Energy & Buildings 259 (2022) 111872

Contents lists available at ScienceDirect

Energy & Buildings

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

Real vs. simulated: Questions on the capability of simulated datasets on building fault detection for energy efficiency from a data-driven perspective

Jiajing Huang [a], Jin Wen [b], Hyunsoo Yoon [c,*], Ojas Pradhan [b], Teresa Wu [a,*], Zheng O'Neill [d], Kasim Selcuk Candan [a]

[a] School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281, USA
[b] Department of Civil, Architectural & Environmental Engineering, Drexel University, Philadelphia, PA 19104, USA
[c] Department of Industrial Engineering, Yonsei University, Seoul 03722, Republic of Korea
[d] J. Mike Walker '66 Department of Mechanical Engineering, Texas A&M University, College Station, TX 77843, USA

ARTICLE INFO

Article history:
Received 30 October 2021
Revised 5 January 2022
Accepted 17 January 2022
Available online 22 January 2022

Keywords:
Building AFDD
Machine learning
Simulated
Real
Similarity

ABSTRACT

Literature on building Automatic Fault Detection and Diagnosis (AFDD) mainly focuses on simulated system data due to high expenses and difficulties of obtaining and analyzing real building data. There is a lack of validation on performances and scalabilities of data-driven AFDD approaches using simulated data and how it compares to that from real building data. In this study, we conduct two sets of experiments to seek answers to this question. We first evaluate data-driven fault detection strategies on real and simulated building data separately. We observe that the fault detection performances are not affected by fault detection strategies, sizes of training data, and the number of cross-validation folds when training and blind test data come from the same data source, namely, simulated or real building data. Next, we conduct a cross-dataset study, that is, develop the model using simulated data and tested on real building data. The results indicate the model trained on simulated data is not generalized to be applied for real building data for fault detection. Kolmogorov-Smirnov Test is conducted to confirm that there exist statistical differences between the simulated and real building data and identify a subset of features with similarities between the two datasets. Using the subset of the feature, cross-dataset experiments show fault detection improvements on most fault cases. We conclude that even if the system produces simulated data with the same fault symptoms from physical analysis perspectives, not all features from simulated datasets may not be beneficial for AFDD but only a subset of features contains valuable information from a machine learning perspective.

1. Introduction

Buildings are complex and integrated systems consisting of multiple sensors, subsystems, and automatically controlled components. According to the United Nations Environment Programme, 36% of global energy use and 39% of energy-related carbon dioxide emission is attributed to building systems [1]. And, 30% of building energy usage is wasted [2] due to malfunctioning control, operation, and building equipment [3,4]. It is estimated additional energy consumption caused by some key building faults is anywhere between 0.37 and 17.96 EJ each year in the U.S. [5]. One viable solution for an energy-efficient building system is automatic fault detection and diagnosis (AFDD) [5]. From building design to the retrofit and commissioning process, understanding the reliability of a building and its energy faults is critical. Faults that degrade the performance of the entire building should be detected, diagnosed, and rectified, while in practice, significant follow-up and technical assistance to correct faults are required once detected and diagnosed. Over the past decades, many AFDD methods have been developed for component level and whole building level. Katipamula, Brambly, and Kim provide a comprehensive review and classification for methods used for Heating, Ventilation, and Air Conditioning (HVAC) system AFDD [6–8]. Generally speaking, there are two groups of AFDD methods: qualitative and quantitative model-based methods (such as rule-based and physics-model based); and process history-based methods (mostly various data-driven and machine learning-based methods).

* Corresponding authors.
E-mail address: hs.yoon@yonsei.ac.kr (H. Yoon).

https://doi.org/10.1016/j.enbuild.2022.111872
0378-7788/

Qualitative and quantitative models are easy to understand and are popular among building engineers and researchers. However, the issues are the high development cost, low scalability due to their needs to be customized for each specific building/project (such as the associated physics-based models, rules, and thresholds). As a result, the market adoption rate has been low [9]. Process history-based methods have therefore received great attention in recent years for their good scalability and low implementation cost. However, the performance of a process history-based method heavily relies on the data that the method is trained from, and it is recognized that the quality of the training data strongly affects the performance of process history-based AFDD tools [10].

Literature-reported AFDD methods are mostly developed and evaluated using simulated system data [11,12]. This is due to the difficulties of obtaining and analyzing real building data. Implementing faults and obtaining data that contain fault impacts in real buildings are already challenging. Cleaning and analyzing real building data to obtain "ground truth" is even more arduous since unexpected naturally occurred faults could exist in the system and cause abnormalities or complicate (sometimes even eliminate) the fault impacts expected from the artificially implemented faults.

Strategies that are only tested using simulated data might experience difficulties when applied to real buildings due to the two following issues: 1) data quality issues with typical building automation system data (missing data, noise, sensor faults, sensor accuracy, etc.), and 2) inherited differences between simulated data and real data, in terms of fault symptoms and data characteristics. After all, no model is perfect to represent reality completely. These challenges exist for all AFDD strategies but are more significant for process history-based (data-driven) methods since their performance depends on the data quality. To the best of our knowledge, the literature only has a few discussions on how to evaluate the accuracy of simulated fault data and nearly no discussion on how to understand the impact of simulation data's quality on the data-driven AFDD method's performance and scalability.

This study is therefore designed to examine how training data obtained from simulation might affect a data-driven AFDD strategy's performance, in terms of accuracy, false alarm rate, etc., when the developed strategy is used to analyze real building data. Two datasets generated from an American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) project (ASHRAE 1312 [13]) are used. The two datasets include those from a real building and those from the real building's digital twin, i.e., simulation models representing the real building. The performance of the fault detection strategy on each individual dataset and cross-dataset is then examined.

The paper is structured as follows. First, the two datasets are introduced in section 2. In section 3, the data-driven fault detection strategy used in this study is explained. The performances of the fault detection strategies trained using real building data are compared with those using simulated building data. In section 4, the generalization of the fault detection strategies is evaluated by a cross-datasets study, that is, fault detection strategy trained by simulated building data is tested by real building data. The generalization study also investigates the degree of similarity between the two datasets to assess the gaps of fault detection strategies using simulated vs. real building data. Section 5 provides conclusions and future research directions based on this study.

2. Datasets Description

The two datasets used in this study were generated from the ASHRAE 1312 research project [13,14]. These data were collected from a laboratory building that was set up like a small office building. The office building layout is shown in Fig. 1. The building consisted of two variable air volume (VAV) air handling unit (AHU) HVAC systems, each of which served 4 different rooms. The design of the test facility was intended to have each AHU serving room with nearly identical loads. As can be observed, each HVAC system served rooms facing east, west, south, and one interior room.

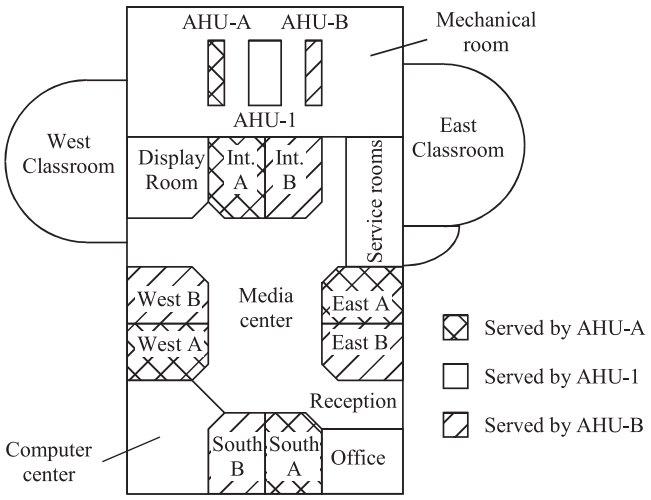


Fig. 1. ERS experimental setup.

