the output of the outlier detection algorithms is not too different. The idea of using transient fault data of RP-1043 dataset as outliers has also been considered. However, the transient chiller fault data themselves are neither too divergent nor distinct from the steady fault data. Experimental results conducted in this study (i.e., Table 2) have shown that when the supervised classifiers like SVM, KNN, RF, etc. are tested with the processed data (after removing artificial outliers and considering both steady-state data and transient data), the accuracy rates reach to 99.8%. This means that with sufficient training data, supervised classifiers can generalize the classification task for both the online steady data and the transient data.

To demonstrate the effectiveness of the proposed supervised DAGMM (S-DAGMM) method, its performance is compared to those of the original unsupervised DAGMM and some state-of-the-art outlier detection algorithms such as One-Class Support

Table 2
Testing diagnostic accuracy rates of different base classifiers in case outlier detection algorithms applied.

Outlier detection algorithm	Severity Level 1				Severity Level 2			
	KNN	SVM	RF	AdaBoost	KNN	SVM	RF	AdaBoost
	Diagnostic Accuracy on Testing Set				Diagnostic Accuracy on Testing Set			
None	86.1	84.3	94.8	63.0	87.6	85.4	96.8	72.3
S-DAGMM	**97.7**	**98.4**	**99.8**	**67.1**	**98.4**	**99.0**	**99.8**	**74.5**
DAGMM	94.1	94.6	97.9	64.4	95.9	95.6	98.9	73.9
OC-SVM	93.0	93.0	97.2	63.8	94.9	94.7	98.9	73.1
IF	91.3	91.0	96.9	63.6	93.5	93.9	97.8	72.5

**None.* *do not apply outlier detection on the data (use the original data).*

Outlier detection algorithm	**Severity Level 3**				**Severity Level 4**			
	KNN	**SVM**	**RF**	**AdaBoost**	**KNN**	**SVM**	**RF**	**AdaBoost**
	Diagnostic Accuracy on Testing Set				**Diagnostic Accuracy on Testing Set**			
None	88.9	86.3	98.8	74.0	91.2	86.6	98.8	85.2
S-DAGMM	**99.4**	**99.6**	**99.9**	**77.5**	**99.2**	**99.4**	**99.8**	**88.9**
DAGMM	96.1	94.8	98.9	76.2	97.8	97.0	99.6	86.5
OC-SVM	95.4	94.7	98.6	76.7	97.2	96.8	99.4	86.0
IF	95.2	94.0	98.8	75.4	96.5	94.2	98.9	85.5

**None.* *do not apply outlier detection on the data (use the original data).*

Vector Machine (OC-SVM) [35] and Isolation Forest (IF) [36]. The parameter needs to be tuned in S-DAGMM is the pre-defined value of votes n. If the pre-defined value is set high, the final decision of S-DAGMM will be more certain (True Positive is high) since the decision is agreed upon by the majority of individual DAGMMs in S-DAGMM, but the probability to have misdetected outlier samples will increase (False Negative is high). On the contrary, if the pre-defined value is set low, more samples will be detected as outliers, but many of these detected outliers may be actually inliers (False Positive is high). In this study, the pre-defined value of votes is set so that the number of samples detected as outliers by S-DAGMM accounts for 20% of the original samples, equal to the proportion of outliers were manually generated. For the experimental datasets, the pre-defined value is set to 5 (i.e., $n = 5$) to satisfy the above requirements.

As an indirect measurement for the comparison, base classifiers are utilized for fault diagnosis following the outlier detection algorithms processing. The procedure to indirectly validate the performance of outlier detection algorithms is illustrated in Fig. 6. To enforce the objectivity and reliability of the comparison, some notable and state-of-the-art algorithms are chosen as base classifiers, including KNN, SVM, RF, and Adaptive Boosting (AdaBoost). Diagnostic accuracy, which is the ratio of the number of correctly testing diagnosed samples over the total number of testing samples, is chosen as the performance evaluation metric. To reduce the variance and make results more credible, the grid search technique with 5-fold cross validation is applied for base classifiers to find their best estimators. In the case of the RF and AdaBoost classifiers, decision tree (DT) is used as a base estimator. Feature scaling is useful and applied for the KNN classifier.

For each of four datasets, the number of samples in training set and testing set is divided equally, so each contains 41,528 / 2 = 20,764 samples. After applying the outlier detection algorithms to remove outlying data samples (i.e., 20% of the dataset), the training set and testing set each contains 0.8 * 20,764 = 16,611 samples. Using these datasets to validate base classifiers, the classification accuracy rates are listed in Table 2.

It can be perceived from Table 2 that diagnostic accuracy can be remarkably improved when sample outliers are detected and discarded from the chiller data. Without outlier detection applied, the testing diagnostic accuracy rates of KNN, SVM, RF, and AdaBoost of severity level 1 are 86.1%, 84.3%, 94.8%, and 63.0%, respectively. When the outlier detection is applied, the diagnostic accuracy rates of these base classifiers can be improved to 97.7%, 98.4%, 99.8%, and 67.1%, respectively. The accuracy rates of KNN, and SVM in case without outlier detection applied (i.e., 86.1%, and 84.3%, respectively) proportionally reflect the percentage of artificial outliers in the analysis data (i.e., 20% in both training data and testing data). While DAGMM has approximate performance as those of the OC-SVM, and IF algorithms, S-DAGMM has a higher performance than the other methods with around 2% $\sim$ 5% diagnostic accuracy rates improvement, regardless of the base classifier used. The reason for this performance improvement is that the proposed S-DAGMM uses a group of individual DAGMMs, that accordingly trained based on label information of training data, to make the final decision on potential outliers. A mechanism like ensemble models helps S-DAGMM model overcome the uncertainty of chiller data and can detect ambiguous outliers in both training data and testing data as well.

The statistical energy values of 1000 data samples estimated by individual DAGMMs are shown in Fig. 7. Out of 1000 data samples, the first 800 samples are picked randomly from the chiller dataset severity level 1, the last 200 samples are outliers generated manually from the normal and seven different faults chiller data of same severity level. From Fig. 7, we can see that an outlier can be simply detected by statistical energy values estimated by individual DAGMMs of S-DAGMM model. The voting scheme enforces the reliability and ensures that S-DAGMM can detect efficiently outliers, even ambiguous ones.

