Sustainable Cities and Society 69 (2021) 102816

Contents lists available at ScienceDirect

Sustainable Cities and Society

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

Application of data-driven attack detection framework for secure operation in smart buildings

Mariam Elnour [a], Nader Meskin [a,*], Khaled Khan [b], Raj Jain [c]

[a] Department of Electrical Engineering, Qatar University, Doha, Qatar
[b] Department of Computer Science and Engineering, Qatar University, Doha, Qatar
[c] Department of Computer Science and Engineering, Washington University in St. Louis, USA

ARTICLE INFO

Keywords:
Building management system (BMS)
Smart buildings
Anomaly detection
Industrial control system (ICS)
HVAC systems
Isolation Forest (IF)
Convolutional Neural Network (CNN)

ABSTRACT

With the rapid advancement in the industrial control technologies and the increased deployment of the industrial Internet of Things (IoT) in the buildings sector, this work presents an analysis of the security of the Heating, Ventilation, and Air Conditioning (HVAC) system which is a major component of the Building Management System (BMS), has become critical. This paper presents a Transient System Simulation Tool (TRNSYS) model of a 12-zone HVAC system that allows assessing the cybersecurity aspect of HVAC systems. The thermal comfort model and the estimated total power usage are used to assess the magnitude of the malicious actions launched against the HVAC system. Simulation data are collected and used to develop and validate a semi-supervised, data-driven attack detection strategy using Isolation Forest (IF) for the system under study. Three schemes of the proposed approach are investigated, which are: using raw data, using Principal Component Analysis (PCA) for feature extraction, and using 1D Convolutional Neural Network (CNN)-based encoder for temporal feature extraction. The proposed approach is compared with standard machine-learning approaches, and it demonstrates a promising capability in attack detection for a range of attack scenarios with high reliability and low computational cost.

1. Introduction

The security of the recently evolving Cyber-Physical Systems (CPSs), such as smart cities and critical infrastructures, has gained increased attention from the research community in the past few years due to the rapid advancement in the industrial control technologies and the increased deployment of the industrial Internet of Things (IoT) (Braun, Fung, Iqbal, & Shah, 2018; Habibzadeh, Nussbaum, Anjomshoa, Kantarci, & Soyata, 2019; Khan, 2018). Smart buildings are integral elements of smart cities and their IoT technologies are becoming additional elements in the complex network of the smart city, starting from the sensors collecting data to the associated smart systems such as lighting, power system, ventilation system, etc. (Ande, Adebisi, Hammoudeh, & Saleem, 2020). The industry predicts that the IoT market will grow from an installed base of 30.7 Billion devices in 2020 to 75.4 Billion in 2025; many of which will be deployed in intelligent buildings (IoT Security Foundation, 2020). These sophisticated technologies help establishing an urban landscape and provide the functionality of unprecedented levels of comfort and convenience. In addition, they improve the operation and capabilities of buildings in smart cities. However, they subject the smart buildings to risks of intrusions due to the increased vulnerabilities and advanced attack vectors. According to Kaspersky Lab, which is a multinational cybersecurity and anti-virus provider, nearly four in ten intelligent buildings were targeted by attacks in the first half of 2019, and it is expected that the impact of cyberattacks on the building and construction industry will be significant in the coming years (Kaspersky, 2019).

There have been several works for securing the operation of smart cities using solutions considering a group of its interconnected subsystems, i.e. smart buildings, power system, transportation system, etc., such as in Rahman et al. (2020) where authors proposed a machine-learning (ML)-based distributed intrusion detection system (IDS) for the IoT network of resource-constrained devices in smart cities using feature extraction/selection ML-based models and neural network-based detection models. Qureshi, Rana, Ahmed, and Jeon (2020) presented an attack detection framework for low power and lossy networks in large scale industrial IoT environments in smart cities consisting of a threshold modulation phase in which the detection

* Corresponding author.
E-mail addresses: me1003659@qu.edu.qa (M. Elnour), nader.meskin@qu.edu.qa (N. Meskin), k.khan@qu.edu.qa (K. Khan), jain@wustl.edu (R. Jain).

https://doi.org/10.1016/j.scs.2021.102816
Received 15 September 2020; Received in revised form 19 February 2021; Accepted 24 February 2021
Available online 2 March 2021

Table 1
Summary of the previous research works in the cybersecurity of building management systems.

Reference	Objective	Approach	Perspective (network/process)	Validation method	Attack detection?
Novak and Gerstinger (2010)	Security analysis	–	Network	–	×
Granzer et al. (2010)	Security analysis	–	Network	–	×
Peacock and Johnstone (2014)	Vulnerabilities analysis	Rule-based	Process	Simulation	×
Wardell et al. (2016)	Vulnerabilities analysis	Rule-based	Process	Testbed	×
Hachem et al. (2020)	Vulnerabilities analysis	Rule-based	Network	Testbed	×
Yoon et al. (2016)	Evaluation of the readiness of cyber-professionals	Rule-based	Process	Simulation	×
D'Innocenzo et al. (2016)	Resilient control	Model-based	Network	Simulation	×
Hernandez-Ramos et al. (2015)	Intrusion detection	Data-driven	Network	Testbed	✓
Singh et al. (2017)	Intrusion detection	Data-driven	Network	Testbed	✓
Paridari et al. (2018)	Attack diagnosis and resilience	Hybrid	Process	Simulation	✓

threshold is set for each attack type followed by the attack detection phase. In Jararweh, Otoum, and Ridhawi (2020), the authors presented a service delivery solution at the edge of the network using a collaborative technique between distributed edge servers and privacy mediator nodes with the support of an intrusion detection system to enhance the availability, reliability, and security of smart city applications. Singh, Jeong, and Park (2020) proposed a deep learning-based IoT-oriented framework for secure smart cities using Blockchain technology to provide a distributed environment at the communication layer, and Software-Defined Networking (SDN) to establish the protocols for network data forwarding.

Research in securing the operation of intelligent buildings is indispensable. That is, they are equipped with building management systems (BMSs), which are computer-based control systems used for monitoring and control of the building's equipment, such as air conditioning and ventilation, lighting, power systems, etc. Specifically, the Heating, Ventilation, and Air Conditioning (HVAC) systems are part of BMSs that are dedicated to providing healthy and comfortable indoor environments for occupants with minimum energy utilization. They are the most extensively operated equipment, and they contribute to about 40% of the total energy consumption and more than 55% of the electricity demand in the buildings (IEA, 2017; UN Environment Programme, 2020). In 2013, two security researchers discovered an exploitable critical vulnerability in the building management system of Google Australia Office and found that the BMS can be easily compromised to gain access to the operating system and any other control systems (Kim Zetter, 2013). At the Security Analyst Summit in 2016, Kaspersky Lab revealed that a hacker could break into HVAC systems across a city to turn them on and cause a blackout due to a power surge (Kate Kochetkova, 2016). There have been a number of security breaches in the HVAC system in the past years such as the case of the Target Corporation hack in 2014 in which the HVAC system was compromised and used to gain access to financial records to steal the credit card information for over 40 million of the store's customers in one of the biggest data breaches in history (KrebsonSecurity, 2014). Another incident took place in 2016 in Finland in which the operation of the heating systems of two residential buildings was disrupted leaving the occupants without heating due to a DDoS attack (Lee Mathews, 2016).

The global energy demand is increasing as well as the pressure on managers to reduce the expense incurred in operating buildings due to the increased electricity prices (Vishwanath, Chandan, & Saurav, 2019). Malfunctioning of an HVAC system would result in a significant rise in the building's energy usage and reduction in its energy efficiency. Moreover, it can result in interference in the execution of safety supervision schemes, and may impact their effectiveness and correctness by resulting in executing unnecessary tasks based on falsified decisions. Nevertheless, the productivity, health, and comfort of the buildings' occupants are also other aspects of importance that are influenced by the operation of HVAC systems. Even though there have been several works for optimizing and managing the operation and energy utilization in smart buildings, such as Iqbal et al. (2018), Rodriguez-Trejo et al. (2017), Safa, Safa, Allen, Shahi, and Haas (2017), Zhu et al. (2019) and Ghofrani, Nazemi, and Jafari (2019), investigating and developing solutions for the cybersecurity of building management systems is essential (Fisk, 2012). There have been limited research works assessing and evaluating the cybersecurity aspect of smart buildings, e.g. D'Innocenzo, Smarra, and Domenica (2016), Granzer, Praus, and Kastner (2010), Hachem, Chiprianov, Babar, Khalil, and Aniorte (2020), Hernandez--Ramos, Moreno, Bernabe, Carrillo, and Skarmeta (2015), Novak and Gerstinger (2010), Paridari et al. (2018), Peacock and Johnstone (2014), Wardell, Mills, Peterson, and Oxley (2016), Yoon, Dunlap, Butts, Rice, and Ramsey (2016) and SSingh, Sharma, and Park (2017). A thorough analysis of the security of building automation systems was presented in Novak and Gerstinger (2010) and Granzer et al. (2010), while the authors in Peacock and Johnstone (2014) presented a threat analysis to identify the security-related challenges in the building automation domain. Wardell et al. (2016) presented a rule-based approach for analyzing the cybersecurity vulnerabilities in industrial control systems, which was demonstrated using a simple HVAC system.

Hachem et al. (2020) proposed a software-based approach to address the challenges of security modeling and analysis for cyber-physical systems and it was applied to a real-life smart building. Yoon et al. (2016) proposed adopting the NFPA 1410 standards to investigate the readiness of first responders in real-world scenarios to attacks launched against critical infrastructures, and it was validated using a simulation-based HVAC system. In D'Innocenzo et al. (2016), the idea of the co-design of a resilient control system and communication protocol for networked control systems against node failures and attacks was addressed and applied to an HVAC system simulation environment. Hernandez-Ramos et al. (2015) proposed a network-based security framework for smart buildings based on an anomaly behavior analysis intrusion detection system, while a secure network architecture utilizing Multivariate Correlation Analysis (MCA) to detect DoS attacks in the real-time network traffic for a smart home was presented in Singh et al. (2017). Paridari et al. (2018) proposed a hybrid cyber-physical-security framework for building management systems using a model-based attack mitigation strategy using Kalman filters whenever an attack is detected using a hybrid detection framework combing expert knowledge with One-Class Support Vector Machine (OCSVM) approach.

As summarized in Table 1, the analysis of the most previous works was conducted to assess the security and the cyber-related vulnerabilities in intelligent buildings as in Granzer et al. (2010), Hachem et al. (2020), Novak and Gerstinger (2010), Peacock and Johnstone (2014), Wardell et al. (2016) and Yoon et al. (2016) while intrusion resilience control system was developed in D'Innocenzo et al. (2016). On the other hand, detection mechanisms from the network perspective were proposed in Hernandez-Ramos et al. (2015) and Singh et al. (2017). However, we believe that examining the physical dynamics of the system can be very useful and can yield promising outcomes. Even though the authors in Paridari et al. (2018) proposed a process-based attack detection framework, the detection model requires the knowledge of the physical rules and relationships between the system's variables which might be complex or difficult to obtain for nonlinear processes in the system.

In this paper, we present a semi-supervised, data-driven Isolation Forest (IF)-based attack detection approach for a multi-zone HVAC system in which the normal operation data are used to develop an IF-based detection model. Isolation forests are characterized by their low computational requirements, and they are based on the concept of isolation, which means pointing out anomalies. This improves the attack detection capability, and the use of isolation forests for attack detection has shown promising performance for water treatment plants (Elnour, Meskin, Khan, & Jain, 2020), smart grid networks (Ahmed, Lee, Hyun, & Koo, 2019), and information security (Vartouni, Kashi, & Teshnehlab, 2018). Three schemes of the proposed approach are examined such that in Scheme 1, the detection model is developed using the raw data; while in Scheme 2 and Scheme 3, feature extraction is performed using Principal Component Analysis (PCA) and 1D Convolutional Neural Network (CNN) Encoder model, respectively. The difference between the last two schemes is that the temporal features in the data are taken into consideration in Scheme 3.

The main contribution of this work is as follows:

1. We developed and presented a detailed simulator for a 3-floor, 12-zone HVAC system using TRNSYS, which is a reliable tool for simulating the behavior of HVAC systems because its modules have been developed to be consistent with practical data. The simulator is useful for simulating false injection attacks, which is advantageous to study and assess the security aspect of the HVAC systems from the process perspective.
2. We proposed a strategy for assessing the level of severity of the attacks launched against the HVAC system in terms of the occupants' thermal comfort model and the estimated overall power usage of the system.
3. We proposed a new approach using Isolation forest algorithm for HVAC systems attack detection. It is a data-driven approach utilizing just the normal data and without the need for the system mathematical model. It is based on the principle of separating-away anomalous observations, resulting in improving the attack detection capability.