While the two systems (A and B) would not generate the same data, the performance was found to be very similar under all operating conditions. During the study, System B (AHU-B and all B rooms) was continuously operated in a fault-free state, while System A (AHU-A and all A rooms) was artificially injected with various commonly occurring faults.

In the same project, dynamic behaviors of the HVAC systems and the four building zones that were served by the AHU, and four VAV boxes, were modeled using HVACSIM + software [15]. The model (called the 1312 model hereafter) was systematically validated using the real building data collected from carefully designed tests to ensure that the 1312 model simulated the dynamic behavior of the test facility for both fault free and faulty operation under three seasons (winter, summer, and spring). It was concluded in the ASHRAE 1312 project that the fault models were able to replicate all major fault symptoms although detailed dynamics between simulated data and real building measured data did not always overlap. This is because the simulation models were physics-based leading to the data generated may exhibit some variations from the real measured data. For example, these simulations most time don't consider the behaviors associated with the latency associated with sensor and control systems.

In this study, two datasets, i.e., real building dataset and their corresponding simulated dataset, are selected from the ASHRAE 1312 project's 2007 summer tests. Each dataset includes fault test data generated from the A system and fault free data from the B system. During a fault test, a fault was artificially implemented into the system for 12 h from 6:00 am to 6:00 pm. There are 16 types of fault tests used in this study as described in Table 1. The data sampling rate for both real and simulated datasets is one minute. ASHRAE 1312 project provides 24 measurements (also referred to as features in later sections) as summarized in Table 2. For this study, Outdoor Air Temperature (OA-TEMP) and (OA-HUMD) are not included since they represent weather conditions, not building/system conditions. Moreover, the simulation testbed does not simulate humidity variations. Therefore, two humidity-related features, i.e., Supply Air Humidity (SA-HUMD) and Return Air Humidity (RA-HUMD) are not included. As a result, there are 20 features considered in this study.

Table 1
Implemented 16 AHU faults during a summer period.

Category	Fault Name
Equipment	AHU Duct Leaking Fault - After Supply Fan
	AHU Duct Leaking Fault - Before Supply Fan
	Return Fan Complete Failure
Controlled Device	Heating Coil Valve Leaking - Stage 1 (0.4GPM)
	Heating Coil Valve Leaking - Stage 2 (1.0GPM)
	Heating Coil Valve Leaking - Stage 3 (2.0GPM)
	Cooling Coil Valve Stuck Fully Closed
	Cooling Coil Valve Stuck Fully Open
	Cooling Coil Valve Stuck 15% Open
	Cooling Coil Valve Stuck 65% Open
	OA Damper Stuck - Fully Closed
	OA Damper Leaking – 45% Open
	OA Damper Leaking – 55% Open
Controller	Cooling Coil Valve Control Unstable
	Cooling Coil Valve Reverse Action
	Return Fan at 30% SPD

Table 2
Description of the 20 Features used in this study.

Category	Features	Abbreviation
Temperature	Supply Air Temperature	SA-TEMP
	Mixed Air Temperature	MA-TEMP
	Return Air Temperature	RA-TEMP
	Heating Coil Discharge Air Temperature	HWC-DAT
	Cooling Coil Discharge Air Temperature	CHWC-DAT
Position	Exhaust Air Damper Position	EA-DMPR
	Return Air Damper Position	RA-DMPR
	Outdoor Air Damper Position	OA-DMPR
	Heating Coil Valve Position	HWC-VLV
	Cooling Coil Valve Position	CHWC-VLV
Pressure	Supply Fan Differential Pressure	SF-DP
	Return Fan Differential Pressure	RF-DP
	Supply Air Static Pressure	SA-SP
Airflow Rate	Supply Airflow Rate	SA-CFM
	Return Airflow Rate	RA-CFM
	Outdoor Airflow Rate	OA-CFM
Fan Speed	Supply Fan Speed	SF-SPD
	Return Fan Speed	RF-SPD
Fan Power	Supply Fan Power	SF-WAT
	Return Fan Power	RF-WAT

Each of the two datasets includes the same test days. Since the measurement sampling rate in the test facility is one minute, for each test day, each feature has 1,440 samples. Considering that a fault is implemented from 6:00 am to 6:00 pm on a test day, each test day contains 720 data points representing faulty operation and 720 data points representing fault-free operation. More detailed descriptions about fault testing facilities and faulty operational conditions can be referred to ASHRAE 1312-RP project reports [13].

3. Automated fault detection Strategy: Simulation vs. Real building data

The main objective of this study is to understand whether there exist performance differences between a AFDD strategy that is developed using simulated building data vs. using real building data. The two building datasets as described in Section 2 are used as the simulated and real building data. To prevent the comparison conclusion only applicable to a specific AFDD strategy, we have designed a series of experiments using different strategies. Considering the complexity of fault diagnosis strategy, in this study, only fault detection strategies are evaluated.

When developing a data-driven fault detection strategy, there are typically three steps [16]: In step 1, a data-driven baseline model that represents fault free status is developed. In step 2, incoming snapshot data from the building is compared with the baseline data, this is often done by comparing the reduced-order snapshot model with the baseline model. Step 3 is to flag the system status to be faulty or fault free based on the differences between the snapshot data and baseline data. Hence two important aspects affecting a data-driven fault detection strategy are: 1) data-driven model (typically based on machine learning strategies), and 2) baseline data and training, which include sample sizes and cross-validation schemes.

As an initial attempt in exploring AFDD strategies on different datasets, we choose to design the experiments in a sequential manner. That is, in Experiment I, we investigate the fault detection strategy performance using different machine learning models with a set of training sample sizes and cross-validation. With the machine learning model being chosen from the first experiment, we then set the cross-validation and investigate the impact of the training sample size on fault detection strategy performance. In the last experiment, we set the machine learning model and training sample size to investigate the impact of cross-validation on fault detection strategies. By systematically varying the above-mentioned fault detection strategy characteristics for both simulated and real building data, we attempt to examine whether the conclusions are scalable to different data-driven fault detection strategies.

Notice that in this section, for all experiments, both training and testing data are from the same dataset. In another word, if the fault detection strategy is trained using the simulated dataset, then it is tested using the simulated dataset. A cross-dataset comparison is performed in Section 4.

3.1. Automated fault detection Data-Driven models

Extensive efforts have been dedicated to investigating machine learning models for building AFDD including random forest (RF) [17–21], support vector machine (SVM) [21–26], decision tree (DT) [21,26–28], K-nearest neighbors [21], neural networks [29,30]. In this research, RF and SVM are of interest because RF shows a strong ability to deal with flexible and overlapping decision boundaries, tolerate noisy data [21], and improve classification performance by reducing overfitting on decision trees [31]; SVM has shown great power for solving nonlinear and high-dimension classification problems [32].

Random forest (RF) is an ensemble learning algorithm proposed by Breiman [33] for classification and regression [34,35]. RF consists of an ensemble of M trees, $T_i\{f\}, i = 1, \cdots, M$, where $F = \{f_1, \cdots, f_n\}$ is a set of n features from the data (e.g., room temperature, sensor position, fan speed). The ensemble of M trees generates M predictions $Y_1, \cdots, Y_M$, where Y_i is the predicted value by i^{th} tree $T_i\{F\}, i = 1, \cdots, M$ (e.g., in fault detection,Y_i = 1 indicating faulty condition being detected while Y_i = 0 being fault-free). The final prediction $\widehat{Y}$ is obtained by taking the majority results from the M trees. For example, let$M$ = 10, if a data is predicted to be faulty by 6 out of 10 trees, the data is predicted to be in a faulty condition.