F1-score is one of the common statistics besides accuracy. This metric, which is calculated from the precision (P) and recall (R) of the test, reflects the comprehensive diagnostic performance of classification models. Precision is the ratio of correctly predicted samples over prediction samples, while recall refers to the fraction of correctly predicted samples in the real samples. In some cases, high recall is obtained by sacrificing precision. The F-score is calculated as in Eq. (7). The larger the F-score is, the better the comprehensive performance of the model gets.

$$F1 - score = \frac{2P * R}{P + R} \tag{7}$$

Table 3 and Fig. 8 shows F1-score for all categories of the base classifiers on the testing set of severity level 1 when S-DAGMM algorithm is applied, in terms of statistical and graphical perspectives, respectively. From Table 3, we can see that F-score for all categories of some base classifiers mostly reached 100% (i.e., RF, SVM). This result proves that the application of S-DAGMM algorithm comprehensively improves the diagnostic performance of the base classifiers. In addition, Fig. 8 shows that for all classifiers, especially AdaBoost, F1-scores for normal state and system-level faults (RefLeak, RefOver, ExcsOil) are lower than those of element-level faults (ReduCF, ReduEF, ConFoul, NonCon). While F1-score for component-level faults is higher than 94%, regardless of the base classifier used, the F1-score for system-level faults of the AdaBoost

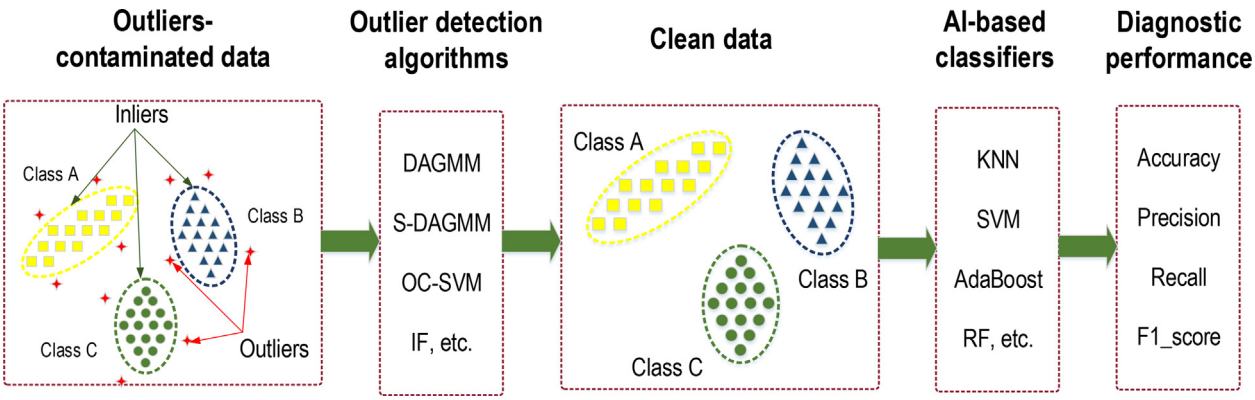

Fig. 6. The experimental procedure to validate the performance of outlier detection algorithms.

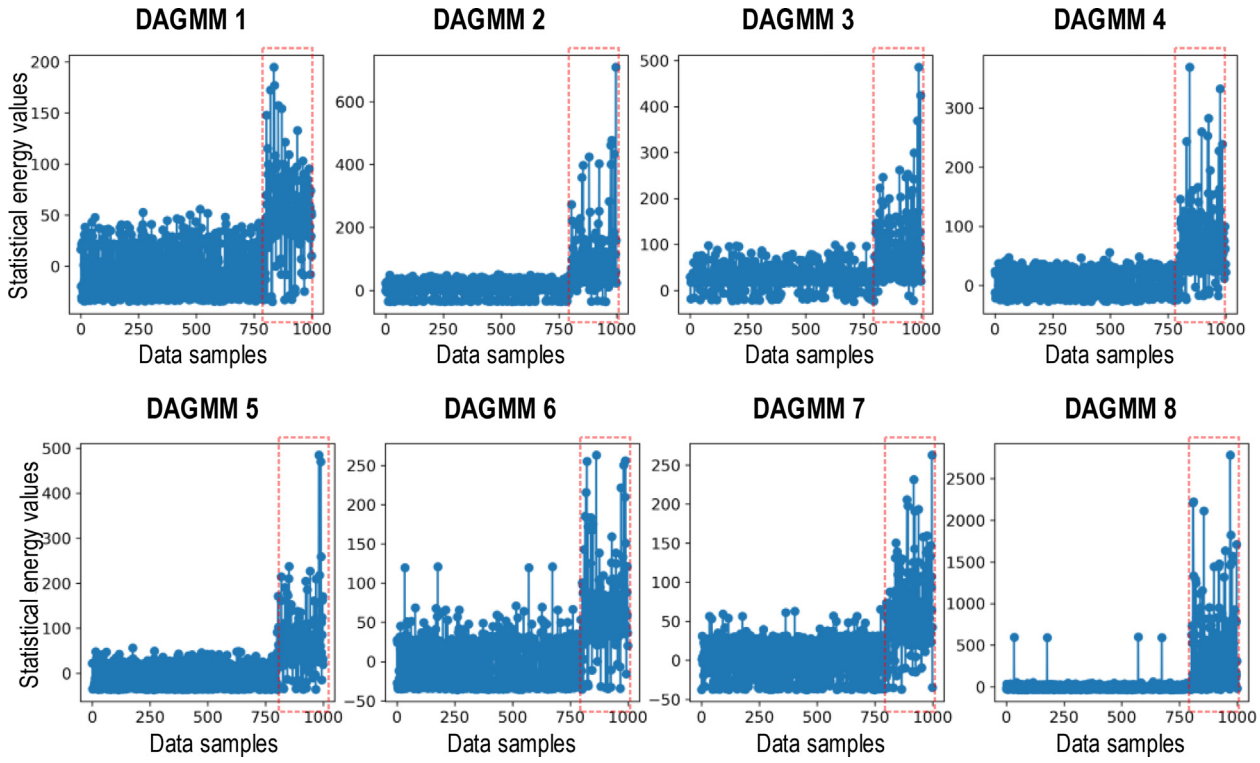

Fig. 7. Statistical energy values of data samples estimated by individual DAGMMs.

Table 3
Testing F1-score of different base classifiers in case S-DAGMM algorithm applied.

Severity level 1 Fault types	KNN			SVM			RF			AdaBoost		
	P	R	F	P	R	F	P	R	F	P	R	F
F0 - Normal	94.1	97.8	**95.9**	97.8	97.9	**97.8**	99.4	99.8	**99.6**	51.2	15.2	**23.5**
F1 - ReduCF	99.3	98.2	**98.8**	100	98.1	**99.0**	100	100	**100**	100	100	**100**
F2 - ReduEF	99.9	98.9	**99.4**	100	99.7	**99.8**	100	100	**100**	100	100	**100**
F3 - RefLeak	94.7	95.5	**95.1**	97.7	96.4	**97.0**	99.8	99.2	**99.5**	28.5	43.3	**34.3**
F4 - RefOver	96.8	95.9	**96.4**	93.6	98.5	**96.0**	99.9	100	**99.9**	42.7	94.7	**58.9**
F5 - ExcsOil	97.8	97.2	**97.5**	98.9	98.8	**98.9**	99.9	100	**99.9**	13.9	1.0	**1.9**
F6 - ConFoul	99.6	98.9	**99.2**	99.8	99.0	**99.4**	99.9	100	**99.9**	100	90.2	**94.8**
F7 - NonCon	99.9	99.0	**99.4**	99.8	98.8	**99.3**	100	99.9	**99.9**	99.3	97.0	**98.1**