The paper is organized as follows. In Section 2, the details of the development of the HVAC system simulation model are provided, which includes the description of the system under study, the description of the injected attacks' models, and the specifications of the collected dataset. In Section 3, the details of the proposed approach and the model training procedure are presented. Results evaluation and discussion are demonstrated in Section 4, and the conclusion is summarized in Section 5.

2. Development of the HVAC system simulation model and attack models

It is a strenuous task to obtain actual data or gain access to real building management systems due to reasons such as confidentiality, unfeasibility, etc. Thus, the use of simulation tools is common and convenient to provide flexible means to conduct the research and analysis with high fidelity. In this work, the HVAC system is simulated using Transient System Simulation Tool (TRNSYS), which is a graphical software environment that allows the simulation of transient systems' behavior through energy and mass balance equations (Klein, G.M., & Sherrill, 2017). TRNSYS has been widely used for HVAC systems simulation for research and development purposes as in Du, Fan, Jin, and Chi (2014), Elnour, Meskin, and Al-Naemi (2020), Every, Rodriguez, Jones, Mammoli, and Martinez-Ramon (2017), Qiu et al. (2020), Sun, Hu, and Spanos (2017) and Darure, Yamé, and Hamelin (2016), and it has proven to be a reliable tool.

A public TESS HVAC library was established with TRNSYS and it is consistent with practical data. The accuracy and fidelity of TRNSYS models may be attributed to the fact that the modules have been developed by an authoritative department – the Thermal Energy Systems Specialists of the United States; and the software adopts Component Object Method (COM) technology so that it can reproduce the HVAC system to a large extent (Qiu et al., 2020). As presented in Qiu et al. (2020), many of the existing TRNSYS modeling works tend to carry out follow-up work directly without verification as in Alibabaei, Fung, Raahemifar, and Moghimi (2017), Cutillas, Ramírez, and Miralles (2017), Diallo et al. (2017), Seo, Ooka, Kim, and Nam (2014) and Li, Joe, Hu, and Karava (2015).

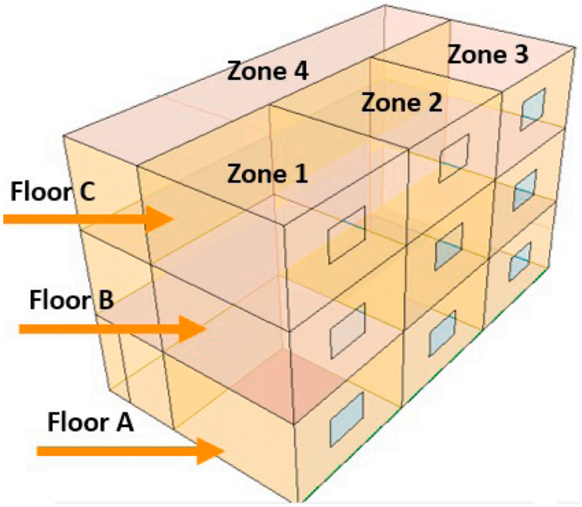


Fig. 1. A sketch of the simulated 12-zone building.

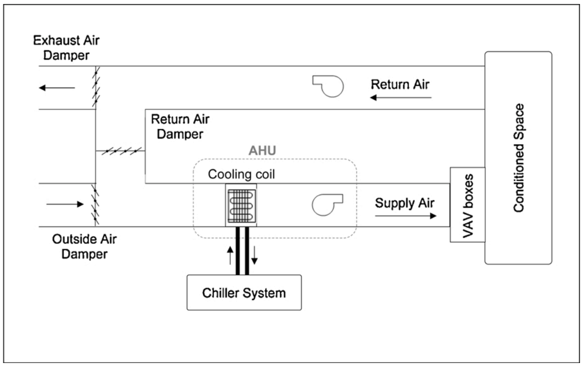


Fig. 2. The diagram of a typical HVAC system using the Variable Air Volume (VAV) system (Elnour et al., 2020).

2.1. Description of the HVAC system simulator

The building under this study is a 3-floor office building operating from 6 AM to 6 PM. The floors are labeled A, B, and C. Each floor consists of four zones with a total floor area of 120 m^2 as shown in Fig. 1. Zones 1–3 are office rooms with a volume of 75 m^3 per zone, while Zone 4 is a hall with a volume of 135 m^3. The building is equipped with a simple HVAC system for the cooling application as shown in Fig. 2. The temperature at each zone is controlled using Proportional Integral Derivative (PID) controllers. The HVAC system is simulated using TRNSYS in which the cold output air from the Air Handling Unit (AHU) is supplied using a supply air fan to the zones through the Variable Air Volume (VAV) boxes terminals. The zones' temperature controllers modulate the position of the air dampers according to the heat gains/losses within the zones to achieve the desired setpoints. The return air of the zones is fed

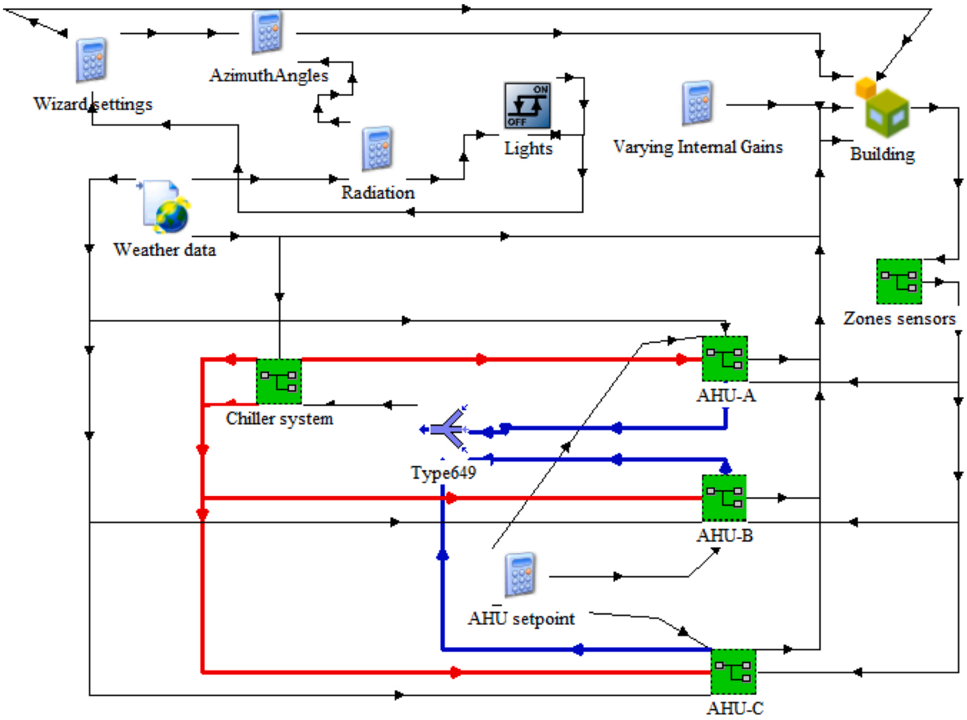


Fig. 3. Overview of the TRNSYS simulator of the 12-zone HVAC system.

back to the AHU through the return air ducts using the return fan. The exhaust air dampers (EA), outside air (OA) dampers, and return air (RA) dampers are operated simultaneously to control the proportions of the recirculated air and the ventilation air in order to maintain the indoor air quality.

Each floor is equipped with a dedicated AHU which supplies the zones with the cold air at a constant temperature of 13 °C, and a variable flow rate controlled by the VAV box terminals. The water chiller and the cooling coil are connected via the chilled water tank that supplies chilled water to the cooling coil using a pump. The temperature of the chiller supply water is set at 9 °C. Using a PID controller, the water tank temperature is controlled at 11 °C via a water valve regulating the flow of chilled water from the chiller to the tank. The system's variables of interest for the current study are: the 12 zones temperatures, T_{zAi}, T_{zBi}, and T_{zCi}, for $i = 1, \ldots, 4$, the temperature of the chilled water tank T_t, the temperature of the chiller supply water T_{chiller}, the temperatures of the AHUs' supply air T_{aoA}, T_{aoB}, T_{aoC}, the temperatures of AHUs' output water T_{woA}, T_{woB}, T_{woC}, the ambient temperature T_{amb}, the zones VAV boxes control signals U_1–U_{12}, and the water tank valve control signal U_{13}.

The details of the TRNSYS simulator are shown in Figs. 3 and 4 . The overall configuration of the HVAC system is demonstrated in Fig. 3 – with its main elements being presented as green blocks. The four diagrams presented in Fig. 4 show the content of those green blocks. In addition, the building occupancy and the internal loads such as the equipment and lighting are taken into account as presented in Table 2.

The main objective of the HVAC system is to provide healthy and comfortable indoor conditions for occupants at minimum energy utilization. The HVAC system energy usage can be estimated by the amount of power consumption of the extensively operated equipment such as the chiller, fans, and pumps. The thermal comfort is the condition of mind that expresses the satisfaction of occupants with the indoor thermal environment and is assessed by subjective evaluation. It involves factors such as the occupants' activity, clothing, indoor air temperature, and velocity, etc. The Predicted Mean Vote (PMV) index is used to predict the mean response of a larger group of people according to the ASHRAE thermal sense scale (Ogoli, 2007) in which feeling hot = +3, feeling warm = +2, feeling slightly warm = +1, neural = 0, feeling slightly cool = − 1, feeling cool = − 2, and feeling cold = − 3.

Considering that the system under study is an office building, the thermal comfort model is simulated using TRNSYS for the factors presented in Table 3.

2.2. *Modeling of the HVAC system attacks*

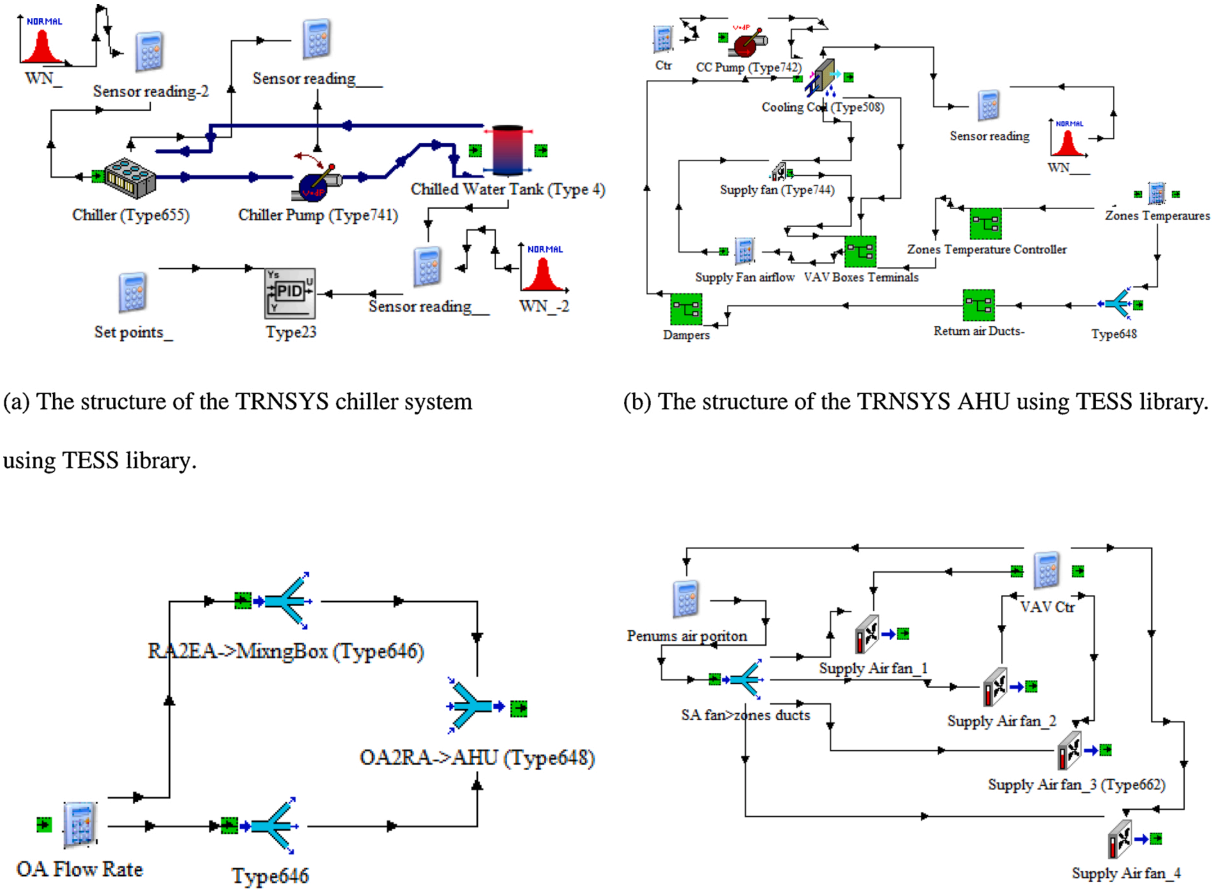
The objective of a malicious agent compromising an HVAC system is mainly related to two factors: the building energy consumption, and the thermal comfort of occupants. An attack can be launched to degrade the building efficiency resulting in unnecessary energy consumption or to alter the level of satisfaction and comfort of the building occupants. In this work, the impact of the attack is evaluated according to the following definition.