Support vector machine (SVM) is proposed by Boser, Guyon, and Vapnik [36]. In fault detection, there are only faulty and fault-free cases, the problem is constructed as a binary SVM classification. Let (F_i, y_i) describe a datapoint i, $i = 1, \cdots, m$, where F_i representing the set of its features, y_i representing its label, y_i= 1 if it is faulty andy_i = −1 if fault-free. The SVM is solved as an optimization problem:

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i \tag{1}$$

$$s.t. y_i(\boldsymbol{w}^T F_i + b) \geq 1 - \xi_i \tag{2}$$

$$\xi_i \geq 0, i = 1, \cdots, m \tag{3}$$

where ξ_i is a loss function, and C is a penalty parameter on the training error. A commonly used loss function for the SVM is the hinge function:

$$\max\left(0, 1 - y_i\left(\boldsymbol{w}^T \phi(\boldsymbol{F_i}) + b\right)\right)^2 \tag{4}$$

where ϕ is a function mapping training data to higher dimensional spaces. A linear kernel function is commonly used in the SVM training procedure. That is, given kernel function $K(\boldsymbol{F_i}, \boldsymbol{F_j}) = \phi(\boldsymbol{F_i})^T \phi(\boldsymbol{F_j}), let \phi(\boldsymbol{F_i}) = \boldsymbol{F_i}$, $K(\boldsymbol{F_i}, \boldsymbol{F_j}) = \boldsymbol{F_i}^T \boldsymbol{F_j}$. For the testing, the result is predicted by:

$$\widehat{y}_i = sgn\left(\boldsymbol{w}^T \phi(\boldsymbol{F_i}) + b\right) \tag{5}$$

3.2. Evaluation metrics

The performance metrics used for fault detection strategy evaluation are accuracy, sensitivity, specificity [37], and F1-score [38]. Accuracy measures a strategy's ability to identify both abnormal and normal samples correctly. Sensitivity, known as the true positive rate (TPR), measures a strategy's ability to identify abnormal samples. Specificity, known as the true negative rate (TNR), measures a strategy's ability to identify normal samples. F1-score, alternative to accuracy, measures a strategy's ability to identify both abnormal and normal samples correctly. Accuracy, TPR, TNR and F1-score are calculated by equations as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{6}$$

$$\text{Sensitivity (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{Specificity (TNR)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{8}$$

$$\text{F1-score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} \tag{9}$$

In these equations, true positive (TP) denotes the number of abnormal samples correctly identified, the false negative (FN) denotes the number of abnormal samples incorrectly identified as normal; true negative (TN) denotes the number of normal samples correctly identified as normal, while false positive (FP) denotes the number of normal samples incorrectly identified as anomalous.

3.3. Fault detection strategy evaluation experiments

To comprehensively assess the fault detection data-driven strategies (see Table 3), for both real and simulated datasets, three experiments are conducted to investigate the impact of different (1) fault detection data-driven models; (2) sizes of training data; (3) cross-validation (CV) folds. Because 80% size of training data and 10-fold CV are commonly adopted in statistical classification problems [39], three experiments are conducted as follows. In Experiment I, we focus on the performances of two fault detection models - RF and SVM (linear kernel) under a fixed 80% size of training data and a 10-fold CV. In the second experiment, we evaluate the performances of different sizes of training data (90%, 80%, 70%, 60%, 50%, 40%, 30% and 20%) by choosing RF (supported by the literature) and a 10-fold CV. In the third experiment, we compare the performances of different cross validation folds (3-fold, 5-fold, and 10-fold) under RF and an 80% size of training data. Again, these variations are designed to examine if the conclusions are scalable to different AFDD strategies.

Table 3
AFDD Data-Driven Strategy.

Design parameters	
AFDD data-driven models	Random forest (RF), Support vector machine (SVM)
Size of training data	90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%
Cross-validation folds	3-fold, 5-fold, 10-fold

3.3.1. Experiment I: Data-Driven models evaluation

The objective of this experiment is to evaluate whether two different AFDD models, RF and SVM, result in performances of differences between fault detection strategy trained by a simulated building data and that by a real building data. Here the training data size is fixed at 80% and the number of CV folds is fixed at 10.

Results under RF are firstly compared between training using real and simulated building data (see Fig. 2A). It is observed that there is no significant difference on fault detection performances, using three metrics defined in Section 3.2, between the fault detection strategy trained using real data and that using simulated data, as long as that they are also tested using the same category of data, i.e., simulated or real. Among all fault tests, one fault case, i.e., *Cooling Coil Valve Control Unstable case*, shows slightly different performance. For this case, the accuracy is 0.99 with real data, and 0.91 with simulated data. Additionally, the sensitivity is 0.99 with real data and 0.86 with simulated data. The specificity is 0.99 and 0.95, respectively.

Similar conclusions are drawn for SVM (see Fig. 2B). That is, there is no significant difference on fault detection performances between the fault detection strategy trained using real data and using simulated data for all fault tests except for *Cooling Coil Valve Control Unstable case*, whose accuracy is 1.00 with real data, and 0.77 with simulated data; sensitivity 0.99 with real data, and 0.68 with simulated data; specificity 1.00 with real data, and 0.85 with simulated data; F1-score 1.00 with real data, and 0.75 with simulated data.

We conclude under the fixed size of training data (80%) and the number of cross validation folds (10-fold CV), SVM and RF using simulated vs. real building data reach similar performance for fault detection except for *Cooling Coil Valve Control Unstable.* Since RF has been adopted in many AFDD literatures [17–21] due to its advantage of making no assumptions on data distribution [33], we choose to use RF as the AFDD model in the following experiments. RF, as an ensemble of decision trees, can be visually presented to demonstrate the decision logics. Here for illustration, we applied the RF method on real building data from the fault test case AHU Duct Leaking Fault - Before Supply Fan, as shown in Fig. 3

3.3.2. Experiment II: Training data size evaluation

In Experiment II, different sizes of training data (20% to 90% with 10% increment) incorporating RF with 10-fold CV are compared using real and simulated building data. It is observed the experiments using training data of 20% to 60% have similar performance as those from 70% to 90%. Since it is common practice in machine learning studies that more data is used in training with remaining is reserved for testing, the results from 70%, 80% and 90% are presented in Fig. 4 for illustration. As seen, there is no significant difference in the performances under different sizes of training data for all cases, but under each size of training data, there is a significant difference of performances between using real and simulated building data for *Cooling Coil Valve Control Unstable*.