***P.** Precision; ***R.** Recall; ***F.** F1-score.*

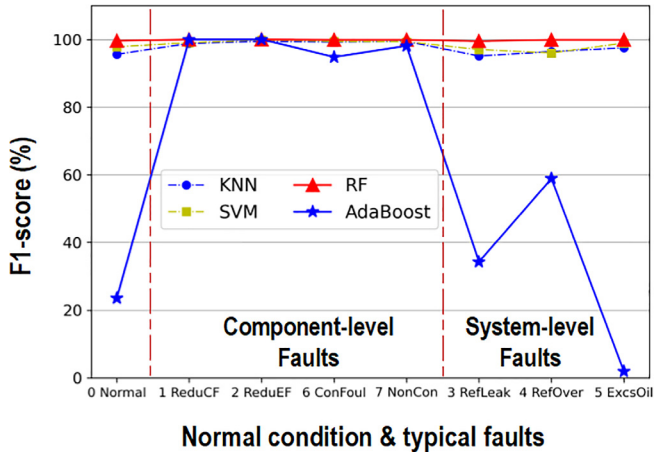

Fig. 8. F1-score curves of different base classifiers in case S-DAGMM algorithm applied (severity level 1).

classifier is less than 60%. This means system-level faults such as RefLeak, RefOver, ExcsOil are more difficult to distinguish than component-level faults.

Fig. 9 is the graphical representation of Table 2. Fig. 9 shows that performance of the outlier detection algorithms has the same trend when tested on different severity levels (i.e., severity level 1, 2, 3, and 4). Another noticeable point is that the diagnostic accuracy rate of the base classifiers increases when the severity level becomes higher. However, even the faults are in the incipient stage (i.e., severity level 1), the faults can be accurately diagnosed from outliers non-contaminated chiller data. From the statistics shown in Table 2, we can observe that unlike base classifiers such as the KNN and SVM that are sensitive to outliers, the accuracy performance of RF and AdaBoost classifiers is less affected by outliers mixed in training data. For the case of SVM, the reason is that outliers cause SVM to misidentify the separating hyperplane. While KNN is the distance-based method and outliers dramatically change its class boundaries. In the contrary, for all 4 severity levels, the diagnostic accuracy rates of RF and AdaBoost classifiers on outliers-contaminated data are close to those of RF and AdaBoost classifiers when tested on non-contaminated data.

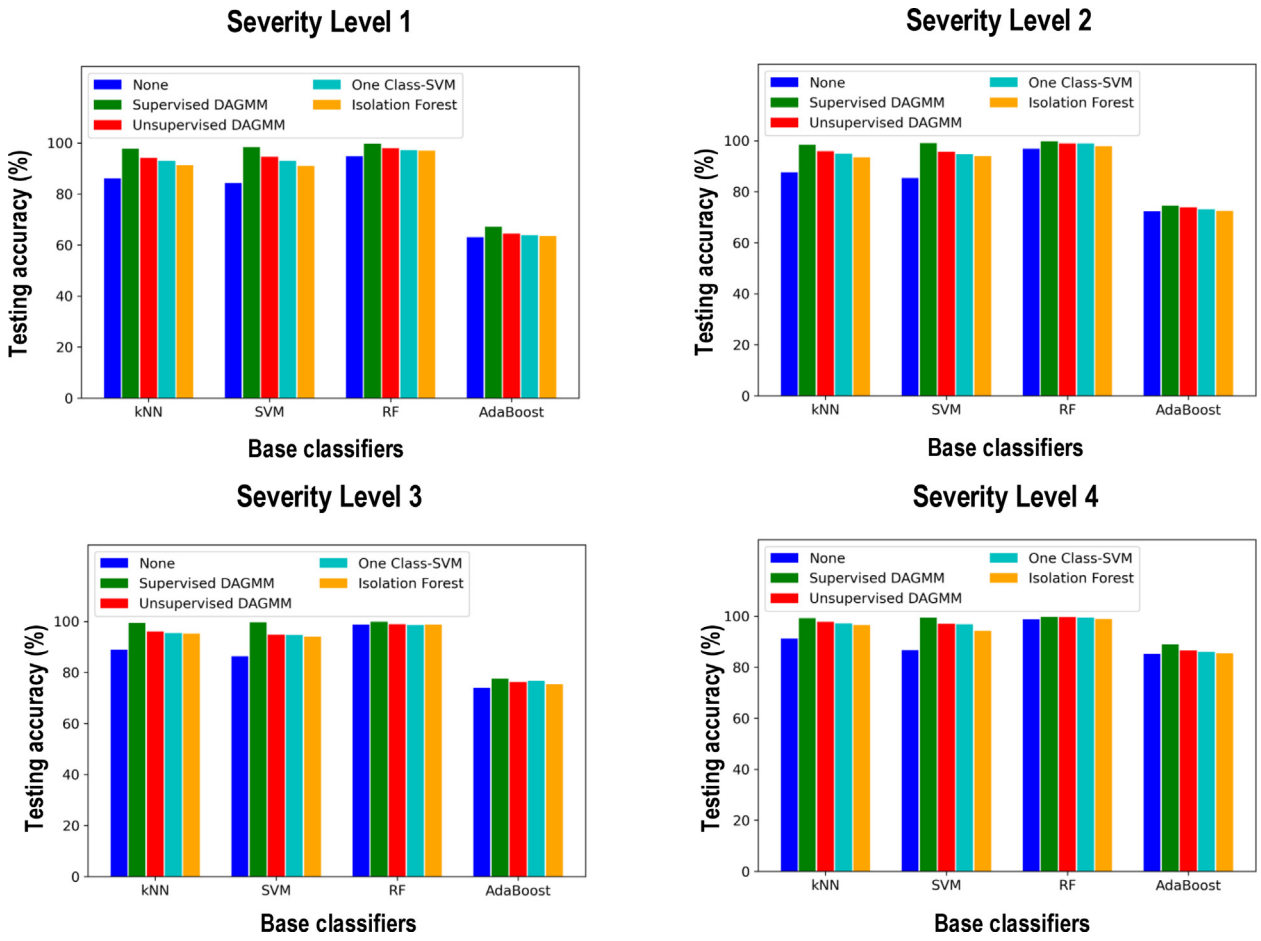

Fig. 9. Testing accuracy rates of different base classifiers at 4 severity levels in case outlier detection algorithms applied.