Definition 2.1. An attack is considered **critical** if at least one of the following conditions is met:

- **Condition 1:** PMV $>+1$ or PMV < -1,
- **Condition 2:** $P_{\text{actual}} >> P_{\text{nominal}}$,Otherwise, the attack is considered a **failed attack** where P_{actual} (kJ/hr) and P_{nominal} (kJ/hr) are the actual total power utilization under attack, and the nominal total power utilization, respectively, under normal operation.

That is, a **critical attack** causes unnecessary energy consumption and/or alters the level of satisfaction and comfort of the building occupants. Additionally, it is assumed that the intent of an adversary launching attacks against the HVAC system is to achieve at least one of these two conditions. Hence, if an attack is critical, the goal of the adversary is achieved.

The attacks demonstrated in this section are presented in Wardell et al. (2016) as the various types of expected attacks against industrial control systems and supervisory control and data acquisition systems. The possible malicious actions that can be launched against an HVAC system are:



(a) The structure of the TRNSYS chiller system using TESS library.

(b) The structure of the TRNSYS AHU using TESS library.

(c) The details of the dampers block in the AHU.

(d) The structure of the VAV boxes in the AHU using TESS library, which are modeled using variable speed fans (Type662).

Fig. 4. The details of the TRNSYS-based HVAC system simulator. (For interpretation of the references to color in this figure citation, the reader is referred to the web version of this article.)

Table 2
The details of the building's internal heat gain sources.

Space	Details
Reception hall	Occupation: 6 ± 1 persons (6 AM to 6 PM) Person: Standing, light work, 185 W Computer: 140 W Lights: 15 W/m^2
Halls	Occupation: 5 ± 1 persons (6 AM to 6 PM) Person: Standing, light work, 185 W Lights: 15 W/m^2
Office rooms	Occupation: 5 ± 1 persons (6 AM to 6 PM) Person: Seated, light work, typing, 150 W Computer: 140 W per person Lights: 5 W/m^2

Table 3
The factors of the thermal comfort model in the building.

Factor	Value	Description
Clothing	1.0 col	Typical business suits
Air velocity	0.1 m/s	Nominal air velocity
Metabolic rate	1.2 mets	Seated, light work
	1.6 mets	Standing, light work

2.2.1. ***Attack 1:*** *Changing system setpoints*

Setpoints are used to determine the controller's actions towards regulating the operation of the controlled system. Only authorized users can change the setpoints by accessing the controller from an engineering or operator console. The attack points for the HVAC system under study are the zones temperature setpoints, the chiller outlet temperature setpoint, the water tank temperature setpoint, and the AHU supply air temperature setpoint.

2.2.2. ***Attack 2:*** *Falsifying sensor measurements*

False measurements are received by the controller in man-in-the-middle attacks when sensors fail to provide real-time measurements to the control system due to malicious actions. The control system then fails to maintain a reliable system operation. The possible attack points are the sensors associated with the control loops, which are the zones' temperature sensors and the tank temperature sensor. The attack model can be expressed as follows:

(1) Frozen sensor measurement:

$$x_{\mathrm{a}}(t) = x_{\mathrm{h}}(t^{s}_{\mathrm{a}}), \quad \text{for} \quad t^{s}_{\mathrm{a}} < t < t^{e}_{\mathrm{a}}, \tag{1}$$

(2) Biased sensor measurement:

$$x_{\mathrm{a}}(t) = x_{\mathrm{h}}(t) + a_{t}, \quad \text{for} \quad t^{s}_{\mathrm{a}} < t < t^{e}_{\mathrm{a}}, \tag{2}$$

Table 4
List of injected attacks on the HVAC system simulator.

Attack index	Description	Attack time
1.1	Changing the setpoint of the chiller to 14 °C	Day 1, 12:00
1.2	Changing the setpoint of the water tank to 16 °C	Day 2, 06:00
1.3	Changing the setpoint of the AHU to 20 °C	Day 2, 10:00
1.4	Changing the setpoint of Zone A1 to 26 °C **(failed attack)**[a]	Day 20, 11:00
1.5	Changing the setpoint of Zone C4 to 18 °C **(failed attack)**[a]	Day 1, 03:00
2.1	Freezing Zone B1 reading **(failed attack)**[a]	Day 5, 17:00
2.2	Freezing Zone C4 reading	Day 7, 07:30
2.3	Freezing Zone A2 reading	Day 9, 05:30
2.4	Freezing Zone C3 reading	Day 10, 06:00
2.5	Introducing a bias of 3 °C to Zone B3	Day 3, 08:00
3.1	Freezing the control signal of Zone C2 **(failed attack)**[a]	Day 10, 15:00
3.2	Freezing the control signal of Zone B3 **(failed attack)**[a]	Day 13, 18:00
3.4	Freezing the control signal of Zone B1	Day 15, 06:00
3.5	Setting control signal of Zone B2 to 0 **(failed attack)**[a]	Day 19, 06:00
3.6	Setting control signal of Zone A3 to 1	Day 19, 20:00
4.1	Reducing the AHU-B water pump to 1/3 of its speed	Day 18, 12:00

[a] **Failed attack** is when the adversary fails to achieve its objective that is causing unnecessary energy consumption and/or altering the thermal comfort of the occupants.

where $x_a(t)$ and $x_h(t)$ are the falsified and healthy sensor measurements, respectively, t_a^s and t_a^e are the attack start and end times, respectively, and a_t is the amount of bias introduced in the falsified sensor measurements.

2.2.3. *Attack 3: Falsifying control signals*

A malicious agent gaining access to the controller-plant communication link can modify the control signals before sending them to the plant with the objective of harming the controlled equipment, e.g. a fan, a pump, etc. or perturbing the system operation. Similar to Attack 2, the possible attack points are the control signals for the control loops of the zones' temperatures and the tank temperature. The attack model is described by:

(3) Frozen control signal:

$$u_a(t) = u_h(t_a^s), \quad \text{for} \quad t_a^s < t < t_a^e, \tag{3}$$

(4) False control signal:

$$u_a(t) = u_h(t) + b_t, \tag{4}$$

where $u_a(t)$ and $u_h(t)$ are the falsified and valid control signals, respectively, $u_h(t_a^s)$ is the valid control signal at time t_a^s when the attack is launched, and b_t is the amount of bias introduced in the falsified control signal.

2.2.4. *Attack 4: Modifying command signals to components*

Some components of the HVAC system operate on a specified command by an authorized user or according to the specification of the system operation without the need for continuing regulation. A malicious agent can tamper with the command signals sent to these components such as turning off a fan or a pump, forcing them to operate at a slower speed than they should, or manipulating the positioning of dampers, etc. The potential HVAC system attack points are the cooling coil pump, supply and return air fans, and exhaust and outside dampers. The component's command signal c can be expressed as:

$$c = \begin{cases} c_h, & \text{for} \quad t \le t_a^s \quad \text{or} \quad t \ge t_a^e, \\ c_a, & \text{for} \quad t_a^s < t < t_a^e, \end{cases} \tag{5}$$

where c_h and c_a are the valid and falsified command signals, respectively, provided that $c_h \neq c_a$.

2.3. *Description of the HVAC system dataset*

The dataset consists of two logs in which the first log contains normal operational data collected for four months – from June to September-with a total of 194301 samples that was used for the models' development in the training and validation phases. The second log represents data collected in a span of 20 days of system operation with a total of 8840 samples, during which the 16 attacks listed in Table 4 were injected by modifying the sensor's setpoint, sensor's reading, or actuator's control signal as described in Section 2.2, comprising about 50% of the data log, 10 of them were critical according to Definition 2.1. The subset of the second log containing only the critical attack scenarios was used for the models' testing phase. The reason for excluding the samples corresponding to the failed attacks is to achieve unbiased evaluation given that those attacks are unlikely to be detected as they do not reflect significantly on the system operation. Table 4 lists the description of the attacks, the time they were injected, and their indices such that an attack

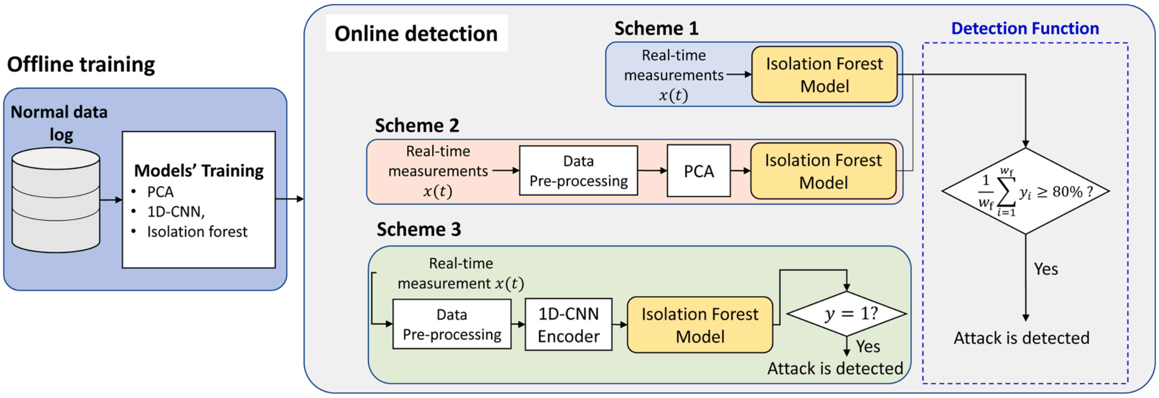
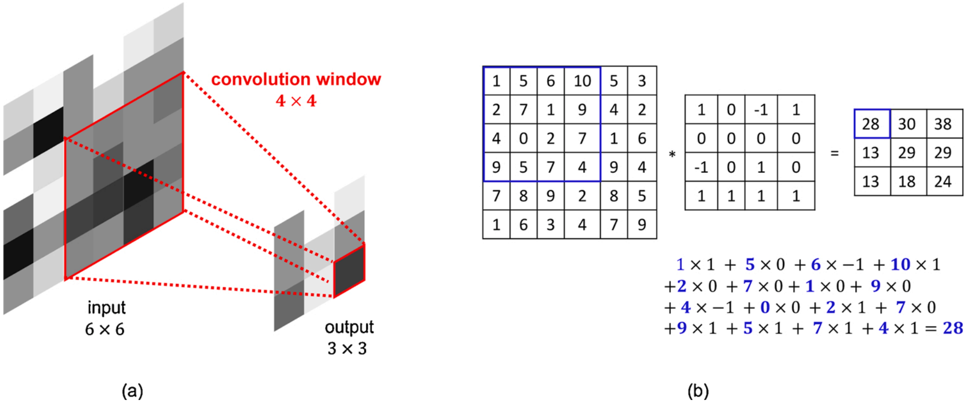


Fig. 5. The diagram of the proposed isolation forest-based HVAC system attack detection framework. The framework consists of an offline training phase during which the models are developed and an online detection phase in which the developed models are used to perform attack detection. Three schemes are investigated, which are Scheme 1, using raw data; Scheme 2, using PCA for feature extraction; and Scheme 3, using 1D CNN-based encoder for temporal feature extraction. Data pre-processing is required for Schemes 2 and 3 to eliminate the effect of the diverse ranges of the different data features. The pre-processing step in Scheme 3 includes data segmentation to convert the data samples to 1-dimensional vectors form. Scheme 1 and Scheme 2 have the same detection function.



$$\begin{array}{|c|c|c|c|c|c|}\hline 1 & 5 & 6 & 10 & 5 & 3\\ \hline 2 & 7 & 1 & 9 & 4 & 2\\ \hline 4 & 0 & 2 & 7 & 1 & 6\\ \hline 9 & 5 & 7 & 4 & 9 & 4\\ \hline 7 & 8 & 9 & 2 & 8 & 5\\ \hline 1 & 6 & 3 & 4 & 7 & 9\\ \hline\end{array} * \begin{array}{|c|c|c|c|}\hline 1 & 0 & -1 & 1\\ \hline 0 & 0 & 0 & 0\\ \hline -1 & 0 & 1 & 0\\ \hline 1 & 1 & 1 & 1\\ \hline\end{array} = \begin{array}{|c|c|c|}\hline 28 & 30 & 38\\ \hline 13 & 29 & 29\\ \hline 13 & 18 & 24\\ \hline\end{array}$$

$$\begin{aligned}&1\times 1 + 5\times 0 + 6\times -1 + 10\times 1\\ &+2\times 0 + 7\times 0 + 1\times 0 + 9\times 0\\ &+4\times -1 + 0\times 0 + 2\times 1 + 7\times 0\\ &+9\times 1 + 5\times 1 + 7\times 1 + 4\times 1 = 28\end{aligned}$$

Fig. 6. The operation of a typical convention layer is performed across every channel of the input. (a) The sliding of the convolution window (filter) over the input. (b) The convolution operation (valid convolution).

is identified by its type and number. For example, Attack 1.2 was launched on Day 2 at 6:00, and it presents the second injected attack of type 1, changing the setpoint. It is worth noting that the attacks were launched at diverse times throughout the day to cover the different system's operational times, but they are not listed in Table 4 according to the order they were launched. The dataset was collected for a sampling duration of 1 min. It consists of 50 variables, which are the hour of the day, the measurements of the zones' temperatures, the AHUs' supply air temperatures, the cooling coils' return water temperatures, the water tank temperature, the chiller's output water temperature, the ambient temperature, the control signals, the hour of the day, and the temperature setpoints of the zones, the AHU's supply air, the water tank, and the chiller's output water.