We conclude that under fixed AFDD models (RF in this experiment) and 10-fold CV, using real or simulated building data can

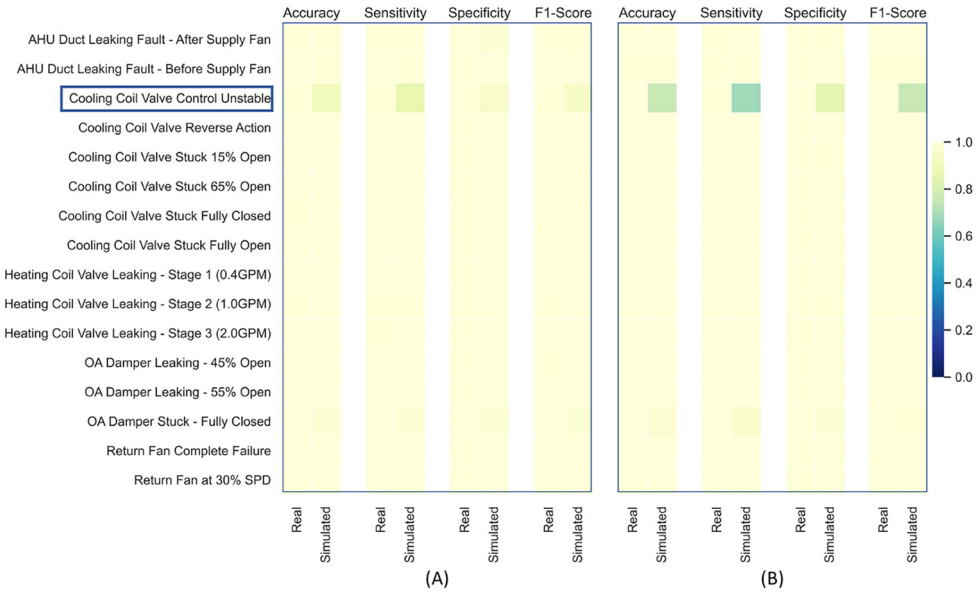


Fig. 2. Performances comparisons between random forest (Fig. 2A) and SVM (Fig. 2B) with 80% training size and 10-fold CV. "Real" means real building data and "Simulated" means simulated building data.

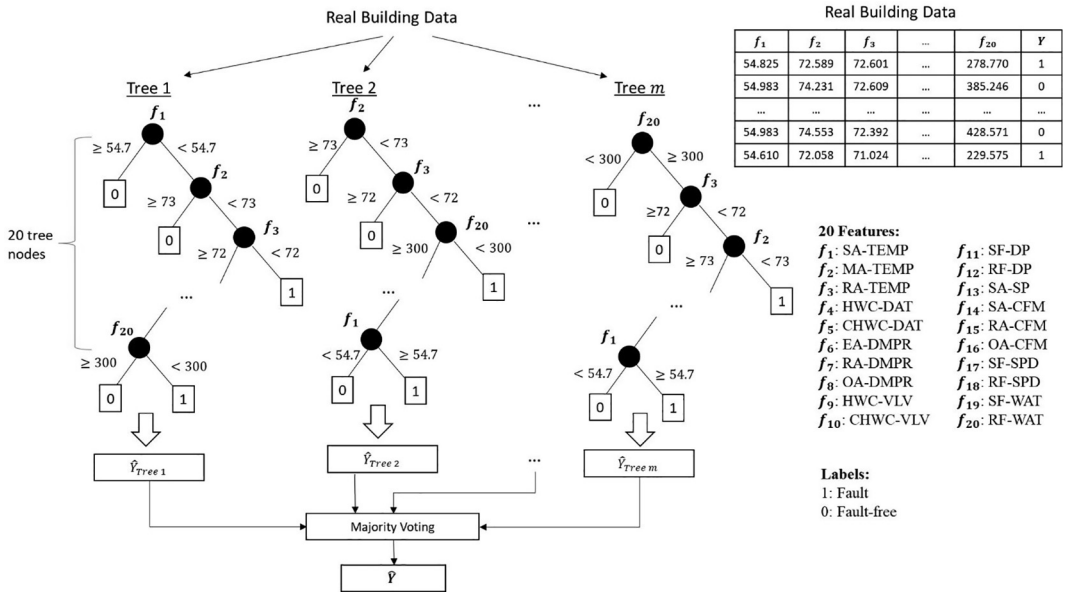


Fig. 3. Random forest for fault detection using real building data of the fault test case *AHU Duct Leaking Fault - Before Supply Fan*. There are m different decision trees in the random forest, and each decision tree contains 20 nodes, each node representing one feature. Each decision tree provides a prediction, and the final prediction of the random forest is determined by the majority voting.

reach comparable performances for all 16 faults with different training sizes.

3.3.3. Experiment III Results: CV folds evaluation

Experiment III compares different CV folds (3-fold, 5-fold, and 10-fold) under RF and with an 80% training size. Similar to the results from Experiments I and II (Fig. 5), under different CV folds, the performances of fault detection trained by real or simulated data are similar. Equivalently, under each CV fold, there is no significant difference between using real and simulated building data for all fault tests except for *Cooling Coil Valve Control Unstable*. We conclude that under the fixed AFDD model and size of training data, fault detection performances using real or simulated building data do not change with respect to the number of CV folds.

3.4. Conclusion on AFDD strategy evaluation

In summary, we comprehensively evaluate AFDD data-driven strategies trained and tested on real building data vs. those trained and tested on simulated building data. We observe when the training data and blind test data are from the same data sources, although the training and test data are not overlapped, the fault detection performance is similar between using real building data and using simulated building data. This conclusion is not affected by fault detection strategies, sizes of training data, and the number of cross-validation folds. However, under any circumstances described above, there are some gaps of fault detection performances between using real and simulated building data for the fault case *Cooling Coil Valve Control Unstable*. This case is the only

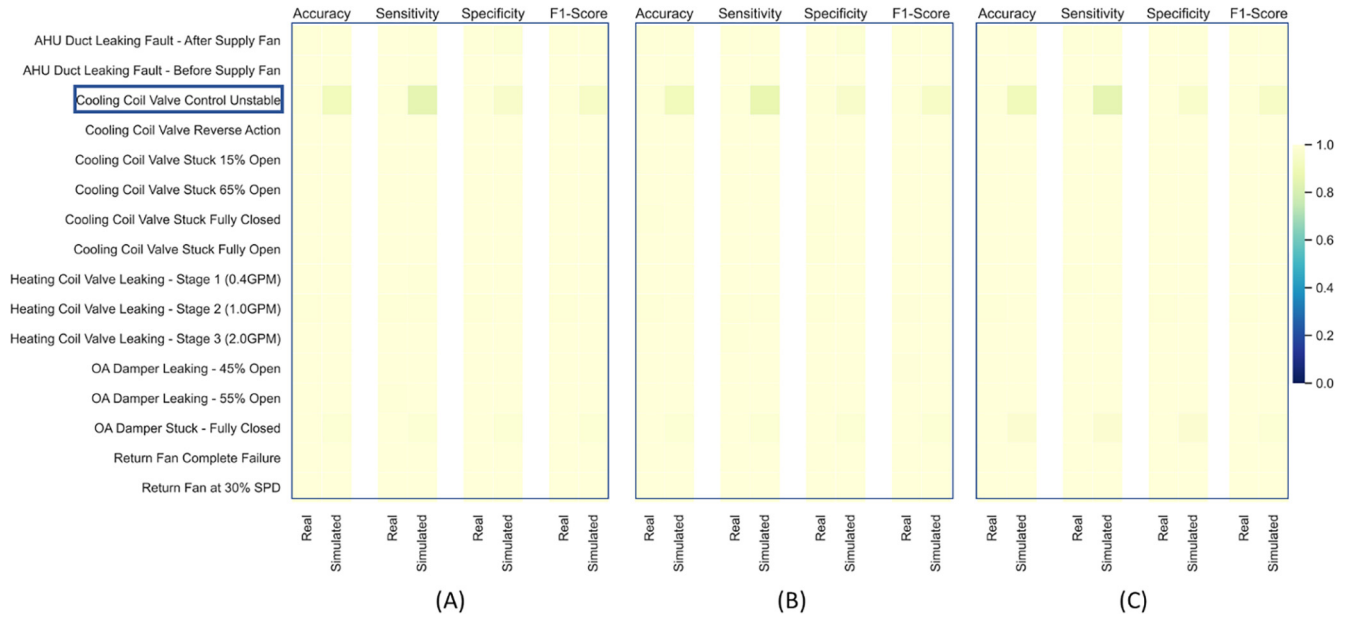


Fig. 4. Performances comparisons among different training sizes of 70% (Fig. 4A), 80% (Fig. 4B), 90% (Fig. 4C) using real and simulated building data under random forest and 10-fold CV. "Real" means real building data and "Simulated" means simulated building data.