4.2. The efficacy of DAGMM-based pretraining technique

Our experiment aims to create a scenario as close to the real-life situation as possible, in which only a small fraction of available historical data is labeled, while the rest are not. The labeled training dataset is formed from only a small number of chiller training samples, while the unlabeled training dataset contains training samples that their labels are neglected and assumed unknown. This unlabeled training set is used to pretrain DAGMM for finetuning DNN classifier. To show the advantage of the proposed DAGMM-based finetuned DNN classifier (DAGMM-DNN), a comparative study was conducted between the proposed DAGMM-DNN and the Stacked Autoencoder (SAE)-based finetuned DNN classifier (SAE-DNN) in terms of diagnostic accuracy metric. The DNN classifiers without using pretraining technique and the base classifiers in the above experiments are also the objects of this comparison.

The information in the unlabeled training samples can be harnessed by our proposed DAGMM-DNN to obtain optimal weights via pretraining DAGMM. Instead of starting with randomly initialized weights, pretraining helps the DNN escape from less attractive local optima and gain relatively high accuracy even when the number of labeled training samples are limited. For each severity level, the size of the unlabeled training set and the testing set is fixed and is half the size of the original dataset (i.e., 41,528 / 2 = 20,764 samples). Meanwhile, the size of the labeled training set changes from 0.5% (103 samples), 1% (207 samples), 3% (622 samples), and 10% (2076 samples) the size of the unlabeled training set. This setting aims to validate the performance of supervised classifiers in case the labeled training data' volume is small. Because the labeled training data are randomly chosen from the training set, different random seeds will result in different labeled training sets that lead to different results. Therefore, the final accuracy rate of each classifier in this experiment is calculated by taking the average of 5 repeated trials. In each trial, a random selection of the labeled training samples is performed. To make fair comparison, the structure of DAGMM-DNN, SAE-DNN, and DNN is the same with the input layer, the output layer, and some hidden layers of diminishing sizes. Since the experimental data contain 65 features, then the input layer size is 65. The output layer has 8 neurons, according to 8 different data classes. Logically, the neural network with more hidden layers is capable of extracting more complex patterns from the input. Through many numerical tests, however, the patterns of the chiller dataset are found relatively simple [3]. For the case in this study, when the depth of hidden layers increases to more than three, the DNN classifier's classification performance does not increase proportionally. Therefore, the DNN classifier's hidden structure is optimized with 3 layers of diminishing sizes (50 units, 30 units, then 10 units). The procedure to validate the performance of pretraining techniques is demonstrated in Fig. 10.

The fault diagnosis results of the classifiers are shown in Table 4. From Table 4, we can see the diagnostic improvement of the DAGMM-DNN and SAE-DNN classifiers compared with the DNN classifier without applying pretraining technique. The difference is significant when the size of the labeled training set is small. In case the number of training samples is 103, the improvement in terms of testing accuracy rate is about 6% $\sim$ 10%, regardless severity level of chiller data. The reason for this improvement is that with abundant unlabeled training data, pretrained networks like

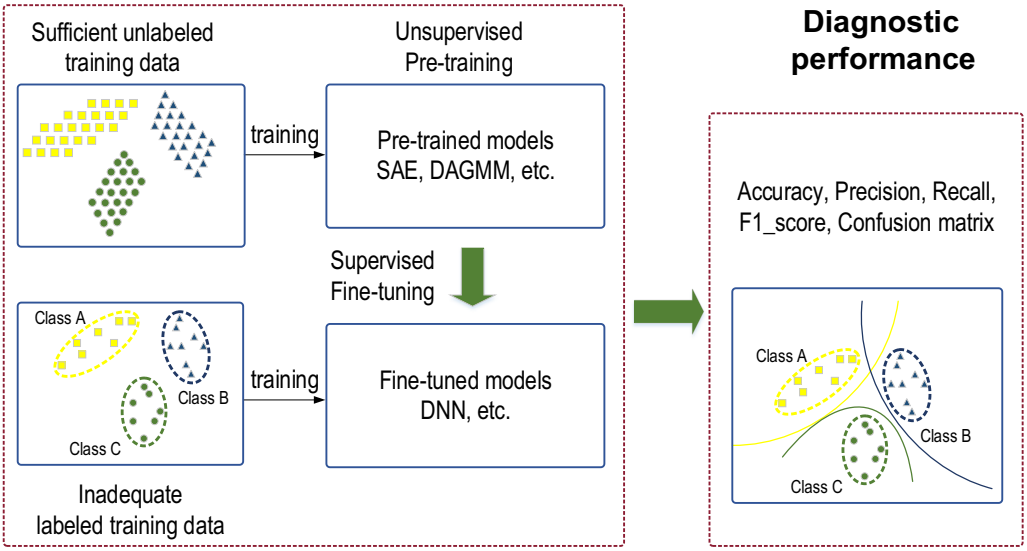

Fig. 10. The experimental procedure to validate the performance of pretraining techniques.

Table 4
Testing diagnostic accuracy rates of different classifiers according to different sizes of labeled training set.

Diagnostic models	Severity Level 1				Severity Level 2			
	Size of the labeled training set				Size of the labeled training set			
	2,076	**622**	**207**	**103**	**2,076**	**622**	**207**	**103**
	Diagnostic Accuracy on Testing Set				Diagnostic Accuracy on Testing Set			
SAE-DNN	92.7	90.8	80.4	72.5	95.8	91.8	88.4	74.9
DAGMM-DNN	**95.4**	**92.2**	**84.2**	**73.4**	**97.0**	**96.3**	**89.2**	**76.2**
DNN	95.4	91.4	77.9	64.6	96.9	96.0	85.6	69.1
KNN	86.7	64.5	25.3	23.7	93.6	74.9	43.2	31.8
SVM	91.6	74.5	45.2	30.0	95.5	81.7	52.0	32.8
Random Forest	97.6	91.2	79.2	69.0	98.1	91.1	81.5	70.6
AdaBoost	75.3	74.2	72.3	63.8	72.9	70.1	66.2	56.4

**SAE-DNN.* *DNN classifier uses SAE to pretrain;* **DAGMM-DNN.* *DNN classifier uses DAGMM to pretrain.*

Diagnostic models	**Severity Level 3**				**Severity Level 4**			
	Size of the labeled training set				Size of the labeled training set			
	2,076	**622**	**207**	**103**	**2,076**	**622**	**207**	**103**
	Diagnostic Accuracy on Testing Set				Diagnostic Accuracy on Testing Set			
SAE-DNN	98.5	97.4	93.9	88.0	98.2	97.4	96.3	87.8
DAGMM-DNN	**98.6**	**97.8**	**96.3**	**89.9**	**98.7**	**97.5**	**96.4**	**87.6**
DNN	98.8	97.8	92.8	79.0	98.4	97.5	93.9	81.7
KNN	94.9	78.1	48.7	38.3	96.3	86.2	64.3	53.0
SVM	95.3	81.0	50.5	33.8	96.7	90.2	59.6	39.3
Random Forest	99.2	96.2	90.2	82.5	98.9	97.1	93.8	86.6
AdaBoost	82.1	78.5	71.0	64.6	93.1	92.0	84.0	81.4

**SAE-DNN.* *DNN classifier uses SAE to pretrain;* **DAGMM-DNN.* *DNN classifier uses DAGMM to pretrain.*

DAGMM and SAE prevent the DNN classifier from getting stuck on local optima, especially in case the number of labeled training samples is not enough for the DNN classifier to generalize a fault diagnostic task. As the size of the labeled training set is 622 or higher, the testing accuracy rates between the DNN classifiers with pretraining and the DNN classifier without pretraining are almost similar. This is because as the labeled training data are sufficient, supervised classifiers like DNN and the base classifiers can obtain enough generalization capability to discriminate different faults via training.