Table 5
Details of the 1D CNN-based auto-encoder network.

	Layer identifier	Dimension of the layer's output
Encoder	Input	(w, n)
	Conv 1	(w, F_1)
	Down-sampling 1 (Pooling 1)	$\left(\frac{w}{p1}, F_1\right)$
	Conv 2	$\left(\frac{w}{p1}, F_2\right)$
	Down-sampling 2 (Pooling 2)	$\left(\frac{w}{p1 \times p2}, F_2\right)$
Bottleneck	Conv 3	$\left(\frac{w}{p1 \times p2}, F_3\right)$
Decoder	Conv 4	$\left(\frac{w}{p1 \times p2}, F_2\right)$
	Up-sampling 1	$\left(\frac{w}{p1}, F_2\right)$
	Conv 5	$\left(\frac{w}{p1}, F_1\right)$
	Up-sampling 1	(w, F_1)
	Conv 6	(w, n)

3. Development of the proposed attack detection approach

Isolation forest algorithm has been proven to be excellent in anomaly detection as it is based on the principle of separating-away observations that are anomalous (Liu, Ting, & Zhou, 2012). As shown in Fig. 5, three independent schemes of the proposed isolation forest-based approach are examined. In Scheme 1, a detection model is developed using the raw data. Feature extraction is a popular and effective strategy for improving the performance of the machine-learning algorithms and potentially reducing the computational complexity of ML models. Hence, feature extraction models are investigated in two additional configurations to complement the isolation forest model, which are Scheme 2 and Scheme 3 using Principal Component Analysis (PCA) and 1D Convolutional Neural Network (1D CNN) Encoder model, respectively. The difference between the last two schemes is that the temporal features in the data are taken into consideration in Scheme 3. Unlike the isolation forest algorithm, data normalization is a necessary step before applying feature extraction using PCA and 1D CNN to eliminate the effect of the diverse ranges of the different data features, which can cause model's sub-optimality during the training phase.

3.1. Background

3.1.1. Principal Component Analysis (PCA)

PCA is a multivariate statistical analysis method that is widely used in data dimensionality reduction. The projection matrix $\mathrm{P} \in \mathbb{R}^{n\times l}$ is used to transform the data onto the new feature subspace where n is the number of variables, and l is the number of principal components. It is composed of the first l eigenvectors of the correlation matrix of the data that are associated with the largest eigenvalues.

The data transformation of a normalized measurement vector $\mathrm{x} \in \mathbb{R}^{1\times n}$ to the new data vector $\widehat{\mathrm{x}} \in \mathbb{R}^{1\times l}$ is expressed as:

$$\widehat{\mathrm{x}} = \mathrm{xP}. \tag{6}$$

3.1.2. 1D Convolutional Neural Network Auto-Encoder

An auto-encoder is a neural network composed of an encoder and a decoder parts and is trained to reproduce its input at the output. The encoder produces a compressed representation of the input that is fed to the decoder to reconstruct the input at the output layer (Goodfellow, Bengio, & Courville, 2016). The objective of the auto-encoder's training is to minimize the error between the input and the output. However, attentive design of its internal structure is required to avoid buffering the input to the output layer. The 1D CNN-based auto-encoder model consists of a total of 7 layers, which are the input layer, the output layer, and five hidden layers. They are convolution layers (Conv) in which several convolution windows called filters or kernels slide over the input as demonstrated in Fig. 6. They are characterized by several hyper-parameters such as:

1 The activation function,
2 The size of filter $k \in \mathbb{R}$, which defines the field of view of the convolution,
3 The number of filters $F_c \in \mathbb{R}$ representing the number of feature maps,
4 The stride $\in \mathbb{R}$ representing the size of the convolution step,
5 The type of convolution, which can be same, valid, or dilated. In the same convolution, zero-padding is performed on the input to produce an output of the same size as the input, whereas in the valid convolution, zero-padding is not performed and the output shape

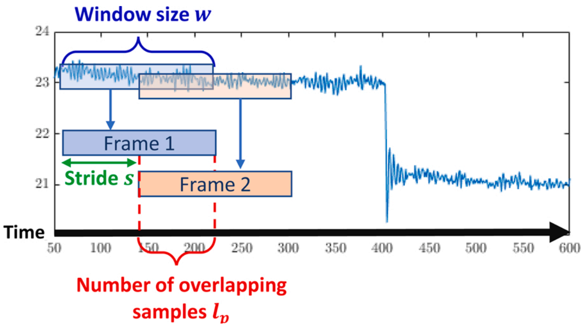


Fig. 7. Data segmentation of a time-series into frames of length w with a stride of $s = w - l_p$.

corresponds to $(\text{Input size} - k)/\text{stride} + 1$. Dilated convolution is used to increase the receptive field of the layer without increasing the computations. It is determined by the dilation factor $f_{\text{D}} \in \mathbb{R}$. The size of the dilated convolution's output is $(\text{Input size} - ((k-1) \times f_{\text{D}} + 1))/\text{stride} + 1$.

The dimension of the output of each layer is listed in Table 5. The dimension is characterized by the pair (d_1, d_2) where d_1 represents the number of time instants of the 1D data, and d_2 represents the number of channels/features. For example, the output of Conv 1 layer is (w, F_1) with F_1 channels and each channel containing w points. The first four layers represent the encoder part of the network. The dimension of the input to the auto-encoder network is $w \times n$, where w is the frame (sample) size, and n is the number of variables. The number of filters used in the layers Conv 1, Conv 2, and Conv 3 are F_1, F_2, and F_3, respectively, such that $n > F_1 > F_2 > F_3$. Each Conv layer is followed by a pooling layer to perform sub-sampling along the first dimension by a rate of p_i. The decoder performs the decompression in which layers Conv 4, Conv 5, and Conv 6 have F_2, F_1, and n filters, respectively, provided that the up-sampling is performed after each Conv layer.

Time-series signals are sequences of data points of successive measurements or observations made over a time interval. Since the input to the 1D CNN is expected to be in the form of 1-dimensional vectors, data segmentation is performed by slicing the time-series into frames of equal lengths as demonstrated in Fig. 7. The segmentation step of a time-series involves two parameters, which are the frame size w, representing the number of data points in a single slice, and the stride $s = w - l_p$, reflecting the number of overlapping data points between the successive frames.

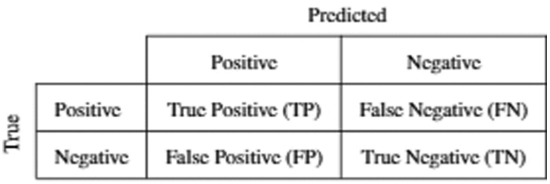
3.1.3. Isolation Forest

Isolation Forest (IF) is an unsupervised Machine Learning algorithm that is used for anomaly detection (Liu et al., 2012). It is an ensemble regressor encompassing several isolation trees in which each tree is trained using a random subset of the training data. For a dataset with n number of features and m data samples, the parameters associated with an isolation forest are the number of trees $n_{\text{estimators}}$, the size of the data subset used to train each tree $m_{\max} \leq m$, and the subset of the data features used to train each tree $n_{\max} \leq n$. The isolation forest uses the concept of isolation to separate-away anomalies. Recursive binary splitting is performed by each isolation tree for a random subset of the data until all samples are isolated.

Anomalies are different from normal observations, and they can be easily isolated. It is expected that they will be closer to the root, and hence have a shorter path. The anomaly detection for a given data sample x is made upon the score $s(x)$ relative to the detection threshold ϵ as follows:

$$s(x) = 2^{-\frac{\bar{h}(x)}{\text{H}}}, \tag{7}$$

Table 6
Table of confusion for a 2-class problem.



		Predicted	
		Positive	Negative
True	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

where H is the average expected path length of trees in the forest provided that anomalies are labeled as 1 while normal observations are labeled with 0, and $\bar{h}(x)$ denotes the average path length on all trees. The anomaly is detected using the following function:

$$y(x) = \begin{cases} 0 & \text{if } s(x) > \epsilon, \\ 1 & \text{if } s(x) \leq \epsilon \end{cases}. \tag{8}$$

3.2. Detection function

As shown in Fig. 5, the outputs of the isolation forest model in Schemes 1 and 2 are evaluated using a detection function. An observation window of frame size w_{f} is checked such that an attack is detected if the output of the detection model y is 1 for at least 80% of the observation period. The stride between the successive observation windows is 1, maintaining the maximum rate of the detection procedure. The decision function is not required in Scheme 3 since feature extraction is performed on a sequence of data points of length w.

3.3. Performance evaluation metrics

The confusion matrix is typically used to evaluate the performance of the classification model. It is a form of contingency table with two dimensions identified as True and Predicted, and set of classes in both dimensions as presented in Table 6. The following performance metrics are derived from the confusion matrix (Ting, 2010):

(9) Precision: It is also called Positive Predictive Value (PPV), which is a measure of the closeness the set of predicted results and it is expressed as,

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{9}$$

(10) Recall: It is called True Positive Rate (TPR) or Recall and is calculated by,

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{10}$$

3.4. Models training

As mentioned previously, the system's dataset that was used to develop the models consists of 50 features of the time, sensors' measurements, setpoints, and actuators control signals. For PCA-based feature extraction in Scheme 2, PCA was performed to retain 95% of the cumulative explained variance resulting in 9 principal components. For developing the 1D CNN model for Scheme 2, data segmentation was performed on the system data, which consist of 50 channels as mentioned in Section 2.3, with each channel representing a time-series. Data segmentation involves two parameters, which are the frame size w, representing the number of data points in a single slice, and the stride s, reflecting the number of overlapping data points between the successive frames. The two parameters were optimized to capture temporal events

Table 7
The ranges of the hyper-parameter values for the 1D CNN-based auto-encoder network and the isolation forest model.

1D CNN			IF		
w	k	F_1	$n_{\text{estimators}}$	$m_{\max}$	$n_{\max}$
12–20	$3–\frac{w}{2}$	10–30	50–400	$2^7–2^{14}$	2–50

Table 8
The performance of the 1D CNN-based auto-encoder models across the hyper-parameter space.