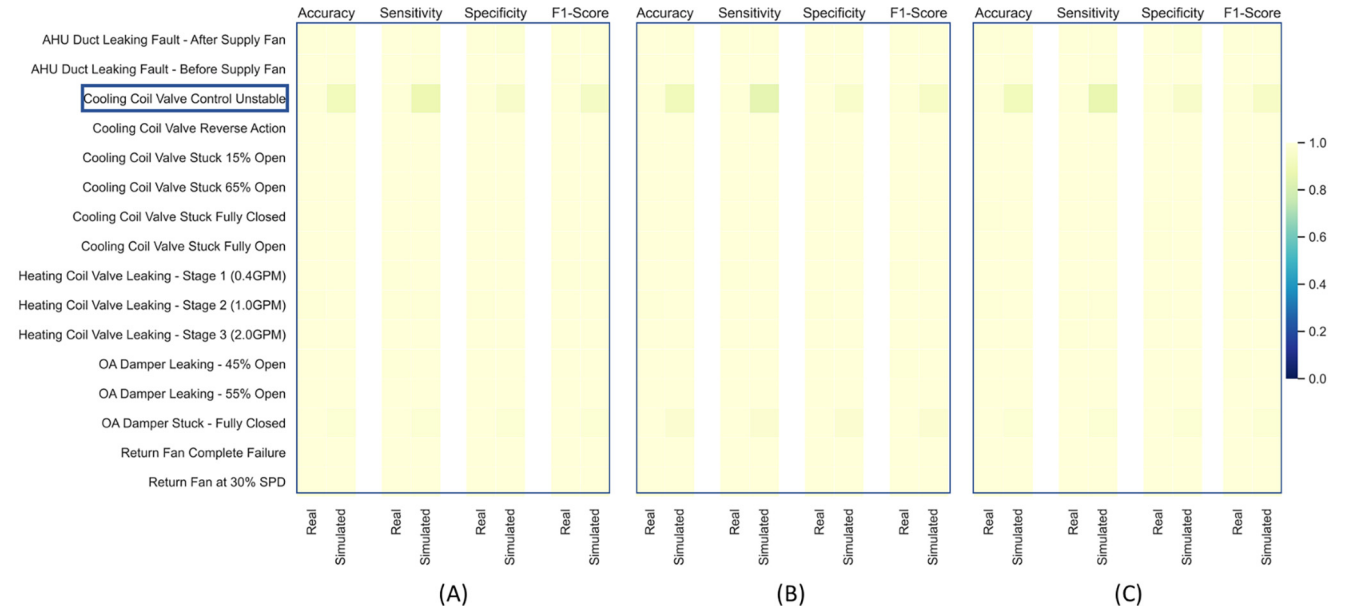


Fig. 5. Performances comparisons among a 3-fold CV (Fig. 5A), 5-fold CV (Fig. 5B), and 10-fold CV (Fig. 5C) using real and simulated building data under random forest and 80% training size. "Real" means real building data and "Simulated" means simulated building data.

control-related fault case tested in this study. Fig. 5 illustrates the original data, and it is observed that the simulated fault symptom for this case, i.e., unstable control valve position, has a different oscillation frequency from the real symptom. Notice that valve oscillation frequency is the main fault symptom of unstable control. Hence, such frequency differences are not typically observed in other fault cases. In other fault cases, although simulated data could be different from real data, the variables often follow similar frequency and differences are typically time-varying biases. We contend that this could be a reason that causes the fault detection strategies' performance differences for this fault case. One might also wonder why the cooling coil valve positions from simulated and experimental data vary by about 10%. The simulation model accuracy for cooling coil valve position was not directly reported by the 1312 research project [13,14]. However, simulated valve position accuracy is highly related to coil and zone energy modeling accuracies, which were reported to be about 10%. Considering that energy meters generally have an accuracy of +/- 5%, it was considered acceptable for energy indices to be within 10% accuracy. Such 10% accuracy is also reflected in Fig. 6, when comparing simulated cooling coil valve position vs. the real values.

Please note in this section, the experiments are conducted on real and simulated building data independently. It is interesting to explore the generalization of the AFDD strategies. In the next section, we conduct cross-dataset studies using RF, an 80% training size, and a 10-fold CV as the AFDD strategy.

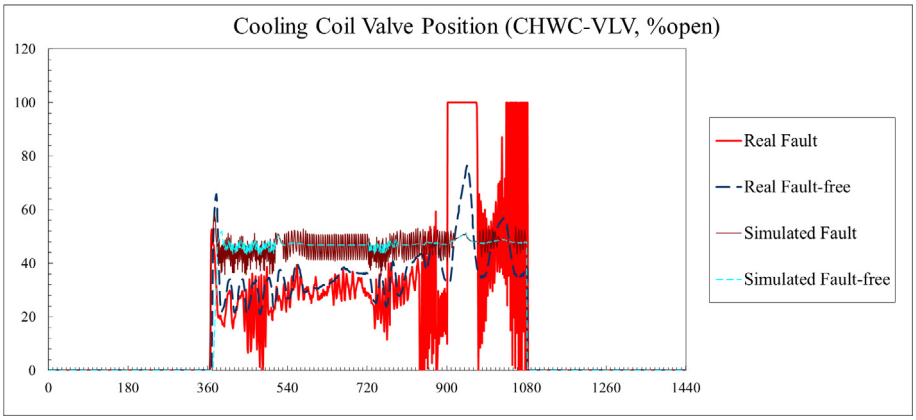


Fig. 6. Symptom plot for the fault case *Cooling Coil Valve Control Unstable*.[1–43].

4. Automated fault detection and diagnosis (AFDD) Strategy: Generalization evaluation

4.1. Cross-dataset AFDD strategy evaluation

In the cross-dataset study, we train the fault detection strategy using simulated data and test it on the real building data. Given the interest of the study is fault detection, we focus the discussions on overall accuracy and sensitivity. As seen in Table 4, out of 16 faults, only 2 fault cases (*Return Fan Complete Failure, OA Damper Stuck Fully Closed*) achieve over 0.90 accuracy and 0.85 sensitivity, and 3 fault cases (*Cooling Coil Valve Stuck Fully Closed, Heating Coil Leaking - Stage 2 (1.0 GPM), Return Fan at 30% SPD*) achieve over 0.85 accuracy and 0.75 sensitivity. The remaining 11 cases were performed with accuracy or sensitivity below 0.5. Thus, in most fault cases, fault detection strategy trained by simulated building data and tested on real building data cannot reach comparable fault detection performances to those of models trained and tested by simulated building data.

We conclude the RF strategy shows a degradation of performance when compared to its performance in Section 3, where both training and test data are from the same source. It is hence of interest to investigate further as to what the cause is for this performance degradation. A hypothesis is that, although the simulation model is validated from a physics perspective, i.e., the absolute values of key measurements and the fault symptoms are similar to those from a real building, the features generated from a simulation model, from a data science perspective (e.g., distribution), differ from those in the real building dataset. We conduct a statistical test in the next section to evaluate the differences.