Comparing between two fine-tuned models, the diagnostic accuracy performance of SAE-DNN is always inferior to DAGMM-DNN with about 1% $\sim$ 5% lower in accuracy rate. Especially when the number of labeled training samples is 622 or higher, the accuracy rate of the SAE-DNN classifier is even lower than that of the DNN classifier without pretraining. The better performance of DAGMM is due to the regularization from estimation network, which is outstandingly useful in terms of local optima escaping. Meanwhile, a stacked autoencoder without the regularization from the estimation network has high chance to converge to the local minimum. Besides, the diagnostic accuracy rate of DAGMM-DNN is also superior to that of the base classifiers such as KNN, SVM, RF, and AdaBoost as the size of the labeled training set is less than 622 (i.e., 103 or 207). The improvement in terms of testing accuracy rate is about 5% $\sim$ 59%, regardless severity level of chiller data. In case the size of the labeled training set is 622 or higher (i.e., 622 or 2076), the DAGMM-DNN's accuracy rate is not much better and even lower in some cases. For example, as the size of the labeled training set is 2,076, the accuracy rate of the RF classifier is highest, reaching over 97.6%, regardless severity level of chiller data.

Table 5
Testing F1-score of different classifiers with 207 labeled training samples (severity level 1).

Severity level 1 Fault types	SAE-DNN			DAGMM-DNN			DNN			RF		
	P	R	F	P	R	F	P	R	F	P	R	F
F0 - Normal	66.3	68.5	**67.4**	72.5	71.5	**72.0**	70.5	64.4	**67.3**	71.3	59.7	**65.0**
F1 - ReduCF	99.9	97.8	**98.8**	99.8	98.4	**99.1**	93.7	98.9	**96.2**	100	99.7	**99.8**
F2 - ReduEF	93.9	99.9	**96.8**	96.0	98.2	**97.1**	99.0	98.1	**98.5**	96.5	100	**98.2**
F3 - RefLeak	40.4	39.5	**39.9**	53.3	48.2	**50.6**	39.4	38.9	**39.1**	45.3	49.5	**47.4**
F4 - RefOver	76.2	81.2	**78.6**	77.5	83.1	**80.2**	67.2	76.9	**71.7**	66.0	78.4	**81.6**
F5 - ExcsOil	68.7	60.1	**64.1**	74.0	78.6	**76.2**	61.2	60.6	**60.9**	58.3	51.9	**54.9**
F6 - ConFoul	97.0	97.5	**97.2**	99.3	97.8	**98.6**	98.4	90.7	**94.3**	98.5	97.6	**98.1**
F7 - NonCon	98.7	98.4	**98.6**	99.3	97.1	**98.2**	94.6	94.0	**94.3**	99.8	96.2	**98.0**

***P.** Precision; ***R.** Recall; ***F.** F1-score;*

Table 5 shows F1-score for all categories of different classifiers with 207 training labeled samples. These classifiers include DNN, SAE-DNN, DAGMM-DNN and RF, which has the highest accuracy rate between the base classifiers. Fig. 11 is the graphical representation of statistical results in Table 5. From Fig. 11, it is clear that F1-scores of all the classifiers for system-level faults are lower than those of element-level faults. Compared with SAE-DNN and DNN, we can see that when the volume of training labeled set is small, DAGMM-DNN achieves the best F1-score for all categories. Especially for system-level faults such as Refrigerant leak (RefLeak), Refrigerant overcharge (RefOver) and Excess oil (ExcsOil), the difference in F1-score of DAGMM-DNN and DNN is even more significant (at about 13% $\sim$ 25%). In addition, the F1-score of DAGMM-DNN is also higher than that of RF for almost the categories.

A confusion matrix is another metric that is often used to describe the performance of classification models on a testing set for which the true values are known. For more in-depth diagnostic performance comparison, confusion matrices of the classifiers are also studied as shown in Fig. 12. In Fig. 12, the rows represent the true labels of the faults, while the columns represent the predicted labels. The reports with a great number of misclassified samples are marked with red dot circles. From Fig. 12, we can see that with small number of labeled training samples (i.e., 207 samples), system-level faults (RefLeak, RefOver, ExcsOil) are misclassified at a great scale. One reason for this great misclassification is that faults samples are in the incipient stage (i.e., severity level 1) so that many of them are misclassified as normal states (and vice versa). More specifically, many normal samples in row 1 are misclassified as RefLeak, RefOver, or ExcsOil. While the samples of RefLeak state are mainly misclassified as RefOver, ExcsOil, or normal state, the high proportion of ExcsOil samples are misdiagnosed as normal state, RefLeak and RefOver.

In comparison with SAE-DNN and DAGMM-DNN, the DNN classifier does not perform well and reports the largest number of misdiagnosed samples, especially in case of normal state and system-level faults. More specifically, for 2595 testing samples of each of the typical faults, the number of correctly diagnosed samples of the DNN classifier in case of normal state, RefLeak, RefOver, and ExcsOil are only 1662, 997, 2041, and 1572, respectively. Meanwhile, the figures for DAGMM-DNN are higher with the numbers of 1844, 1236, 2207, and 2037, respectively. These results reinforce the conclusion that using DAGMM to pretrain DNN significantly increases the diagnostic performance of DNN.

Fig. 13 is the graphical representation of Table 4. From subfigures in Fig. 13, we can see that the diagnostic rate curves of the base classifiers have a same trend. These curves of the classifiers sharp in the beginning and tend to be constant with the increase of unlabeled training samples. Regardless the size of labeled training set, the accuracy rates of the DAGMM-DNN classifier are higher than those of base classifiers, except for very few exceptions (i.e., the accuracy rates of Random Forest classifier are highest among compared classifiers in case the size of labeled training set is 2076 samples). Another noticed point is that when the number of labeled training samples increases, the classification accuracy difference between the finetuned networks and the supervised base classifiers becomes smaller. While the SVM and KNN classifiers are sensitive to the size of the labeled training sets as their accuracy rates are almost lower than 50% in case the labeled training sample is 207, the accuracy rates of the Random Forest and AdaBoost classifiers are not affected too much by the size of training set. In both experiments, Random Forest classifier proved it to be a robust model that is resistant to outliers and unaffected by the size of training set.