CNN name	w	k	F_1	F_2	F_3	MSE
CNN-AE 1	12	3	13	6	3	0.161
CNN-AE 2		3	14	7	3	0.142
CNN-AE 3		4	16	8	4	0.152
CNN-AE 4		5	14	7	3	0.154
CNN-AE 5	16	3	16	8	4	0.170
CNN-AE 6		4	11	5	2	0.164
CNN-AE 7		4	15	7	3	0.134
CNN-AE 8		5	15	7	3	0.154
CNN-AE 9	20	3	12	6	3	0.153
CNN-AE 10		3	30	15	7	0.145
CNN-AE 11		5	15	7	3	0.128
CNN-AE 12		9	12	6	3	0.177

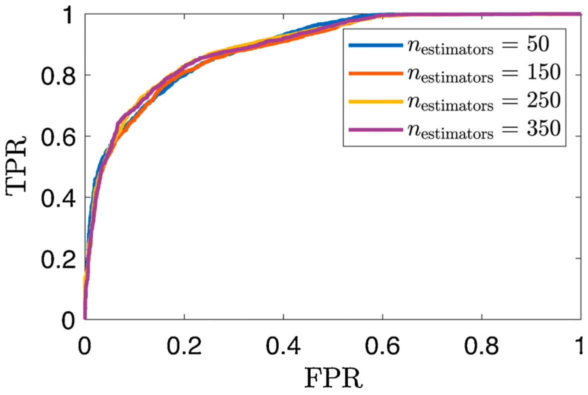


Fig. 8. ROC curves for models of PCA-IF with $m_{\max} = 2000$, $n_{\max} = 2$, and varying $n_{\text{estimators}}$.

or features effectively. The training of the 1D CNN-based auto-encoder network was conducted using Keras library, which is an open-source neural-network library written in Python for the hyper-parameters ranges presented in Table 7. Same convolution with a stride of 1 was used in all the Conv layers, and the Rectified Linear Unit (ReLU) activation function was used since it is the most efficient and commonly used function with CNNs.

The training of the isolation forest models in the three schemes was conducted using Scikit-learn library, which is an open-source Machine Learning library for the Python programming language (Pedregosa et al., 2011). The training was conducted using 5-fold cross-validation, which is a well-known approach used for the validation of Machine Learning-based models to assess the model's generalization ability especially when the amount of data is limited. The dataset was divided equally into 5 random subsets – also called folds. Then each IF model was trained five times, and each time one fold (20% of the data) was used as the validation set and the remaining ones for training (80% of the data). Grid search was utilized for model tuning given the limited number of hyper-parameters associated with the isolation forest model for the ranges presented in Table 7 and to achieve a maximum false alarm rate of 5% on the training dataset. The computer used for the training has 64 GB RAM and 12-cores AMD Ryzen 9 3900X CPU with 3.8 GHz speed using 64 bit Windows 10 Pro OS.

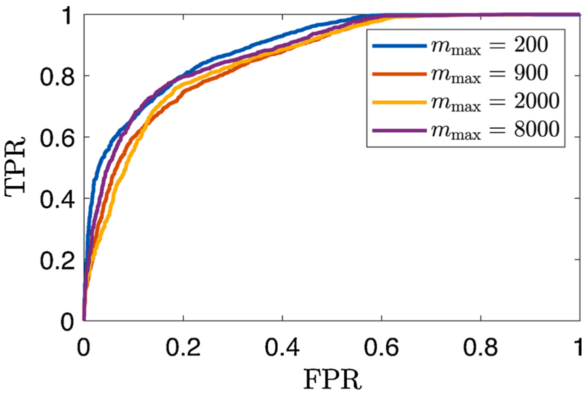


Fig. 9. ROC curves for models of PCA-IF with $n_{\text{estimators}} = 50$, $n_{\max} = 2$ and varying $m_{\max}$.

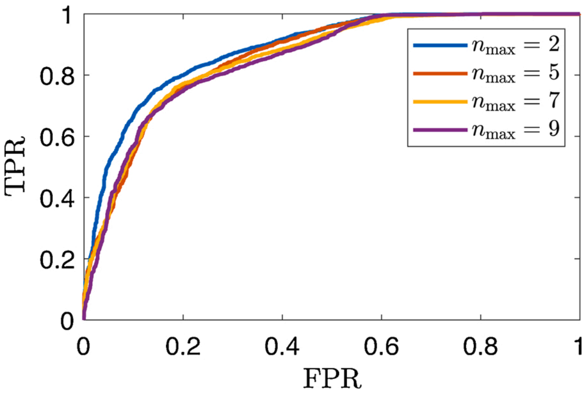


Fig. 10. ROC curves for models of PCA-IF with $n_{\text{estimators}} = 50$, $m_{\max} = 200$, and varying $n_{\max}$.

Table 8 represents the validation dataset's reconstruction error of some of the 1D CNN-based auto-encoder models, which is the Mean-Squared-Error (MSE) between the input and the output of the model of 12 of the trained CNNs. Overly, it was found that the reconstruction error was improved using a larger frame size w and an average kernel size k of 5. The CNN-based auto-encoder with the least MSE was chosen, which was CNN-AE 11 using $w = 20$, $k = 5$, $F_1 = 15$, $F_2 = 7$, and $F_3 = 3$.

A sample of the effect of the hyper-parameters choices for the isolation forest model in Scheme 2 for the detection framework is demonstrated in Figs. 8–10 by the Receiver Operating Characteristic (ROC) curves. As shown in Fig. 8, the number of trees $n_{\text{estimators}}$ used for the isolation forest model did not have a significant effect on the model's detection capability. However, the computational requirement of the IF model is directly proportional to $n_{\text{estimators}}$. In Figs. 9 and 10, the performance of the isolation forest is improved for low values of $m_{\max}$ and $n_{\max}$ as the model's ability to identify anomalous observations will be improved when examining relativity small subset of the dataset across the several trees in the isolation forest model in terms of the data features n and data samples m.

4. Evaluation and discussion

The evaluation of the performance of the proposed framework is demonstrated and compared with standard ML-based approaches. The three isolation forest-based attack detection schemes were evaluated using a subset of the second data log consisting of the critical attack scenarios only. As mentioned previously, the reason for excluding the samples corresponding to the failed attacks is to achieve unbiased

Table 9
Descriptions of the standard Machine Learning algorithms used.

Algorithm	Description	Main parameter
PCA	• It is a linear dimensionality reduction technique used to project the data to a lower-dimensional space using Singular Value Decomposition. • The Principal Components with high eigenvalues capture most of the variance in the data. • A low-dimensional representation constructed by *p* Principal Components can capture most of the variance in a normal data sample.	*p*: number of Principal Components.
OCSVM	• It works by mapping the data into the feature space corresponding to the kernel and finding the hyperplane that separates them from the origin with maximum margin. • The circumstance of the sample is determined by evaluating which side of the hyperplane it falls on in the feature space	Kernel function.
kNN	• It is a global distance-based algorithm. • It depends on the measure of the distance from a sample to its *k*th nearest neighbor.	*k*: number of nearest neighbors.
LOF	• It is a local density-based algorithm. • It depends on the measure of the local deviation of the density of a sample with respect to its neighbors.	*k*: number of nearest neighbors.

evaluation given that those attacks are unlikely to be detected as they do not reflect significantly on the system operation. As presented in Table 10, detection models utilizing One-Class Support Vector Machine (OCSVM), k-Nearest Neighbors (kNN), Local Outlier Factor (LOF), and PCA were included in the comparison using the implementation in (Zhao, Nasrullah, & Li, 2019) as they are widely accepted for anomaly detection applications. Brief descriptions of each algorithm are presented in Table 9.

Most of the attack samples were detectable using kNN, LOF, and PCA as indicated by the recall value, which represents the percentage of correct positive predictions among all positive cases. However, their performance was unsatisfactory due to the relatively high false alarms inferred from the low precision, which was about 43% on average for the three approaches. The precision of the OCSVM-based detection model was the highest with about 98%, but its ability to detect attacks was poor, with a recall of 17%, which is attributable to the complexity and multidimensionality of the system's data such that the two classes; normal and attack, are not separable by a hyperplane. The IF-based model was able to detect around 50% of the attack instances with an overall precision of about 81%. This performance is due to the underlying theory of the isolation forests, which works based on separating-away samples that are anomalous.

The performance was improved with the use of PCA for data dimensionality reduction resulting in a precision of 90% and a recall of about 61%. The PCA-IF attack detection approach scored the best performance with an overall increase of 8% in accuracy compared to the IF-based and the CNN-IF-based schemes. Even though the CNN-IF detection scheme's performance in identifying attack instances was the best as indicated by the recall value of 74%, the precision was the lowest at about 69%, meaning a high rate of false alarms. This can be explained by the fact that the 1D CNN auto-encoder was employed to extract abstract representations using a set of 1-dimensional kernels from signals that

Table 10
Comparison results between the different approaches.

Method	kNN	LOF	PCA	OCSVM	IF	PCA-IF	1D CNN-IF
Precision	37.33%	38.07%	47.40%	97.78%	80.93%	90.01%	68.85%
Recall	100.00%	100.00%	93.85%	17.10%	50.40%	60.49%	74.28%

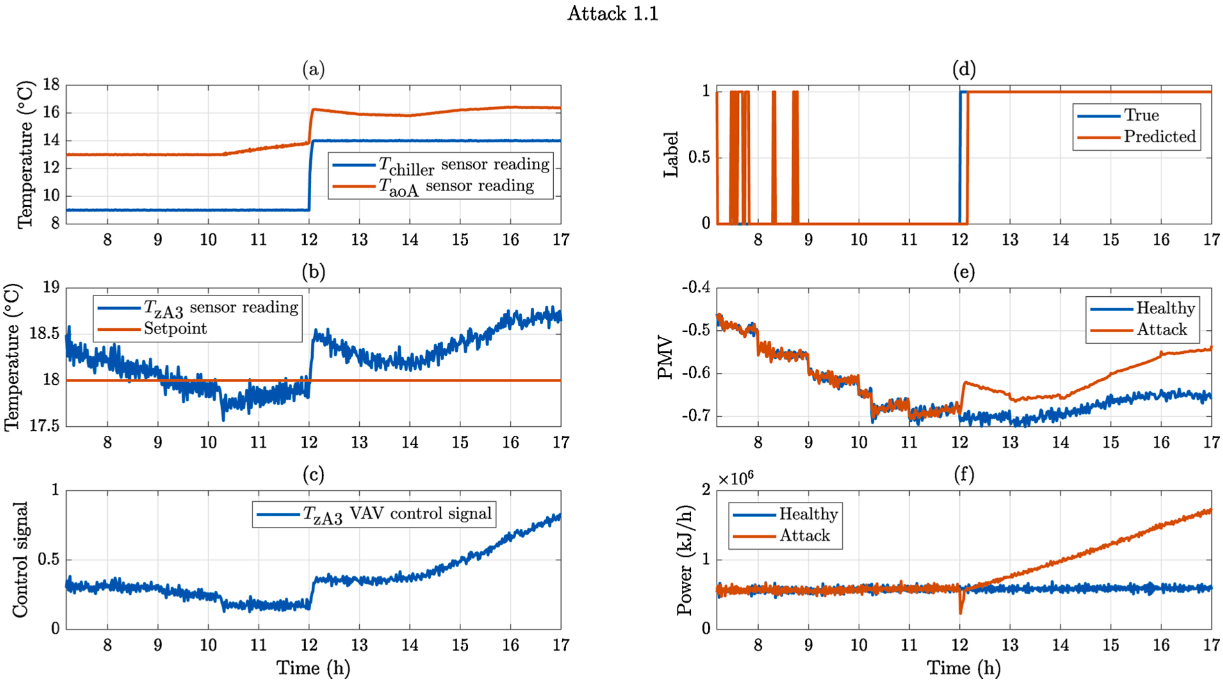


Fig. 11. The performance of the PCA-IF detection framework on Attack 1.1: Changing the setpoint of the chiller to 14 °C. (a) The temperature of chiller's outlet water and AHU-A's supply air, (b) the temperature of Zone A3, (c) the control signal of Zone A3, (d) the predictions of the detection framework for Attack 1.1, (e) the thermal comfort index of Zone A, (e) the total power usage of the HVAC system.