4.2. Cross-dataset feature evaluation

We conduct Kolmogorov-Smirnov Test (KS test) [40], a non-parametric statistical test to determine how significantly different real dataset vs. simulated datasets is for fault detection. KS test is widely used since it does not need assumptions on the data distribution. The KS two-sample test hypothesis is defined as:

H_0 : Two samples collected from different sources /datasets come from the same distribution.

H_a : Two samples collected from different sources /datasets do not come from the same distribution.

The test statistic is defined as $D = |E_1(i) - E_2(i)|$, where E_1 and E_2 are the empirical functions for the two samples.

Considering fault detection, in either simulated or the real building dataset, we have faulty and fault-free conditions. We conduct two KS tests on the 20 features (see Table 2): the first KS test is on a simulated dataset to identify a subset of the features that mostly differ in the faulty vs. fault-free condition; (2) given the subset features, the second KS test is to identify the similar features in comparing simulated data (fault and fault-free combined)

Table 4
Random Forest trained by simulated data and tested by real data (*: faults being detected with > 0.90 accuracy and > 0.85 sensitivity. **: faults being detected with > 0.85 accuracy and > 0.75 sensitivity).

Fault Type	Accuracy	Sensitivity	Specificity	F1-Score
AHU Duct Leaking Fault - After Supply Fan	0.51	0.79	0.22	0.62
AHU Duct Leaking Fault - Before Supply Fan	0.50	0.01	0.99	0.03
Return Fan Complete Failure*	**0.94**	**0.88**	**1.00**	**0.93**
Heating Coil Valve Leaking - Stage 1 (0.4 GPM)	0.50	0.00	1.00	0.01
Heating Coil Valve Leaking - Stage 2 (1.0 GPM)*	***0.89***	***0.79***	***1.00***	***0.88***
Heating Coil Valve Leaking - Stage 3 (2.0GPM)	0.50	0.00	1.00	0.00
Cooling Coil Valve Stuck Fully Closed*	***0.92***	***0.83***	***1.00***	***0.91***
Cooling Coil Valve Stuck Fully Open	0.50	0.00	1.00	0.00
Cooling Coil Valve Stuck 15% Open	0.54	0.33	0.76	0.35
Cooling Coil Valve Stuck 65% Open	0.50	0.00	1.00	0.00
OA Damper Stuck - Fully Closed*	**0.96**	**1.00**	**0.93**	**0.97**
OA Damper Leaking – 45% Open	0.50	0.00	1.00	0.00
OA Damper Leaking – 55% Open	0.50	0.00	1.00	0.00
Cooling Coil Valve Control Unstable	0.49	0.77	0.21	0.60
Cooling Coil Valve Reverse Action	0.62	0.24	1.00	0.36
Return Fan at 30% SPD*	***0.90***	***0.81***	***1.00***	***0.89***

vs. real data (fault and fault-free combined). For both KS tests, 0.4 is used as a cut-off threshold here as in [41].

The final subset features selected by the two KS tests for each fault test are summarized in Table 5. It is interesting to observe there are no common features selected for *Cooling Coil Valve Control Unstable* (see Section 3.4 for discussion). We contend that for an unstable control, the fault indicator is typically the frequency of a control device position change. That is, the control device oscillates with a much higher frequency than that from the baseline. Such feature may need wavelet-based approaches instead statistical KS test to be selected. It is also observed fewer numbers of features are selected from the 5 *Position* category compared to other categories. Next, we use the selected subset features for the cross-datasets studies again to explore whether these features can help to improve fault detection strategy performances. Please note this is not to develop a better fault detection strategy, instead, this is to understand how training data affects a fault detection method's performance (accuracy and scalability). By examining more closely the subset features for the cross-dataset studies, we attempt to see if the accuracy of certain subset of a training dataset would affect the developed fault detection strategy.

4.3. Cross-dataset AFDD strategy revisit: Using selected subset features

The subset features identified by the two KS tests for faulty data in Table 5 are used by the random forest strategy. Since there are no subset features identified by the KS test for the case *Cooling Coil Valve Control Unstable*, meaning that features from simulated datasets cannot capture any characteristics of those from real datasets, this case is excluded for the comparison study. As a result, we are interested in investigating the 10 cases (see Table 4) with accuracy or sensitivity below 0.75 when the full feature set was used. The RF strategy is trained using simulated building data on the subset features with respect to each fault test. The model is then tested on the real building data. The results show that the fault detection performance improved for 5 out of the 10 fault tests. These include (1) *AHU Duct Leaking Fault - Before Supply Fan,* (2) *Heating Coil Valve Leaking - Stage 1 (0.4GPM),* (3) *Heating Coil Valve Leaking - Stage 3 (2.0GPM),* (4) *Cooling Coil Valve Stuck 15% Open,* (5) *Cooling Coil Valve Stuck 65% Open* (see Fig. 7). We conclude the RF strategy using the selected features identified by KS tests may improve the accuracy and sensitivity for some fault cases. Fig. 8 illustrates the example of RF to detect the fault on real building data from the fault test case *AHU Duct Leaking Fault - Before Supply Fan* using 11 selected features.

In summary, cross-dataset study between real and simulated building data indicates that simulated building data differ from real ones in terms of statistical learning, although they are validated to be similar by real building data from a physical perspective. KS test assists in identifying similar features between real and simulated building data, which indirectly indicates that simulated building data are not always similar to real ones because a fraction of similar features is less than 30%. However, AFDD strategy incorporating these identified features show promises to improve the sensitivity for some fault cases, meaning that these features are an important component in true fault instances. Specifically, for the 3 fault cases from the *Equipment* category, cross-dataset experiment on the full feature set was able to detect 1 fault, and cross-dataset experiment on the selected feature set was about to detect the additional 1 fault. For the 10 fault cases from *Controlled Device* category, cross-dataset experiment on the full feature set was able to detect 3 faults, and using the selected feature sets, additional 4 fault cases were able to be detected. For the 3 fault cases from *Controller* category, KS selected features have no improvements for detection. This cross-dataset study raises a warning for data-driven AFDD strategy development using simulated fault data. Clearly, under the statistical learning lens, simulated data often contain different information (e.g. distribution difference of Heating Coil Discharge Air Temperature in *Return Fan at 30% SPD*) from real building data. Different learning strategies, such as transfer learning, may need to be explored for this purpose.

5. Conclusions and future directions

The purpose of this research is to answer two questions in AFDD development for building systems: (1) do simulated or real datasets affect AFDD strategy performance, and (2) to what degree of similarity is between simulated and real building datasets from a machine learning perspective. In the first place, we evaluate data-driven fault detection strategies on real and simulated building data, respectively, from which we observe that the fault detection performances are not affected by fault detection strategies, sizes of training data, and the number of cross-validation folds when training and blind test data come from the same data source.

Table 5
Feature Subsets by two KS tests for each Fault Test.