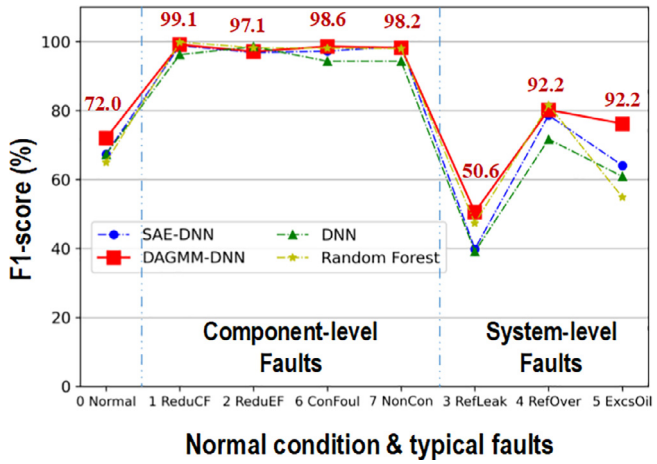

Fig. 11. F1-score curves of DNN classifiers and RF classifier with 207 labeled training samples (severity level 1).

5. Conclusion

With a focus on outlier detection problem, this paper proposed a supervised DAGMM (S-DAGMM) model that can effectively detect outliers of multiclass chiller data by profiling the label information of data during training process. A mechanism like ensemble models helps S-DAGMM model overcome the uncertainty of chiller data and can detect even ambiguous outliers in both training data and testing data as well. The experimental results have shown that the effectiveness of S-DAGMM is superior to that of DAGMM, OC-SVM, and IF. More specifically in case the data is processed by S-DAGMM, the diagnostic accuracy rate of the base classifiers such as KNN, SVM, RF, and AdaBoost is improved by 4% $\sim$ 11 %, compared to the case outliers-contaminated data is directly used. Compared with the case of contaminated data processed by DAGMM, OC-SVM, and IF, the improvement is around 2% $\sim$ 5%, regardless of the base classifier used.

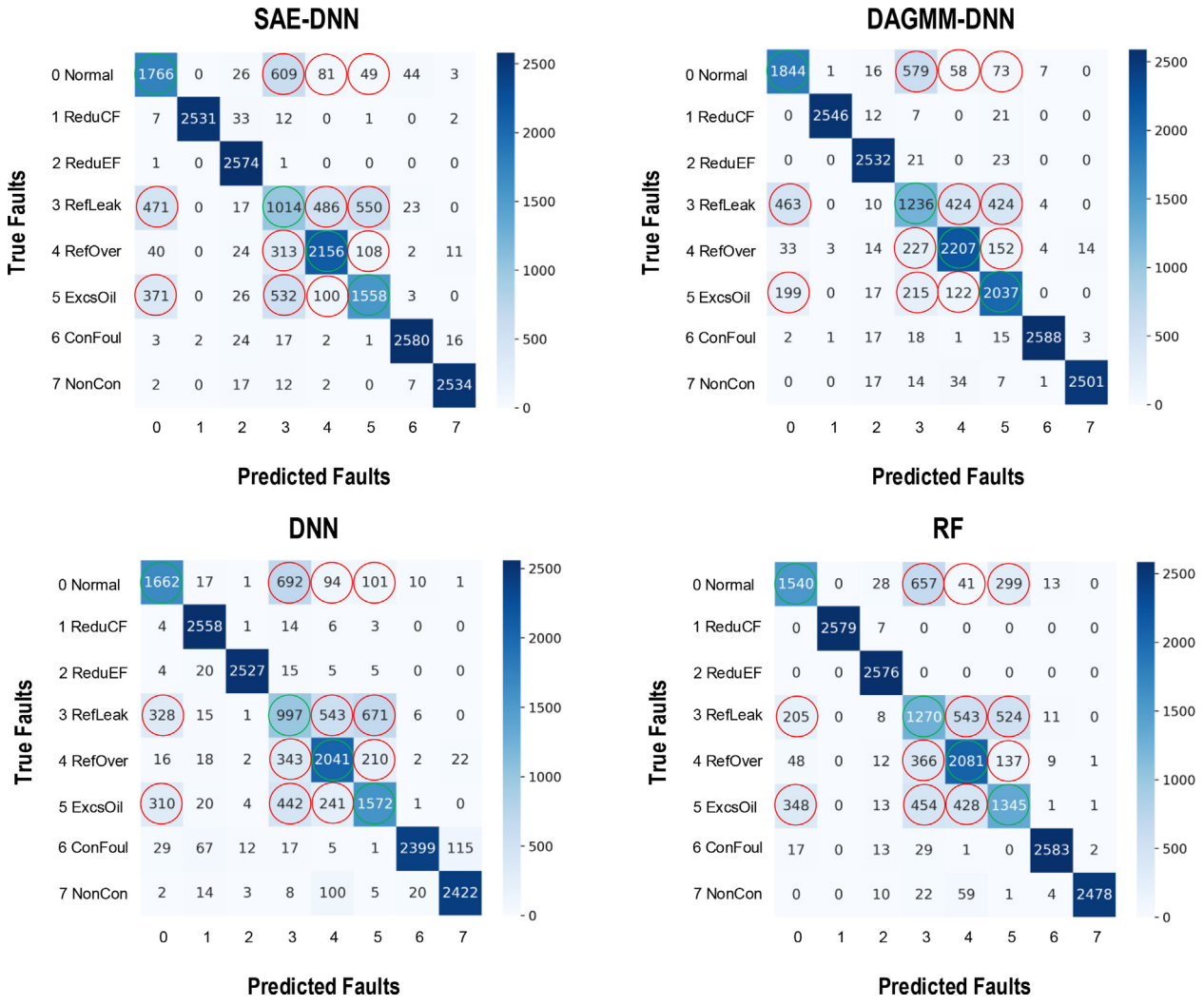

Fig. 12. Confusion matrices of DNN classifiers and RF classifier with 207 labeled training samples (severity level 1).

Besides the proposal of S-DAGMM, this paper also proposed to use DAGMM to pretrain the DNN classifier (DAGMM-DNN). To prove the effectiveness of the estimation network in the structure of DAGMM, we conducted a study by comparing the diagnostic performance of DAGMM-DNN with SAE-DNN in terms of diagnostic accuracy rate, F1-score, and confusion matrix. The DNN classifier without using pretraining and the base classifiers such as KNN, SVM, RF, and AdaBoost are also the objects of this comparison. The experimental results have shown that the diagnostic accuracy performance of the DAGMM-DNN classifier is always superior to SAE-DNN with about 1% $\sim$ 5% higher accuracy rate. In case the size of training labeled data is small (i.e., 207 samples) DAGMM-DNN achieves the best F1-score for all categories as compared with SAE-DNN, and DNN. Especially for system-level faults such as Refrigerant leak (RefLeak), Refrigerant overcharge (RefOver) and Excess oil (ExcsOil), the difference in F1-score of DAGMM-DNN and DNN is even more significant (at about 13% $\sim$ 25%).