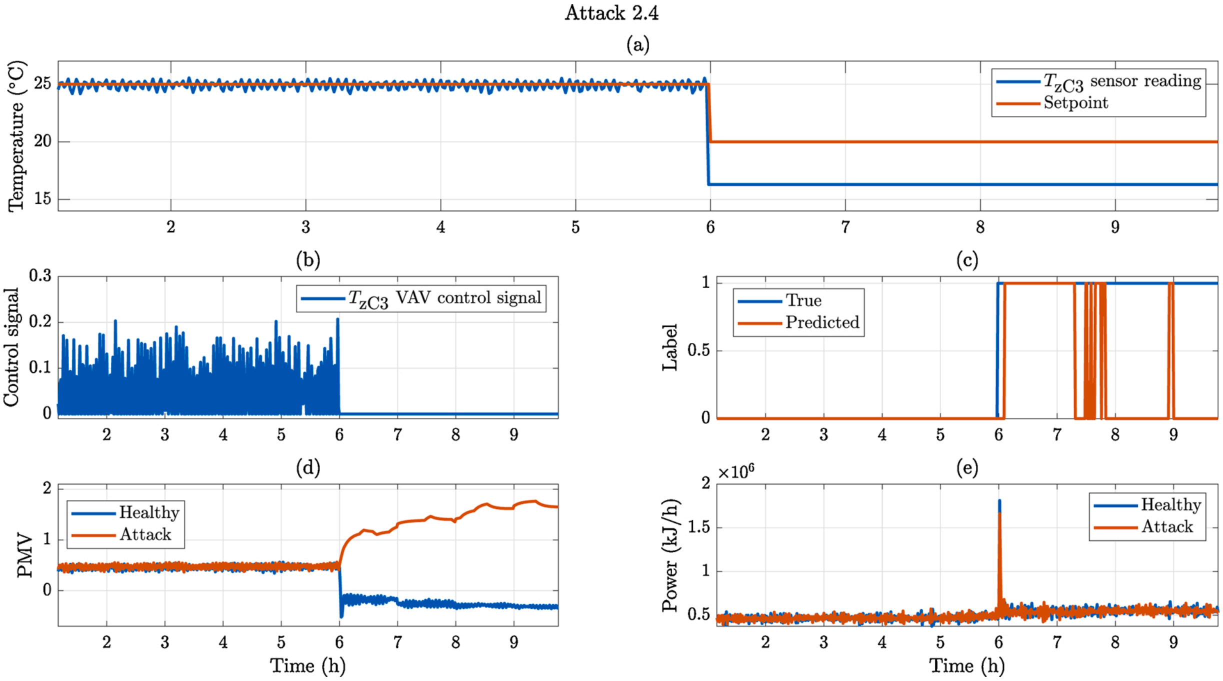


Fig. 12. The performance of the PCA-IF detection framework on Attack 2.4: Freezing Zone C3 reading. (a) The temperature of Zone C3, (b) the control signal of Zone C3, (c) the predictions of the detection framework for Attack 2.4, (d) the thermal comfort index of Zone C3, (e) the total power usage of the HVAC system.

exhibit slow and steady dynamics overly. The HVAC system is characterized by its slow-moving processes concerning temperature and flow rate change. Therefore, the use of the 1D CNN-based temporal feature extraction model was not advantageous.

Using the attack list in Table 5, Figs. 11–17 demonstrate the performance of the PCA-IF-based detection framework on examples of each attack type described in Section 2.2. Overly, it is evident that there are a couple of incidents of false alarms, but they are minimal and not frequent. In **Attack 1.1**, the setpoint of the chiller was increased from 9 to 14 °C at 12:00 as shown in Fig. 11(a). This attack resulted in increasing the water tank temperature and consequently, rising the AHUs' supply air temperatures. It caused a significant impact on the power utilization as demonstrated in Fig. 11(f) due to the system's attempt to meet the cooling load of all the zones by operating the fans and pumps at full speed. As an example, the effect of this attack on Zone A3 was demonstrated in which the control signal of the zone's VAV box started to increase to maintain the zone's at the desired setpoint – which was 18 °C. It is increased gradually to over 100% of its expected value by

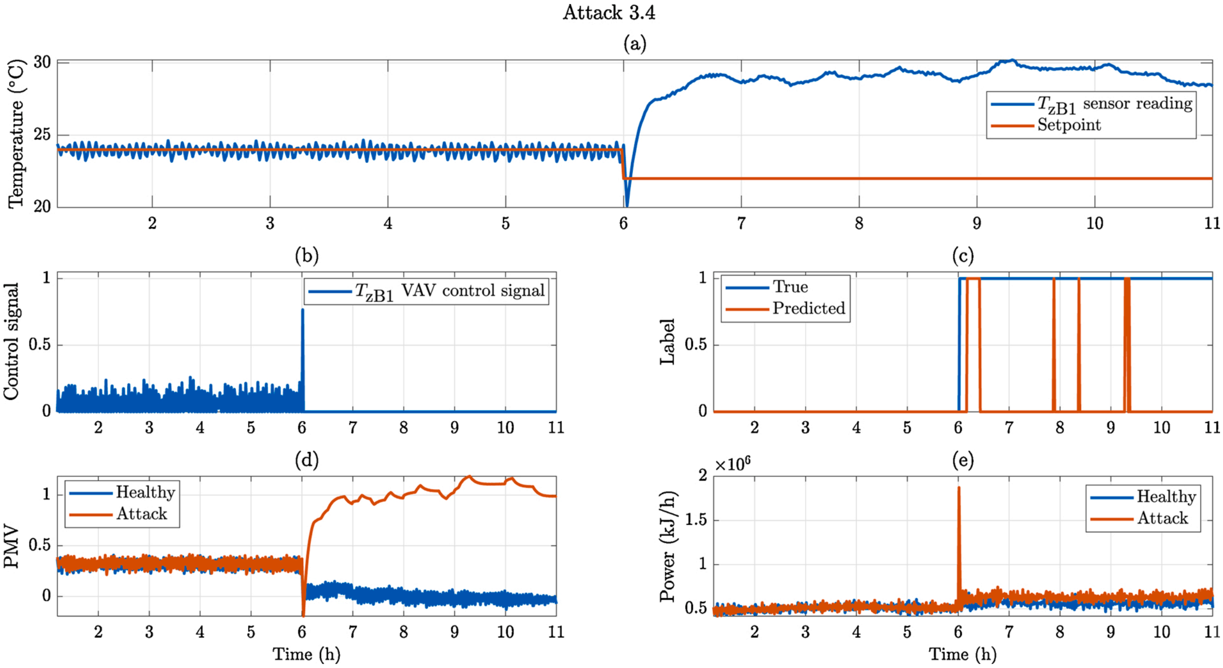


Fig. 13. The performance of the PCA-IF detection framework on Attack 3.4: Freezing the control signal of Zone B1, (a) the temperature of Zone B1. (b) the control signal of Zone B1, (c) the predictions of the detection framework for Attack 3.4, (d) the thermal comfort index of Zone B1, (e) the total power usage of the HVAC system.

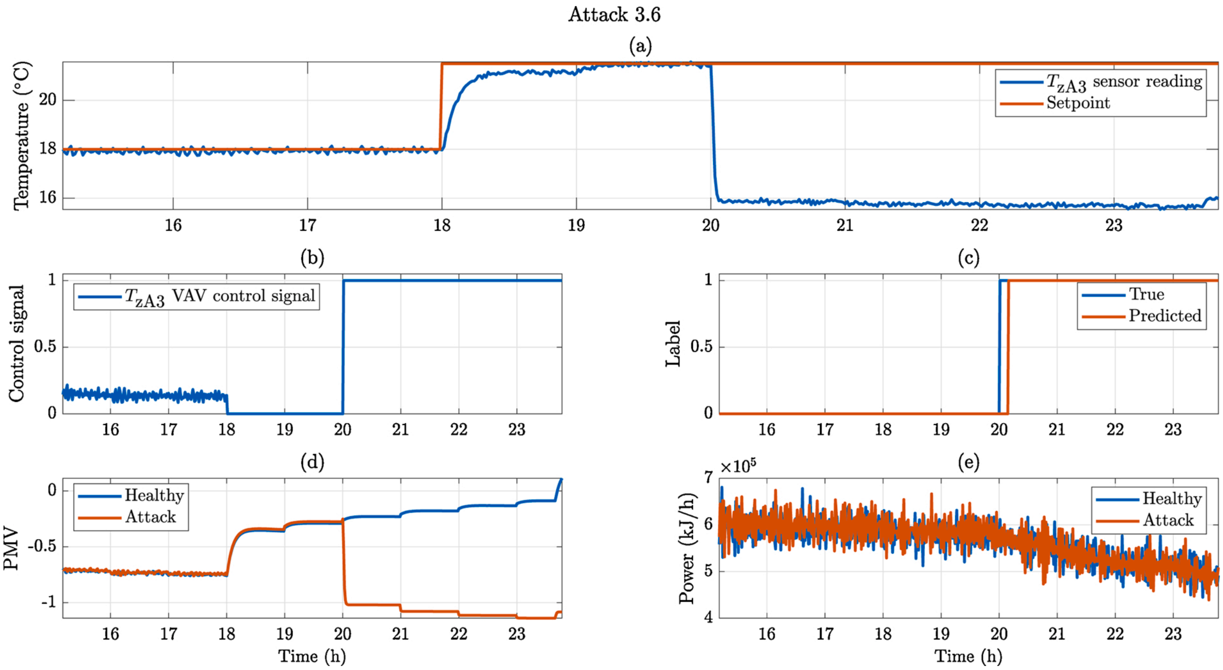


Fig. 14. The performance of the PCA-IF detection framework on Attack 3.6: Setting control signal of Zone A3 to 1. (a) The temperature of Zone A3, (b) the control signal of Zone A3, (c) the predictions of the detection framework for Attack 3.6, (d) the thermal comfort index of Zone A3, (e) the total power usage of the HVAC system.

17:00 as shown Fig. 11(c). The zone temperature was not significantly altered by the attack with only 3% deviation from the setpoint and hence the thermal comfort was not alarmed – indicated by the PMV index (Fig. 11(e)). It is worth noting that the same impact was observed on the remaining zones as the chiller system is a central unit supplying to all the AHUs; hence, the conspicuous increase in the total power utilization in the system shown in Fig. 11(f).

In **Attack 2.4**, the reading of Zone C3 temperature sensor was frozen at the start of the day operation at 5:59 (Fig. 12(a)) and consequently, the control system operated using the frozen sensor reading resulting in a fully closed VAV damper (Fig. 12(b)). Failing to accommodate for the actual cooling load during the daytime operation, it reflected on the occupants' thermal comfort dreadfully as indicated by the PMV index >1 (Fig. 12(d)) indicting hot indoor conditions.

Fig. 13 represents the attack detection performance on **Attack 3.4** in which the control signal of Zone B1 was stuck at the end of the night

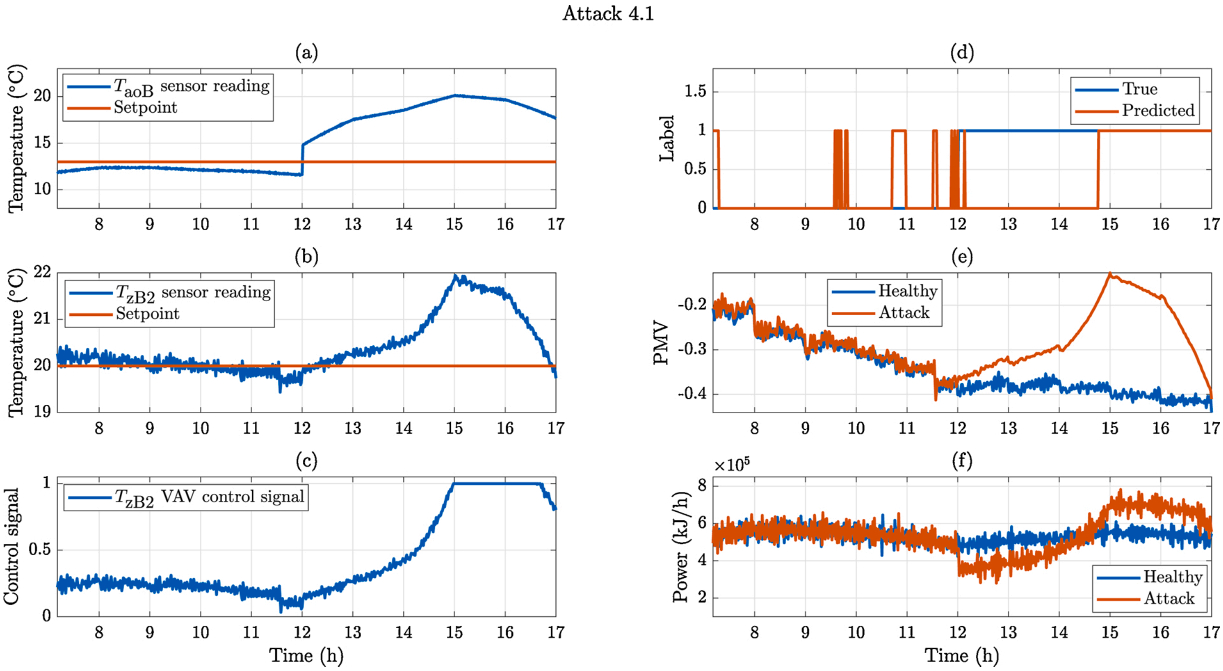


Fig. 15. The performance of the PCA-IF detection framework on Attack 4.1: Reducing the AHU-B water pump to 1/3 of its speed. (a) The temperature of AHU-B's supply air, (b) the temperature of Zone B2, (c) the control signal of Zone B2, (d) the predictions of the detection framework for Attack 4.1, (e) the thermal comfort index of Zone B. (e) the total power usage of the HVAC system.