	Temperature					Position					Pressure			Airflow Rate			Fan speed		Fan Power	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
AHU Duct Leaking After Supply Fan					X					X	X	X		X	X	X	X	X	X	X
AHU Duct Leaking Before Supply Fan		X		X	X						X	X			X	X	X	X	X	X
Return Fan Complete Failure											X		X		X	X	X	X	X	
Heating Coil Leaking Stage 1 (0.4 GPM)		X		X	X					X	X	X	X		X	X	X	X	X	X
Heating Coil Leaking Stage 2 (1.0 GPM)				X													X	X		
Heating Coil Leaking Stage 3 (2.0 GPM)				X						X	X	X	X	X	X	X	X	X	X	X
Cooling Coil Valve Stuck Fully Closed	X	X	X	X	X								X							
Cooling Coil Valve Stuck Fully Open	X				X						X	X	X	X	X	X	X	X	X	X
Cooling Coil Valve Stuck 15% Open	X	X	X	X	X						X	X	X	X					X	
Cooling Coil Valve Stuck 65% Open	X		X								X	X	X	X	X				X	X
OA Damper Stuck - Fully Closed																				
OA Damper Leaking – 45% Open	X	X	X	X	X					X	X	X	X	X	X	X			X	X
OA Damper Leaking – 55% Open		X		X	X					X	X	X			X		X	X	X	X
Cooling Coil Valve Control Unstable																				
Cooling Coil Valve Reverse Action		X	X	X							X	X	X	X	X	X	X	X	X	
Return Fan at 30% SPD												X			X	X	X			X

*See Table 2 for details on the 20 features, 1: SA-TEMP, 2: MA-TEMP, 3: RA-TEMP, 4: HWC-DAT, 5: CHWC-DAT, 6: EA-DMPR, 7: RA-DMPR, 8: OA-DMPR, 9: HWC-VLV, 10: CHWC-VLV, 11: SF-DP, 12: RF-DP, 13: SA-SP, 14: SA-CFM, 15: RA-CFM, 16: OA-CFM, 17: SF-SPD, 18: RF-SPD, 19: SF-WAT, 20: RF-WAT

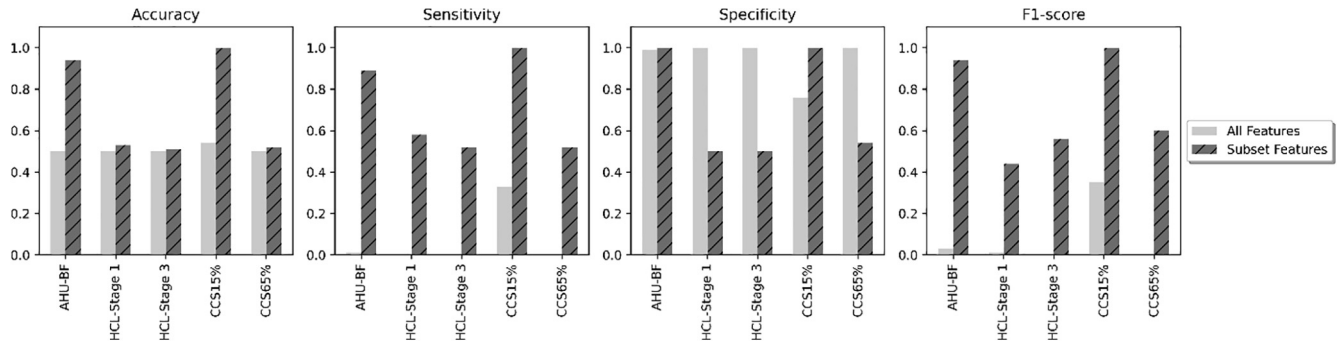
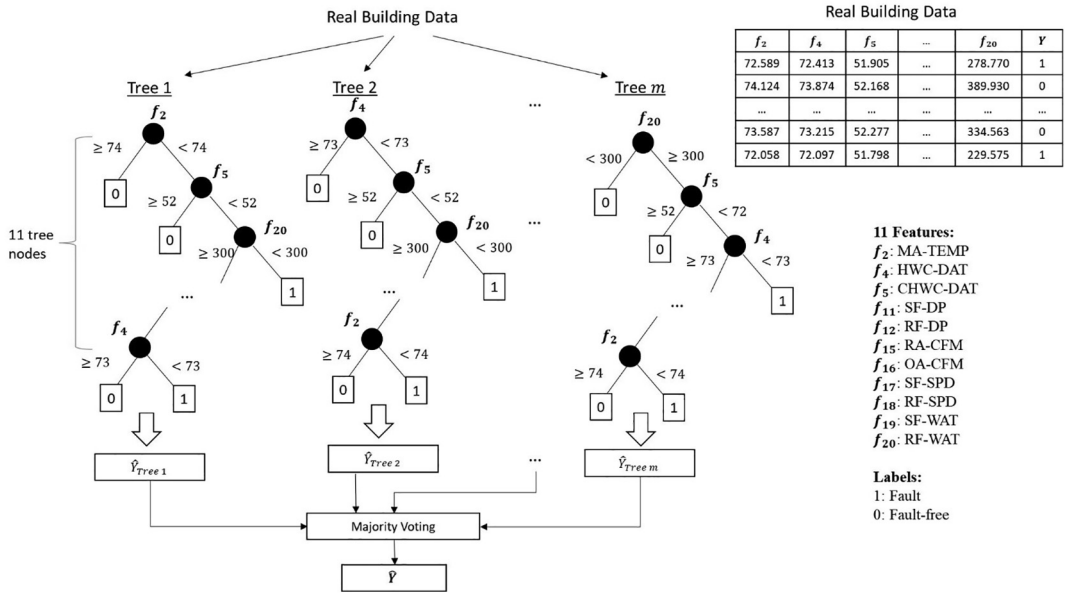


Fig. 7. Performances comparisons between using all features and sub features. (1) AHU-BF: AHU Duct Leaking Fault - Before Supply Fan, (2) HCL-Stage 1: Heating Coil Valve Leaking - Stage 1 (0.4GPM), (3) HCL-Stage 3: Heating Coil Valve Leaking - Stage 3 (2.0GPM), (4) CCS15%: Cooling Coil Valve Stuck 15% Open, (5) CCS65%: Cooling Coil Valve Stuck 65% Open.



f_2	f_4	f_5	...	f_{20}	Y
72.589	72.413	51.905	...	278.770	1
74.124	73.874	52.168	...	389.930	0
...	...	...	...	...	...
73.587	73.215	52.277	...	334.563	0
72.058	72.097	51.798	...	229.575	1

Fig. 8. Random forest for fault detection on real building data of the fault test case *AHU Duct Leaking Fault - Before Supply Fan* using 11 selected features. There are *m* different decision trees in the random forest, but each decision tree contains only 11 nodes, each node representing one feature. Similarly, each decision tree provides a prediction, and the final prediction of the random forest is determined by the majority voting.

In the second place, the cross-dataset study indicates the performance of the AFDD strategy developed from simulated data directly applied to the real building data is less than satisfactory. This indicates even though the simulation model was carefully developed and calibrated with a high degree of similarity with real building data in terms of physical perspective, the two datasets may not share the same statistical distributions from the data science perspective. With the help of the KS test, we confirm the distribution differences of the two datasets and identify similar features between simulated and real building data that could be used to improve detecting some but not all the fault cases. This cross-data study explains the poor performances often observed when applying AFDD strategies developed solely on simulated data in the field. Clearly, even if a simulated dataset produces the same fault symptoms from physical analysis perspectives, it contains different information from a machine learning perspective.