In summary, the proposed methods have more advantages in the practical use as the chiller data are easily contaminated by outliers and the cost to obtain labeled fault data is expensive. The general ability of the proposed methods is guaranteed by several factors. First, DAGMM and S-DAGMM are deep neural networks that work on the platform of Gaussian Mixture Mode (GMM) whose application is very diverse and flexible. Second, the experimental RP-1043 dataset is one of the most common chiller datasets used in chiller FDD studies, in which both fault-free and various fault data are recorded at ten-second intervals on the 90-ton chiller. This chiller is a representative of chillers used in larger installations. Third, all chiller faults were introduced at four severity levels (SL1-SL4, from slightest to severest). The data were collected not only when system has reached steady states, but also at transient states. Such dataset ensures that the proposed methods have been comprehensively tested under different operating conditions and system states. Future works could be devoted to the compression network of DAGMM model. Instead of using SAE as the compression network of DAGMM, some autoencoder variants, for example a variational autoencoder (VAE), can be used instead to better extract high-level features for the estimation network.

To ensure reproducibility of the results by the research community and a potential future improvement of our framework by other researchers the complete source code is provided in the following repository: https://github.com/viettra-xai/S-DAGMM.

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.

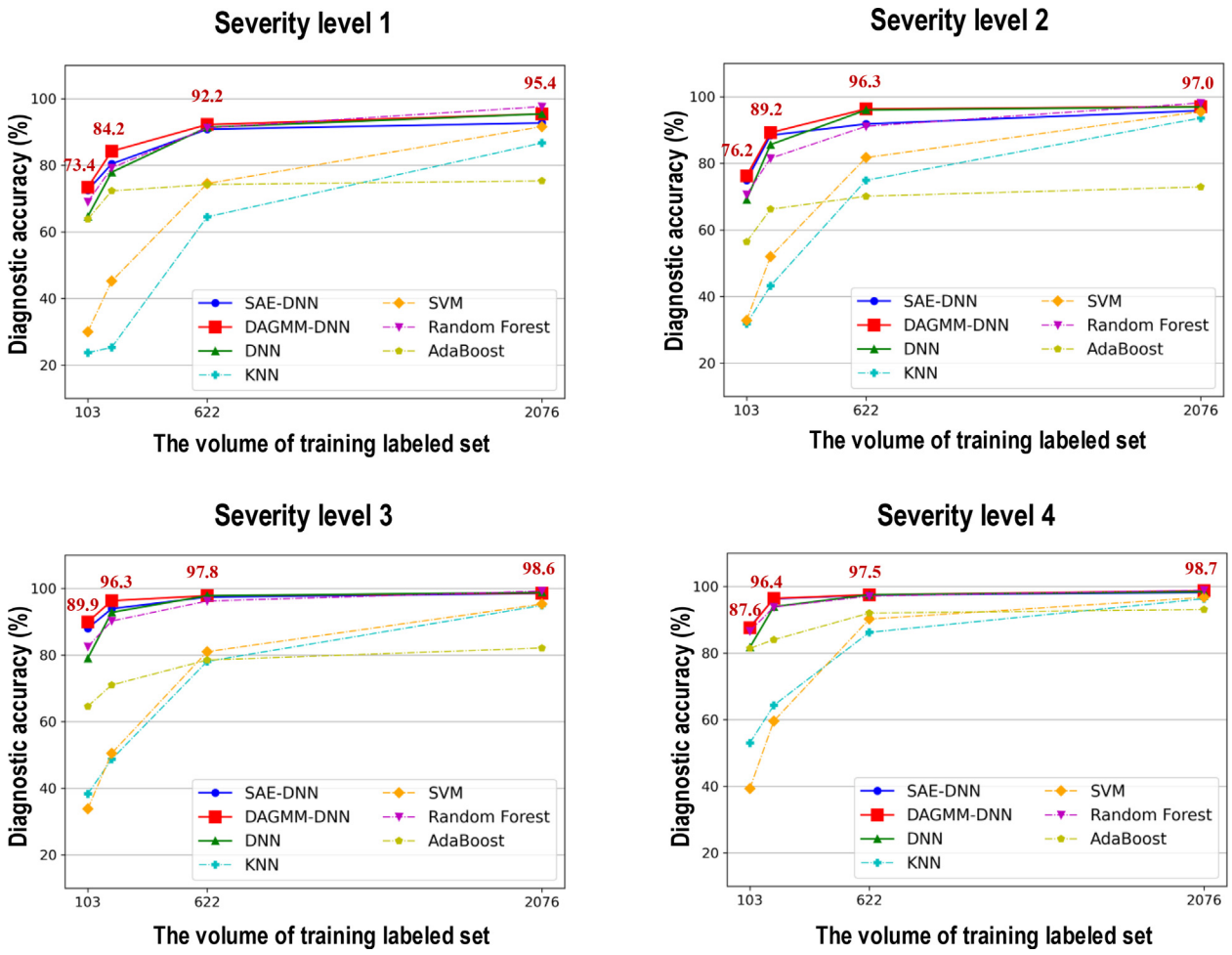

Fig. 13. Testing accuracy rate curves of different classifiers according to different sizes of labeled training set.

Acknowledgment

The completion of this research was made possible thanks to Natural Sciences and Engineering Research Council of Canada (NSERC) and the INVOLVED ANR-14-CE22-0020-01 project of the French National Research Agency. The authors would like to thank the associate editor and the reviewers for their helpful comments and suggestions.