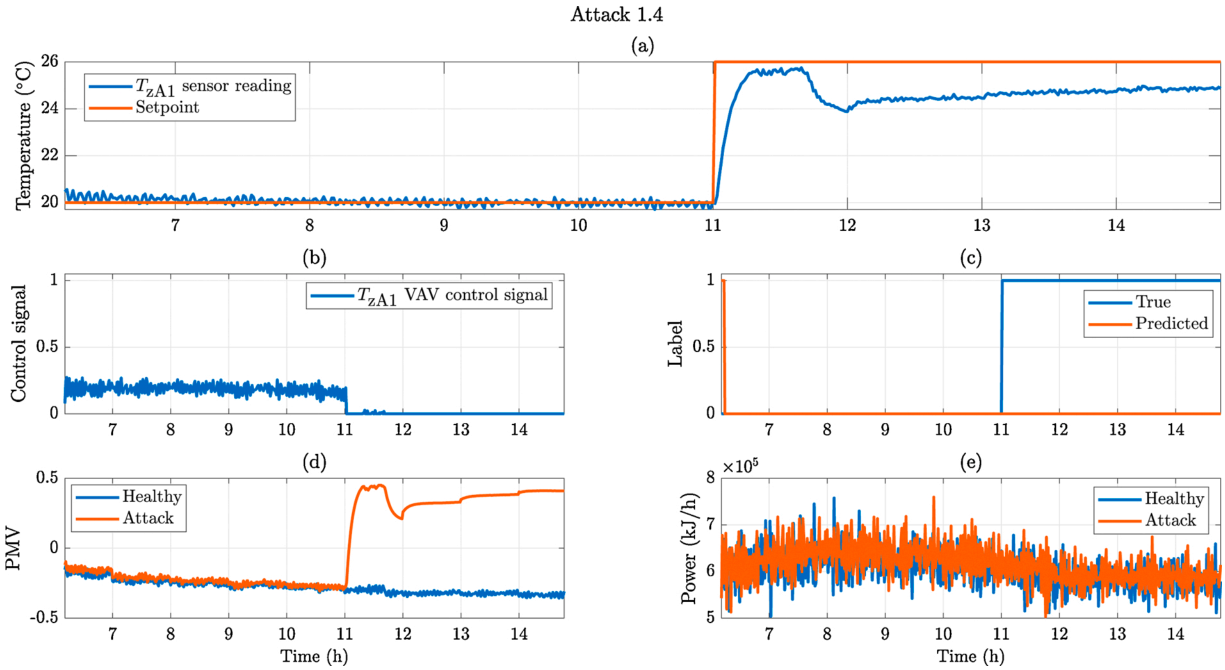


Fig. 16. The performance of the PCA-IF detection framework on Attack 1.4: Changing the setpoint of Zone A1 to 26 °C, (a) the temperature of Zone A1, (b) the control signal of Zone A1, (c) the predictions of the detection framework for Attack 1.4, (d) the thermal comfort index of Zone A1. (e) the total power usage of the HVAC system.

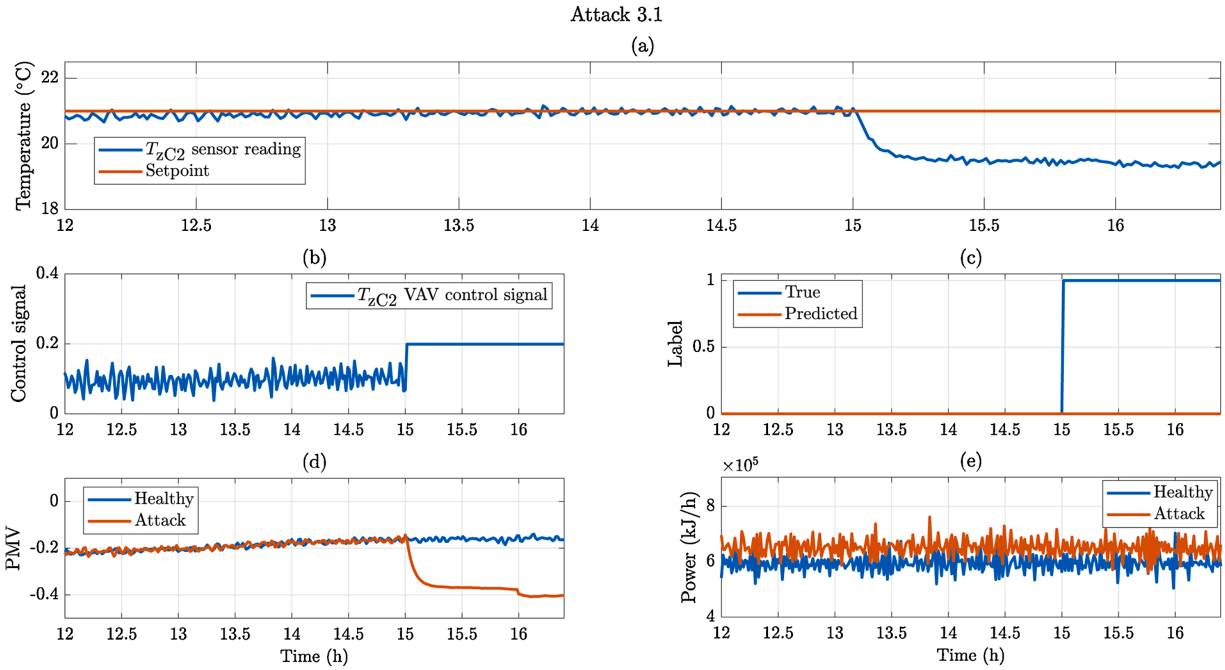


Fig. 17. The performance of the PCA-IF detection framework on Attack 3.1: Freezing the control signal of Zone C2. (a) The temperature of Zone C2. (b) The control signal of Zone C2. (c) The predictions of the detection framework for Attack 3.1. (d) The thermal comfort index of Zone C2. (e) The total power usage of the HVAC system.

operation as shown in Fig. 13(b). It resulted in failing to control the zone's temperature at the setpoint (Fig. 13(a)) and hence the thermal comfort of Zone B1's occupants was altered as indicated by the PMV index (Fig. 13(d)).

While in **Attack 3.6**, the control signal of Zone A3 was set to 1 during the night operation at 20:00 (Fig. 14(b)). This resulted in a negative thermal discomfort of the zone's occupants with PMV < -1 as demonstrated in Fig. 14(d) due to the extreme cold indoor environment as the temperature of Zone A3 decreased to 16 °C at the instant the attack was launched (Fig. 14(a)). It is worth noting that the total power utilization of the system was not acutely affected as shown in Fig. 14(e) because only a single zone was affected by this attack resulting in an inconsequential power usage increase relative to the overall power utilization of the system.

Attack 4.1 demonstrates falsifying the component's command to operate at a lower speed; in this case, it was the water pump of AHU-B's

Table 11
Comparison between the proposed IF-based HVAC attack detection and the recent works from the literature.

Reference	
Hernandez-Ramos et al. (2015)	**Objective:** Detection of network security breaches **Description of proposed framework:** 1 – An IoT-based security system is proposed by integrating coherent data as fundamental components 2 – It utilizes the available localization data to implement the access control for the building devices 3- It employs authentication and authorization mechanisms for service access protection **Evaluation metric:** N/A **Limitation:** It was not integrated into the constrained IoT environments for defining alternative approaches to validate its suitability
Singh et al. (2017)	**Objective:** Detection of network security breaches **Description of proposed framework:** 1 – Multivariate correlation analysis technique and the known attack database are utilized to analyze the flow packets in the network layer 2 – The received network traffic data is analyzed as follows: (a) It is compared with the attack database for a match, if exists, an alarm is raised and the packet is dropped b) Unmatched traffic data is analyzed using the correlation extracted features using data flow diagrams – if vulnerability is detected, the attack database gets updated, alarm is raised, and the packet is dropped – if vulnerability is not detected, the packet gets forwarded **Evaluation metric:** throughput, round-trip-time, accuracy **Limitation:** It requires the availability of attack historical network data and it has unpredictable performance for unseen attacks.
Our work	**Objective:** Detection of sensor and actuator FDI attacks **Description of proposed framework:** 1 – The detection model is used to detect the abnormal system's operation data of the sensors and actuators 2 – It is developed utilizing just the normal data and without the need for the system mathematical model 3 – Isolation forest algorithm is used that utilizes the principle of separating-away anomalous observations and hence improves the attack detection capability **Evaluation metric:** precision, recall **Limitation:** It can only detect attacks that impact the system operation by causing excessive energy consumption and/or altering the thermal comfort levels

coil resulting in deviating the temperature of the supply air of AHU-B (T_{aoA}) from its setpoint (Fig. 15(a)). It is considered a low-impact attack as the temperature of the zones on floor B was not extremely influenced. An example is presented in Fig. 15(b) as the attack caused the temperature of Zone B2 to increase to 22 °C. Even though it was shifted from the setpoint by +2 °C, it was still acceptable as indicated by the PMV index in Fig. 15(c). Hence, the attack was detected after 3 hours after a noticeable increase in power consumption.

Figs. 16 and 17 demonstrate samples of the failed attacks. In **Attack 1.4**, the setpoint of Zone A1 was increased to 26 °C at 11:00 AM, but the attack did not influence the system's energy efficiency nor the thermal comfort significantly as indicated by the PMV index that was still below +1 (Fig. 16(d)). In addition, in **Attack 3.1**, the control signal of Zone C2 was frozen at 0.2 during the day (Fig. 17(b)) resulting in shifting the zone's temperature by about − 2 °C from its setpoint, but without altering the thermal comfort severely or increasing the energy usage immensely.

As mentioned previously, the degree of attack severity is determined by the measure of its impact on the system's power usage and the occupants' thermal satisfaction as in Definition 2.1. Therefore, the PCA-IF detection framework is sufficient for identifying the occurrence of attacks that can reduce the efficiency of the HVAC system or cause extreme thermal discomfort. Moreover, it was observed that detection is quicker and more robust for attacks with greater impact or severity. For instance, **Attack 1.1** was detected within less than 15 min. It is considered a high-impact attack as it targets a central unit as the chiller system. If its operation is perturbed, inevitably the function of the other components of the HVAC system is adversely affected, unlike the slow detection of **Attack 4.1** which was identified after about 3 hours.

The computational overhead varies among the different Machine Learning algorithms, and it is highly dependent on the choice of the algorithm's hyper-parameters. It is defined by the amount of time and space resources required to run the algorithm, and it can be approximated based on the number of computational operations performed using the algorithm. For instance, CNNs involve extensive computations, and hence, they are characterized by their high computational complexity. The key advantage of isolation forests is their low computational complexity, which is in the order of $log(m_{max}) \times n_{estimators} \times n$, where $log(m_{max})$ reflects the depth of the trees, $n_{estimators}$ is the number of trees, and n is the number of features (Liu et al., 2012). The number of features in the data is a key element affecting the computations involved in the isolation forest model directly and indirectly. That is, as the dimension and complexity of the dataset increase, larger values of isolation forest's hyper-parameters are used to achieve adequate detection performance, leading to increased computational overhead. This can be tackled by means of dimensionality reduction, which can improve the computational complexity by removing the redundancy in the data, and consequently reducing its feature space.

4.1. *Comparison with the existing works*

As previously demonstrated in Table 1, Hernandez-Ramos et al. (2015) and Singh et al. (2017) presented data-driven attack detection frameworks for smart buildings. The primary difference between those works and the proposed IF-based approach is the detection objective, which is towards securing the network of the building management systems against security breaches with no regard to the dynamics of the system/process. That is, the developed approaches in Hernandez-Ramos et al. (2015) and Singh et al. (2017) work by analyzing the network traffic data for abnormality in the network packets flow. They rely on network data and IoT strategies for attack detection. Since examining the physical dynamics of the system can be very useful and can yield promising outcomes, in this paper, we proposed the isolation forest-based attack detection framework in which the detection is made based on analyzing the process data, i.e. sensors and actuators signals, to examine the behavior of the HVAC system and identify the signs of attacks impacting the system operation. An analytical comparison is presented in Table 11 highlighting the contrast between those works and the proposed framework and presenting a summary of the frameworks' descriptions.