While this research does not propose new AFDD methods, this study is much needed to help understand the challenges and issues in using simulated data to support the development of data-driven fault detection method development. The impacts are multi-fold: it will help develop new transfer learning mechanism, for example, pre-train fault detection on simulation and fine-tune on the real building data; it will help construct and validate building baseline to support the fault detection, just to name a few. It is our intention to develop new methods based on the insights gathered from this research. Specifically, our interest in the future study may lie in answering two questions: (1) is it possible to develop a more robust AFDD strategy to resolve issues in the cross-dataset study since two sets of building data significantly differ from each other statistically while they are indeed validated to be similar from building domain knowledge? (2) is it likely to enhance the current simulation model to generate more reliable simulated building data for real building AFDD? We expect that the key to these two questions will provide great potentials for building AFDD. Additionally, it is noted that deep learning methods such as LSTM on temporal data has been extracted great attention for building fault detection and diagnosis. For example. Fan's [42] and Marino's works [43] investigate LSTM methods for building energy predictions respectively. We plan to incorporate deep learning methods in the comparison study as our immediate next step.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by funds from the National Science Foundation award under grant number IIP #1827757. The U.S. Government is authorized to reproduce and distribute for governmental purposes notwithstanding any copyright annotation of the work by the author(s). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

References

[1] International Energy Agency and the United Nations Environment Programme, Global Status Report: towards a zero-emission, efficient and resilient buildings and construction sector, International Energy Agency, Paris, France, 2018, p. 2018.
[2] L. Pérez-Lombard, J. Ortiz, C. Pout, A Review on Buildings Energy Consumption Information, Energy Build. 40 (3) (2008) 394–398.
[3] Energy Conservation in Buildings and Communities programme. Real time simulation of HVAC systems for building optimisation, fault detection and diagnostics. International Energy Agency, Paris, France, 1996, Report No.:IEA ECBCS Annex25.
[4] M.R. Brambley, S. Katipamula, Commercial Building Retuning, ASHRAE J. 51 (2009) 12–23.
[5] K.W. Roth, D. Westphalen, P. Llana, M. Feng, The energy impact of faults in us commercial buildings, in, in: I.N. West Lafayette (Ed.), Proceedings of International Refrigeration and Air Conditioning Conference, 2004, p. 665.
[6] 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 (1) (2005) 3–25.
[7] 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 (2) (2005) 169–187.
[8] W. Kim, S. Katipamula, A review of fault detection and diagnostics methods for building systems, Sci. Technol. Built Environ. 24 (1) (2018) 3–21.
[9] S. Frank, X. Jin, D. Studer, A. Farthing, Assessing barriers and research challenges for automated fault detection and diagnosis technology for small commercial buildings in the United States, Renew. Sust. Energ. Rev. 98 (2018) 489–499.
[10] N. Omri, Z. Al Masry, N. Mairot, S. Giampiccolo, N. Zerhouni, Towards an adapted PHM approach: Data quality requirements methodology for fault detection applications, Comput. Ind. 127 (2021) 103414.
[11] Y. Li, Z. O'Neill, A critical review of fault modeling of HVAC systems in buildings, Build. Simul. 11 (5) (2018) 953–975.
[12] Z. Shi, W. O'Brien, Development and implementation of automated fault detection and diagnostics for building systems: A review, Autom. Constr. 104 (2019) 215–229.
[13] J. Wen, S. Li, Tools for evaluating fault detection and diagnostic methods for air-handling units, ASHRAE Research Project (2011).
[14] S. Li, J. Wen, Description of fault test in Summer of 2007. Final report, ASHRAE Research Project 1312, 2007
[15] M. Galler, Users guide to the HVACSIM+ configuration tool, Technical note, National Institute of Standards and Technology, Gaithersburg MD, 2020, Report No.:2110.
[16] S. Li, J. Wen, Application of pattern matching method for detecting faults in Air Handling Unit system, Autom. Constr. 43 (2014) 49–58.
[17] A.A. Bhattacharya, D. Hong, D. Culler, J. Ortiz, K. Whitehouse, E. Wu, Automated metadata construction to support portable building applications, in: in: BuildSys 2015 – Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built, 2015, pp. 3–12.
[18] B. Balaji, C. Verma, B. Narayanaswamy, Y. Agarwal, Zodiac: organizing large deployment of sensors to create reusable applications for buildings, in: in: BuildSys 2015 – Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built, 2015, pp. 13–22.
[19] D. Hong, H. Wang, J. Ortiz, K. Whitehouse, in: The Building Adapter: Towards Quickly Applying Building Analytics at Scale, in: BuildSys 2015, 2015, pp. 123–132.
[20] T. Mulumba, A. Afshari, K.e. Yan, W. Shen, L.K. Norford, Robust model-based fault diagnosis for air handling units, Energy Build. 86 (2015) 698–707.
[21] J. Gao, M. Bergés, A large-scale evaluation of automated metadata inference approaches on sensors from air handling units, Adv. Eng. Inform. 37 (2018) 14–30.
[22] J. Liang, R. Du, Model-based fault detection and diagnosis of HVAC systems using support vector machine method, Int. J. Refrig. 30 (6) (2007) 1104–1114.
[23] H. Han, Z. Cao, B.o. Gu, N. Ren, PCA-SVM-based automated fault detection and diagnosis (AFDD) for vapor-compression refrigeration systems, HVAC&R Res. 16 (3) (2010) 295–313.
[24] K.-Y. Chen, L.-S. Chen, M.-C. Chen, C.-L. Lee, Using SVM based method for equipment fault detection in a thermal power plant, Comput. Ind. 62 (1) (2011) 42–50.
[25] K.e. Yan, W. Shen, T. Mulumba, A. Afshari, ARX model based fault detection and diagnosis for chillers using support vector machines, Energy Build. 81 (2014) 287–295.
[26] J. Gao, J. Ploennigs, M. Bergés, A data-driven meta-data inference framework for building automation systems, in: in: BuildSys 2015 – Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built, 2015, pp. 23–32.
[27] 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.
[28] 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.
[29] B.o. Fan, Z. Du, X. Jin, X. Yang, Y. Guo, A hybrid FDD strategy for local system of AHU based on artificial neural network and wavelet analysis, Build. Environ. 45 (12) (2010) 2698–2708.
[30] Y. Zhu, X. Jin, Z. Du, Fault diagnosis for sensors in air handling unit based on neural network pre-processed by wavelet and fractal, Energy Build. 44 (2012) 7–16.
[31] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[32] N. Cristianini, J.S. Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, U.K., 2000.
[33] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[34] 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.
[35] A. Kusiak, M. Li, F. Tang, Modeling and optimization of HVAC energy consumption, Appl. Energy 87 (10) (2010) 3092–3102.
[36] B. Boser, I. Guyon, V. Vapnik, A training algorithm for optimal margin classifiers, In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992, Pittsburgh, PA.
[37] A. Baratloo, M. Hosseini, A. Negida, G. El Ashal, Part 1: simple definition and calculation of accuracy, sensitivity and specificity, Emerg. 3 (2) (2015) 48–49.
[38] G. Hripcsak, A.S. Rothschild, Agreement, the F-measure, and reliability in information retrieval, J. Amer. Med. Informat. Assoc. Jamia 12 (3) (2005) 296–298.
[39] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, New York, 2013.
[40] F.J. Massey, The Kolmogorov-Smirnov test for goodness of fit, J. Am. Stat. Assoc. 46 (253) (1951) 68–78.
[41] D. Steinskog, D.B. Tjøtheim, N.G. Kvamstø, A cautionary note on the use of the Kolmogorov-Smirnov test for normality, Mon. Weather Rev. 135 (3) (2007) 1151–1157.
[42] C. Fan, J. Wang, W. Gang, S. Li, Assessment of deep recurrent neural network-based strategies for short-term building energy predictions, Appl. Energy 236 (2019) 700–710.
[43] D.L. Marino, K. Amarasinghe, M. Manic, Building energy load forecasting using deep neural networks, in: in: Proceedings of the IECON (Industrial Electronics Conference), 2016, pp. 7046–7051.