References

[1] M.S. Mirnaghi, F. Haghighat, Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review, Energy Build. (2020) 110492.
[2] 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.
[3] 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.
[4] K. Yan, A. Chong, Y. Mo, Generative adversarial network for fault detection diagnosis of chillers, Build. Environ. 172 (2020) 106698.
[5] M.C. Comstock, J.E. Braun, E.A. Groll, A survey of common faults for chillers/ Discussion, Ashrae Transactions 108 (2002) 819.
[6] X. Zhu, K. Chen, B. Anduv, X. Jin, Z. Du, Transfer learning based methodology for migration and application of fault detection and diagnosis between building chillers for improving energy efficiency, Build. Environ. (2021) 107957.
[7] Y. Zhao, F. Xiao, S. Wang, An intelligent chiller fault detection and diagnosis methodology using Bayesian belief network, Energy Build. 57 (2013) 278–288.
[8] X. Luo, K. Fong, Novel pattern recognition-enhanced sensor fault detection and diagnosis for chiller plant, Energy Build. 228 (2020) 110443.
[9] Z. Wang, L. Wang, Y. Tan, J. Yuan, X. Li, Fault diagnosis using fused reference model and Bayesian network for building energy systems, J. Build. Eng. 34 (2021) 101957.
[10] J. Liu et al., Data-driven and association rule mining-based fault diagnosis and action mechanism analysis for building chillers, Energy Build. 216 (2020) 109957.
[11] F. Xiao, C. Zheng, S. Wang, A fault detection and diagnosis strategy with enhanced sensitivity for centrifugal chillers, Appl. Therm. Eng. 31 (17–18) (2011) 3963–3970.
[12] X. Zhao, M. Yang, H. Li, A virtual condenser fouling sensor for chillers, Energy Build. 52 (2012) 68–76.
[13] 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.
[14] Z. Wang, L. Wang, K. Liang, Y. Tan, Enhanced chiller fault detection using Bayesian network and principal component analysis, Appl. Therm. Eng. 141 (2018) 898–905.
[15] G. Li, Y. Hu, An enhanced PCA-based chiller sensor fault detection method using ensemble empirical mode decomposition based denoising, Energy Build. 183 (2019) 311–324.
[16] G. Li, Y. Hu, Improved sensor fault detection, diagnosis and estimation for screw chillers using density-based clustering and principal component analysis, Energy Build. 173 (2018) 502–515.
[17] 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.
[18] Z. Wang, Z. Wang, X. Gu, S. He, Z. Yan, Feature selection based on Bayesian network for chiller fault diagnosis from the perspective of field applications, Appl. Therm. Eng. 129 (2018) 674–683.
[19] 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.
[20] H. Han, X. Cui, Y. Fan, H. Qing, Least squares support vector machine (LS-SVM)-based chiller fault diagnosis using fault indicative features, Appl. Therm. Eng. 154 (2019) 540–547.

[21] K. Sun, G. Li, H. Chen, J. Liu, J. Li, W. Hu, A novel efficient SVM-based fault diagnosis method for multi-split air conditioning system's refrigerant charge fault amount, Appl. Therm. Eng. 108 (2016) 989–998.
[22] D. Li, Y. Zhou, G. Hu, C.J. Spanos, Fault detection and diagnosis for building cooling system with a tree-structured learning method, Energy Build. 127 (2016) 540–551.
[23] 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.
[24] 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.
[25] K. Yan, J. Su, J. Huang, Y. Mo, Chiller Fault Diagnosis Based on VAE-Enabled Generative Adversarial Networks, IEEE Trans. Autom. Sci. Eng. (2020).
[26] Z. Wang, Y. Dong, W. Liu, Z. Ma, A novel fault diagnosis approach for chillers based on 1-D convolutional neural network and gated recurrent unit, Sensors 20 (9) (2020) 2458.
[27] J. Gao, H. Han, Z. Ren, Y. Fan, Fault diagnosis for building chillers based on data self-production and deep convolutional neural network, Journal of Building Engineering 34 (2021) 102043.
[28] B. Jin, D. Li, S. Srinivasan, S.-K. Ng, K. Poolla, A. Sangiovanni-Vincentelli, Detecting and diagnosing incipient building faults using uncertainty information from deep neural networks, in: 2019 IEEE International Conference on Prognostics and Health Management (ICPHM), IEEE, 2019, pp. 1–8.
[29] H. Han, L. Xu, X. Cui, Y. Fan, Novel chiller fault diagnosis using deep neural network (DNN) with simulated annealing (SA), Int. J. Refrig 121 (2021) 269–278.
[30] E. Klevak, S. Lin, A. Martin, O. Linda, and E. Ringger, "Out-Of-Bag Anomaly Detection," *arXiv preprint arXiv:2009.09358,* 2020.
[31] G. Pang, L. Cao, L. Chen, H. Liu, Learning representations of ultrahigh-dimensional data for random distance-based outlier detection, in: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 2041–2050.
[32] K. Bhaduri, B.L. Matthews, C.R. Giannella, Algorithms for speeding up distance-based outlier detection, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, 2011, pp. 859–867.
[33] V. Hautamaki, I. Karkkainen, P. Franti, Outlier detection using k-nearest neighbour graph, in: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, IEEE, 2004, pp. 430–433.
[34] K. Zhang, M. Hutter, H. Jin, A new local distance-based outlier detection approach for scattered real-world data, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2009, pp. 813–822.
[35] M.S. Sadooghi, S.E. Khadem, Improving one class support vector machine novelty detection scheme using nonlinear features, Pattern Recogn. 83 (2018) 14–33.
[36] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation-based anomaly detection, ACM Trans. Knowledge Discov. Data (TKDD) 6 (1) (2012) 1–39.
[37] S. Serneels, T. Verdonck, Principal component analysis for data containing outliers and missing elements, Comput. Stat. Data Anal. 52 (3) (2008) 1712–1727.
[38] Y.-J. Lee, Y.-R. Yeh, Y.-C.-F. Wang, Anomaly detection via online oversampling principal component analysis, IEEE Trans. Knowl. Data Eng. 25 (7) (2012) 1460–1470.
[39] J. Chen, S. Sathe, C. Aggarwal, D. Turaga, Outlier detection with autoencoder ensembles, in: Proceedings of the 2017 SIAM international conference on data mining, 2017: SIAM, 2017, pp. 90–98.
[40] T. Kieu, B. Yang, C. Guo, C.S. Jensen, Outlier detection for time series with recurrent autoencoder ensembles, IJCAI (2019) 2725–2732.
[41] K. Yamanishi, J.-I. Takeuchi, G. Williams, P. Milne, On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms, Data Min. Knowl. Disc. 8 (3) (2004) 275–300.
[42] M. Goldstein A. Dengel "Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm," KI-2012: Poster and Demo Track 2012 59 63
[43] E.J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, J ACM (JACM) 58 (3) (2011) 1–37.
[44] B. Zong et al., Deep autoencoding gaussian mixture model for unsupervised anomaly detection, International Conference On Learning Representations, 2018.
[45] V. Tra, J. Kim, S.A. Khan, J.-M. Kim, Bearing fault diagnosis under variable speed using convolutional neural networks and the stochastic diagonal levenberg-marquardt algorithm, Sensors 17 (12) (2017) 2834.
[46] V. Tra, T.-K. Nguyen, C.-H. Kim, J.-M. Kim, Health indicators construction and remaining useful life estimation for concrete structures using deep neural networks, Appl. Sci. 11 (9) (2021) 4113.
[47] V. Tra, B.-P. Duong, J.-M. Kim, Improving diagnostic performance of a power transformer using an adaptive over-sampling method for imbalanced data, IEEE Trans. Dielectr. Electr. Insul. 26 (4) (2019) 1325–1333.
[48] Y. Fan, X. Cui, H. Han, H. Lu, Chiller fault diagnosis with field sensors using the technology of imbalanced data, Appl. Therm. Eng. 159 (2019) 113933.
[49] 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.
[50] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[51] M. Comstock, J.E. Braun, Fault Detection And Diagnostic (FDD) Requirements And Evaluation Tools For Chillers, ASHRAE, West Lafayette, IN, 2002.