5. Conclusion

A semi-supervised, data-driven isolation forest-based attack detection approach for a multi-zone HVAC system was proposed in which the normal operation data were used to develop the detection model. Using the data generated from the TRNSYS model, three schemes of the proposed approach were examined based on the representation of the data used to develop the model as 1) the raw data for Scheme 1, PCA-extracted features for Scheme 2, and 3) 1D CNN autoencoder-extracted features for Scheme 3. Feature extraction using PCA was found to be useful as the redundancy and the uncorrelated components are removed, unlike the 1D CNN-based model due to the steady and slow nature of the HVAC system dynamics. The performance of the proposed approach was compared with standard Machine Learning-based approaches, which are kNN, LOF, PCA, and OCSVM, and it was found promising with low computational complexity, quick and reliable detection, and a relatively low false alarm rate scoring a precision of 90% and a recall of about 61%.

It was found that the detection framework was capable of detecting critical attacks launched against the HVAC system, preciously the ones that resulted in notable system inefficiency and/or thermal discomfort. Moreover, the detection time for critical attacks with low impact on the HVAC system was longer unlike the attacks causing considerable perturbation to the system. It is worth noting that even though Denial-of-Service (DoS) attacks are common in industrial control systems, we have not addressed them since we evaluated the system operation from the process point of view. However, the nature of the DoS attack and its modeling is different and the main issue for DoS is the vulnerability analysis of the system to this type of attack and from the attack detection point of view, DoS attacks can be easily detected by monitoring the received information in each node. Hence, the DoS attack is out of the scope of this work which is mainly focusing on the attack detection problem. Moreover, the effect of the DoS attack is generally identical to a frozen sensor reading or a frozen control signal as indicated by Sánchez, Rotondo, Escobet, Puig, and Quevedo (2019), and both types of attacks have been addressed in the presented evaluation (Attacks 2.1–2.4, and 3.1–3.4).

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

This work was supported by the Qatar National Research Fund (a member of Qatar Foundation) through the National Priorities Research Program (NPRP) under Grant NPRP 10-0206-170360. The findings achieved herein are solely the responsibility of the authors.

References

Ahmed, S., Lee, Y., Hyun, S., & Koo, I. (2019). Unsupervised machine learning-based detection of covert data integrity assault in smart grid networks utilizing isolation forest. *IEEE Transactions on Information Forensics and Security, 14*, 2765–2777.

Alibabaei, N., Fung, A. S., Raahemifar, K., & Moghimi, A. (2017). Effects of intelligent strategy planning models on residential HVAC system energy demand and cost during the heating and cooling seasons. *Applied Energy, 185*, 29–43.

Ande, R., Adebisi, B., Hammoudeh, M., & Saleem, J. (2020). Internet of things: Evolution and technologies from a security perspective. *Sustainable Cities and Society, 54*, 101728.

Braun, T., Fung, B. C., Iqbal, F., & Shah, B. (2018). Security and privacy challenges in smart cities. *Sustainable Cities and Society, 39*, 499–507.

Cutillas, C. G., Ramírez, J. R., & Miralles, M. L. (2017). Optimum design and operation of an HVAC cooling tower for energy and water conservation. *Energies, 10*, 299.

Darure, T., Yamé, J., & Hamelin, F. (2016). Model-based fault-tolerant control of VAV damper lock-in place failure in a multizone building. *2016 14th international conference on control, automation, robotics and vision (ICARCV)*, 1–6.

Diallo, T. M., Zhao, X., Dugue, A., Bonnamy, P., Javier Miguel, F., Martinez, A., et al. (2017). Numerical investigation of the energy performance of an opaque ventilated Façade system employing a smart modular heat recovery unit and a latent heat thermal energy system. *Applied Energy, 205*, 130–152.

D'Innocenzo, A., Smarra, F., & Domenica, M. (2016). Resilient stabilization of multi-hop control networks subject to malicious attacks. *Automatica, 71*, 1–9.

Du, Z., Fan, B., Jin, X., & Chi, J. (2014). Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis. *Building and Environment, 73*, 1–11.

Elnour, M., Meskin, N., & Al-Naemi, M. (2020). Sensor data validation and fault diagnosis using auto-associative neural network for HVAC systems. *Journal of Building Engineering, 27*, 100935.

Elnour, M., Meskin, N., Khan, K., & Jain, R. (2020). A dual-isolation-forests-based attack detection framework for industrial control systems. *IEEE Access, 8*, 36639–36651.

Every, P. M. V., Rodriguez, M., Jones, C. B., Mammoli, A. A., & Martinez-Ramon, M. (2017). Advanced detection of HVAC faults using unsupervised SVM novelty detection and Gaussian process models. *Energy and Buildings, 149*, 216–224.

Fisk, D. (2012). Cyber security, building automation, and the intelligent building. *Intelligent Buildings International, 4*, 169–181.

Ghofrani, A., Nazemi, S. D., & Jafari, M. A. (2019). HVAC load synchronization in smart building communities. *Sustainable Cities and Society, 51*, 101741.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press.

Granzer, W., Praus, F., & Kastner, W. (2010). Security in building automation systems. *IEEE Transactions on Industrial Electronics, 57*, 3622–3630.

Habibzadeh, H., Nussbaum, B. H., Anjomshoa, F., Kantarci, B., & Soyata, T. (2019). A survey on cybersecurity, data privacy, and policy issues in cyber-physical system deployments in smart cities. *Sustainable Cities and Society, 50*, 101660.

Hachem, J. E., Chiprianov, V., Babar, M. A., Khalil, T. A., & Aniorte, P. (2020). Modeling, analyzing and predicting security cascading attacks in smart buildings systems-of-systems. *Journal of Systems and Software, 162*, 110484.

Hernandez-Ramos, J. L., Moreno, M. V., Bernabe, J. B., Carrillo, D. G., & Skarmeta, A. F. (2015). SAFIR: Secure access framework for IoT-enabled services on smart buildings. *Journal of Computer and System Sciences, 81*, 1452–1463.

IEA. (2017). *Energy technology perspectives 2017* (Accessed 28 June 2020) https://www.iea.org/etp2017/.

IoT Security Foundation (n.d.). Smart Cities – The emergence of the CyberSafe building. IoT Security Foundation. https://www.iotsecurityfoundation.org/smart_cities_the_emergence_of_the_cyber_safe_building/ (Accessed 28 June 2020).

Iqbal, J., Khan, M., Talha, M., Farman, H., Jan, B., Muhammad, A., et al. (2018). A generic internet of things architecture for controlling electrical energy consumption in smart homes. *Sustainable Cities and Society, 43*, 443–450.

Jararweh, Y., Otoum, S., & Ridhawi, I. A. (2020). Trustworthy and sustainable smart city services at the edge. *Sustainable Cities and Society, 62*, 102394.

Klein, J. D., G.M, M., & Sherrill, J. B. (2017). *TRNSYS 17: Transient system simulation program, solar energy laboratory*.

Kaspersky. (2019). *Nearly four in ten smart buildings targeted by malicious attacks in H1 2019* (Accessed 28 June 2020) https://www.usa.kaspersky.com/about/press-releases/2019_smart-buildings-threat-landscape/.

Kate Kochetkova. (2016). *Hacking air conditioners leads to the whole block's blackout* (Accessed 15 November 2020) https://www.kaspersky.com/blog/air-conditioner-hack/11348/.

Khan, Z. A. (2018). Using energy-efficient trust management to protect IoT networks for smart cities. *Sustainable Cities and Society, 40*, 1–15.

Kim Zetter. (2013). *Researchers hack building control system at Google Australia office* (Accessed 15 November 2020) https://www.wired.com/2013/05/googles-control-system-hacked/.

KrebsonSecurity. (2014). *Target hackers broke in via HVAC company* (Accessed 15 November 2020) https://krebsonsecurity.com/2014/02/target-hackers-broke-in-via-hvac-company/.

Lee Mathews. (2016). *Hackers use DDoS attack to cut heat to apartments* (Accessed 15 November 2020) https://www.forbes.com/sites/leemathews/2016/11/07/ddos-attack-leaves-finnish-apartments-without-heat/?sh=1e87e5ec1a09/.

Li, S., Joe, J., Hu, J., & Karava, P. (2015). System identification and model-predictive control of office buildings with integrated photovoltaic-thermal collectors, radiant floor heating and active thermal storage. *Solar Energy, 113*, 139–157.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery on Data, 6*, 3:1–3:39.

Novak, T., & Gerstinger, A. (2010). Safety- and security-critical services in building automation and control systems. *IEEE Transactions on Industrial Electronics, 57*, 3614–3621.

Ogoli, D. (2007). Thermal comfort in a naturally-ventilated educational building. *Enquiry: A Journal for Architectural Research*, 4.

Paridari, K., O'Mahony, N., El-Din Mady, A., Chabukswar, R., Boubekeur, M., & Sandberg, H. (2018). A framework for attack-resilient industrial control systems: Attack detection and controller reconfiguration. *Proceedings of the IEEE 106*, 113–128.

Peacock, M., & Johnstone, M. (2014). An analysis of security issues in building automation systems. *12th Australian information security management conference*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Qiu, A., Yan, Z., Deng, Q., Liu, J., Shang, L., & Wu, J. (2020). Modeling of HVAC systems for fault diagnosis. *IEEE Access, 8*, 146248–146262.

Qureshi, K. N., Rana, S. S., Ahmed, A., & Jeon, G. (2020). A novel and secure attacks detection framework for smart cities industrial internet of things. *Sustainable Cities and Society, 61*, 102343.

Rahman, M. A., Asyhari, A. T., Leong, L., Satrya, G., Hai Tao, M., & Zolkipli, M. (2020). Scalable machine learning-based intrusion detection system for IoT-enabled smart cities. *Sustainable Cities and Society, 61*, 102324.

Rodriguez-Trejo, S., Ahmad, A. M., Hafeez, M. A., Dawood, H., Vukovic, V., Kassem, M., et al. (2017). Hierarchy based information requirements for sustainable operations of buildings in Qatar. *Sustainable Cities and Society, 32*, 435–448.

Safa, M., Safa, M., Allen, J., Shahi, A., & Haas, C. T. (2017). Improving sustainable office building operation by using historical data and linear models to predict energy usage. *Sustainable Cities and Society, 29*, 107–117.

Seo, J., Ooka, R., Kim, J. T., & Nam, Y. (2014). Optimization of the HVAC system design to minimize primary energy demand. *Energy and Buildings, 76*, 102–108.

Singh, S., Sharma, P. K., & Park, J. H. (2017). SH-SecNet: An enhanced secure network architecture for the diagnosis of security threats in a smart home. *Sustainability*, 9.

Singh, S. K., Jeong, Y.-S., & Park, J. H. (2020). A deep learning-based IoT-oriented infrastructure for secure smart city. *Sustainable Cities and Society, 60*, 102252.

Sánchez, H. S., Rotondo, D., Escobet, T., Puig, V., & Quevedo, J. (2019). Bibliographical review on cyber attacks from a control oriented perspective. *Annual Reviews in Control, 48*, 103–128.

Sun, J., Hu, G., & Spanos, C. J. (2017). Development and verification of a multizone building HVAC model with TRNSYS. *2017 12th IEEE conference on industrial electronics and applications (ICIEA)*, 887–894.

Ting, K. M. (2010). Confusion matrix. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning* (p. 209). Boston, MA: Springer US.

UN Environment Programme (n.d.). Sustainable buildings. https://www.unenvironment.org/explore-topics/resource-efficiency/what-we-do/cities/sustainable-buildings/ (Accessed 28 June 2020).

Vartouni, A. M., Kashi, S. S., & Teshnehlab, M. (2018). An anomaly detection method to detect web attacks using stacked auto-encoder. *2018 6th Iranian joint congress on fuzzy and intelligent systems (CFIS), 2018*, 131–134.

Vishwanath, A., Chandan, V., & Saurav, K. (2019). An IoT-based data driven precooling solution for electricity cost savings in commercial buildings. *IEEE Internet of Things Journal, 6*, 7337–7347.

Wardell, D. C., Mills, R. F., Peterson, G. L., & Oxley, M. E. (2016). A method for revealing and addressing security vulnerabilities in cyber-physical systems by modeling malicious agent interactions with formal verification. *Procedia Computer Science, 95*, 24–31.

Yoon, J., Dunlap, S., Butts, J., Rice, M., & Ramsey, B. (2016). Evaluating the readiness of cyber first responders responsible for critical infrastructure protection. *International Journal of Critical Infrastructure Protection, 13*, 19–27.

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). PyOD: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research, 20*, 1–7.

Zhu, J., Shen, Y., Song, Z., Zhou, D., Zhang, Z., & Kusiak, A. (2019). Data-driven building load profiling and energy management. *Sustainable Cities and Society, 49*, 101587.